IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 787
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 11 2022

A Framework to Assess Added Sugar Levels in food nutrition using KNN and XG boost algorithm

M. Sangavi [1]

Assistant Professor, Computer Science and Engineering,

J.N.N Institute of Engineering, Chennai, Tamil Nadu, India

P. Priya [2]***, R. Sathish*** [3]***, A. Krishnaveni*** [4]***, R. Subraman i*** [5]

Assistant professor, Computer Science and Engineering,

J.N.N Institute of Engineering, Chennai, Tamil Nadu, India

Abstract:

Consuming additional sugar can exacerbate the development of obesity and diabetes, two conditions that are becoming more and more dangerous to the general public's health. While knowledge of nutritional content might significantly impact consumption patterns, not all nations now mandate the disclosure of added sugar. Nonetheless, an increasing number of people worldwide have access to portable devices, which makes it possible to develop technological tools to support actions and decisions linked to health. This article gathered detailed nutritional data, including added sugar content data, for 69,769 foods in order to investigate whether developments in computational science can be used to create a scalable and precise model to forecast the added sugar content of foods based on their nutrient profile. A gradient boosted tree model estimating added sugar concentration was trained using 80% of the data, with the remaining 20% kept out to evaluate the model's predictive power. The final model's performance revealed that 94.89% of the variation was explained for each default portion size. For each preset portion size, the mean absolute error of the estimate was 0.90 g. Because of this, this approach can be used to provide precise added sugar estimates via digital devices in nations where the data is not given on packaged foods, allowing customers to be informed of the added sugar amount of a wide range of meals.

Keyterms:diabetes; added sugar; informed decisions; obesity;

Introduction

Ultra-processed food, and especially its added sugar content, is thought to be a major cause of poor diet and nutrition, which can result in the development of diabetes and obesity [1-3]. The average daily intake of added sugar is more than 13 percent, with sugar-sweetened beverages being the main source [4] of this added sugar, despite the 2020–2025 dietary recommendations for Americans recommending a maximum of 10 percent of daily calories be taken with added sugar. Coffee,tea and sandwiches, candies, breakfast cereals and bars, higher-fat milk and yoghurt, desserts and savoury foods, and tea and coffee are additional prevalent forms of added sugar [4]. The availability of items with added sugar rose dramatically between 2000 and 2013, especially in the beverage industry [5-6]. For the first time in more over 20 years, the Nutrition Facts label was updated with this [4].

Although many nations have begun to label packaged goods with sugar information, very few of them currently list added sugar in nutrition [7]. In most parts of the world, people struggle to make an informed choice about which items to buy and/or consume because they lack access to information on

added sugars. But as more people gain admittance to cell phone, the possibility for digital solutions that assist in making health-related decisions increases [8].

It can be difficult to integrate previously published methods for estimating the added sugar content of meals into an automated process since they are frequently labor-intensive and require many, frequently manual procedures [9–11]. Recently, techniques based on fully automated supervised machine learning have proven effective in estimating nutritional dashboard. Most notably, [15] showed how a method employing the k nearest neighbours (kNN) algorithm may be effectively tested using a curated dataset that has category labels and thorough ingredient information. However, the kNN algorithm, which automatically predicts a variable's value from its closest neighbours in the feature space, is known to be susceptible to outliers in the raw data and fails to scale well because it must store every example used for training [12–15] and locate the nearest neighbours at prediction time. [16].

The XGboost algorithm, a more contemporary gradient boosted tree-based technique, has gained widespread recognition for its exceptional performance in ML. For instance, all 10 of the top submissions in the annual KDD Cup data modelling competition in 2015 included XGboost into their models [17]. Since XGboost is a tree-based approach, the dependent variable's value is predicted by building classification and regression trees (CART). A model that incorporates the results of numerous individual trees is produced at the gradient boosting stage by adding more trees in an effort to reduce the mistakes from earlier trees. Numerous optimisations have been made to the XGBoost algorithm to reduce the likelihood of overfitting the training set and increase scalability [17].

