ORIGINAL ARTICLE

Characterization of the estimated glyoxal concept via machine learning-based transformation strategies

YANKI BAŞARAN[1], SEMA ÖZDEMIR[2], MUHAMMED HAKAN YORULMUŞ[3], BÜŞRA YUSUFOĞLU[1]

[1]Department of Chemistry, Faculty of Science and Letters, Istanbul Technical University, Istanbul, Turkey; [2]Department of Chemistry, Faculty of Science and Letters, Yıldız Technical University, Istanbul, Turkey; [3]Department of Management Engineering, Istanbul Technical University, Istanbul, Turkey

ABSTRACT

Background and aim: Glyoxal (GO), formed during food processing, is a highly toxic α-dicarboxyl compound and a precursor in the formation of advanced glycation end products (AGEs). Estimating the GO concentration in food products plays a pivotal role in improving food safety.

Methods: In this study, only studies with similar analytical methods, measurement protocols, reporting standards, and high comparability were included in the dataset; studies showing methodological inconsistencies were excluded. This study used machine learning (ML) regression models to estimate the GO content (μg/100 g) in foods via information on nutrients such as carbohydrates, protein, fat, and sugars obtained from studies in the literature. Fourteen algorithms, including tree-based, ensemble, and regularized linear methods, were tested under different target transformation strategies, such as the Yeo–Johnson, quantile, standard, root-mean-square, and logarithmic shift strategies. Coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE) metrics were used to compare the model findings.

Results: LightGBM presented the lowest MAE value (13.63) under the Yeo–Johnson transformation, whereas the square-root-transformed CatBoost model presented the highest prediction accuracy (R^2 = 0.53, RMSE = 22.24) among all the configurations. Preprocessing significantly improved prediction performance, and model performance was sensitive to the chosen transformation type. On the basis of these findings, the estimated eGO (eGO) concept was introduced into the literature. Fat content was the most influential variable in the CatBoost models, whereas LightGBM exhibited a more balanced feature contribution, with sugars and carbohydrates prominent under certain transformations.

Received: 12 December 2025 | Accepted: 26 May 2026

Correspondence: Bürşa Yusufoğlu, Asst. Prof. Dr. Istanbul / Technical University, Faculty of Science and Letters Ayazağa, 34485 Sarıyer, İstanbul, Turkey Phone: +905452091591 / Email: yusufoglu@itu.edu.tr
ORCID: 0000-0002-9158-9732

Conclusion: These findings provide useful methodological guidance for data science and food safety professionals when selecting appropriate modeling techniques for chemical prediction in food science.

Key words: precursors of advanced glycation end products, processed packaged food, glyoxal formation, machine learning, data transformation techniques

Introduction

Today, changing dietary habits have led to an increase in the demand for processed foods (1). This leads to the formation of advanced glycation end products (AGEs), which affect parameters such as aroma, taste, and color and are known for their impact on human health as a result of these processes. Schiff bases, unstable compounds, are formed as a result of reactions between the carbonyl groups of reducing sugars and the amine groups found in nucleic acids, lipids, and proteins (2). These reactions result in the formation of glyoxal (GO) molecules. GO occurs not only in food products but also endogenously in metabolism. GO formation is important for food safety because it has the potential to undergo advanced oxidation reactions (3, 4). GO accumulation is associated primarily with prolonged processing and the application of parameters such as high temperatures to food products. One of the most important effects of GO is its ability to lead to the formation of AGEs. The resulting accumulation of AGEs directly predisposes individuals to cellular dysfunction, oxidative stress, and various types of chronic inflammation (5). According to current epidemiological findings, the clinical significance of GO intake due to food consumption can be explained by the fact that the amount of AGEs ingested through the diet may be greater than the level of endogenous production (6, 7). Basic approaches, such as careful management of temperature and time adjustments during heat treatment and control of pH and water activity values, are known to limit GO formation by optimizing strategies to minimize GO exposure. Furthermore, recent studies have yielded promising results in reducing GO levels through colloid systems, hydrogels, amino acids, probiotic applications, and natural antioxidants (8, 9). However, these findings also emphasize the need for a detailed examination of their impacts across various matrices, such as their industrial scalability and nutritional value. Notably, meaningful GO quantification is critical for the accuracy and reliability of studies. In this context, although advanced analytical methods such as high-performance liquid chromatography (HPLC), which is widely used for GO quantification, provide high sensitivity, they are limited in screening because of their high cost and long analysis time (10). The challenges and unresolved issues noted in earlier research, together with the high cost and time demands of experimental processes, highlight the need for more efficient predictive strategies. In this context, machine learning (ML)-based modeling offers a compelling solution, and ML utilizes several techniques, including support vector machines (SVMs), artificial neural networks (ANNs), random forests (RFs), and multiple linear regression (MLR). These methods aim to reliably predict sugar, protein, fat, and carbonyl levels (11-13). ML models provide much faster results than laboratory analyses do and offer the opportunity to monitor the dynamics of reactions occurring step by step during the production process (14). In this way, model outputs make significant contributions to both production optimization and understanding the formation mechanisms of GO. Furthermore, the adaptable and scalable structure of these models, supported by chromatographic data, increases the prediction accuracy by retraining with small datasets and can be easily applied to different food matrices (15). In conclusion, combining the accuracy of classical analytical techniques with the speed and cost advantages of AI-based modeling is a strategic approach that supports food safety and public health. This study aimed to estimate the GO content of foods via carbohydrate, protein, fat, and sugar components via ML regression models and to

introduce the concept of "*eGO*" (estimated GO) to the literature. In this context, four different algorithms and various target transformation strategies were evaluated to achieve the highest accuracy and demonstrate the importance of nutrients. Furthermore, this study offers a methodological contribution to ML-based prediction models for improving food safety and optimizing processing processes.

Methods

Overall workflow of the study

The overall workflow of the current study is shown in Figure 1. The framework was designed to integrate both endogenous and exogenous perspectives of glyoxal formation, along with statistical and ML-based predictive modeling approaches to estimate GO levels in foods. Glyoxal is considered a highly reactive α-dicarbonyl compound that can be produced both within biological systems and via food processing pathways (16).

From an endogenous perspective, glyoxal is generated in vivo through multiple metabolic and oxidative pathways, including glucose autoxidation, lipid peroxidation, and the degradation of glycated proteins under conditions characterized by oxidative and carbonyl stress (16, 17). Increased endogenous glyoxal production has been associated with diabetes, inflammation, aging, and metabolic disorders because of its contribution to the formation of AGEs (18).

