Machine Learning Approaches for Accurate Energy Content Prediction in Foods Using Nutritional Data

Nishan Wickramasinghe
Department of Science and Engineering
Solent University
Southampton, United Kingdom
2senan77@solent.ac.uk

Raza Hasan
Department of Science and Engineering
Solent University
Southampton, United Kingdom
raza.hasan@solent.ac.uk

Shakeel Ahmad
Department of Science and Engineering
Solent University
Southampton, United Kingdom
shakeel.ahmad@solent.ac.uk

Salman Mahmood
Department of Computer Science
Nazeer Hussain University
Karachi, Pakistan
salman.mahmood@nhu.edu.pk

***Abstract*—This study marks a step toward more effectively translating nutritional information to inform public health policy as well as individual dietary choices. Motivated by the increase in diet-related health issues, this research aims to analyze a comprehensive nutritional dataset to uncover valuable insights. Using the USDA National Nutrient Database, the study employs data preprocessing to clean the data, exploratory data analysis to identify hidden patterns, and various machine learning models to predict nutritional values. The results demonstrate the usefulness of these models in explaining the composition data and highlight a range of trends and relationships within the observed amounts. The discussion emphasizes that these findings could be instrumental in guiding health professionals and policymakers toward healthier dietary guidelines. The significance of this research lies in its potential to advance public health through more sophisticated nutritional recommendations.**

Keywords—***Data Analytics, Food Nutrition, Machine Learning*.**

I. INTRODUCTION

The analysis of nutritional data is becoming increasingly relevant to public health concerns and individual dietary alleys. This, amidst the increasing prevalence of diet-related health conditions such as obesity, diabetes, and cardiovascular diseases, has heightened nutritional interest in an individual food commodity. Because nutritional diseases, such as obesity and diabetes, are becoming increasingly prevalent in society, there has been an increasing need for the nutritional values of various individual foods. The proposed study will seek to apply newer machine learning methods to predict energy content using the data on macronutrients for which nutritional content is already available from the USDA National Nutrient Database.

This study aimed to use the USDA National Nutrient Database in analyzing different foods and their nutrients that included energy, macronutrients (fat, protein), vitamins & minerals as well fiber. The machine learning (ML) methods were chosen for this study were Linear Regression, Support Vector Regression (SVR), K-Nearest Neighbors (KNN), and Deep Neural Networks (DNN) because they handle the variable relationships of macronutrients and the energy content well. For example, DNN is one option because it can model complex nonlinear associations; SVR was tested to see how it behaved to predict over small deviations in energy content. We have chosen Linear Regression due to its simplicity in modeling linear relationships, which extends a useful baseline. SVR was used to explore its strength in high-dimensional spaces and small-sample learning, whereas KNN has been employed because it can capture nonlinear relationships based on proximity. Finally, DNN is a multi-layer complex architecture that might help in modeling nonlinear dependencies between macronutrients and energy.

This study has several main aims in exploring the work covered issues related to nutrition information. In the first part of this study, we will perform data preprocessing to cleanse and prepare the dataset for algorithmic analysis, removing any errors present. After this, Exploratory Data Analysis (EDA) will be done to find out hidden patterns or trends which are not so visible. From here on the study is moved towards modelling where we will have models created and implemented with respect to nutrient values to predict them, accuracy testing of these predictions will be done. Therefore, careful attention should be given to the presentation and interpretation of research findings to highlight their relevance for public health and dietary guidelines.

This study makes an important advance toward translating nutrition research into public health. Using data analytics, and machine learning this research will provide an insight into the food nutrition profiles for healthier decision-making, facilitating informed public health measures.

II. LITERATURE REVIEW

Recent advances in machine learning (ML) enables to support human experts and policy makers with developing evidence-based dietary guidelines. ML offer a potentially solution to solve the problems of controlled trial in diet where challenges and issues prevailed including complexity, compliance steaming [1]. Lastly, our reliance on observational studies also raises questions about the extent to which causal claims can or should be inferred and calls for improved correction of biases.

