ISSN 2063-5346

MACHINE LEARNING-BASED APPROACH FOR ENERGY CONTENT PREDICTION IN PACKAGED FOODS

Saureng Kumar[1*], S. C. Sharma[2]

Article History: Received: 01.02.2023 **Revised:** 07.03.2023 **Accepted:** 10.04.2023

Abstract

Energy is inextricably tied to the health condition. Various countries do not require the labelling of nutrition information on packaged food products. This lack of information leaves consumers and policymakers unaware of the energy content (low energy content, high energy content) in these products. To address this issue, we have created a machine learning-based approach that employs nutrition facts labels to estimate the energy content of packaged food products. We obtained 204 samples of nutrition information, of which (n=152) samples were utilized for the training dataset (and n=52) samples were utilized for testing. This approach enables complete traceability also enhance the clarity and precision of food product labeling. Utilizing of various machine learning algorithm (SVM, KNN, RF, DT, MLR) and measure the performance KNN performed the highest accuracy of 97.06%. our research emphasizes the potential of applying machine learning to effectively forecast the energy value of packaged food in a large scale.

Keywords: Energy content, Food product, Machine Learning, Prediction, Packaging

[1,2]Electronics and computer discipline, IIT Roorkee, Saharanpur Campus,Saharanpur, India.

Corresponding E-mail address: skumar@pp.iitr.ac.in

DOI: 10.31838/ecb/2023.12.s1.101

1. Introduction

Energy content is a crucial aspect of packaged food products that is closely linked to their nutritional value. The energy content of food is measured in calories or joules and represents the amount of energy that a food item can provide to the body when consumed. Accurately estimating the energy content of packaged food items is essential for ensuring that consumers are aware of their nutritional value and can make informed decisions about their dietary choices[1]Moreover, food manufacturers must provide accurate information about the energy content of their products to comply with regulatory requirements and to maintain consumer trust.[2] Recent advances in machine learning and other data-driven techniques have opened up exciting possibilities for accurately predicting the energy content of packaged food items.[1]. According to Dunford et al. (2014), monitoring packaged foods for food safety and using a machine learning algorithm to look at the relationship between the composition of macronutrients and the amount of energy they contain is crucial. However, the accuracy of these models can be influenced by various factors such as the quality and quantity [3] of data used for training, the selection of features, and the choice of algorithm. Furthermore, the accuracy of predictions can vary depending on the type of packaged food, storage condition [4] and the methods used to measure energy content.

One possible application of this technique is in food labelling and regulatory compliance. Precise estimation of energy content can ensure that packaged foods are correctly labelled and that their nutritional information is reliable and consistent.

Packaged food energy prediction through machine learning is an essential research area with the potential for many applications in the food packaging industry[5] and public health[6]. However, there is a need for further research to develop more accurate and robust models that can account for the complex factors that can affect the energy content of packaged foods.

Our contributions of this research work are as follows:

- To address this challenge, an innovative machine learning approach has been developed to predict the energy content of packaged foods accurately. The approach is based on a supervised learning algorithm that leverages a vast dataset of nutritional information on various packaged foods to train a predictive model.

- The model uses a range of inputs, including the food's ingredient list, nutritional information, and packaging details, to predict the energy content accurately. The algorithm considers various factors that can impact the food's energy content, such as cooking methods, serving size, and storage conditions.

- To validate the effectiveness of this machine learning approach, a comprehensive evaluation was conducted, comparing the predictions of the algorithm with actual measurements of energy content. The results show that the machine learning approach can predict the energy content of packaged foods with high accuracy, outperforming traditional methods.

- Perform high accuracy and reliable result.

The remainder of the paper is arranged as follows, section 1 gives a brief introduction, including the research question and contribution of the research work. In section 2, we have discussed the research design, methodology, data collection, and analysis. Our research findings, data visualization, and result discussion, including performance comparison of different ML algorithm, has discussed in section 3. In the next section we conclude the research work and suggested the future research direction of this study.