On the other hand, exogenous glyoxal is commonly produced during thermal food processing, particularly via Maillard reaction pathways involving reducing sugars and amino-containing compounds (19). Processed foods subjected to baking, roasting, frying, or prolonged heat treatment are considered significant dietary sources of glyoxal and related reactive carbonyl species (20). During Maillard reactions, glucose-derived intermediates such as Schiff bases, enaminol intermediates, and Amadori products contribute to the formation of glyoxal and other α-dicarbonyl compounds (21). On the basis of this background biochemical and food chemistry, processed food products were selected as the primary data sources for the study. Nutritional composition values of packaged foods were collected per 100 g and integrated with experimentally reported glyoxal-related information

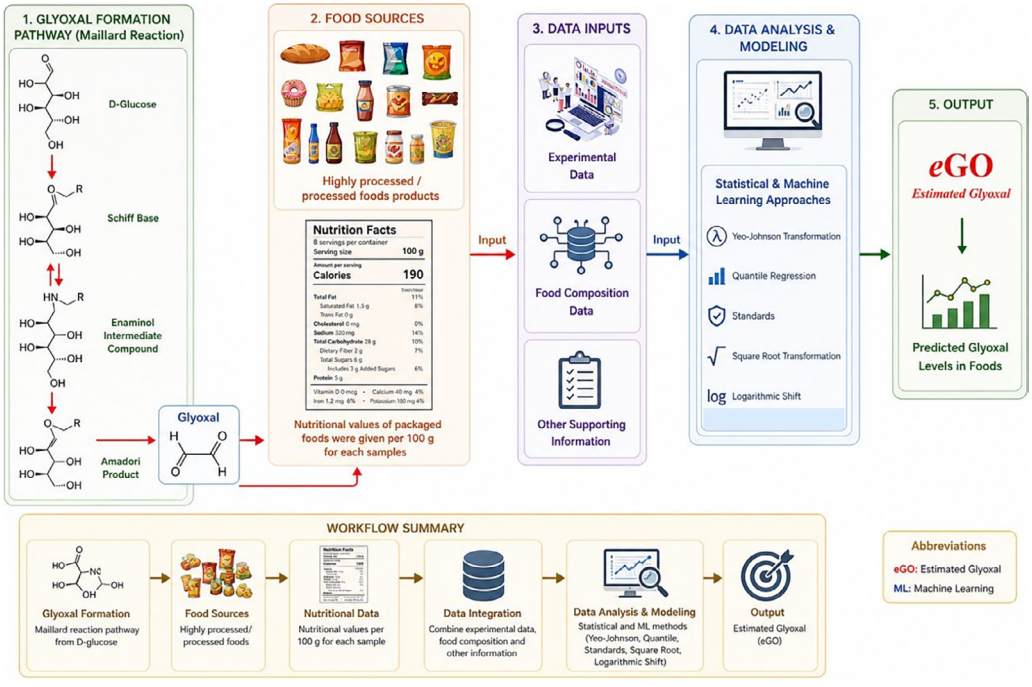

Figure 1. Conceptual workflow of glyoxal formation and machine learning-based prediction of estimated glyoxal (eGO) levels in foods.

obtained from the literature. The compiled dataset includes additional supporting analytical information regarding glyoxal formation behavior as well as food composition variables.

Preprocessing and transformation methods, including the Yeo–Johnson transform, logarithmic transform, square root transform, standardization, and quantile-based approaches, were subsequently applied to improve the data distribution characteristics and optimize model performance. Then, statistical and machine learning approaches were used to investigate the relationships between food composition parameters and glyoxal formation in food matrices. Finally, the proposed framework provides a data-driven strategy for predicting glyoxal levels in processed foods by generating estimated glyoxal (eGO) values for food samples.

As summarized in Figure 1, the workflow combines the biochemical mechanisms of glyoxal formation with the integration of food composition data and predictive modeling approaches.

Data collection

The data used in this study were obtained from the Web of Science Core Collection (WoS), Google Scholar and ResearchGate databases. To identify the relevant literature, keyword searches using the terms "glyoxal," "AGEs," "glycation," and machine learning" were supported by bibliometric studies. This approach allows us to determine both which concepts are being studied and how the connections between these concepts are formed (22, 23). The bibliometric analysis in this study was created by using algorithms extracted from the Scopus database in the VOSviewer application. In particular, the dataset on the Web of Science platform was created with a topic-based search strategy and focused on studies applying modern machine learning techniques to glyoxal analysis in food matrices. The dataset was limited to peer-reviewed articles published in English between 2020 and 2025, and the data selection and curation process was completed on October 20, 2025. Studies were included if they (i) reported quantitative glyoxal concentrations in processed packaged foods; (ii) used validated chromatographic methods; or (iii) provided the nutritional composition (carbohydrate, protein, fat, sugars) of the analyzed foods or sufficient information to recover it. Studies were excluded if they reported only relative values, if glyoxal concentrations were below the limit of detection without numerical reporting, or if the food matrix could not be assigned to a clear processed food category. When the same food product appeared in more than one source, the most recent and methodologically detailed study was retained. All glyoxal values were harmonized to a common unit (μg/100 g) prior to analysis. The keyword map shown in Figure 2 shows the relationships of the keywords with different

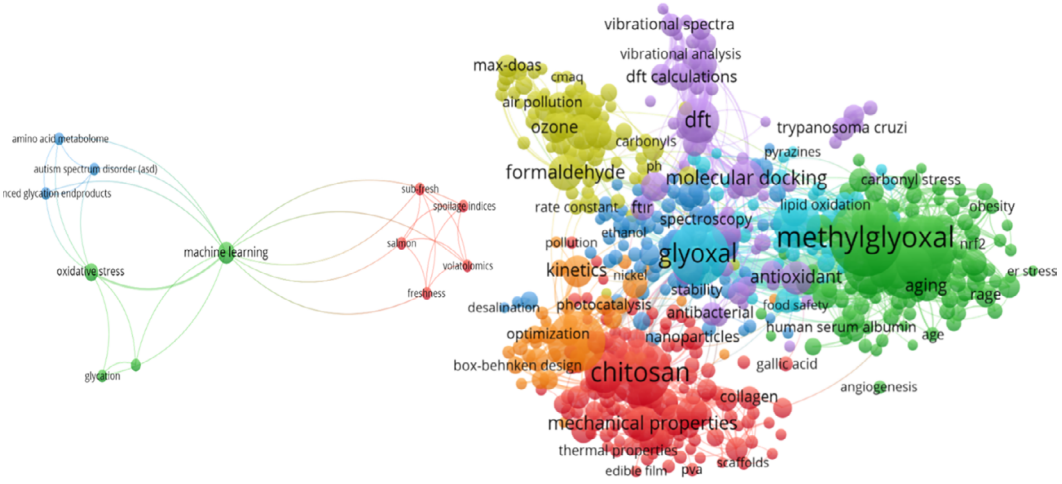

Figure 2. Glyoxal-Machine learning-centered bibliometric keyword map (VOSviewer).

research areas in the literature. Machine learning (oxidative stress, glycation, and AGE formation) and food science applications (predictions of freshness, quality, and spoilage).

In this context, our study will make a significant contribution to the field in terms of machine learning-glycation integration due to the use of comprehensive data. The keyword-view map shows the thematic distribution of the top 1,000 studies selected from 20,000 publications associated with the term "glyoxal" in the literature. Each circle (node) on the map represents a keyword, its size represents its frequency of use, and colors represent research themes (24). The analysis revealed that GO research is focused primarily on the fields of biomedicine, materials, and environmental chemistry. Its association with terms such as oxidative stress, carbonyl stress, and RAGE suggests that glyoxal is a critical intermediate in the formation of AGEs and contributes to cellular damage through oxidative stress (25). Overall, glyoxal is involved in both biological glycation mechanisms and food and environmental chemistry. It appears to play a central role in carbonyl reactions in systems and is therefore considered a key molecule of reactive carbonyl chemistry (23).

