Cogent Food & Agriculture

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/oafa20

High accuracy prediction of Thai rice glycemic index using machine learning

Yusuf Durmus

To cite this article: Yusuf Durmus (2024) High accuracy prediction of Thai rice glycemic index using machine learning, Cogent Food & Agriculture, 10:1, 2411032, DOI: 10.1080/23311932.2024.2411032

To link to this article: https://doi.org/10.1080/23311932.2024.2411032

Published online: 07 Oct 2024.

COGENT FOOD & AGRICULTURE
2024, VOL. 10, NO. 1, 2411032
https://doi.org/10.1080/23311932.2024.2411032

FOOD SCIENCE & TECHNOLOGY | RESEARCH ARTICLE

OPEN ACCESS Check for updates

High accuracy prediction of Thai rice glycemic index using machine learning

Yusuf Durmus

Department of Gastronomy and Culinary Arts, Faculty of Tourism, Artvin Çoruh University, Artvin, Türkiye

ABSTRACT

This study investigated the effectiveness of machine learning (ML) models in estimating the glycemic index (GI) of Thai rice starches from their physicochemical characteristics. Three models, XGBoost, CatBoost and RandomForest, were employed on a dataset comprising various starch properties. All models yielded high R^2 values exceeding 0.957, signifying their accuracy in GI prediction. Moreover, CatBoost exhibited superior performance with the highest test R^2 value (0.977) and the lowest test root mean square error (RMSE) and mean absolute error (MAE) values (1.674 and 1.302, respectively). The CatBoost model was optimized to explore the impact of gelatinization parameters on GI prediction. Optimization revealed that maximizing the GI requires lower onset and peak temperatures, moderate conclusion temperature, and higher enthalpy, while minimizing the GI necessitates higher onset temperature and conclusion temperature, slightly higher peak temperature and slightly lower enthalpy. These findings provide valuable insights into tailoring starch properties to achieve desired GI levels in rice.

ARTICLE HISTORY
Received 10 April 2024
Revised 27 August 2024
Accepted 25 September 2024

KEYWORDS
Thai rice; XGBoost; RandomForest; CatBoost; glycemic index

SUBJECTS
Food Chemistry; Food Engineering; Agriculture & Environmental Sciences; Health Conditions

1. Introduction

Rice (*Oryza sativa* L.) serves as a staple food for over three billion people globally. Thailand, Vietnam, India, China and Pakistan stand as the top five leading net exporters of rice worldwide (Teye et al., 2019). The Thai commercial rice cultivars primarily exported to international markets include high-quality, long-grain aromatic Jasmine rice, as well as glutinous rice and Kor Kho 15 varieties (Sangwongchai et al., 2021). The glycemic response associated with rice is recognized to be relatively high compared to other starchy foods (Frei et al., 2003). The glycemic index (GI) is characterized by the incremental blood glucose observed after consuming the test food, expressed as a percentage of the corresponding area following the consumption of a carbohydrate-equivalent load of a reference product (Bharath Kumar & Prabhasankar, 2014). Evidence substantiates direct correlations between the intake of a low GI diet and a decreased risk of conditions such as obesity, coronary heart disease, diabetes, breast cancer and colon cancer (Thiranusornkij et al., 2019). Numerous studies on Thai rice have also addressed GI values. Miller et al. (1992) also categorized rice as a high GI food, with values ranging from 64 to 93. GI values, derived from 10 distinct rice varieties, were stated to range between 62 (Sangyod Muang Phatthalung) and 82 (Kalasin Kaowong).

The starch structure of rice has a direct influence on its GI. The properties of rice starch are impacted by factors such as its crystalline structure, ratio of amylose to amylopectin and the fine amylopectin structure (Pinkaew et al., 2017). There exists a reported negative correlation between the amylose content and the GI. Frei et al. (2003) noted a GI value of 92.9 for a rice variety with 0% amylose, whereas another starch variety containing 26.9% amylose exhibited a GI of 68. The effect of the mean diameter on the GI was relatively low. Nevertheless, the glycemic load from the consumption of whole grains, raw grains or low-moisture foods such as cookies or tortillas can be reduced by choosing grains with a larger starch granule size (Li

CONTACT Yusuf Durmus yusufdurmus@artvin.edu.tr Department of Gastronomy and Culinary Arts, Faculty of Tourism, Artvin Çoruh University, Artvin, Türkiye

et al., 2023). Crystallinity refers to the ordered arrangement of starch molecules within the granule. The decrease in digestibility may be due to the increase in relative crystallinity. It was mentioned that materials with enhanced crystallinity exhibited greater resistance to enzymatic degradation (Korompokis et al., 2021; Tan & Kong, 2020). In addition, the swelling power also has an effect on enzymatic hydrolysis. Limited swelling power has been reported to contribute to resistance against enzyme hydrolysis (Hoover et al., 2010). The characteristics of starch gelatinization serve to reveal retrogradation and establish correlations with the GI. Elevated gelatinization temperatures, including conclusion, peak and onset temperatures, typically signify a slower breakdown of starch attributable to a more compact molecular arrangement. These higher temperatures can yield a resistant gel structure that is less susceptible to digestive enzymes (Boers et al., 2015). On the other hand, the pasting temperature and final viscosity exhibit a negative correlation with rapidly digestible starch (RDS), indicating a corresponding negative correlation with the GI (Chung et al., 2011). There are many factors that influence the GI, such as dietary fiber, lipid, protein, starch properties, phytic acid, phenolic compounds, cooking and processing. Soluble dietary fibers, such as β-glucan, can form viscous gels that slow down the digestion and absorption of carbohydrates, leading to a reduced GI. Resistant starch shown potential in reducing the GI of foods (Tsitsou et al., 2023). Lipids are known to form inclusion complexes with amylose, resulting in amylose-lipid inclusions that prolong enzymatic hydrolysis. The reduced digestibility of these complexes has been reported to lower the GI. The presence of protein within starchy foods has the potential to attenuate the glycemic response. Protein denaturation and hydrolysis may act as inhibitors of starch hydration and enzymatic digestion (Lal et al., 2021). Phytic acid is primarily found in cereals and legumes. Phytic acid inhibits the activity of α-amylase. This inhibition results in a slower rate of starch hydrolysis, leading to a lower GI (Kumar et al., 2020). Processing techniques employed during food preparation can influence the GI of foodstuffs. Boiling, frying, microwaving and pressure cooking of potatoes results in significantly lower GI (Lal et al., 2021).