It also looked at whether clinically useful predictions could be made given the achievable prediction accuracy. This hypothesised that foods with considerable added sugar content might have altered macro- and micronutrient profiles compared to whole and/or unprocessed foods. Therefore, we assumed that other nutritional data, including protein and sugar, on the nutrition label might be utilised to approximate the amounts of added sugar. This also expected that the connection between additional sugar and other nutrients would be nonlinear and involve interactions. As a result, we predicted additional sugar quantities using the XGBoost method, which functions well in these conditions [17], and we contrasted the outcomes with a kNN approach that had been applied before [15].

Materials and Methods

First, we connected the Nutrient-related data from the Indian organization Food and Safety and Indian Health Organization, which have important divisions. It have given below figure 1 .the Nutrition Symbols table which contains more than 71,000 levels of sugar content which can be classified as important diabetes for indices such as alcohol where these measurements are important. The properties of these tests do not vary slightly due to the multidimensionality of the tests required for research. Because it involves sampling from 100% to 20%, the measurements are randomly selected depending on the errors and tables of the data. In addition, the measurements we can tell over time continuously show that sugars, carbohydrates, and other important nutrients such as fiber and minerals vary, and human metabolism depends on nutrient deficiencies.

ISSN PRINT 2319 1775 Online 2320 787
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 11 2022

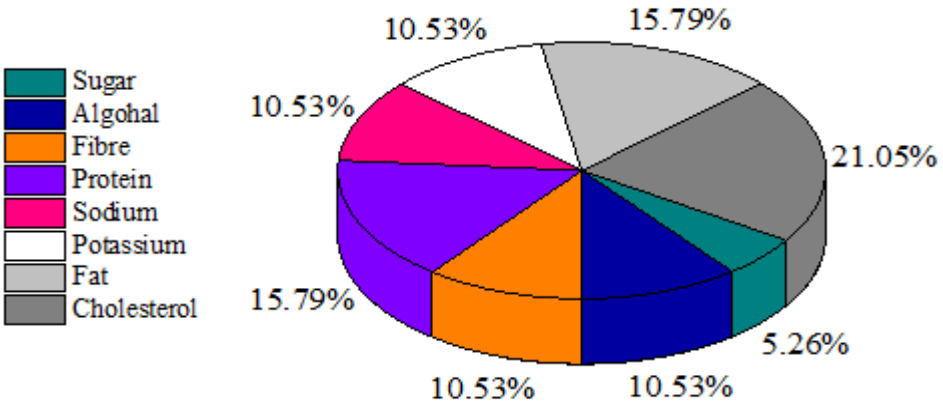

Figure 1 : nutrition chart

Research related to nutrition has been found in many literatures and thus this technique plays a very important role in our research by using it to examine the outputs of similar foods in XG boost method and evolve it through individual tree and firmly understand it and it can also be developed in Four-cast method. Because in nutrition generally this model has 20% data accuracy by using these models in reporting method and this technique is called in serve measurement. Using the serve technique in the Chut and assuming 100 calories plus deficit data for light training data while interpolating this method to achieve 0% sugar content in nutrition as per the Intraexon method, we report that the measurements do not vary by applying this method to this study. Because by using this method, we have also found out through this table how the nutrition measurements are done and we have collected the data of this report through the Indian Food Data Set and shown you the tab. This research objective is to apply this method as data.

Figure 2 Food tree with Kgal { Test data set in to 4.26g in SD =8.37 in the quantity of 4.88 (SD = 5.39) per 1000Mcal }

ISSN PRINT 2319 1775 Online 2320 787
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 11 2022

Also known as sugar-containing products, piper potassium is one of the key metabolic factors such as zinc. When ordering foods without sugar content from Hero Sugar, you can find the classifieds of sugar-free yield content available as a number of supplements through this bar chart pin.