The coauthorship map in Figure 3 depicts the collaboration network among scientists publishing in their field, encompassing the top 1,000 studies out of 20,000. Each circle (node) on the map represents an author, and the size of the circle represents the author's number of publications or citations on the relevant topic. The lines between the circles indicate collaborative relationships between authors on the basis of joint publications. Different colors represent different research groups or collaboration networks within which publications are clustered. A general overview of this map reveals that the field is clustered around leading researchers, publication production is collaborative, and the research network is international. These analyses visually reveal influential researchers in the literature, powerful collaboration groups, and the structure of the research network.

Dataset and variables

The dataset used in the study included literature data obtained from laboratory analyses of samples from different product categories, such as packaged snacks and processed foods. The dependent variable was defined as the concentration (μg/100 g) per 100 grams of product. The independent variables were carbohydrate (g/100 g), protein (g/100 g), fat (g/100 g), fructose (g/100 g), glucose (g/100 g), and sucrose (g/100 g) contents. Because GO formation is closely related to biochemical processes such as sugar autoxidation and lipid peroxidation, carbohydrate and fat contents were considered critical independent variables in the estimation of the GO concentration in this study.

This study also reviewed several properties of auxiliary components, such as salts and fibers, whose relationships with GO and AGE formation are occasionally discussed in the literature. One study examined the relationships between salt level and AGE accumulation and α-dicarbonyl AGE accumulation. In this study, such auxiliary features were not directly included in the basic regression models (26). Instead, their effects on the model setup were observed qualitatively and evaluated within the scope of exploratory analysis.

Target transformations

Various transformation techniques have been applied to the target variable, GO, on the basis of its distribution properties. These included log-shift, square root (sqrt), Yeo–Johnson, quantile, and standard scaling, with untransformed data (none) serving as the primary benchmark. A summary and parametric details of the transformations are presented in Table 1.

The statistical basis of the relevant transformation techniques summarized in Table 1 and their use in regression modeling have also been thoroughly discussed in the literature. For example, one study described transformation techniques to approximate right-skewed distributions to normality, whereas another study methodologically discussed the application of these transformations in the context of regression (27, 28). One study evaluated the functions of log, square root, and Yeo–Johnson transformations in adjusting the data distribution (29). These studies demonstrate the methodological contributions of transformations, particularly in improving distributions.

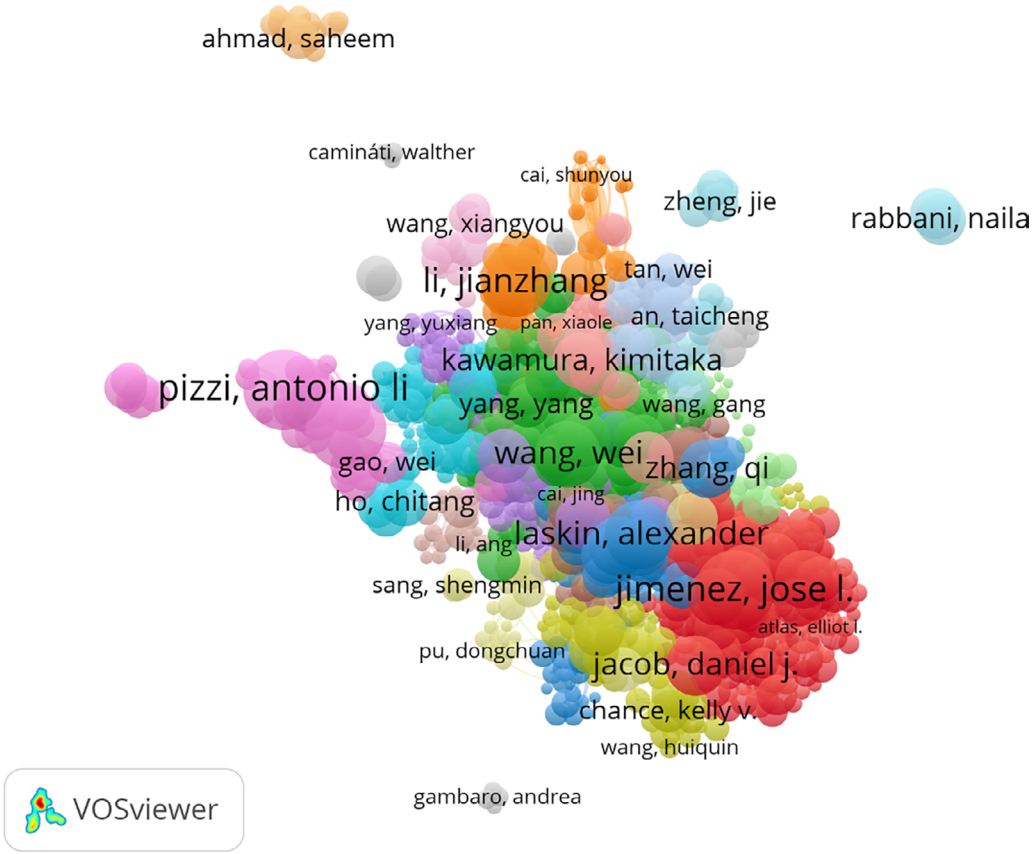

Figure 3. Authorship bibliometric data (VOSviewer).

Table 1. Applied target transformation types, definitions and usage areas

Conversion	Description	Usage Area
Yeo–Johnson	A power transformation that makes skewed (asymmetric) distributions more symmetric. It also works with negative values.	It normalizes the distribution so that the model can learn the target variable better.
Standard	StandardScaler scales data so that the mean is 0 and the standard deviation is 1.	Suitable for algorithms that assume a normal distribution.
Quantile	With QuantileTransformer, values are forced to a uniform or normal distribution.	Ideal for reducing the impact of outliers.
Sqrt	The square root is applied to the target variable; it only works with positive values.	Normalizes right-skewed data, reducing variance.
Log-shift	It is the log transformation in the form np.log (y + c); the constant c prevents negative values.	It symmetricizes exponentially growing or skewed distributions.
None	The target variable is not transformed.	The target variable is not transformed.

Log-shift; Logarithmic shift transformation, None; No transformation, Quantile; Quantile transformation, Sqrt; Square root transformation, Standard; Standard scaling, Yeo–Johnson; Yeo–Johnson transformation.

Model training and validation

Model performance was assessed via three complementary metrics: R^2, RMSE, and MAE. Reporting the RMSE and MAE together provides a more balanced assessment, where a single metric may not be sufficient in all cases, depending on the nature of the error distribution (30). Given the dataset size (N=87), a 5-fold cross-validation strategy was used instead of a single static training-test split to ensure the reliability and robustness of the model predictions. For this purpose, the dataset was randomly shuffled and split into 5 folds. In each iteration, 80% of the data were used for training, and 20% were used for validation. This process was repeated 5 times to ensure that each sample was used only once for validation. Hyperparameter tuning was performed within these folds via GridSearchCV to optimize the model for the lowest root mean square error (RMSE).

Models and their parameter grids

In this study, various regression algorithms based on linear (Linear, Ridge, Lasso, ElasticNet), kernel-based (SVR), neighborhood (KNN), tree/ensemble (decision tree, random forest, gradient boosting, AdaBoost, XGBoost, LightGBM, CatBoost), and neural network (MLP) methods were evaluated. The defined ranges of the applied models are summarized in Table 2.