Various relationships exist between starch properties and the GI. Employing machine learning (ML) methods appropriately can elucidate these connections, fostering a better understanding and facilitating the production of products with the desired GI. There are limited number of studies in the literature employing ML methods for GI determination (Faradji, 2021; Khan et al., 2022; Li & Arandjelovic, 2017). Hence, there is a need for increasingly comprehensive studies. Although numerous ML methods exist, XGBoost, RandomForest and CatBoost were employed in this study because of their notable accuracy rates. These ML methods offer powerful tools for navigating the complexities of GI prediction. These algorithms can learn from vast datasets of experimental data, identifying subtle relationships between diverse input variables (such as the amylose content, crystallinity and gelatinization parameters of starch) and the desired output variable (GI).

XGBoost is based on gradient boosting decision trees and uses a novel sparsity-aware algorithm for sparse data and a weighted quantile sketch for approximate learning. XGBoost iteratively improves the model by learning from the mistakes of previous models and offers various regularization techniques to prevent overfitting and improve generalizability (Chen & Guestrin, 2016).

RandomForest belongs to the ensemble learning family and is a decision tree-based method that combines multiple decision trees to build a more robust and accurate model. The algorithm works by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Each tree is trained on a random subset of the training data and a random subset of the features. This randomness helps to reduce overfitting and improve the generalization performance of the model (Breiman, 2001).

CatBoost is a gradient boosting method that is designed to handle categorical data and can automatically convert categorical features into numerical features. CatBoost uses a novel algorithm called ordered boosting that optimizes the learning objective and can result in faster convergence and better model accuracy, especially for datasets with a large number of features. It also has an overfitting detector that stops the training when it observes overfitting, which helps to reduce overfitting on small or noisy datasets (Dorogush et al., 2018).

The objective of this study was to investigate the efficacy of ML models in predicting the GI of Thai rice starches based on their physicochemical properties. It was specifically sought to identify the most accurate ML model and the key starch properties impacting GI. Additionally, this study aimed to develop a practical tool for predicting and optimizing GI values using minimal input starch properties.

2. Materials and methods

2.1. Data collection

The data used in this study were obtained from the Thai Rice Starch Database (https://thairicestarch.kku.ac.th/en/browse.html – ID numbers between 1 and 1084) provided by Wanichthanarak and Thitisaksakul (2020). There are currently 1084 entries, each containing composition, chemical and physical measurements of Thai rice. A total of 112 research papers were accessible at the time of data collection, and slight variations in sample preparation and GI analysis methods may exist among them. According to the preliminary examination of the data, 436 of the 1084 entries could be used for GI estimation. Waxy or non-waxy type, amylose content (%), amylose:amylopectin ratio, mean diameter (µm), crystallinity (%), solubility (%), swelling power (g/g), onset temperature (°C), peak temperature (°C), conclusion temperature (°C), enthalpy (J/g), pasting temperature (°C), peak viscosity (°C), final viscosity (°C) and hot paste viscosity (°C) were selected as independent variables, and the GI was defined as the dependent variable to be estimated. As the Thai Rice Starch Database was compiled from many different research articles, missing data were imputed because these independent variables were not measured in each study. Missing values were imputed by multiple imputation by chained equations (MICE) explained by Van Buuren (2018) using the miceforest library published in the Python package repository (https://pypi.org/project/miceforest/).

2.2. Machine learning (ML) modeling

Three different ML methods were used: extreme gradient boosting (XGBoost) (Chen & Guestrin, 2016), RandomForest (Breiman, 2001) and gradient boosting with categorical features (CatBoost) (Dorogush et al., 2018).

The XGBoost model was built with a maximum tree depth of 9, learning rate of 0.1371, number of boostings of 82, minimum loss reduction value of 0.08363, sample rate for bagging of 0.7605, percentage of features at each split selection of 0.9431, L1 regularization value of 0.7781 and L2 regularization value of 0.1097.

The RandomForest model was built with 31 trees, maximum depth with 45, maximum features of 4, minimum number of samples of 3 and minimum number of samples for leaf with 2.

The CatBoost model was built with a tree depth of 4, learning rate of 0.1159, maximum number of trees of 380, minimum number of samples in a leaf of 15, sample rate for bagging of 0.7005, percentage of features at each split selection of 0.3693, and L2 regularization value of 4.4312. Each ML method supports variable importance calculation, whereby the importance of independent variables on GI prediction within each model was determined. All three models were validated using the train-test split method, with 20% of the data used for combined testing and validation and the remaining 80% for training. The data were randomly partitioned into training and testing subsets using the train_test_split function from the Scikit-learn Python library. The XGBoost model was built with the xgboost library, RandomForest was built with the scikit-learn library, and the CatBoost model was built with the CatBoost library using Python 3 version 3.9.5.

2.3. Performance metrics

Three different performance metrics were used to evaluate the XGBoost, RF and CatBoost models; coefficient of determination (R^2), root mean square error (RMSE) and mean absolute error (MAE). These metrics are shown in Equations (1–3).

$$R^2 = 1 - \frac{\sum_i^n (y_i - \check{y}_i)^2}{\sum_i^n (y_i - \bar{y})^2} \tag{1}$$

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

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^n |y_i - \check{y}_i| \tag{3}$$

where y_i is the ith observed value; $\bar{y}$ is the average of the observed values; $\check{y}_i$ is the predicted value for the ith observed value; and n is the number of observed values.

2.4. Principal component analysis (PCA) and Hierarchical cluster analysis (HCA)

The independent variables of the starch samples were analyzed by principal component analysis (PCA) to establish linear relationships among the variables. The data were standardized and decomposed into principal components (PCs) and a biplot was drawn using the first 2 PCs (Jolliffe & Cadima, 2016).

Hierarchical cluster analysis (HCA) was performed to group together the independent variables, and the outcome was represented as a dendrogram. The distance between two clusters was determined using the nearest point algorithm presented in Equation (4). This algorithm determines the distance by considering the distance between the closest objects in two clusters.

$$D(X,Y) = \min_{x \in X\, y \in Y} d(x,y) \tag{4}$$

where X and Y are the first and second clusters, respectively, and x and y are the objects belonging to each cluster.

The standardized Euclidean metric (seuclidean) shown in Equation (5) was used to determine the distance between two points.

$$D(X,Y) = \sqrt{(}\left(\Sigma(xi - yi)^2 / s^2\right) \tag{5}$$

where X and Y are the two vectors being compared, xi and yi are the corresponding elements of the vectors, and s^2 is the variance of the entire dataset. PCA and HCA analyses were performed with Python 3 version 3.9.5.