Research happens more in all war cast and its range comes to 0 to 100% value if there is a sugar content then how much it varies depending on the accuracy and algorithm. Call it hyperparameter. The level at which absolute value to real value can be Predicted. By using this method, researchers can describe and extract the variance and percentage of the test data with the use of irresponsible questions and this article can describe and extract it in detail comparison in the result discussion .

Results

Predicting Added Sugar Accuracy with the kNN Algorithm

It was found that 90.6% of the variation as a default serving size, or 85.5% of the variance per 100 kcal, was explained by this model. The expected values for added sugar were 2.45 g per 100 kcal and 4.32 g per default portion size, with a mean absolute error of 4.80 g per default piece size and 4.3 g (SD = 4.89) per 100 kcal. Upon examination of Figure 1A more closely, it can be seen that the majority of errors were in underestimating the quantity of additional sugar was tilted towards bigger errors larger than 0 g. With a mean inaccuracy of 0.20 g, the "Beverages" category had the highest MAE per 200 kcal, as seen in Figure 2A.The mean errors in every other relevant category were negative, suggesting that the amounts of added sugar were understated.

The XGBoost Algorithm's Accuracy in Predicting Added Sugar

Our grid search revealed that the XGboost model was the most accurate in predicting the added sugar amounts of foods. It had the following characteristics: a learning rate of 0.006, a maximum tree depth of 19, a sampling ratio of 0.9 for columns, a minimal child weight of 4, a subsample proportion of 0.5 for rows, and 2000 trees. By modifying these hyperparameters, this article were able to predict the additional sugar levels in the test dataset using an XGBoost model that researchers had trained using our training data.

ISSN PRINT 2319 1775 Online 2320 787
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 11 2022

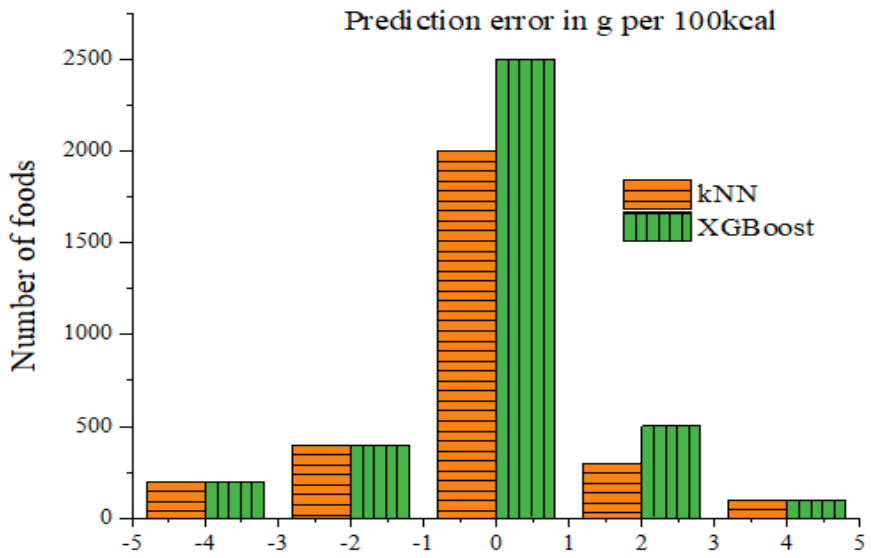

Figure 3. Added sugar quantities error calculated as a function of 100 kcal/g. KNN and XGBoost

Significant results were obtained from the rank correlation between the actual and predicted added sugar values. It was discovered that errors expressed in grammes per 100 kcal were rather symmetrically dispersed about zero, with a slight bias in favour of inflated the added sugar. Figure 3 kNN provides an example of this. As shown in Figure 3, XGBoost the "Bever-ages" group had the largest mean absolute inaccuracy, at 3.4 g per 200 kcal. Examining the error more closely shows that the drinks' added sugar content was usually exaggerated. This is particularly prevalent in beverages such as 100% fruit juices, which are high in sugar that exists naturally yet lack added sugar. In every other category of relevance, the mean error was negative, suggesting that the amount of added sugar in those groups was underestimated.