In discussing opportunities and challenges of ML applications in nutritional epidemiology, [2] identify the capacity for non-linear modelling features and to control confounding among its strengths. Nevertheless, they observe that the small scale and circumscribed nature of these studies as well as the dearth of research in particular modalities across existing literature could impair their view to what ML can accomplish. Also, but most likely researchers do not have the technical resources and computing power to use of all that ML has to offer.

[3] mentioned the application of ML in personalized dietary recommendations. These authors have demonstrated better improvement in diet adherence. [4] show that in nutritional epidemiology, ML applications result in significant enhancements in nonlinear modeling, which helps control confounding variables in large datasets. [3] suggested that ML would help to analyze the large, complex datasets for nutritional research. But the attitude that prevents machine learning from being widely used is also present in the nutritionist profession. Within nutritional genomics, [5] explain how ML may be combined with traditional statistical methods to interpret multi-omics data while addressing issues of dealing variable importance and missing data that can confound analyses. [6] also remind us that ML may be an attractive tool to improve dietary measurement and accommodate the complexities of diet as a multidimensional exposure, but caution should exercise when applying these purely data-driven approaches; otherwise, overfitting could occur. It needs a lot of training and computational power to achieve this.

As for recent directions in the application of ML, newer models have been developed to identify calorie-dense items internally from food images [7] and predict dietary patterns by means of regression analysis approaches [8]. However, also [9] highlighted the benefits of ML in pediatric nutrition concerning risks of malnutrition and obesity but appropriately called for established ethical frameworks and standardized practices.

Additional studies demonstrate the potential for prediction of ML across different contexts, including food processing levels [10] and micronutrient profiles in processed foods [11]. The potential impact of these advancements on public health strategies and future dietary recommendation based upon ML is being recognized, although the challenges related to data quality as well complexity of interactions between genetic and environmental fabric are not certainly addressed [12].

Overall, while the literature reveals ML's transformative potential in nutrition science, it also emphasizes the necessity for further research to navigate existing challenges, enhance interpretability, and optimize its application for public health improvements.

III. METHODOLOGY

The methodology for this study encompasses several key steps, from data collection and preparation to model training and evaluation as shown in Fig. 1. Each step is aimed at comprehensively analyzing the USDA National Nutrient Database to predict energy content in kilocalories based on macronutrient values.

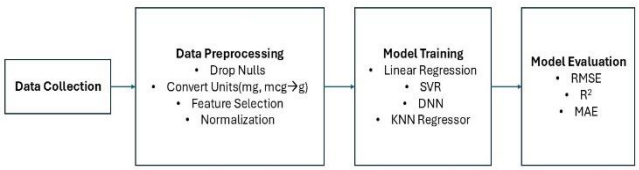

Fig. 1. Methodology workflow

A. Dataset

The primary dataset utilized in this research is the "USDA National Nutrient Database," sourced from data. World [13]. This dataset includes detailed nutritional information for a variety of food items, encompassing features such as macronutrients (fat, protein, carbohydrates) and minerals, with energy in kilocalories as the target variable. Gathered directly from the USDA National Nutrient Database, it contains a wide range of foods along with their macronutrient data. This will ensure good generalization of the models on various kinds of foods since the dataset contains huge variation in food types and their macronutrient profiles. Such comprehensiveness of the dataset, with high granularity of nutritional components involved, makes this dataset quite promising for more complex food prediction models in the future.

B. Dataset Preparation

- Certain columns contained significant null values (more than 80%). Given their unique nature, these columns were excluded from further analysis. All remaining columns were devoid of null values and were thus deemed suitable for analysis.
- Nutritional values were converted to consistent units (grams) to mitigate any potential bias in analysis.

C. Feature Selection

Feature selection was performed using three methods:

- Identifying relationships between features, with a focus on those showing correlations greater than 0.2.
- Employing a Random Forest Regressor to determine feature importance, revealing that fat, carbohydrates, and protein significantly contribute to energy content.
- Incorporating expert insights to validate the relevance of selected features such as Fat (g), Carbohydrates (g), Protein (g), Sugar (g), Vitamin E (g) and Magnesium (g).