2.5. Mathematical optimization of the CatBoost model

Mathematical optimization is the process of finding the maximum or minimum point of an objective function by optimally adjusting values across a set of inputs (Zomorrodi et al., 2012). In this study, two objective functions were produced with the CatBoost model. Since the highest R^2 value and the lowest RMSE and MAE values were achieved with the CatBoost model, it was found to be appropriate to optimize this model rather than XGBoost and RF. Variable importance measurements were performed for all three models and it was found that gelatinization parameters were very important in predicting the GI in all three models. Therefore, the CatBoost model was optimized in two different ways, maximization and minimization, to determine what the gelatinization parameters should be to reach the highest GI value and the lowest GI value. The onset temperature (°C), peak temperature (°C), conclusion temperature (°C) and enthalpy (J/g), are the independent variables of the CatBoost model whose optimum values were determined, while the GI is the dependent variable whose value was maximized or minimized. The tree-structured Parzen Estimator algorithm described by Bergstra et al. (2011) was used as optimization algorithm with the optuna library using Python 3 version 3.9.5. The number of trials for optimizing the CatBoost model objective function was 4000.

3. Results and discussion

3.1. Machine learning (ML) models

The Thai rice database contains many starch analysis results, some of which were used in this study. The analysis results used in the study included waxy or non-waxy type, amylose content, amylose:amylopectin ratio, mean diameter, crystallinity, solubility, swelling power, gelatinization parameters (onset temperature, peak temperature, conclusion temperature and enthalpy), pasting properties (pasting temperature, peak viscosity, final viscosity and hot paste viscosity) as independent variables and the GI as the dependent variable. Therefore, using the independent variables described above, it was aimed to determine the GI value. The GI is a direct measure of a food's impact on blood sugar levels, making it the ideal dependent variable for understanding the influence of starch properties on carbohydrate digestion and the glycemic response. For this purpose, three distinct ML methods were utilized: XGBoost, RandomForest and CatBoost.

The results of the three models are shown in Figure 1. The highest training R^2 value of 0.999 was obtained with the XGBoost model. The test R^2 value was 0.973 for XGBoost. The best R^2 value among the test sets was reached with the CatBoost model with 0.977. Considering that R^2 ranges from 0 to 1 and is an indicator of how well the model fits the data, with higher values indicating greater reliability of the model, it can be concluded that the CatBoost model is more promising than the others. It is a significant fact that all three models yielded very high R^2 values, indicating that the GI can be predicted with high accuracy. It is also important to examine the RMSE and MAE values for a better evaluation of the models. RMSE and MAE offer valuable insights, as they express model error in the same units as the target variable, while R^2 facilitates the comparison of diverse target variables. In scenarios where dataset deviations are anticipated to be substantial, particularly in the presence of outliers, the RMSE proves more sensitive in detecting these errors. Conversely, the MAE yields more easily interpretable results when deviations exhibit minimal variance. The combination of the MAE and RMSE enables the identification of the variance in the errors. It is noteworthy that the RMSE

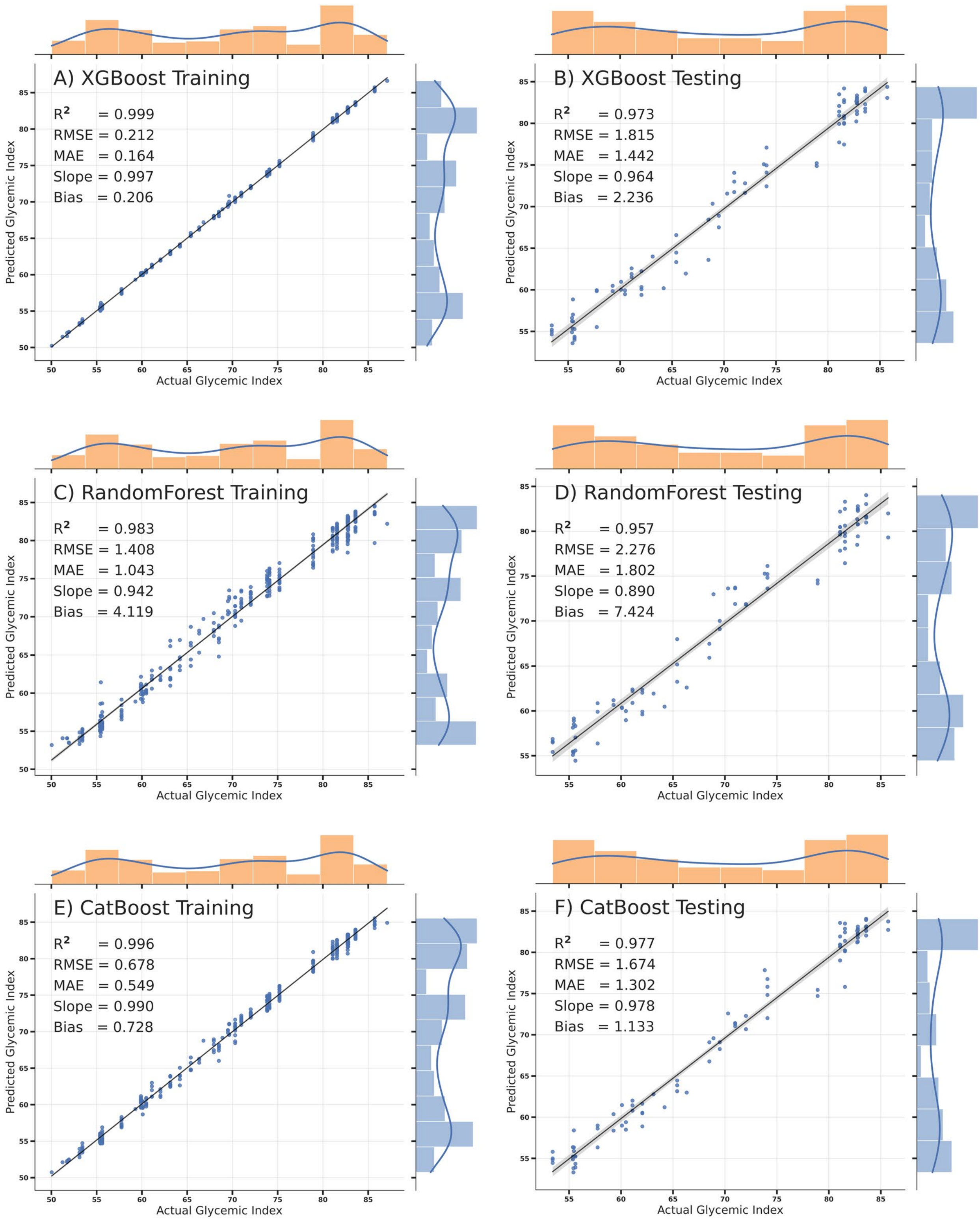

Figure 1. R-squared (R^2), root mean square error (RMSE) and mean absolute error (MAE) values of the XGBoost training (A), XGBoost testing (B), RandomForest training (C), RandomForest testing (D), CatBoost training (E) and CatBoost testing (F).