2. Materials and methods

2.1 Dataset

The George Institute's FoodSwitch[6] developed a smartphone application called "FoodSwitch India" which allows users to scan barcodes of food products and get information about their nutritional content. We have downloaded the application from the google play store and This FoodSwitch app utilizes your mobile phone's camera to scan the barcode of packaged food items, and then uses science-based algorithms to analyze and display essential nutritional details of the product. By scanning the barcode of a packaged food item, the information and calculations stored in our database are retrieved, and then presented in a clear and simple manner as either a Health Star Rating (HSR) [7]or traffic-light colored icons[8] for crucial nutrients and energy. Through this we have generated the dataset for our research work a total of 205 dataset has been generated with the different product group (Chicken breast, Beans, Peanut butter, Greek yogurt, Quinoa, Avocado, Nuts, Eggs, Whole grain bread, Milk) through this application.

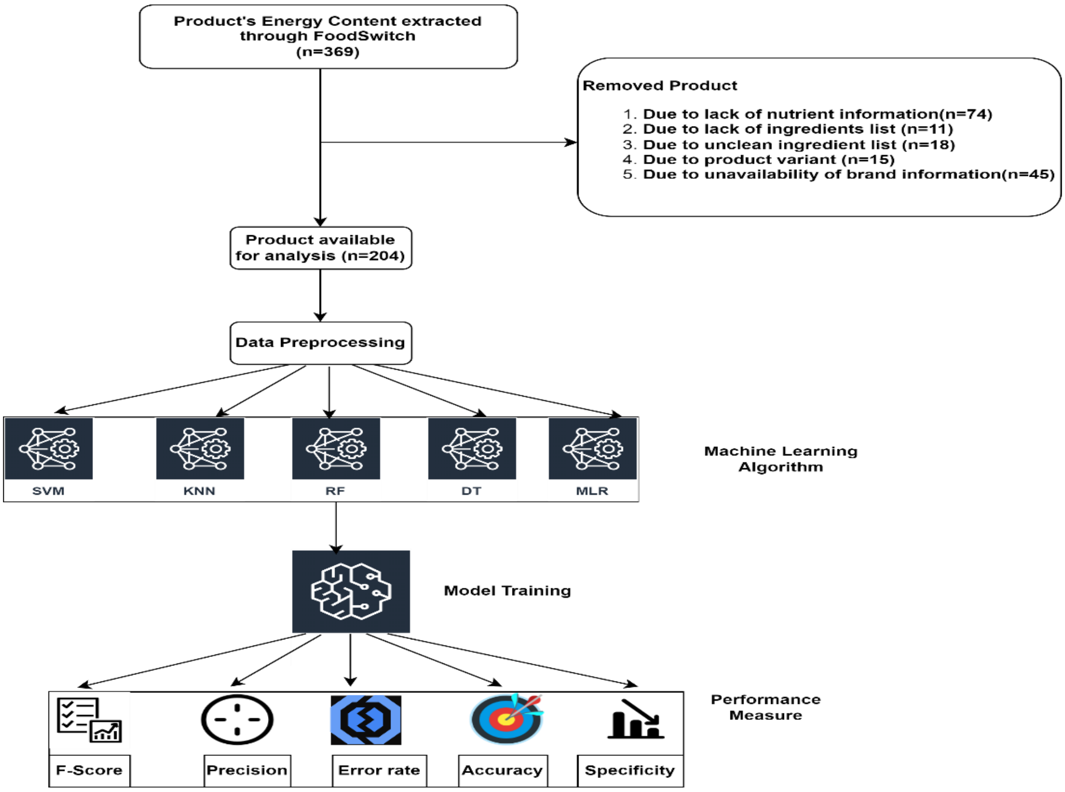
Product's Energy Content extracted through FoodSwitch (n=369)
Removed Product
1. Due to lack of nutrient information(n=74)
2. Due to lack of ingredients list (n=11)
3. Due to unclean ingredient list (n=18)
4. Due to product variant (n=15)
5. Due to unavailability of brand information(n=45)
Product available for analysis (n=204)
Data Preprocessing
SVM
KNN
RF
DT
MLR
Machine Learning Algorithm
Model Training
F-Score
Precision
Error rate
Accuracy
Specificity
Performance Measure

Figure 1: Screening criteria and machine learning model

2.2 Screening Criteria and machine learning model

We collected a total of 369 unique barcode-assigned foods from FoodSwitch between January 2023 and March 2023. We removed products that lacked a nutrition information panel (n = 74), an ingredients list (n = 11), or had an unclean ingredients list (n = 18) (e.g., unequal number of opening and closing parentheses). We also excluded product variants (n = 15) and those without brand information (n = 45).