D. Predictive Modeling

The following machine learning models were employed to predict the energy content based on selected macronutrients:

- Linear Regression which assumes a linear relationship between independent and dependent variables.
- Support Vector Regression (SVR) which focuses on finding a function that minimizes deviation within a specified margin.
- K-Nearest Neighbors (KNN) which estimates output based on the average of the nearest K training instances.
- Deep Neural Network (DNN) which utilizes multiple hidden layers to learn complex patterns in the data.

Modelling was conducted using Python's scikit-learn library, which facilitated the implementation and evaluation of each model iteratively.

E. Evaluation Metrics

It was measured with the R squared (R^2), that reflects a measure of how much variance is accounted for by this model. Root Mean Squared Error or RMSE is the square root of mean squared error and measures average (calculated on per item basis) deviation from actual value. Mean Absolute Error (MAE) shows distance in absolute sense between predicted and true values.

F. Limitations

While using diverse regression models endows substantial potential for effective prediction, there are downsides. It should be noted the assumptions of constant relationships, dependence on data quality and hyperparameter tuning etc. These models also do not explain the nutritional interaction in biological mechanistic terms; thus, results should be interpreted with caution.

IV. Results

A. Exploratory Data Analysis

The EDA provided valuable insights into the nutritional dataset's characteristics and interrelationships among variables.

1) Descriptive Statistics

A descriptive statistical analysis of this dataset reveals the following. The average nutritional composition of the 8,618 items is as follows: 21.8 grams of carbs; 11.5 grams of protein 10.6 g of fat and 226.4 kilocalories of energy. All these values have a large variation. The maximum values of these four nutrients are 88.3g for protein, 100g for fat and carbohydrates, and 902 kcal for energy. Micro-nutritional values on the other hand have more differences between them. The average amount of vitamin A is 0.000094g and vitamin C is 0.008g. Phosphorus has the highest average among minerals of 0.156g while copper has the lowest of 0.00000017g. Most of the items in this dataset have a minimum of 0 for almost all nutrients. Only one item lacks energy information. Many values, especially the standard deviation, are significantly higher than the mean indicating large variations. Furthermore, for almost all the nutritional variables, the mean is greater than the median which implies a right-skewed distribution in most. This dataset is varied with the interquartile ranges of all nutrients varying significantly. This variability does not only make this dataset descriptive of the full variety, but also this variability increases its value for any suitable analysis.

2) Univariate Analysis

Univariate analysis examines the distribution of individual variables separately. For this nutritional dataset, fat, carbohydrates, and protein are identified as key components. Their distributions, as shown in the descriptive statistics table, reveal right-skewed patterns. This skewness is evident from the mean values exceeding the median (50th percentile) for each nutrient:

- Fat: mean of 10.6g vs median of 5.2g as shown in Fig. 2.
- Carbohydrates: mean of 21.8g vs median of 8.9g as shown in Fig. 3.
- Protein: mean of 11.5g vs median of 8.3g as shown in Fig. 4.