consistently exceeds or equals the MAE; a greater disparity between the metrics indicates a greater variance in individual errors within the dataset (Chai & Draxler, 2014). During the model training process, XGBoost demonstrated the lowest RMSE (0.212) and MAE (0.164) values. The XGBoost model yielded RMSE and MAE values of 1.815 and 1.442, respectively, for the test dataset. Considering the model test results, the RandomForest exhibited higher RMSE (2.276) and MAE (1.802) compared to the other models. In the

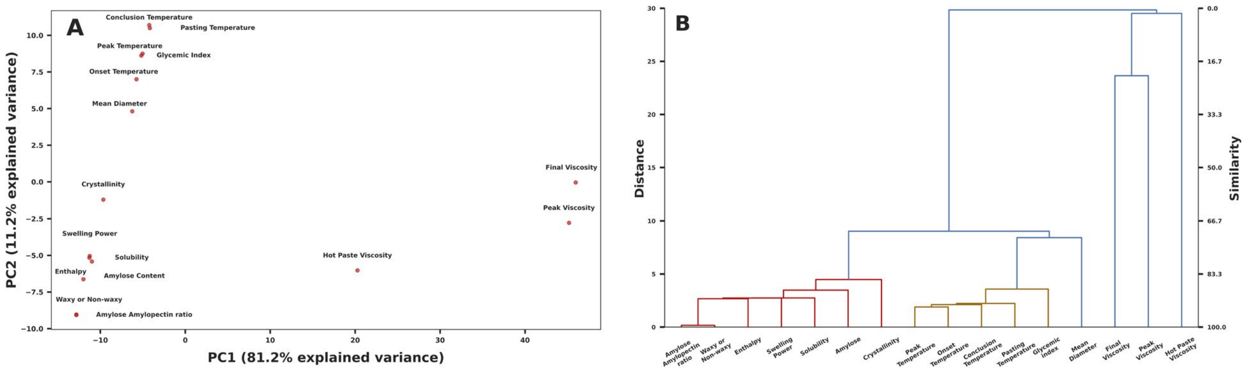

Figure 2. Biplot of PCA illustrating the overall variation among the experiments (A) and dendrogram of cluster analysis based on experiments (B).

assessment of test performance, CatBoost yielded the most favorable RMSE (1.674) and MAE (1.302) scores. CatBoost presents several potential advantages that could lead to superior performance compared to other algorithms. CatBoost implements ordered boosting to reduce overfitting, while XGBoost employs regularization techniques for overfitting prevention. RandomForest inherently exhibits robustness against overfitting due to its ensemble nature comprised of multiple decision trees. CatBoost builds balanced decision trees, unlike the asymmetric trees in XGBoost. CatBoost has been shown to exhibit competitive or superior accuracy, particularly in the context of datasets containing a high number of categorical features. The presence of a categorical feature indicating the waxy or non-waxy characteristic of the sample likely contributed to improved model performance.

The GI is a measure of how a carbohydrate-containing food increases blood glucose. The GI scale ranges from 0 to 100. The GI was classified such that values below 55 were designated as low GI, those falling within the range of 56–69 were considered moderate GI, and values equal to or exceeding 70 were characterized as high GI (Ahmed et al., 2021). Considering the diverse range of GI values and their potential variations, the remarkably low MAE (1.302) exhibited by the CatBoost model in GI prediction highlights the substantial success attained through ML modeling in this study.

3.2. Principal component analysis (PCA) and hierarchical cluster analysis (HCA)

PCA was employed to elucidate the variation observed among the experiments. PCA serves as a method to reduce the dimensionality of data by calculating PCs along the directions exhibiting the greatest variance from the original dataset (Achten et al., 2019). A biplot illustrating PCA across the experiments is shown in Figure 2. The first two PCs accounted for 92.4% of the total variance.

There exists a negative correlation between both the amylose content and the amylose-to-amylopectin ratio with the GI (Figures 2 and 3). Starches with an amylose content less than 1% are categorized as waxy (Lal et al., 2021). Waxy starches lack amylose, contributing to higher solubility and potentially higher GI compared to non-waxy starches (Frei et al., 2003). In addition, higher amylose content and amylose:amylopectin ratios likely reflect more tightly packed starch structures and restricted water interaction, leading to a lower GI (Sissons et al., 2020). Conversely, lower amylose content, crystallinity, smaller granule size, lower onset and peak temperatures and lower pasting viscosities tend to increase the GI (Boers et al., 2015; Ek et al., 2014; Sissons et al., 2020). This is in line with the correlation results shown in Figure 3. The crystallinity of starch influences the GI, with more crystalline regions being digested at a slower rate, thereby resulting in a lower GI (Ek et al., 2014). High ordered structure of crystalline regions hinders the accessibility of digestive enzymes to starch molecules, resulting in slower digestion and a slower release of glucose into the bloodstream (Boers et al., 2015).

The PCA plot shown in Figure 2 reveals a negative association between the amylose content and peak viscosity. Increased amylose content facilitates the development of more complex starch structures, leading to a reduced peak viscosity at higher pasting temperatures. This negative correlation has been observed to adversely impact the GI. The amylose content and the amylose:amylopectin ratio exhibited a negative association with the peak, final, and hot paste viscosities according to Figure 2. Higher viscosity values indicate the formation of more robust gel