Additionally, we eliminated food product categories that contributed less than 1% of carbohydrate (e.g., Chicken breast), or had less than 1% of products reporting fat content (e.g., Greek yogurt) (n = 60) (Supplementary Table1). After applying these exclusions, we obtained a sample of 204 products from 10 different food and beverage categories for our analysis.

Energy Calculation:

We utilized three nutrient features from the nutrition information panel [9]to represent all products. These features were protein value (g), fat value (g) and carbohydrate (g), To avoid multicollinearity[10] in the algorithm, we ensured that all chosen nutrients were mutually exclusive. Energy content was calculated by

Energy content (in kilocalories or Calories) = (grams of protein x 4) + (grams of carbohydrate x 4) + (grams of fat x 9)

Note that the energy content is expressed in kilocalories or Calories, and the values for protein, carbohydrate, and fat are expressed in grams.

To enable the algorithm to consider nutrients of different ranges equally, we perform data normalization[11] to all the nutrients by dividing them by the minimum and maximum values in the training dataset.

$$i.e;\, Xo = \frac{X - Xmin}{Xmax - Xmin}$$

Table 1: Description of energy content in packaged food

Packaged Food	Carbohydrates (g)	Protein (g)	Fat (g)
Chicken breast	0	31	3.6
Beans	22	15	1
Peanut butter	6	7	16
Greek yogurt	6	17	0.2
Quinoa	21	4	2
Avocado	9	2	15
Nuts	6	7	14
Eggs	1	6	5
Whole grain bread	12	3	1.5
Milk	12	8	8

3. Research and analysis for the proposed model

We employ the following strategies in the context of this research. In order to remove missing values and outliers that could increase data complexity in the dataset, we employed a data prepossessing strategy[12] for cleaning the data. First, we fed the data to different machine learning algorithm [3]ie; Support vector machine (SVM), k-nearest neighbors' algorithm (KNN), Random Forest (RF), Decision Tree (DT), Multiple linear regression (MLR). Second, in order to increase the quality of the data, it may be necessary to remove features that are redundant or unimportant. This can be especially helpful in instances where the selection of the machine learning model was influenced by deeper factors. The best model must be chosen from a group of candidate models in order to provide the highest performance on unobserved data, prevent overfitting[13], enhance interpretability, and guarantee scalability. During the training process, the machine learning model develops the ability to recognize patterns and relationships in the input data. The model refines its parameters depending on the input data over numerous iterations on a sizable dataset until it can make precise prediction and measure the performance in terms of the following parameter:

1. F-Score: It indicate the harmonized mean of precision and recall. When we have an unbalanced dataset with considerably varied numbers of cases for each class, the F-score is quite helpful. In certain situations, accuracy might not be the appropriate statistic to assess the model because it might give a false sense of how well it performs. Instead, we make sure the model works effectively for both positive and negative classes by using the F-score.

$$\text{F-Score} = 2\times\frac{Percision\times \mathrm{Re}\,call}{Percision+\mathrm{Re}\,call}$$

2. Precision: Precision is defined as the ratio of true positives (positives that were successfully identified) to all expected positives. A high precision suggests that the model rarely generates incorrect positive predictions, that means true positive predictions are accurate.

$$\mathrm{Pr}\,ecision = \frac{TruePositive}{TruePositive+FalsePositive}$$

3. Error rate: Error rate is the measurement of the model's prediction error.

$$ErrorRate = \frac{FalsePositive+FalseNegative}{Positive+Negative}$$

4. Accuracy: Accuracy treats all error types (false positives and false negatives) equally. However, equality is not always preferred. The following formula can calculate the accuracy of the model. The accuracy will be biased in favor of the bigger class if there are much more positive samples than negative ones.

$$Accuracy = \frac{TruePositive+TrueNegative}{FalsePositive+FalseNegative+TrueNegative+TruePositive}$$