This refers to Table 1 for a summary of how much each feature's value affected the prediction. Figure 3A illustrates how the estimated amount of added sugar was linearly correlated with the food's overall sugar level. Still, other characteristics affected the prediction as well. As seen in Figure 3D, the prediction was affected by a food's protein composition for high total sugar values. Lower protein content in that instance suggested a higher added sugar load. Figure 3E demonstrates that, particularly in cases where the overall sugar level was high, reduced fibre content also increased the projections of added sugar.

ISSN PRINT 2319 1775 Online 2320 787
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 11 2022

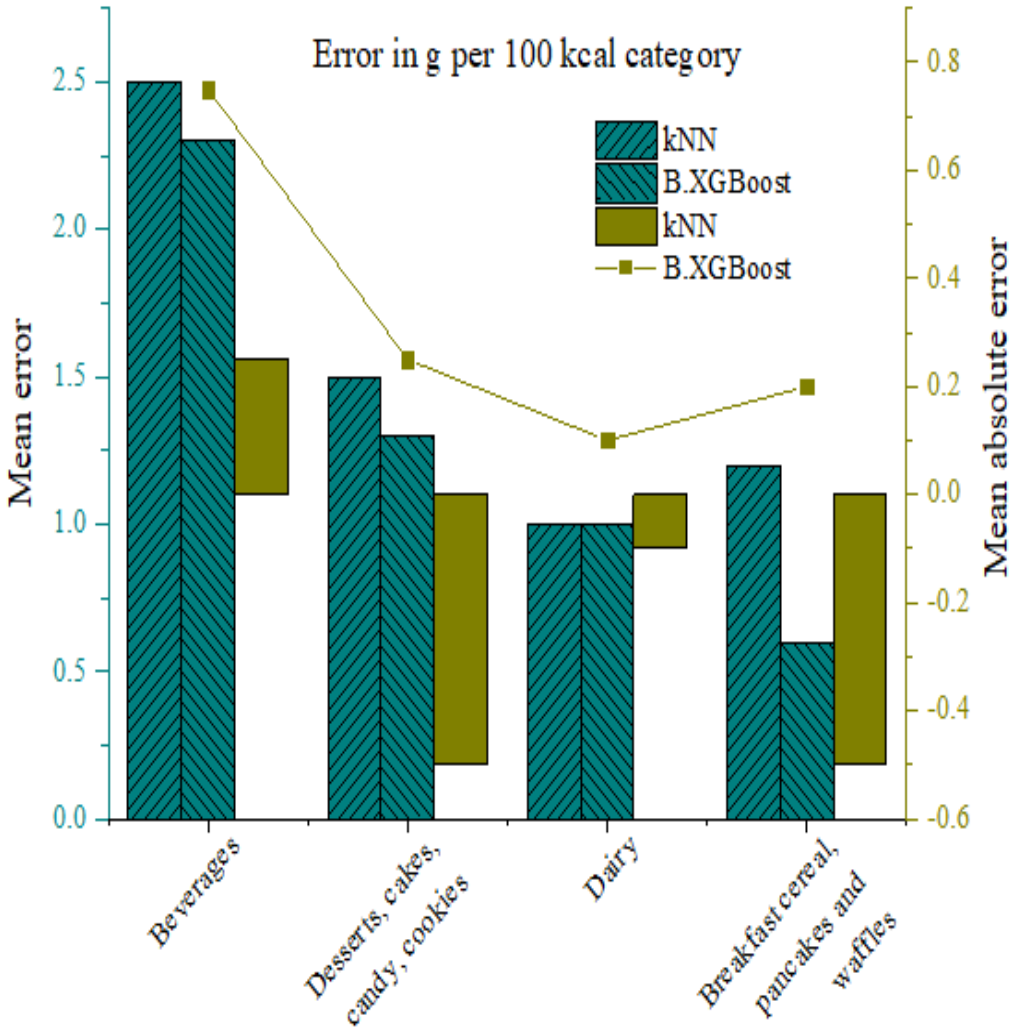

Figure 4. Mean absolute and mean error for (A) the kNN and (B) the XGBoost algorithms in the high-sugar categories when predicting additional sugar quantities in grammes per 100 kcal.

Table 1. Each feature's mean absolute average impact on the output magnitude of the XGBoost algorithm.

	Mean Absolute SHAP
Fiber	**0.56**
Sodium	**0.76**
Sugar	**4.45**

Discussion

Our hypothesis was that the XGBoost method, which has been developed recently in computational science, might be used to create a scalable and precise model to estimate the added sugar content of food items by analysing their nutritional profile. Furthermore, we anticipated that the XGBoost method will outperform earlier methods utilising the kNN algorithm [15] in terms of accuracy as well as scalability when predictions are to be produced based on dietary data that the user directly provides. In order to estimate the added sugar content of foods based on eight nutrients (carbohydrates, sugar, sugar alcohols and without fibre), protein, fibre, saturated fat, fat, calories, and sodium, we trained both a kNN and an XGBoost model. We used the added sugar amounts from the nutrition label for about seventy thousand typical foods from the U.S. section of the WW International