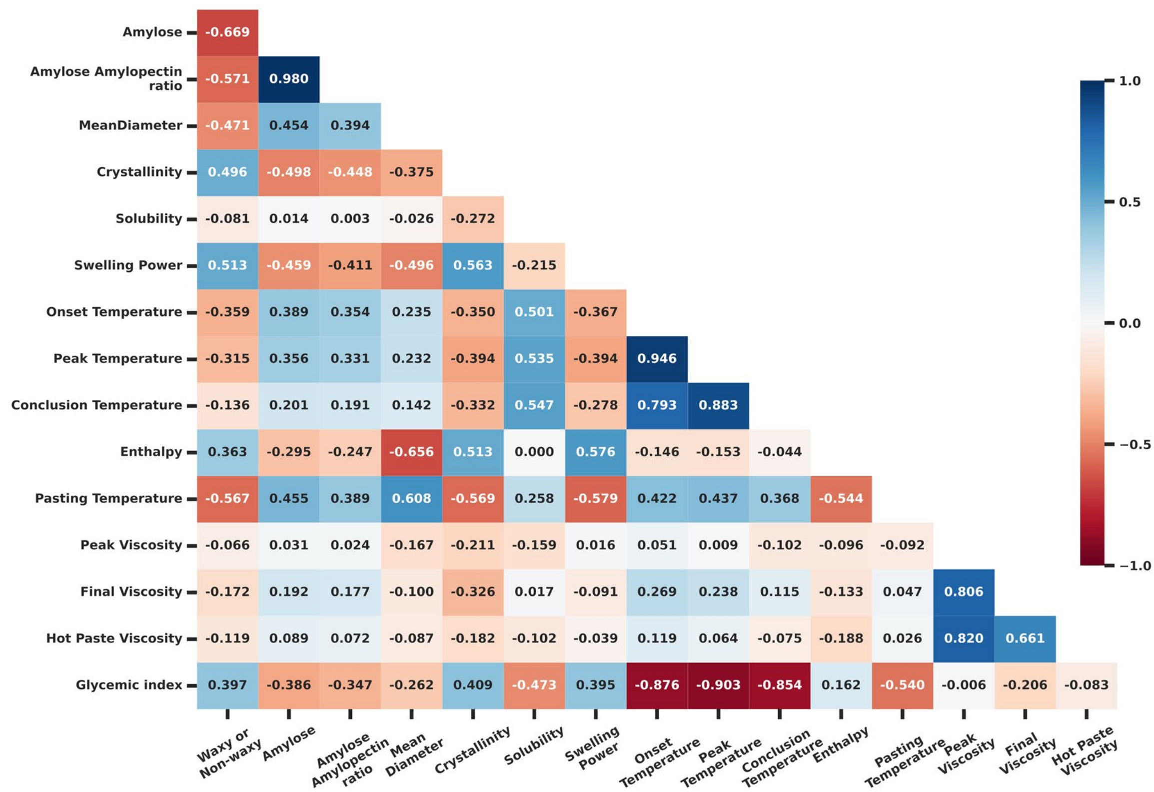

Figure 3. Correlation results for the experiments.

networks in starches, a characteristic that is likely linked to amylopectin content and contributes to the generation of thicker pastes. A decrease in amylose content typically leads to an increase in the number of pores within the starch structure, which correlates positively with faster digestion kinetics. This is because amylopectin, with its branched structure, is more readily accessible to digestive enzymes than the linear chains of amylose (Martens et al., 2018).

The proximity of the swelling power and solubility in the PCA plot (Figure 2) underscores their strong relationship. Senanayake et al. (2013) stated that there was a strong relationship between swelling power and solubility and that as starch granules swell, they tend to dissolve slightly more in water. Increased swelling power can lead to a more rapid gelatinization of starch during cooking, potentially resulting in a higher GI (Thiranusornkij et al., 2019). This could explain the moderate positive correlation observed between the swelling power and GI in Figure 3. Figure 2 shows that there is a relationship between the amylose content and enthalpy. This may be because a higher proportion of amylose or a more crystalline structure typically requires more energy (higher enthalpy) to gelatinize (Ratnayake & Jackson, 2007). The correlation plot presented in this study (Figure 3) shows that there is a negative correlation between the amylose content of starch and the enthalpy of gelatinization, crystallinity and swelling power, which is in agreement with the findings of Yan et al. (2022).

The PCA plot illustrates a strong correlation between the conclusion temperature, peak temperature, onset temperature, pasting temperature and mean diameter variables with the GI. Higher gelatinization temperatures (conclusion, peak, and onset) often indicate slower starch breakdown due to tighter molecular packing. This slower digestion results in a more gradual release of glucose into the bloodstream, resulting in a lower GI. Additionally, higher gelatinization temperatures can result in a more resistant gel structure that is less accessible to digestive enzymes (Boers et al., 2015). Starches with higher pasting temperatures generally require higher temperatures to form a paste, suggesting potentially stronger molecular structures and slower breakdown leading to lower GI. The mean diameter of starch granules can indirectly affect the GI through its impact on the surface area available for enzymatic digestion. Smaller granules tend to have a larger surface area, which may enhance the accessibility of digestive enzymes and potentially increase the rate of digestion. However, the overall impact of granule size on the GI is relatively minor compared to other factors such as amylose content (Chung et al., 2011; Li et al., 2023).

A slight negative correlation exists between the final viscosity and the GI as shown in Figure 3. This correlation may be attributed to the fact that higher a final viscosity indicates a more stable paste that is resistant to breakdown, which in turn delays glucose release and contributes to a lower GI (Ek et al., 2014). Amylose tends to form firmer gels and increase viscosity. The amylose:amylopectin ratio also plays a role, with a higher amylopectin content typically resulting in lower viscosities due to its branched structure. According to Ek et al. (2014), high DSC transition temperatures were associated with a high degree of crystallinity or the presence of more ordered crystalline regions. These characteristics contribute to enhanced structural stability, rendering the granules more resistant to gelatinization. Starch granules with larger mean diameters tend to have lower crystallinity, while smaller granules have higher crystallinity. The crystallinity of starch can affect its functional properties, such as its digestibility and gelatinization behavior (Subzwari et al., 2019).

The relationships between experiments were also investigated using HCA. Higher cutting points indicate more distant associations between the clusters in the dendrogram, while lower cutting points indicate closer links in HCA (Malalgoda et al., 2017). A dendrogram can be used to visually count the number of clusters that are present. Clusterings are evident in both the PCA and HCA (Figure 2) plots. In both plots, whether the starch is waxy or non-waxy, the amylose content, amylose:amylopectin ratio, crystallinity, solubility, swelling power and enthalpy results form one cluster. Moreover, the peak temperature, onset temperature, conclusion temperature, pasting temperature and GI results constitute another cluster. In both plots, there is a notable proximity between peak viscosity and final viscosity, with both parameters situated at a distance from other variables. Consequently, it can be asserted that the PCA and HCA analyses validate each other, producing consistent outcomes for this particular dataset.

3.3. Optimization of the CatBoost model

In this study, three distinct ML models were implemented, yielding highly successful outcomes with R^2 values above 0.95 in all instances. However, the estimation of the GI employed 15 different parameters in these models. Instead of relying on such an extensive set of parameters for GI estimation, a more reasonable approach would involve direct analysis of the GI itself. Hence, variable importance analysis was conducted for all three models to identify the parameters that play the most crucial role in GI estimation. Subsequently, in the second stage, considering the that highest accuracy was achieved with CatBoost, additional modeling was conducted using this ML method. In the third stage, the CatBoost model was optimized, and a detailed examination was conducted to assess the impacts of these parameters on the model. The variable importance results are given in Figure 4. In all three models, the most significant parameters were the onset temperature, peak temperature, and conclusion temperature, all of which pertain to the gelatinization properties. In addition to these factors, the swelling power, pasting temperature, and solubility were also identified as significant in the estimation of the GI. Accordingly, a new CatBoost model was built using four gelatinization parameters (onset temperature, peak temperature, conclusion temperature, and enthalpy) and the corresponding results are illustrated in Figure 5. The R^2 values for the training and test sets were 0.974 and 0.931, respectively. The RMSE and MAE for the test dataset were found to be 2.883 and 2.212, respectively. Paired sample T-test was performed for the CatBoost test model and the difference between actual and predicted values was statistically inspected. Accordingly, the P value was calculated as 0.9136 and it was determined that there was no statistical difference between the actual and predicted values. Even with the exclusion of 11 out of the initial 15 parameters, the model maintains a notably high level of accuracy.