Model evaluation

The performance of the models was evaluated via R^2, RMSE, and MAE under each target transformation. Owing to their boosting-based structure, CatBoost and LightGBM demonstrate high capacity to capture nonlinear relationships and interactions between variables, and the results obtained are consistent with the expected performance superiority of these models (8, 11). The prominence of fat content

Table 2. Models and defined ranges

Model	Defined Range(s)
Linear	fit_intercept ∈ {True, False}
Ridge/Lasso	alpha ∈ {0.01, 0.1, 1, 10}; (Ridge) solver ∈ {auto, svd, cholesky}
ElasticNet	alpha ∈ {0.01, 0.1, 1}; l1_ratio ∈ {0.1, 0.5, 0.9}; max_iter ∈ {1000, 2000}
SVR	C ∈ {0.1, 1, 10}; kernel ∈ {linear, rbf}; gamma ∈ {scale, auto, 0.1}
Decision Tree	max_depth ∈ {3, 5, 10, None}; min_samples_split ∈ {2, 5, 10}; min_samples_leaf ∈ {1, 2, 4}
Random Forest	n_estimators ∈ {50, 100, 200}; max_depth ∈ {3, 5, 10, None}; min_samples_split ∈ {2, 5}
KNN	n_neighbors ∈ {3, 5, 7, 9}; weights ∈ {uniform, distance}; p ∈ {1, 2}
MLP (ANN)	hidden_layer_sizes ∈ {(50), (100), (50,50)}; alpha ∈ {1e-4, 1e-3, 1e-2}; learning_rate_init ∈ {1e-3, 1e-2}; learning_rate ∈ {constant, adaptive}
XGBoost	max_depth ∈ {3, 5, 7}; learning_rate ∈ {0.01, 0.05, 0.1}; n_estimators ∈ {50, 100, 200}; subsample ∈ {0.8, 1.0}
CatBoost	depth ∈ {4, 6, 8}; learning_rate ∈ {0.01, 0.05, 0.1}; iterations ∈ {50, 100, 200}; l2_leaf_reg ∈ {1, 3}
LightGBM	max_depth ∈ {3, 5, 7}; num_leaves ∈ {20, 31, 40}; learning_rate ∈ {0.01, 0.05, 0.1}; n_estimators ∈ {50, 100, 200}
AdaBoost	n_estimators ∈ {50, 100, 200}; learning_rate ∈ {0.01, 0.1, 1.0}; loss ∈ {linear, square}
Gradient Boosting	n_estimators ∈ {50, 100, 200}; max_depth ∈ {3, 5, 7}; learning_rate ∈ {0.01, 0.05, 0.1}; subsample ∈ {0.8, 1.0}

AdaBoost; Categorical Boosting, Decision Tree; Decision Tree Regression, ElasticNet; Elastic Net Regression, Gradient Boosting; Gradient Boosting, KNN; K-Nearest Neighbors Regression, LightGBM; Light Gradient Boosting Machine, Linear; Linear Regression, MLP (ANN); Multilayer Perceptron (Artificial Neural Network), Random Forest; Random Forest Regression, Ridge/Lasso; Ridge Regression and Lasso Regression, SVR; Support Vector Regression ve XGBoost; Extreme Gradient Boosting.

in feature importance analysis is in line with studies reporting that lipid oxidation is one of the main sources of small reactive carbonyl compounds (4, 31). The significant contributions of carbohydrate and sugar variables are consistent with the literature, supporting the simultaneous effects of sugar degradation and MR processes on GO formation (8). The increased error at high GO values is associated with the limited representation of extreme samples in the dataset and the distribution skewness. Previous studies on the extreme value sensitivity of the RMSE and MAE report that error metrics may exhibit instability in such cases (32, 33).

Results

Packaged food products

The predictive modeling framework used in this study is based on a heterogeneous dataset consisting of 87 processed food samples systematically compiled from peer-reviewed literature sources. All the data were input into the system as a continuous target variable, the GO molecules (μg/100 g), as shown in Table 3. The input feature space consisted of six independent nutritional variables, all measured in g/100 g: carbohydrate, protein, fat, fructose, glucose, and sucrose. Descriptive statistical analysis revealed that the dataset represents a high-energy, processed food profile. Specifically, the total fat content across all the samples was 19.73 g/100 g on average, with the highest values observed in the instant coffee mixes because of their nondairy creamer content. Notably, the dataset does not include fried food products; therefore, lipid levels reflect the internal formulation of bakery and ready-to-eat products rather than fat intake from deep-frying processes. Data cleaning was performed before model training, and an outlier related to carbohydrate content was corrected to ensure data integrity. To provide a representative analysis of bakery and prepared food products, the data were divided into five separate categories: cookies (34.5%), coffee mixtures (34.5%), biscuits (21.8%), crackers (5.7%) and puddings (3.4%).

Machine learning-based prediction of estimated glyoxal (eGO)

This study comparatively examined the performance of different ML algorithms and target transformation strategies in predicting GO concentrations in foods. A total of 14 regression algorithms and all combinations of the 6 target transformation strategies were evaluated via R^2 (a measure of how well the model fits the data), RMSE (a measure of the differences between the predicted values and observed values), and MAE (a measure of the average magnitude of the errors in the predictions). Model performance was sensitive to both the algorithm and the transformation strategy. The Yeo–Johnson and square-root transformations, for instance, were effective in reducing skewness in the data distribution and improving variance balance. Transformation strategies affect not only the error metrics but also the model's learning dynamics and chemical interpretability. Specifically, the Yeo–Johnson and square-root transformations reduced the influence of outliers in the data distribution, stabilizing the variance and improving the prediction accuracy by an average of 3–5%. In contrast, the log-shift transformation led to an increase in error by creating a

Table 3. Dataset definition and description (g/100 g) and GO (μg/100 g)

Sample	Carbohydrate	Protein	Fat	Fructose	Glucose	Sucrose	GO
Biscuit	68.51 ± 6.32	7.16 ± 2.54	15.67 ± 3.85	0.65 ± 0.50	1.07 ± 1.55	15.10 ± 7.19	27.48 ± 18.70
Coffee	74.69 ± 9.19	3.27 ± 1.69	17.35 ± 16.34	0.11 ± 0.30	0.65 ± 0.64	35.45 ± 23.06	68.83 ± 52.81
Cracker	67.96 ± 5.15	8.90 ± 2.78	8.34 ± 1.76	0.23 ± 0.19	0.34 ± 0.25	1.61 ± 1.37	11.98 ± 10.96
Cookie	58.38 ± 5.51	7.10 ± 2.58	28.23 ± 4.88	0.77 ± 1.46	1.39 ± 2.66	21.68 ± 12.02	70.90 ± 22.51
Pudding	20.00 ± 0.75	2.90 ± 0.35	3.30 ± 0.30	0.87 ± 0.06	2.10 ± 0.00	3.50 ± 0.79	15.20 ± 2.31
All	65.44 ± 13.04	5.75 ± 3.00	19.73 ± 12.21	0.49 ± 0.95	1.03 ± 1.79	23.21 ± 18.63	55.40 ± 40.93

scale imbalance in some models, illustrating the bidirectional impact of transformation selection on model performance.