Inc. food database. After that, 20% of the data were kept back to evaluate the models' accuracy, while the remaining 80% of the data were utilised to optimise and train the models. While all foods were used to train the models, only items with a sugar content of some were included in the test data. Compared to the previously published kNN model, which explained less than 86% of the variation per default piece size, the final XGBoost model explained over 93% of the variance in added sugar per default portion size on the test dataset, indicating that the XGBoost technique can provide a more accurate model.

Figure 5 The XGBoost model's output can be useful in guiding consumption decisions such that the recommended maximum added sugar intake of 10% is avoided.

There was a little bias towards overestimating additional sugar for items when error was seen in the XGBoost model, allowing for cautious estimation. A small overestimation of added sugar in the model is preferable over underestimating given that consumers frequently underestimate the calories and nutritional value of goods and that average daily added sugar consumption exceeds suggested thresholds. Researchers observed the largest inaccuracies for the assessment of added sugar content for beverages among the food categories that Americans consume the most added sugar from. Visual examination revealed that these mistakes seemed to be mostly caused by overestimating the amount of added sugar in fruit juices, which have a high natural sugar content. This is in line with the model's findings, which show that the best predictor of a food's added sugar content is its overall sugar value. In particular, the expected amount of added sugar is directly correlated with the food's overall sugar level. Low food protein and fibre levels are further predictors of increased added sugar content, especially when total sugar content is taken into account.

ISSN PRINT 2319 1775 Online 2320 787
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 11 2022