The CatBoost model was optimized, and the effects of these parameters on the model were examined. Mathematical optimization can be used to maximize or minimize a target variable, fix it to a certain value or ensure that it is within a certain range. The optimization of the CatBoost model was approached in two different ways: first, by determining the parameter values that maximize the GI, and second, by identifying the values that minimize it. The contributions of the gelatinization parameters to maximizing and minimizing the GI are given in Figure 6. The maximum GI value was 85.76, while an onset temperature of 61.64 °C, a peak temperature of 69.35 °C, a conclusion temperature of 77.38 °C, and an enthalpy of 19.02 J/g were required to reach this value. To achieve the minimum GI value of 52.31, the corresponding parameter values should be an onset temperature of 84.98 °C, a peak temperature of 76.86 °C, a conclusion temperature of 84.33 °C, and an enthalpy of 9.75 J/g. Figure 6

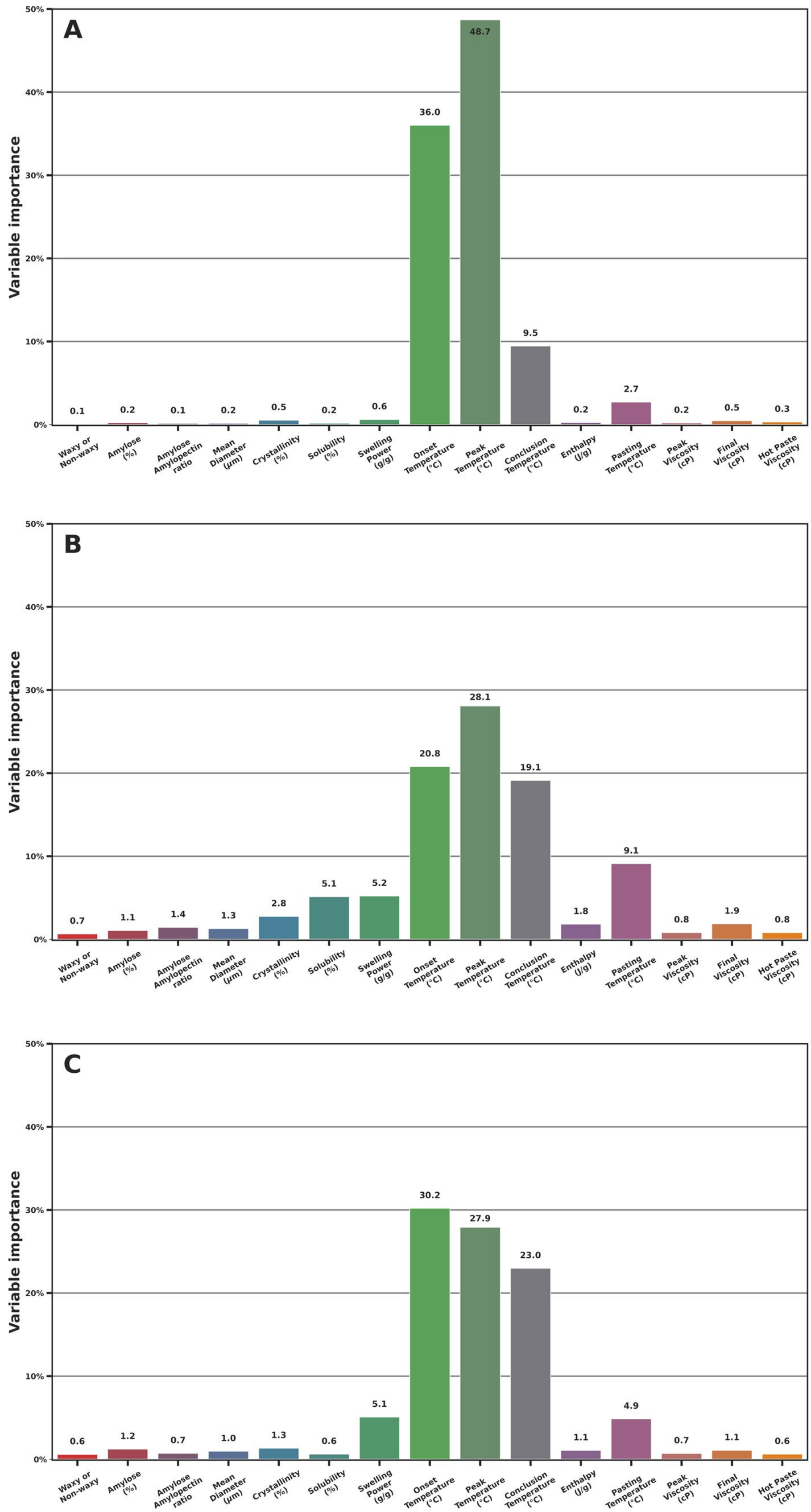

Figure 4. Variable importance measurements based on the XGBoost (A), RandomForest (B), and CatBoost (C) models.

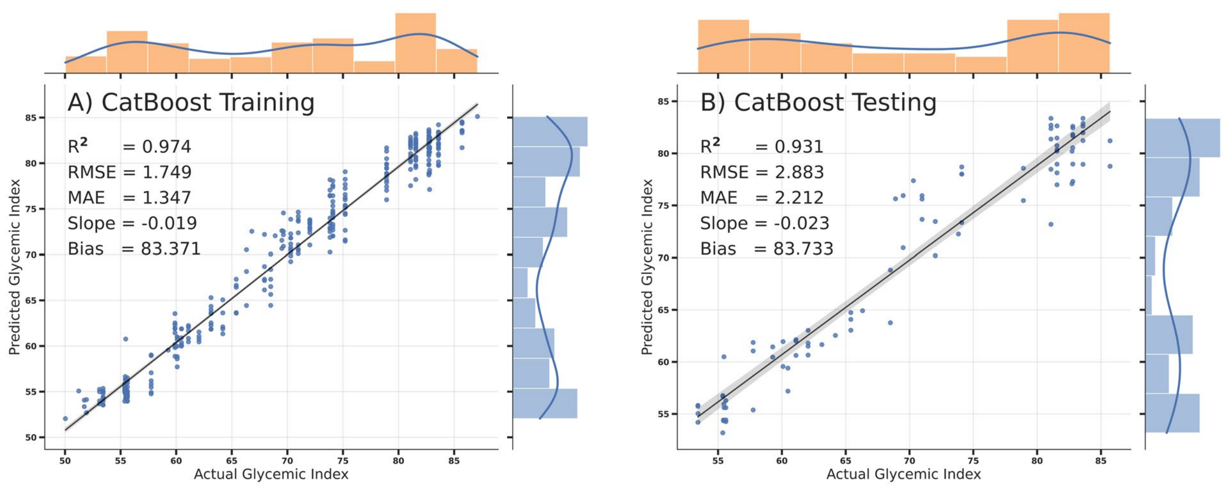

Figure 5. R-squared (R^2), root mean square error (RMSE) and mean absolute error (MAE) values of the CatBoost model built with gelatinization parameter for training (A) and testing (B).