The contributions of the transformation types to model performance are summarized in Table 4. The CatBoost and LightGBM models produced higher R^2 values and lower error values because of their ability to capture nonlinear relationships. These results are consistent with reports suggesting that complex chemical relationships arising from food components can be more successfully modeled with boosting algorithms.

Controlling for skewness and outliers in the data distribution can directly impact model stability and predictive behavior. Among the transformation strategies tested in this study, the Yeo–Johnson and square–root transformations contributed to error reduction by stabilizing the target variable variance. Table 4 presents the best-performing model transformation combinations.

According to the results, the CatBoost–square root transformation model provided the highest fit in terms of overall accuracy, with an R^2 value of 0.53, an RMSE of 22.24, and an MAE of 14.03. In contrast, the LightGBM–Yeo–Johnson transformation yielded the lowest mean error, with an MAE of 13.63. This performance difference can be explained by the

Table 4. Best model–transformation combinations (R^2, RMSE, MAE, and time)

ID	Model	Transform Type	Best Parameters	R^2 Mean	R^2 Std	RMSE Mean	RMSE Std	MAE Mean	MAE Std	Total Time
52	CatBoost	sqrt	{'depth': 4, 'iterations': 200, 'l2_leaf_reg':...	0.53	0.19	22.24	7.03	14.03	4.68	11.31
11	LightGBM	yeo-johnson	{'learning_rate': 0.1, 'max_depth': 3, 'n_esti...	0.52	0.15	22.44	6.36	13.63	4.08	14.67
10	CatBoost	yeo-johnson	{'depth': 4, 'iterations': 200, 'l2_leaf_reg':...	0.51	0.18	22.64	6.74	14.36	4.84	12.69
24	CatBoost	standard	{'depth': 4, 'iterations': 200, 'l2_leaf_reg':...	0.50	0.14	23.00	5.77	14.92	3.89	10.94
53	LightGBM	sqrt	{'learning_rate': 0.05, 'max_depth': 3, 'n_est...	0.50	0.16	22.75	6.29	14.47	3.88	10.89
94	CatBoost	none	{'depth': 4, 'iterations': 200, 'l2_leaf_reg':...	0.50	0.14	23.00	5.77	14.92	3.89	11.25
25	LightGBM	standard	{'learning_rate': 0.05, 'max_depth': 3, 'n_est...	0.49	0.13	23.28	5.72	14.87	3.39	11.52
81	LightGBM	log-shift	{'learning_rate': 0.1, 'max_depth': 3, 'n_esti...	0.49	0.17	23.02	6.12	14.08	3.68	10.67
95	LightGBM	none	{'learning_rate': 0.05, 'max_depth': 3, 'n_est...	0.49	0.14	23.31	5.82	14.93	3.49	10.21
80	CatBoost	log-shift	{'depth': 4, 'iterations': 200, 'l2_leaf_reg':...	0.47	0.17	23.67	6.46	15.27	4.89	11.61
7	KNN Regression	yeo-johnson	{'n_neighbors': 5, 'p': 1, 'weights': 'distance'}	0.46	0.19	23.73	6.22	15.45	5.39	0.16
21	KNN Regression	standard	{'n_neighbors': 5, 'p': 1, 'weights': 'distance'}	0.46	0.17	23.70	5.91	15.65	4.96	0.16
91	KNN Regression	none	{'n_neighbors': 5, 'p': 1, 'weights': 'distance'}	0.46	0.17	23.70	5.91	15.65	4.96	0.15
49	KNN Regression	sqrt	{'n_neighbors': 5, 'p': 1, 'weights': 'distance'}	0.45	0.19	23.73	6.23	15.45	5.40	0.15
20	Random Forest	standard	{'max_depth': None, 'min_samples_split': 2, 'n...	0.44	0.12	24.21	4.97	17.49	4.63	2.76

capacity of boosting-based algorithms to capture nonlinear interactions and learn complex relationships between components. In the datasets without transformation (none), the R^2 values remained low, averaging 0.03–0.05%. This demonstrates that transformation is a critical step in improving model performance, especially in tree-based methods. In contrast, linear (linear, ridge, lasso) models had $R^2 < 0.30$ and were insufficient to capture the multiple interactions involved in GO formation. Consequently, the performance rankings shown in Table 4 clearly demonstrate that the model-transformation interaction is decisive in terms of predictive power. In this study, transformation strategies improved model performance by an average of 3–5%; in particular, Yeo–Johnson and square root transformations significantly reduced the number of error metrics. The model performance ranking (leaderboard) presented in Table 5 is generated via a composite evaluation of the R^2, RMSE, and MAE metrics.

According to Table 5, these results show that the CatBoost-square root model exhibited the best overall performance in terms of the total score (Rank Sum = 4), followed by the LightGBM-Yeo–Johnson model (Rank Sum = 5). These two models showed balanced performance with low error (RMSE < 22.5, MAE < 14.5) and high predictive ability ($R^2 > 0.50$), making them the most suitable methods for predicting the GO content in foods.

Model-Transformation interaction and performance trends

The contribution of the transformation strategies was evaluated by taking the difference (Δ) of each transformation compared with the case. Table 6 summarizes the average performance changes (ΔR^2, ΔRMSE, and ΔMAE) by transformation type.

Yeo–Johnson and square root transformations provided performance improvements, with an average of $\Delta R^2 = +0.04$ and ΔRMSE = −0.53. This result is attributed to limiting the impact of outliers and stabilizing variance. In contrast, the log-shift transformation resulted in a decrease in performance ($\Delta R^2 = -0.02$, ΔRMSE = +0.61), indicating that excessive transformation can weaken model stability by distorting the data scale. Studies in the literature showing that data transformation methods can improve model success are consistent with the performance improvements observed after square root and Yeo–Johnson transformations (23, 34). In their study, (34) successfully completed the model's learning

Table 5. Model ranking on the basis of the highest accuracy (TopR2), lowest error (TopMAE) and overall performance score (leaderboard)

ID	Model	Transform_Type	Rank_R2	rank_RMSE	Rank_MAE	Rank_sum
52	CatBoost	sqrt	1	1	2	4
11	LightGBM	yeo-johnson	2	2	1	5
10	CatBoost	yeo-johnson	3	3	4	10
53	LightGBM	sqrt	4	4	5	13
24	CatBoost	standard	4	5	7	16
94	CatBoost	none	4	5	7	16
81	LightGBM	log-shift	7	7	3	17
25	LightGBM	standard	7	8	6	21
95	LightGBM	none	7	9	9	25
80	CatBoost	log-shift	10	10	10	30
21	KNN	standard	11	11	13	35
91	KNN	none	11	11	13	35
7	KNN	yeo-johnson	11	13	12	36
49	KNN	sqrt	14	14	11	39
20	Random Forest	standard	15	15	15	45

Table 6. Transformation effect summary (ΔR2, ΔRMSE, and ΔMAE)