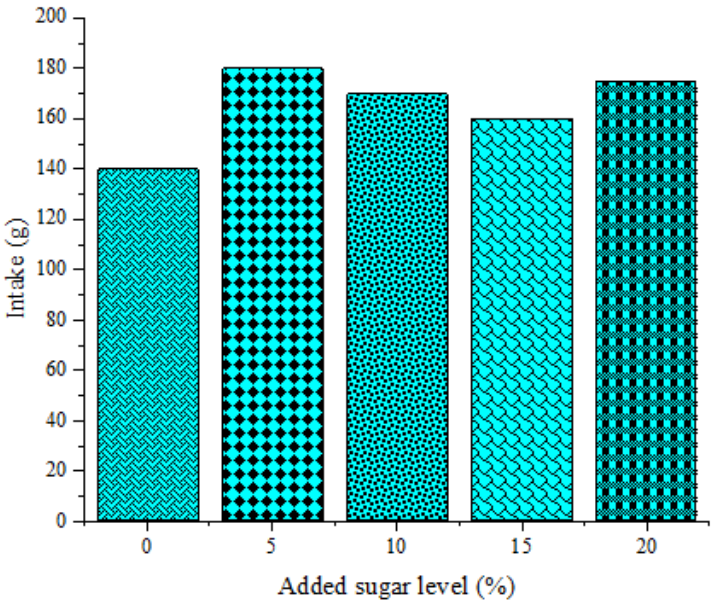

Fig 5.Impact of salt, fat and sugar levels on toddler food in take

The presented study includes numerous limitations in addition to a prediction error rate that was greater than zero. Although the general method described can be used with any dataset, in this case we only used the U.S. section of the WW International Inc. food database to train the model because, at the time the study was conducted, we could only obtain trustworthy information about the actual added sugar content of foods in the U.S. Nevertheless, it should be noted that any machine learning model trained on a particular dataset can only extract the information contained within that particular dataset. As such, it cannot be believed that trends seen in packaged foods sold in the United States will fully transfer to other nations without additional research.

The current recommended intake of fewer than ten percent of daily calories is not being met by the average consumption of added sugar. Most people on the planet do not have access to added sugar because added sugar is only listed in a small number of countries' Nutrition Facts labels. Furthermore, it was shown that including direct words and visuals with the sugar content information in grammes was more beneficial than just including the Nutrition Facts label on its own. The model predictions offered here can be used to develop precise and useful information using a mobile application that helps educate users about added sugar content. As a result, it illustrates a potential use case for data and predictive modelling in digital solutions that guide lifestyle decisions that can help prevent chronic noncommunicable diseases.

Conclusions

The findings reported here are noteworthy, which is helpful considering that current research indicates that detailed knowledge of food composition can effectively influence eating patterns. This highlights

the possibility for developing digital solutions, such as an app that allows users to scan traditional nutrition labels to determine an item's estimated added sugar level and share that information with others to help them make healthy decisions, or a website that estimates the additive content of food. Without these tools, customers are left to judge a food's healthiness based only on their own judgement or the nutrition claims made on the product's packaging in nations where the added sugar level is not disclosed on the nutrition label. This is significant because customer perceptions can be greatly impacted by how well different sugar label formats work. For instance, a systematic review that looked at how nutrition claims about sugar content affected dietary choices and energy consumption revealed that customers who are concerned about their health could make the mistake of believing that an item is healthier just because it says "reduced-sugar" on the label.