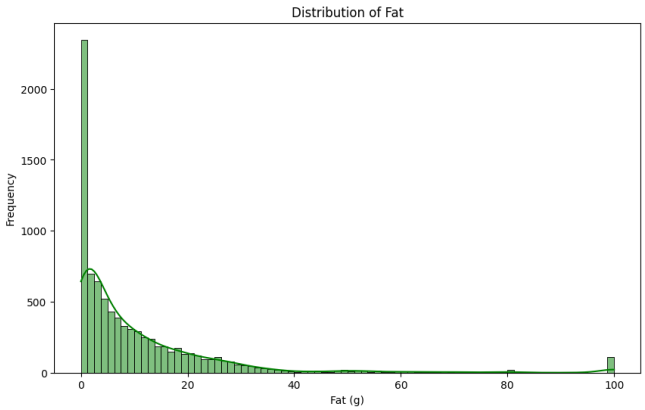

Fig. 2. Distribution of Fat

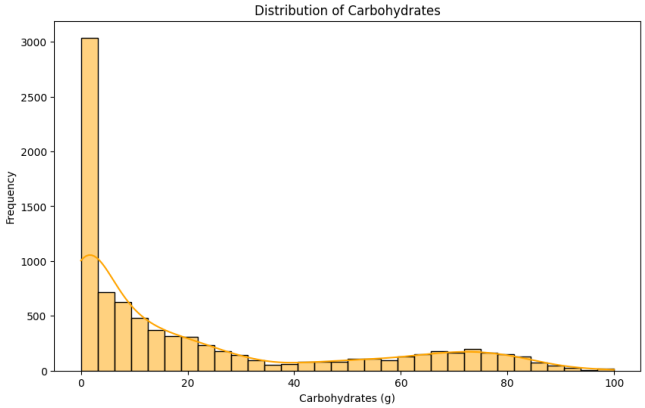

Fig. 3. Distribution of Carbohydrates

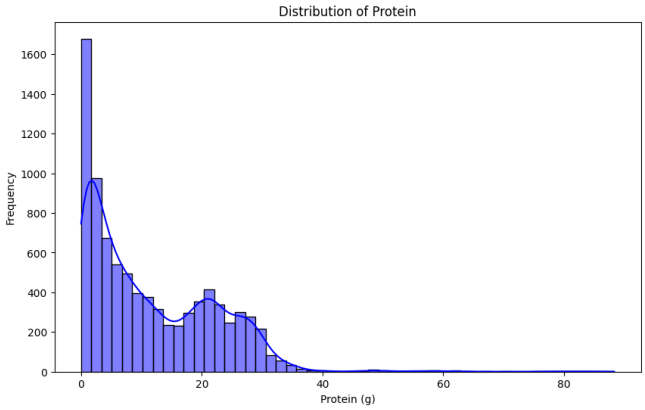

Fig. 4. Distribution of Protein

These statistics indicate that while some foods contain high amounts of these macronutrients (as seen in the maximum values), many items in the dataset have relatively lower amounts. The right skew suggests that most foods contain less than the mean values of these nutrients, with a smaller number of high-nutrient items pulling the averages upward. The violin plot in Fig. 5 further illustrates that the median values for these nutrients cluster around 10 grams, with carbohydrates displaying a wider interquartile range.

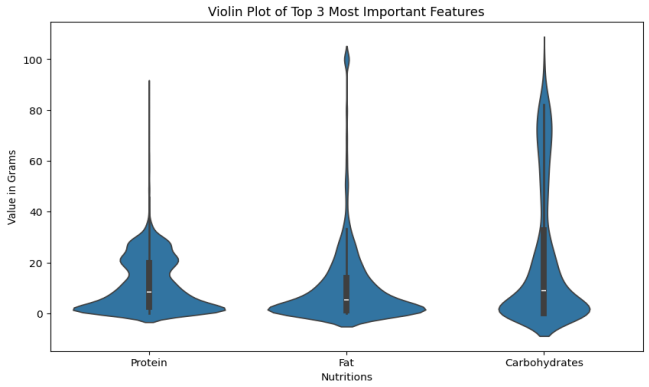

Fig. 5. Violin plot of top 3 most importance features

3) Bivariate Analysis

Bivariate analysis showed strong correlations between energy content and macronutrients, particularly evident in the scatterplots in Fig. 6. The analysis indicates a positive correlation, with protein exhibiting the strongest relationship with energy output.

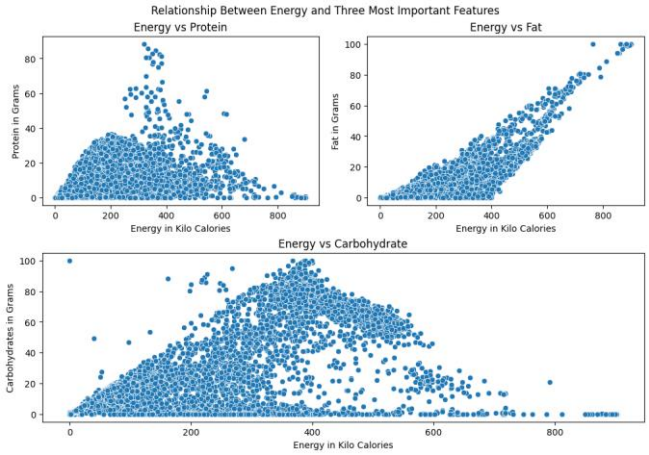

Fig. 6. Energy vs Protein, Fat, and Carbohydrate

4) Multivariate Analysis

A correlation heatmap in Fig. 7 provided a comprehensive view of the interrelationships among all nutritional features, confirming that fat, protein, and carbohydrates are the dominant contributors to energy content. The radar chart in Fig. 8 showcased the average nutritional content across the top five food groups, revealing significant variations in fat, protein, and carbohydrate content.