Model	Transform	Baseline	ΔR^2	ΔRMSE	ΔMAE	Improved
CatBoost	sqrt	none	0.03	-0.76035	-0.89098	3
CatBoost	yeo-johnson	none	0.01	-0.36308	-0.55576	3
CatBoost	standard	none	0	0	0	0
CatBoost	none	none	0	0	0	0
CatBoost	log-shift	none	-0.03	0.668694	0.356728	0
KNN	yeo-johnson	none	0	0.034839	-0.20023	1
KNN	sqrt	none	-0.01	0.035468	-0.20207	1
KNN	standard	none	0	0	0	0
KNN	none	none	0	0	0	0
LightGBM	yeo-johnson	none	0.03	-0.87184	-1.29767	3
LightGBM	sqrt	none	0.01	-0.55945	-0.45717	3
LightGBM	log-shift	none	0	-0.29375	-0.84521	2
LightGBM	standard	none	0	-0.03421	-0.05331	2
LightGBM	none	none	0	0	0	0
Random Forest	standard	standard	0	0	0	0

process by stabilizing transformation methods, improving the model's learning process in skewed and non-normally distributed datasets. Another study reported that appropriate data preprocessing techniques resulted in measurable improvements in model performance in studies investigating ML modeling in food applications (35). The highest R^2 value found in this study, approximately 0.53, is within a suitable range compared with values reported for similar problem types after a comprehensive literature review. In many studies where ML modeling is used in food applications, the R^2 value is mostly reported to be in the range of 0.50--0.70 (36, 37). Another study reported that in models developed by applying sensor-based high-dimensional datasets or spectroscopic data, the R^2 value could exceed 0.90 (38, 39). The observed difference highlights how strongly the choice of data type can influence model performance. In this study, the compositional variables fats, proteins, carbohydrates, and sugars are only indirectly related to GO formation (40). Moreover, GO formation is inherently complex, arising from the interplay of several biochemical pathways, including lipid oxidation, the Maillard reaction, and sugar auto-oxidation (41). Because of this complexity, the literature generally regards it as challenging to achieve highly accurate predictions when relying solely on food composition data (42). In this context, the performance of our models is in line with that of previous studies and is actually an expected outcome, given the dataset's structure and size as well as the intricate nature of the target variable (36). After transformation, a more significant improvement in accuracy was observed for tree-based models such as CatBoost, LightGBM, and gradient boosting. This observation appears to be specific to the characteristics of our dataset and modeling approach. Transformation strategies also influence feature importance distributions. In the post-transformation CatBoost and LightGBM models, fat and carbohydrate variables consistently retained the highest importance coefficients in GO prediction. In general, when transformations were applied to the models, the average R^2 increased by 3–5%, whereas the RMSE and MAE values decreased by approximately 0.4–0.6 points. Although the highest R^2 value obtained in this study (0.53) indicates that the model performance is limited to a certain extent, this can be explained by the biochemical nature of the target variable and the structure of the dataset. GO, which is formed as a result of complex interactions of mechanisms such as the Maillard reaction, autoxidation of sugars, and lipid oxidation, is not a process dependent solely on basic food components (39, 40). Therefore, predicting GO formation with high accuracy via a limited

number of variables, such as fat, carbohydrate, protein, and sugar contents, presents a structural challenge. In addition, the small sample size (n= 87) of the dataset used in the study is another factor limiting the generalization ability and prediction success of ML models (36). The fact that the dataset was compiled directly through a literature review is also an important limitation. GO data obtained from different studies may show heterogeneity depending on processing conditions, variations in analytical methods, and matrix diversity. Although attempts have been made to standardize the methods used, the inability to completely eliminate these differences originating from the literature directly affects model performance (34). The performance data in question prove that there is a moderate but statistically significant relationship between food components and contaminant formation. This relationship also includes the indirect effects of variables such as heating temperature, processing time, and storage conditions, which are not always reported in the literature but play a critical role in the heterogeneous structure of the food matrix. These findings show that the developed model can be used as a practical screening tool that divides foods into "low" and "high" risk groups in terms of glyoxal content rather than as an absolute quantitative determination tool. Nevertheless, the greatest limitation of this study is the limited sample size of the literature. Small datasets increase the risk of overfitting in machine learning models, thus limiting generalizability (43, 44). The scarcity of studies in the literature that present detailed nutrient composition data along with glyoxal concentrations simultaneously has made it difficult to create a more comprehensive data pool. Therefore, the proposed "eGO" approach should not be considered an exact prediction system; it should be evaluated as a methodological starting point and a "proof-of-concept" study. Future research with larger, standardized and multivariate datasets will significantly improve the accuracy and reliability of the model in industrial applications (45).

Feature importance analysis

Figure 4 and Table 7 show the variable importance distributions over the four basic model–transformation combinations. As shown in Figure 4, the fat component has the highest predictability for GO levels in foods, regardless of the type of target conversion utilized. Furthermore, the literature indicates that lipid peroxidation, especially in food matrices with relatively high fat contents, can significantly contribute to the formation of GO and related carbonyl species.

The feature importance analysis results revealed that fat content was the most important determinant, and in this context, addressing the potential bias inherent in the dataset is crucial. Our dataset (n=87) consists of fully processed foods, primarily bakery products (biscuits, cookies, crackers) and instant coffee mixes, with an average fat content of 19.73 g/100 g. The absence of fried foods in the dataset eliminates any bias from deep-fat frying-induced fat absorption. This trend was confirmed in this study, as shown in Table 7. In all model–transformation combinations, fat content emerged as the most dominant determinant, with an average importance rate of 34%.

The carbohydrate, glucose, and sucrose variables presented lower but still significant predictive contributions. These results suggest that GO formation is associated not only with carbohydrate-induced processes but also with lipid oxidation-based mechanisms. These results are also compatible with the feature significance analyses obtained in the current study. The CatBoost and LightGBM models produce more balanced importance distributions under different transformation strategies, which increases the generalization performance by reducing model variance. Thus, both fat- and carbohydrate-based mechanisms could be modeled together in the prediction of GO, which is consistent with the literature explanations of the GO formation mechanism on the basis of lipid–sugar interactions.

Conclusion

This study evaluated 14 machine learning algorithms in combination with six target transformation strategies to determine eGO concentrations in processed packaged foods. Among the tested combinations, CatBoost with square-root transformation provided the highest accuracy (R^2 = 0.53, RMSE = 22.24, MAE = 14.03), whereas LightGBM with the Yeo–Johnson transformation produced the lowest