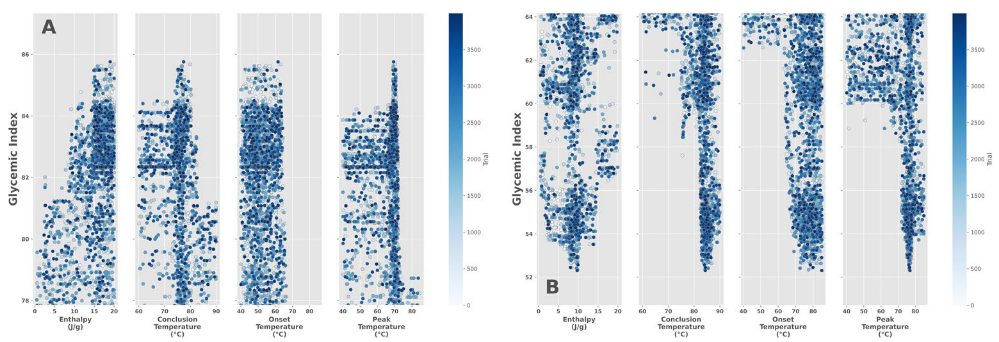

Figure 6. Contribution of gelatinization parameters on maximizing (A) and minimizing (B) the glycemic index.

shows the parameter values for each trial during optimization. To obtain high values for the GI, the onset temperature should be below 70 °C, the peak temperature should be approximately 70 °C, the conclusion temperature should fall within the range of 75–80 °C, and the enthalpy should be greater than 15 J/g. Likewise, to attain a low GI, the onset temperature should exceed 70 °C, the peak temperature should fall within 75–80 °C, the conclusion temperature should be approximately 85 °C, and the enthalpy should range from 7 to 11 J/g. These data hold significant importance because they enable the estimation of GI values based on the available gelatinization parameters for starch. Moreover, employing gelatinization parameters as inputs to the CatBoost model established in this study allows for the determination of the GI with a minimal MAE of only 2.212.

In conclusion, the results demonstrated that all three ML models yielded R^2 values above 0.957, indicating their effectiveness in predicting GI. Notably, the CatBoost model outperformed the others with the lowest RMSE (1.674) and MAE (1.302) for the test dataset, emphasizing its promising potential for accurate GI prediction. Additionally, gelatinization parameters, including onset temperature, peak temperature and conclusion temperature, exhibited strong negative correlations with GI, suggesting their crucial role in determining the glycemic response. The optimization further illustrated the influence of gelatinization parameters on maximizing and minimizing GI values, providing valuable insights for tailoring starch properties to achieve desired glycemic responses. The findings not only contribute to the understanding of carbohydrate digestion but also provide a practical framework for predicting and optimizing the glycemic response of Thai rice based on its inherent starch characteristics.

Acknowledgments

Not applicable.

Disclosure statement

The author declares that there are no conflicts of interest.

About the author

Yusuf Durmus (PhD, Assistant Professor) is currently working at Artvin Çoruh University, Department of Gastronomy and Culinary Arts. He is currently conducting research in Food Engineering. He has experience in research focusing on the application of machine learning algorithms on experimental data in Grain Technology, Food Chemistry and Food Science. His research interests are Cereal Technology, Grain Chemistry, modified starches and machine learning.

Funding

No funding was received.

Data availability

The dataset used in this study are available upon a reasonable request from the author. The dataset are also available in the Thai Rice Starch Database (https://thairicestarch.kku.ac.th/en/browse.html 1084 samples in total. Sample ID numbers between 1–1084).

ORCID

Yusuf Durmus http://orcid.org/0000-0001-8286-4141

References

Achten, E., Schütz, D., Fischer, M., Fauhl-Hassek, C., Riedl, J., & Horn, B. (2019). Classification of grain maize (Zea mays L.) from different geographical origins with FTIR spectroscopy—a suitable analytical tool for feed authentication? *Food Analytical Methods*, *12*(10), 2172–2184. https://doi.org/10.1007/s12161-019-01558-9

Ahmed, J., Riaz, M., & Imtiaz, R. (2021). Glycemic index and glycemic load values. *Pakistan Journal of Medical Sciences*, *37*(4), 1246–1247. https://doi.org/10.12669/pjms.37.4.4555

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in neural information processing systems* (Vol. 24). The MIT Press.

Bharath Kumar, S., & Prabhasankar, P. (2014). Low glycemic index ingredients and modified starches in wheat based food processing: A review. *Trends in Food Science & Technology*, *35*(1), 32–41. https://doi.org/10.1016/j.tifs.2013.10.007

Boers, H. M., Ten Hoorn, J. S., & Mela, D. J. (2015). A systematic review of the influence of rice characteristics and processing methods on postprandial glycaemic and insulinaemic responses. *The British Journal of Nutrition*, *114*(7), 1035–1045. https://doi.org/10.1017/s0007114515001841

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining. ACM.

Chung, H.-J., Liu, Q., Lee, L., & Wei, D. (2011). Relationship between the structure, physicochemical properties and in vitro digestibility of rice starches with different amylose contents. *Food Hydrocolloids*, *25*(5), 968–975. https://doi.org/10.1016/j.foodhyd.2010.09.011

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: Gradient boosting with categorical features support. In *The sixth international conference on learning representations*. April 30 - May 3, 1-7. https://arxiv.org/abs/1810.11363.