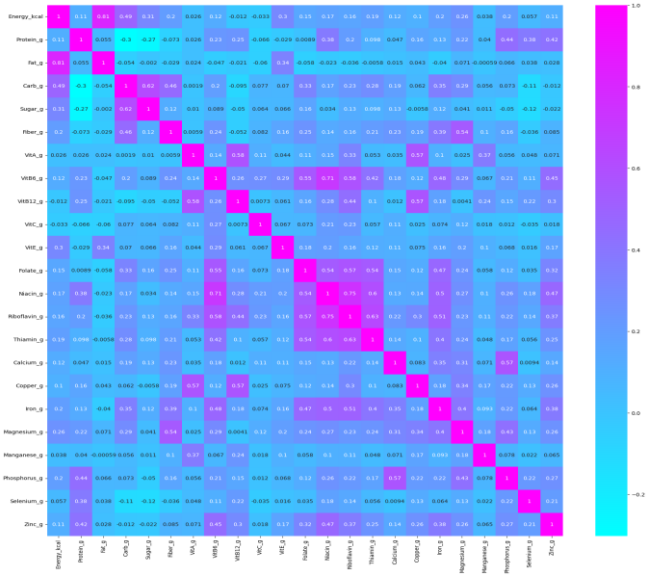

Fig. 7. Correlation heatmap

Fig. 8. Average nutrition content of top five food groups

B. Predictive Analysis

The predictive models employed such as Linear Regression, SVR, KNN, and DNN were evaluated based on their ability to estimate energy content from macronutrient values. The results, displayed in Table I, indicate that all models performed well, with the DNN achieving the highest R^2 score of 0.8573, suggesting that it accounted for approximately 85.73% of the variance in energy content. In comparison with existing studies that employed image analysis techniques for calorie prediction, such as the work by [14], which used convolutional neural networks to estimate macronutrients from food images followed by conventional ML algorithms to predict caloric content, our approach demonstrates a more direct and precise method of prediction by utilizing nutritional data alone. Consequently, relatively lower performance by some of the models is at the expense of previous image-based methods due to the complexity of calorie estimation with only macronutrient data. The methods based on images would also include the portion sizes and textures of food as a visual cue, while in this approach, the data provided relies on the macronutrient information alone.

This leads to more conservative results, indeed; it is a more controlled approach in that it narrows its focus down to nutritional value rather than external factors. This trade-off was in favor of the accuracy inside the specific relationships of macronutrient-energy rather than general caloric predictions.

TABLE I. MODEL PERFORMANCE COMPARISON WITH EXTERNAL WORK

Model	R2 Score		RMSE		MAE	
	Existing	Current	Existing	Current	Existing	Current
LR	0.72	0.98	81.62	22.03	62.20	7.54
SVR (Linear)	0.67	0.98	88.41	23.06	65.84	8.01
SVR (RBF)	0.73	0.90	80.43	49.96	60.24	18.67
KNN	0.80	0.98	69.81	21.27	52.19	9.27
DNN	0.86	0.98	58.50	21.37	32.43	7.67

C. Model Performance

Performance metrics are summarized in Table II, highlighting the effectiveness of each model where the performance of each scaling method for the models. DNN always outperforms the rest in all metrics. Explaining this by the capability of DNN to model complex interactions between macronutrients, whereas linear models are faster to train but lack strengths in capturing the full variance of energy content, as represented by lower R^2 scores.