5. Specificity: The model's specificity refers to the ability of a model to identify true negatives (TN) correctly. The following formula can be used to calculate the specificity.

$$Specificity = \frac{TrueNegative}{TrueNegative + FalsePositive}$$

4. Result and Discussion

In this work, a machine learning system was created to forecast the energy content of particular foods based on their protein, carbohydrate, and fat contents. The dataset included 204 packaged food items. that were split into two sets: a training set of 160 items and a test set of 44 items. Then we perform the size distribution according to their energy content as shown in Fig. 2. To visualize the data pattern, we perform a heat map to identify the outliers and anomalies in the dataset as illustrated in Fig3. The data highlighted by the lighter and darker color, where darker colors represent higher value and the lighter color represent the lower value. It can also identify the feature selection and model-building process. Then we extract the feature by performing the principal component analysis (PCA)[14]. It transforms data space from higher dimension to the lower dimension. The visualization illustrated in Fig.4

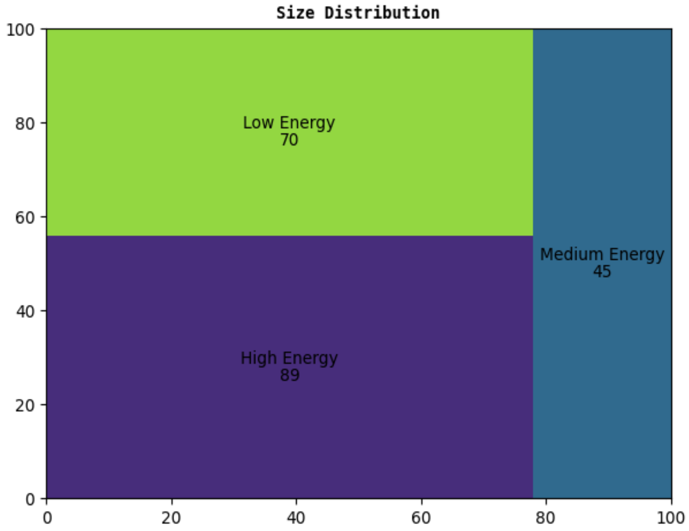

Figure.2. Size distribution of packaged food

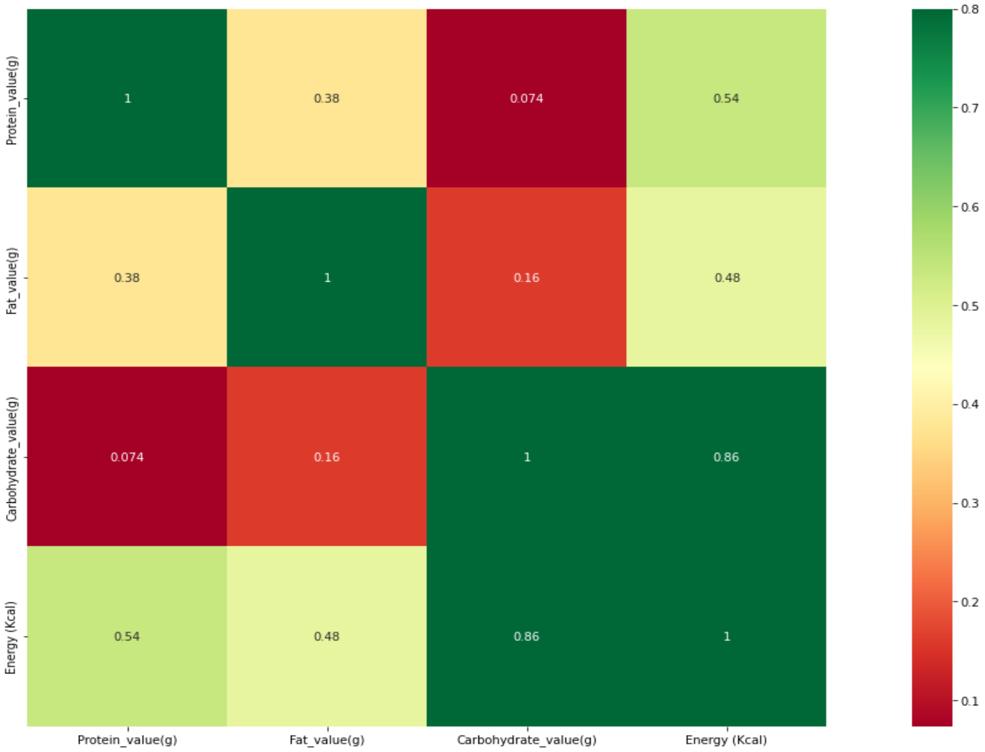

Figure.3. Heatmap for the packaged food dataset

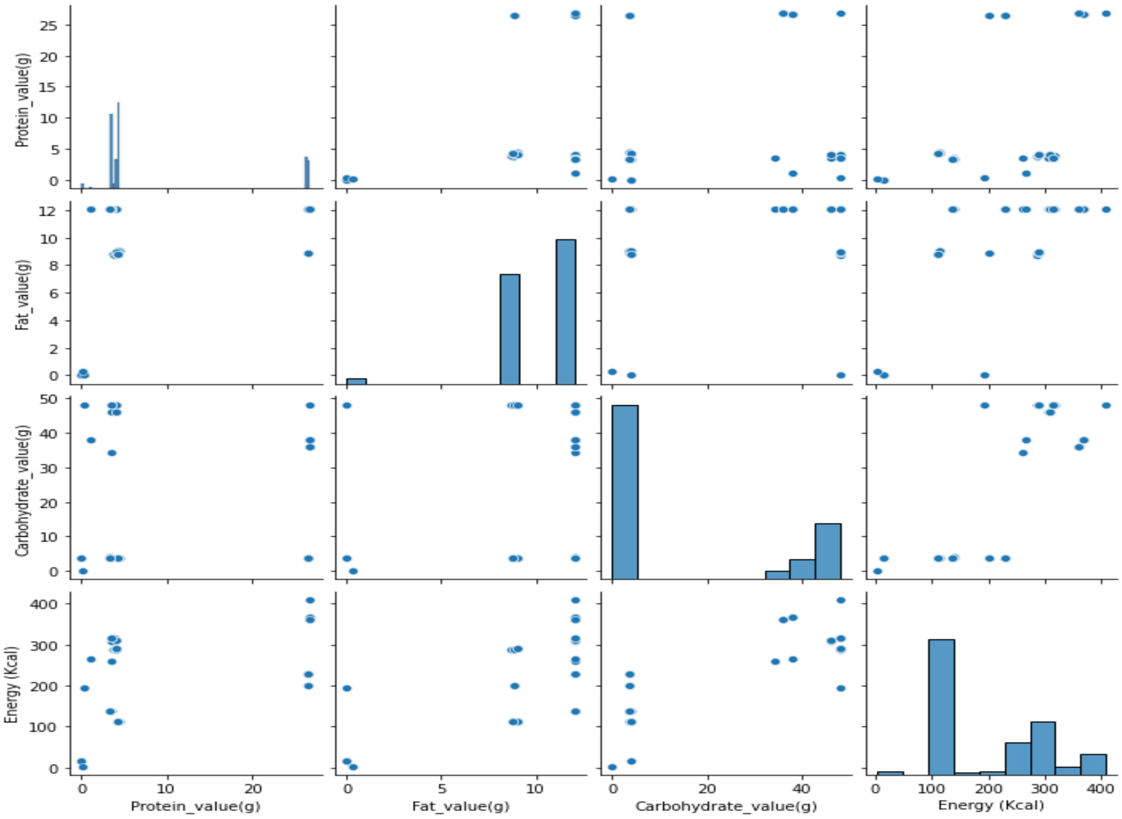

Figure.4. Principal component analysis- visualization of high-dimensional data in a lower-dimensional space.