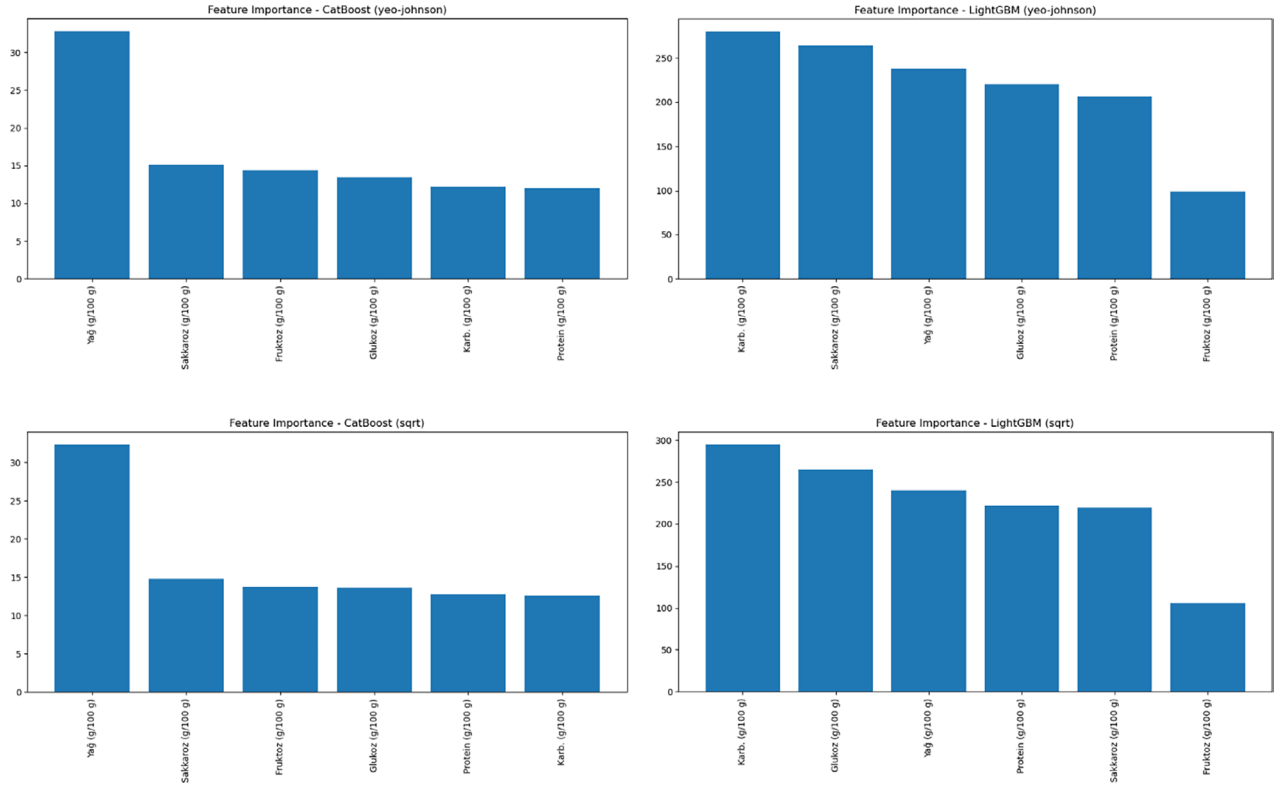

Figure 4. Feature importance distributions for GO estimation in foods under different target transformation strategies in the CatBoost and LightGBM regression models.

Table 7. Relative importance levels of nutritional components in GO estimation according to model-transformation combinations

Variable	Most Important in % of Models	Average Importance Score (estimated)
Fat	85–100%	0.34
Carbohydrate	70–80%	0.29
Glucose	60%	0.18
Protein	50%	0.12
Sucrose	45%	0.10
Fructose	40%	0.09

mean absolute error (MAE = 13.63). The linear models remained below R^2 = 0.30, indicating that GO formation cannot be adequately captured by linear assumptions when only nutritional composition variables are used. Target transformation clearly affected model performance. Yeo–Johnson and square-root transformations reduced skewness and outlier influence, improving R^2 by 3–5% on average, whereas the log-shift transformation increased prediction error in some models. These results indicate that transformation selection should be considered as part of the modeling process rather than as a default preprocessing step. In the feature importance analysis, fat content was the most influential variable across all the model–transformation combinations, followed by carbohydrates, glucose, sucrose, protein and fructose. This finding is consistent with the known contributions of both the lipid oxidation and Maillard reaction pathways

to GO formation. Future studies should expand the dataset to include a wider range of food categories and process-related variables (heating temperature, processing time, and storage conditions). The eGO framework can also be developed into a practical web- or mobile-based screening tool, providing the food industry with a rapid and low-cost preassessment option without the need for chromatographic analysis.

Acknowledgments

We would like to thank the Istanbul Technical University (ITU) Scientific Research Projects Unit. In this study, artificial intelligence tools were used only to a limited and supportive extent for language polishing to improve fluency, restructure some sentences, and ensure consistency of expression. AI tools were not involved in data generation, data analysis, or scientific decision-making. The authors take full responsibility for the accuracy of the data presented and for the final version of the manuscript.

Funding

This research was supported by the Istanbul Technical University Scientific Research Projects Unit with project numbers TYL-2025-47702 and TAB-2023-44987.

Ethics approval: This is a retrospective noninterventional study exempt from the requirement for ethics approval.

Conflict of interest: Each author declares that he or she has no commercial associations (e.g., consultancies, stock ownership, equity interests, patent/licensing, arrangement, etc.) that might pose a conflict of interest in connection with the submitted article.

Author contributions: YB: Conceptualization, methodology, data collection, analysis, manuscript writing. SÖ: Experimental design, data collection, analysis, manuscript writing. MHY: Machine learning methodology, data analysis, manuscript revision. BY: Project management, supervision, manuscript revision. All the authors contributed to the interpretation of the results and the final approval of the manuscript.

Declaration on the use of AI: None

References

1. Banerjee S. Role of food companies to supply nutritious foods as per buyers changing lifestyles, buying habits and the recent trends. International Journal of Innovative Research in Science, Engineering and Technology. 2020;9(3):1062-7. doi:10.15680/IJIRSET.2020.0903127
2. Raczuk E, Dmochowska B, Samaszko-Fiertek J, Madaj J. Different Schiff bases—structure, importance and classification. Molecules. 2022;27(3):787. doi: 10.3390/molecules27030787
3. Zhou X, Zhang Z, Liu X, et al. Typical reactive carbonyl compounds in food products: formation, influence on food quality, and detection methods. Comprehensive Reviews in Food Science and Food Safety. 2020;19(2):503-29. doi: 10.1111/1541-4337.12535
4. Khan MI, Ashfaq F, Alsayegh AA, et al. Advanced glycation end product signaling and metabolic complications: dietary approach. World J Diabetes. 2023;14:995-1012. doi: 10.4239/wjd.v14.i7.995
5. Treibmann S, Widdecke H, Henle T. Glycation reactions of methylglyoxal during digestion in a simulated gastrointestinal model. Molecules. 2024;29(9):2056. doi: 10.1002/fsn3.4118
6. Zhao M, Li Y, Bai X, et al. Inhibitory effect of guava leaf polyphenols on advanced glycation end products of frozen chicken meatballs (-18 °C) and its mechanism analysis. Foods. 2022;11:2509. doi: 10.3390/foods11162509
7. Menichetti G, Leclercq C, Collova C, et al. Machine learning prediction of the degree of food processing. Nature Communications. 2023;14:4562. doi: 10.1038/s41467-023-37457-1
8. El Hosry L, Elias V, Chamoun V, et al. Maillard reaction: mechanism, influencing parameters, advantages, disadvantages, and food industrial applications: a review. Foods. 2025;14:1881. doi: 10.3390/foods14111881
9. Wu X, Yan H, Cao Y, Yuan Y. Prediction acrylamide contents in fried dough twist based on the application of artificial neural network. Food Chemistry: X. 2024;24:102007. doi: 10.1016/j.fochx.2024.102007
10. Samuel HS, Etim EE, Nweke-Maraizu U, Yakubu S. Machine learning in chemical kinetics: predictions, mechanistic analysis, and reaction optimization. Applied Journal of Environmental Engineering Science. 2024;10(1):al-Appl. doi: 10.48422/IMIST.PRSM/ajees-v10i1.47284
11. Li L, Zhuang Y, Zou X, et al. Advanced glycation end products: detection and occurrence in food. Foods. 2023;12(11):2103. doi: 10.3390/foods12112103