Ek, K. L., Wang, S., Brand-Miller, J., & Copeland, L. (2014). Properties of starch from potatoes differing in glycemic index. *Food & Function*, *5*(10), 2509–2515. https://doi.org/10.1039/c4fo00354c

Faradji, D. (2021). A machine learning algorithm to predict the glycemic index of foods. *Available at Social Science Research Network Library as a Preprint:* https://doi.org/10.2139/ssrn.4579233

Frei, M., Siddhuraju, P., & Becker, K. (2003). Studies on the in vitro starch digestibility and the glycemic index of six different indigenous rice cultivars from the Philippines. *Food Chemistry*, *83*(3), 395–402. https://doi.org/10.1016/s0308-8146(03)00101-8

Hoover, R., Hughes, T., Chung, H., & Liu, Q. (2010). Composition, molecular structure, properties, and modification of pulse starches: A review. *Food Research International*, *43*(2), 399–413. https://doi.org/10.1016/j.foodres.2009.09.001

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Khan, M. I., Acharya, B., & Chaurasiya, R. K. (2022). Automatic prediction of glycemic index category from food images using machine learning approaches. *Arabian Journal for Science and Engineering*, *47*(8), 10823–10846. https://doi.org/10.1007/s13369-022-06754-0

Korompokis, K., Verbeke, K., & Delcour, J. A. (2021). Structural factors governing starch digestion and glycemic responses and how they can be modified by enzymatic approaches: A review and a guide. *Comprehensive Reviews in Food Science and Food Safety*, *20*(6), 5965–5991. https://doi.org/10.1111/1541-4337.12847

Kumar, A., Sahu, C., Panda, P. A., Biswal, M., Sah, R. P., Lal, M. K., Baig, M. J., Swain, P., Behera, L., Chattopadhyay, K., & Sharma, S. (2020). Phytic acid content may affect starch digestibility and glycemic index value of rice (Oryza sativa L.). *Journal of the Science of Food and Agriculture*, *100*(4), 1598–1607. https://doi.org/10.1002/jsfa.10168

Lal, M. K., Singh, B., Sharma, S., Singh, M. P., & Kumar, A. (2021). Glycemic index of starchy crops and factors affecting its digestibility: A review. *Trends in Food Science & Technology*, *111*, 741–755. https://doi.org/10.1016/j.tifs.2021.02.067

Li, J., & Arandjelovic, O. (2017). *Glycaemic index prediction: A pilot study of data linkage challenges and the application of machine learning*. IEEE EMBS International Conference on Biomedical & Health Informatics (BHI). 16-19 February, Orland, FL, USA.

Li, M., Daygon, V. D., Solah, V., & Dhital, S. (2023). Starch granule size: Does it matter? *Critical Reviews in Food Science and Nutrition*, *63*(19), 3683–3703. https://doi.org/10.1080/10408398.2021.1992607

Malalgoda, M., Manthey, F. A., & Simsek, S. (2017). Reducing the celiac disease antigenicity of wheat. *Cereal Chemistry*, *95*(1), 49–58. https://doi.org/10.1094/CCHEM-05-17-0100-FI

Martens, B. M., Gerrits, W. J., Bruininx, E. M., & Schols, H. A. (2018). Amylopectin structure and crystallinity explains variation in digestion kinetics of starches across botanic sources in an in vitro pig model. *Journal of Animal Science and Biotechnology*, *9*(1), 91. https://doi.org/10.1186/s40104-018-0303-8

Miller, J. B., Pang, E., & Bramall, L. (1992). Rice: a high or low glycemic index food? *The American Journal of Clinical Nutrition*, *56*(6), 1034–1036. https://doi.org/10.1093/ajcn/56.6.1034

Pinkaew, H., Wang, Y. J., & Naivikul, O. (2017). Impact of pre-germination on amylopectin molecular structures, crystallinity, and thermal properties of pre-germinated brown rice starches. *Journal of Cereal Science*, *73*, 151–157. https://doi.org/10.1016/j.jcs.2016.12.013

Ratnayake, W. S., & Jackson, D. S. (2007). A new insight into the gelatinization process of native starches. *Carbohydrate Polymers*, *67*(4), 511–529. https://doi.org/10.1016/j.carbpol.2006.06.025

Sangwongchai, W., Tananuwong, K., Krusong, K., & Thitisaksakul, M. (2021). Yield, grain quality, and starch physicochemical properties of 2 elite Thai rice cultivars grown under varying production systems and soil characteristics. *Foods (Basel, Switzerland)*, *10*(11), 2601. https://doi.org/10.3390/foods10112601

Senanayake, S., Gunaratne, A., Ranaweera, K., & Bamunuarachchi, A. (2013). Effect of heat moisture treatment conditions on swelling power and water soluble index of different cultivars of sweet potato (Ipomea batatas (L). Lam) starch. *ISRN Agronomy*, *2013*, 1–4. https://doi.org/10.1155/2013/502457

Sissons, M., Sestili, F., Botticella, E., Masci, S., & Lafiandra, D. (2020). Can manipulation of durum wheat amylose content reduce the glycaemic index of spaghetti? *Foods (Basel, Switzerland)*, *9*(6), 693. https://doi.org/10.3390/foods9060693

Subzwari, S., Bryant, G., & Small, D. M. (2019). Characterisation of sorghum starch granules using SAXS: Effects of moisture on crystallinity and structure. *International Journal of Food Science & Technology*, *54*(3), 744–751. https://doi.org/10.1111/ijfs.13989

Tan, L., & Kong, L. (2020). Starch-guest inclusion complexes: Formation, structure, and enzymatic digestion. *Critical Reviews in Food Science and Nutrition*, *60*(5), 780–790. https://doi.org/10.1080/10408398.2018.1550739

Teye, E., Amuah, C. L. Y., McGrath, T., & Elliott, C. (2019). Innovative and rapid analysis for rice authenticity using hand-held NIR spectrometry and chemometrics. *Spectrochimica Acta Part A, Molecular and Biomolecular Spectroscopy*, *217*, 147–154. https://doi.org/10.1016/j.saa.2019.03.085

Thiranusornkij, L., Thamnarathip, P., Chandrachai, A., Kuakpetoon, D., & Adisakwattana, S. (2019). Comparative studies on physicochemical properties, starch hydrolysis, predicted glycemic index of Hom Mali rice and Riceberry rice flour and their applications in bread. *Food Chemistry*, *283*, 224–231. https://doi.org/10.1016/j.foodchem.2019.01.048

Tsitsou, S., Athanasaki, C., Dimitriadis, G., & Papakonstantinou, E. (2023). Acute effects of dietary Fiber in starchy foods on glycemic and Insulinemic responses: A systematic review of randomized controlled crossover trials. *Nutrients*, *15*(10), 2383. https://doi.org/10.3390/nu15102383

Van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). CRC Press. https://doi.org/10.1201/9780429492259

Wanichthanarak, K., & Thitisaksakul, M. (2020). ThRSDB: A database of Thai rice starch composition, molecular structure and functionality. *Database: The Journal of Biological Databases and Curation*, *2020*. 1-9. https://doi.org/10.1093/database/baaa068

Yan, S., Li, Z., Wang, B., Li, T., Li, Z., Zhang, N., & Cui, B. (2022). Correlation analysis on physicochemical and structural properties of sorghum starch. *Frontiers in Nutrition*, *9*, 1101868. https://doi.org/10.3389/fnut.2022.1101868

Zomorrodi, A. R., Suthers, P. F., Ranganathan, S., & Maranas, C. D. (2012). Mathematical optimization applications in metabolic networks. *Metabolic Engineering*, *14*(6), 672–686. https://doi.org/10.1016/j.ymben.2012.09.005