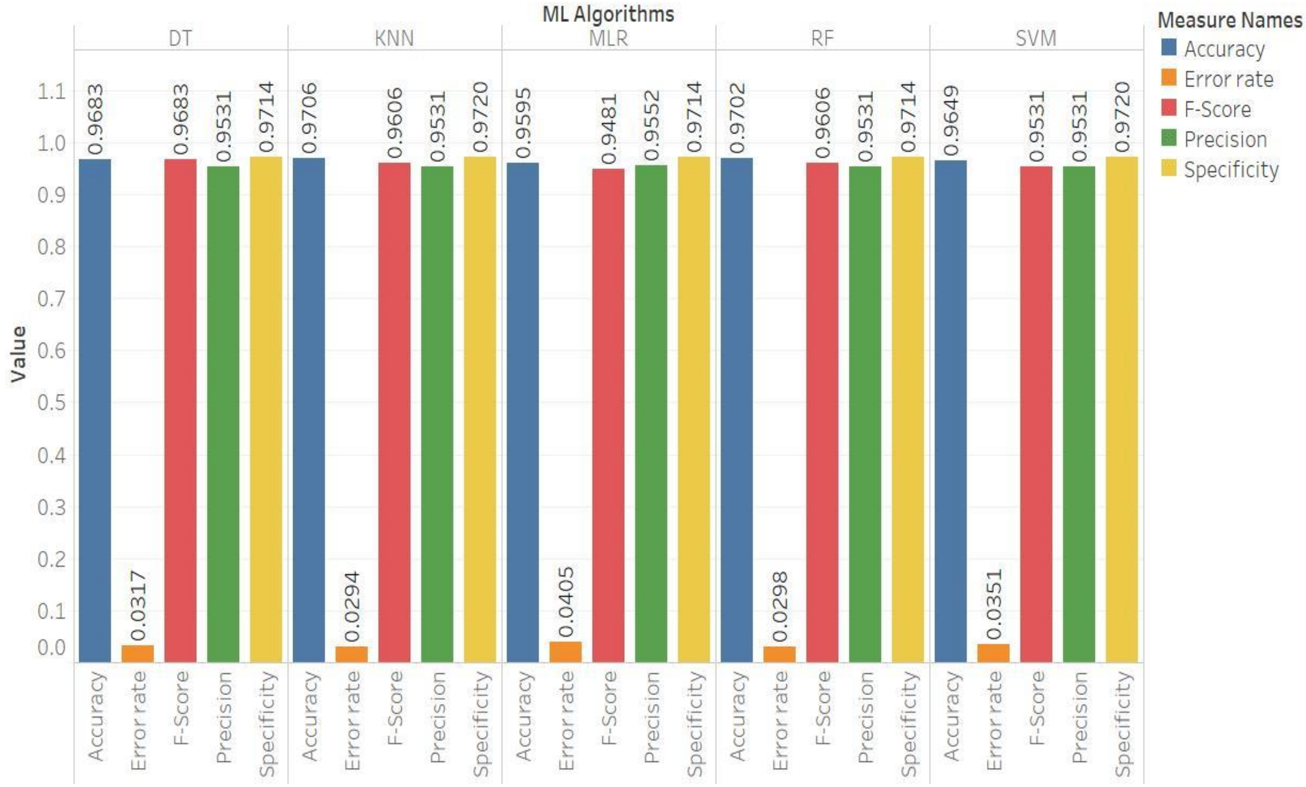

Figure.5. Performance measure of various machine learning algorithms over original data

5. Conclusion

This study successfully developed a machine learning algorithm for predicting the energy content based on the composition of the food's macronutrients. The results showed that the KNN (k-nearest neighbors) algorithm outperformed with the highest accuracy of 97.06% over the support vector machine, random forest multiple linear regression and decision tree algorithms and the three-evaluation metrics, mean absolute error (MAE), mean square error (MSE), and root mean squared error (RMSE) are 7.22733,1.05568,1.02746 respectively which indicates that the model is better at predicting the actual values. The study highlights the potential of machine learning algorithms in improving energy prediction and promoting public health. However, the algorithm's performance can be increased by the availability of large data sets. In addition, considering the more components of food that provide energy

Statements and Declarations

Disclosure of potential conflicts of interest

Funding

The authors declare that no funds, grants, or other support were received during thepreparation of this manuscript.

Conflict of interest

The authors have no relevant financial or nonfinancial interests to disclose.

Data availability

The datasets generated during and/or analyzed during the current study are available fromthe corresponding author on reasonable request.