Reference

1. Davies, T., Louie, J. C. Y., Ndanuko, R., Barbieri, S., Perez-Concha, O., & Wu, J. H. (2022). A machine learning approach to predict the added-sugar content of packaged foods. *The Journal of Nutrition*, *152*(1), 343-349.
2. Peters, S. A., Dunford, E., Jones, A., Ni Mhurchu, C., Crino, M., Taylor, F., ... & Neal, B. (2017). Incorporating added sugar improves the performance of the health star rating front-of-pack labelling system in Australia. *Nutrients*, *9*(7), 701.
3. Menday, H., Neal, B., Wu, J. H., Crino, M., Baines, S., & Petersen, K. S. (2017). Use of added sugars instead of total sugars may improve the capacity of the health star rating system to discriminate between core and discretionary foods. *Journal of the Academy of Nutrition and Dietetics*, *117*(12), 1921-1930.
4. Daniel-Weiner, R., Cardel, M. I., Skarlinski, M., Goscilo, A., Anderson, C., & Foster, G. D. (2023). Enabling Informed Decision Making in the Absence of Detailed Nutrition Labels: A Model to Estimate the Added Sugar Content of Foods. *Nutrients*, *15*(4), 803.
5. Bergenstal, R. M., Johnson, M., Powers, M. A., Wynne, A., Vlajnic, A., Hollander, P., & Rendell, M. (2008). Adjust to target in type 2 diabetes: comparison of a simple algorithm with carbohydrate counting for adjustment of mealtime insulin glulisine. *Diabetes care*, *31*(7), 1305-1310.
6. Fulgoni III, V. L., Keast, D. R., & Drewnowski, A. (2009). Development and validation of the nutrient-rich foods index: a tool to measure nutritional quality of foods. *The Journal of nutrition*, *139*(8), 1549-1554.
7. Bell, K. J., Petocz, P., Colagiuri, S., & Brand-Miller, J. C. (2016). Algorithms to improve the prediction of postprandial insulinaemia in response to common foods. *Nutrients*, *8*(4), 210.
8. Cediel, G., Reyes, M., da Costa Louzada, M. L., Steele, E. M., Monteiro, C. A., Corvalán, C., & Uauy, R. (2018). Ultra-processed foods and added sugars in the Chilean diet (2010). *Public health nutrition*, *21*(1), 125-133.
9. Darmon, N., Sondey, J., Azaïs-Braesco, V., & Maillot, M. (2018). The SENS algorithm—a new nutrient profiling system for food labelling in Europe. *European Journal of Clinical Nutrition*, *72*(2), 236-248.
10. Del Favero, S., Place, J., Kropff, J., Messori, M., Keith- Hynes, P., Visentin, R., ... & Ap@

home Consortium. (2015). Multicenter outpatient dinner/overnight reduction of hypoglycemia and increased time of glucose in target with a wearable artificial pancreas using modular model predictive control in adults with type 1 diabetes. *Diabetes, Obesity and Metabolism*, *17*(5), 468-476.

11. Zhang, Y., Krueger, D., Durst, R., Lee, R., Wang, D., Seeram, N., & Heber, D. (2009). International multidimensional authenticity specification (IMAS) algorithm for detection of commercial pomegranate juice adulteration. *Journal of agricultural and food chemistry*, *57*(6), 2550-2557.

12. Riesenberg, Devorah, Anna Peeters, Kathryn Backholer, Jane Martin, Cliona Ni Mhurchu, and Miranda R. Blake. (2022). "Exploring the effects of added sugar labels on food purchasing behaviour in Australian parents: An online randomised controlled trial." *PloS one* 17, no. 8: e0271435.

13. Parker, Robert S., Francis J. Doyle, and Nicholas A. Peppas. (1999). "A model-based algorithm for blood glucose control in type I diabetic patients." *IEEE Transactions on biomedical engineering* 46, no. 2: 148-157.

14. He, Hong-Ju, Yangyang Wang, Mian Zhang, Yuling Wang, Xingqi Ou, and Jingli Guo. (2022). "Rapid determination of reducing sugar content in sweet potatoes using NIR spectra." *Journal of Food Composition and Analysis* 111: 104641.

15. Cai, Yunying, Mengge Li, Lun Zhang, Jie Zhang, and Heng Su. (2023). "The effect of the modified fat-protein unit algorithm compared with that of carbohydrate counting on postprandial glucose in adults with type-1 diabetes when consuming meals with differing macronutrient compositions: a randomized crossover trial." *Nutrition & Metabolism* 20, no. 1: 43.

16. Garber, Alan J., Martin J. Abrahamson, Joshua I. Barzilay, Lawrence Blonde, Zachary T. Bloomgarden, Michael A. Bush, Samuel Dagogo-Jack et al. (2019). "Consensus statement by the American Association of Clinical Endocrinologists and American College of Endocrinology on the comprehensive type 2 diabetes management algorithm–2019 executive summary." *Endocrine Practice* 25, no. 1: 69-101.

17. Hovorka, Roman. (2006). "Continuous glucose monitoring and closed- loop systems." *Diabetic medicine* 23, no. 1: 1-12.