12. Aria M, Cuccurullo C. bibliometrix: an R-tool for comprehensive science mapping analysis. Journal of Informetrics. 2017;11(4):959-75. doi: 10.1016/j.joi.2017.08.007
13. Donthu N, Kumar S, Mukherjee D, et al. How to conduct a bibliometric analysis: an overview and guidelines. Journal of Business Research. 2021;133:285-96. doi: 10.1016/j.jbusres.2021.04.070
14. Zhang W, Zhang Q, Yu B, Zhao L. Knowledge map of creativity research based on keywords network and coword analysis, 1992-2011. Quality & Quantity. 2015;49(3):1023-38. doi: 10.1007/s11135-014-0032-9
15. Rabbani N. AGEomics biomarkers and machine learning—realizing the potential of protein glycation in clinical diagnostics. International Journal Molecular Science. 2022;23(9):4584. doi: 10.3390/ijms23094584
16. Thornalley PJ. Dicarbonyl intermediates in the Maillard reaction. Ann N Y Acad Sci. 2005;1043:111-7. doi: 10.1196/annals.1333.014
17. Rabbani N, Thornalley PJ. Dicarbonyl stress in cell and tissue dysfunction contributing to aging and disease. Biochem Biophys Res Commun. 2015;458:221-6. doi: 10.1016/j.bbrc.2015.01.140
18. Singh R, Barden A, Mori T, Beilin L. Advanced glycation end-products: a review. Diabetologia. 2001;44:129-46. doi: 10.1007/s001250051591
19. Gokmen V, Senyuva HZ. Study of color and acrylamide formation in coffee, wheat flour and potato chips during heating. Food Chem. 2006;99:238-43. doi: 10.1016/j.foodchem.2005.06.054
20. Uribarri J, Woodruff S, Goodman S, et al. Advanced glycation end products in foods and a practical guide to their reduction in the diet. J Am Diet Assoc. 2010;110:911-6. doi: 10.1016/j.jada.2010.03.018
21. Martins SI, Jongen WM, van Boekel MA. A review of Maillard reaction in food and implications to kinetic modeling. Trends Food Sci Technol. 2000;11:364-73. doi: 10.1016/S0924-2244(01)00022-X
22. Niu L, Kong S, Chu F, et al. Investigation of AGEs, α-dicarbonyl compounds, and their correlations with chemical composition and salt levels in commercial fish products. Foods. 2023;12(23):4324. doi: 10.3390/foods12234324
23. Riani M, Atkinson AC, Corbellini A. Automatic robust Box–Cox and extended Yeo-Johnson transformations in regression. Statistical Methods and Applications. 2023;32(1):75-102. doi: 10.1007/s10260-022-00640-7
24. Hamasha MM, Ali H, Hamasha SD, Ahmed A. Ultrafine transformation of data for normality. Heliyon. 2022;8(5):e09370. doi: 10.1016/j.heliyon.2022.e09370
25. Hodson TO. Root mean square error (RMSE) or mean absolute error (MAE): when to use them or not. Geoscientific Model Development Discussions. 2022:1-10.
26. Onyango AN. Small reactive carbonyl compounds as tissue lipid oxidation products and the mechanisms of their formation thereby. Chemistry and Physics of Lipids. 2012;165(7):777-86. doi: 10.1016/j.chemphyslip.2012.09.004
27. Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. Geoscientific Model Development. 2014;7(3):1247-50. doi: 10.5194/gmd-7-1247-2014
28. Willmott CJ, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research. 2005;30(1):79-82. doi: 10.3354/cr030079
29. Chen X, Chen X, Liu Y, et al. Assessing aromatic-driven glyoxal formation over Eastern China. Remote Sensing. 2025;17(18):3174. doi: 10.3390/rs17183174
30. Jalali MW, Saidi B, Farahmand H, et al. Scalable AI-driven air quality forecasting for public health. Discover Atmosphere. 2025;3(1):25. doi: 10.1007/s44292-025-00052-8
31. Liu BH, Zhang LW, Wei YQ, Chen C. Dual power transformation and Yeo-Johnson techniques for reliability assessments. Buildings. 2024;14(11):3625. doi: 10.3390/buildings14113625
32. Brenning A. Interpreting machine-learning models in transformed feature space. Machine Learning. 2023;112(9):3455-71. doi: 10.1007/s10994-023-06327-8
33. Gyawali A. Utilization of remote sensing, machine learning, and agent-based simulation for biophysical assessment of young boreal forest. Helsinki: University of Helsinki; 2025.
34. Raymaekers J, Rousseeuw PJ. Transforming variables to central normality. Machine Learning. 2024;113(8):4953-75. doi: 10.1007/s10994-021-05960-5
35. Tseng YJ, Chuang PJ, Appell M. When machine learning and deep learning come to the big data in food chemistry. ACS Omega. 2023;8(18):15854-64. doi: 10.1021/acsomega.2c07722
36. Liakos KG, Athanasiadis V, Bozinou E, Lalas SI. Machine learning for quality control in the food industry. Foods. 2025;14(19):3424. doi: 10.3390/foods14193424
37. Wang L, Zhang F, Wang J, et al. Machine learning prediction of dual and dose–response effects of flavone carbon and oxygen glycosides on acrylamide formation. Frontiers in Nutrition. 2022;9:1042590. doi: 10.3389/fnut.2022.1042590
38. Mu F, Gu Y, Zhang J, Zhang L. Milk source identification and milk quality estimation using an electronic nose and machine learning techniques. Sensors. 2020;20(15):4238. doi: 10.3390/s20154238
39. Huang C, Gu Y. A machine learning method for the quantitative detection of adulterated meat using a MOS-based E-nose. Foods. 2022;11(4):602. doi: 10.3390/foods11040602
40. Inan-Eroğlu E, Ayaz A. Formation of AGEs in foods during cooking and underlying mechanisms. Nutrition Research Reviews. 2020;33(2):273-86. doi: 10.1017/S0954422419000209
41. Zhang M, Huang C, Ou J, et al. Glyoxal in foods: formation, metabolism, health hazards, and its control strategies. Journal of Agricultural and Food Chemistry. 2024;72(5):2434-50. doi: 10.1021/acs.jafc.3c08225
42. Naravane T, Tagkopoulos I. Machine learning models to predict micronutrient profile in food after processing. Current Research in Food Science. 2023;6:100500. doi: 10.1016/j.crfs.2023.100500

43. Hawkins DM. The problem of overfitting. J Chem Inf Comput Sci. 2004;44:1-12. doi: 10.1021/ci0342472
44. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med. 2004;66:411-21. doi: 10.1097/01.psy.0000127692.23278.a9
45. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLoS One. 2019;14:e0224365. doi: 10.1371/journal.pone.0224365