References

[1] T. Davies *et al.*, “An Innovative Machine Learning Approach to Predict the Dietary Fiber Content of Packaged Foods,” *Nutrients*, vol. 13, no. 9, p. 3195, Sep. 2021, doi: 10.3390/nu13093195.

[2] X. Deng, S. Cao, and A. L. Horn, “Emerging Applications of Machine Learning in Food Safety,” *Annu. Rev. Food Sci. Technol.*, vol. 12, no. 1, pp. 513–538, Mar. 2021, doi: 10.1146/annurev-food-071720-024112.

[3] Z. Shen, A. Shehzad, S. Chen, H. Sun, and J. Liu, “Machine Learning Based Approach on Food Recognition and Nutrition Estimation,” *Procedia Computer Science*, vol. 174, pp. 448–453, 2020, doi: 10.1016/j.procs.2020.06.113.

[4] S. Kumar and S. C. Sharma, “Two-Level Priority Task Scheduling Algorithm for Real-Time IoT Based Storage Condition Assessment System,” in *Rising Threats in Expert Applications and Solutions*, V. S. Rathore, S. C. Sharma, J. M. R. S. Tavares, C. Moreira, and B. Surendiran, Eds., in Lecture Notes in Networks and Systems, vol. 434. Singapore: Springer Nature Singapore, 2022, pp. 147–158. doi: 10.1007/978-981-19-1122-4_17.

[5] W. Zhu *et al.*, “Development of organic-inorganic hybrid antimicrobial materials by mechanical force and application for active packaging,” *Food Packaging and Shelf Life*, vol. 37, p. 101060, Jun. 2023, doi: 10.1016/j.fpsl.2023.101060.

[6] E. Dunford *et al.*, “FoodSwitch: A Mobile Phone App to Enable Consumers to Make Healthier Food Choices and Crowdsourcing of National Food Composition Data,” *JMIR mHealth uHealth*, vol. 2, no. 3, p. e37, Aug. 2014, doi: 10.2196/mhealth.3230.

[7] A. Jones, A. M. Thow, C. Ni Mhurchu, G. Sacks, and B. Neal, “The performance and potential of the Australasian Health Star Rating system: a four-year review using the RE-AIM framework,” *Australian and New Zealand Journal of Public Health*, vol. 43, no. 4, pp. 355–365, Aug. 2019, doi: 10.1111/1753-6405.12908.

[8] S. Kunz, S. Haasova, J. Rieß, and A. Florack, “Beyond Healthiness: The Impact of Traffic Light Labels on Taste Expectations and Purchase Intentions,” *Foods*, vol. 9, no. 2, p. 134, Jan. 2020, doi: 10.3390/foods9020134.

[9] Evelyn Lau, Hui Jen Goh, Rina Quek, Siang Wee Lim, and Jeyakumar Henry, “Rapid estimation of the energy content of composite foods: the application of the Calorie Answer^(TM),” *Asia Pacific Journal of Clinical Nutrition*, vol. 25, no. 1, Jan. 2016, doi: 10.6133/apjcn.2016.25.1.14.

[10] J. Y.-L. Chan *et al.*, “Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review,” *Mathematics*, vol. 10, no. 8, p. 1283, Apr. 2022, doi: 10.3390/math10081283.

[11] D. Singh and B. Singh, “Investigating the impact of data normalization on classification performance,” *Applied Soft Computing*, vol. 97, p. 105524, Dec. 2020, doi: 10.1016/j.asoc.2019.105524.

[12] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, “A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data,” *Front. Energy Res.*, vol. 9, p. 652801, Mar. 2021, doi: 10.3389/fenrg.2021.652801.

[13] X. Ying, “An Overview of Overfitting and its Solutions,” *J. Phys.: Conf. Ser.*, vol. 1168, p. 022022, Feb. 2019, doi: 10.1088/1742-6596/1168/2/022022.

[14] T. Howley, M. G. Madden, M.-L. O’Connell, and A. G. Ryder, “The Effect of Principal Component Analysis on Machine Learning Accuracy with High Dimensional Spectral Data,” in *Applications and Innovations in Intelligent Systems XIII*, A. Macintosh, R. Ellis, and T. Allen, Eds., London: Springer London, 2006, pp. 209–222. doi: 10.1007/1-84628-224-1_16.