TABLE II. PERFORMANCE COMPARISON WITH SCALERS

	Scaler	LR	SVR (Linear)	SVR (RBF)	KNN	DNN
R^2 Score	MinMax	0.98	0.90	0.92	0.99	0.98
	Normalizer	0.53	-0.08	-1.41	0.46	0.56
	Standard	0.98	0.98	0.98	0.98	0.98
	-	0.98	0.98	0.91	0.98	0.98
RMSE	MinMax	0.02	0.06	0.05	0.02	0.02
	Normalizer	0.04	0.06	0.09	0.04	0.04
	Standard	0.13	0.13	0.14	0.14	0.13
	-	22.03	23.06	49.96	21.27	21.37
MAE	MinMax	0.01	0.05	0.04	0.01	0.01
	Normalizer	0.01	0.04	0.07	0.00	0.01
	Standard	0.04	0.05	0.06	0.06	0.04
	-	7.54	7.58	18.67	9.27	7.67

V. DISCUSSION

The exploratory data analysis of the nutritional dataset encompassing 8,618 food items reveals critical insights into the macronutrient and micronutrient compositions that are vital for dietary planning and public health initiatives. The macronutrient analysis indicates that carbohydrates, with an average content of 21.8 g, dominate the nutritional profiles of these foods, suggesting their fundamental role in dietary intake. This is aligned with current nutritional guidelines emphasizing the importance of carbohydrates as a primary energy source.

The right-skewed distributions observed for fat, protein, and carbohydrates suggest that while a minority of foods are rich in these macronutrients, most items contain lower amounts. This is reflected in the significant differences between the means and medians, indicating that high-nutrient foods may skew average nutrient values upwards. For instance, the median fat content at 5.2 g, contrasted with a mean of 10.6 g, suggests that many foods may provide minimal fat, while a few contribute substantially.

Furthermore, the comprehensive assessment of energy content, averaging 226.4 kcal, underscores the variability inherent within this dataset, with some items providing energy levels as low as 0 kcal. Such findings emphasize the diversity of food types, ranging from low-calorie vegetables to energy-dense snacks, which have implications for calorie management in dietary practices.

The bivariate analysis strengthens the relationship between macronutrients and energy content, especially highlighting protein as the most significant contributor. The correlation heatmap corroborates this finding, illustrating that fat, protein, and carbohydrates are the primary determinants of energy content in foods. This information is essential for nutritional modelling and underscores the need for focused dietary recommendations.

Results of predictive modelling revealed the ability to predict energy content from macronutrient values well predicted by machine learning algorithms, with the highest R^2 score (0.8573) achieved by Deep Neural Network (DNN). In summary, these results show that the macronutrient content influence in a substantial fraction of variance regarding to energy value can be explained by using those predictive models and then support their application for nutritional research/data collection.

Table II shows the performance metrics from various models indicate that the choice of scaling may have a large effect on overall model outcomes (how well a model performs in general). The unchanged excellence of the DNN performances as a function of scaling supports its ability to learn complex statistical relationships within the data, hence encouraging more advanced modelling approaches in nutritional analytics.

In summary, this study demonstrates the significance of varied food groups and for individual nutrition details in dietary advice considering the wide diversity regarding content. These are potential new areas for research using methods developed from machine learning, particularly to develop personalized nutrition plans and public health policies that target enhanced nutritional quality in diets. Using these extensive data sets, key stakeholders will be able to come up with strategies that contribute towards creating a way where healthy eating patterns are recorded more widely throughout populations.

VI. CONCLUSION

Using nutritional information from 8,618 unique food items, the study represents the most robust data set on a wide array of macronutrient and micronutrient distributions. Results emphasize the carbohydrate-rich nature of these food types with marked variability in fat and protein content, hence diversity amongst them. The right-skewed distributions of macronutrients such as Carbohydrate, total and Saturated Fat are indicative that while a few foods may be rich in these nutrients by far the most contain low levels, which is meaningful for dietary planning.

The high correlations found between energy content and macronutrients bring confidence to the known relevance of these components in nutrition, providing important insights for dietetic surveys and public health recommendations. Importantly, application of predictive modelling techniques-

in particular the DNN-exemplifies a pathway for more accurate estimations of energy content as driven by macronutrient composition and could be used to drive future advancements in nutritional research.

In the end, whatever we learn from this analysis will benefit dietary guidelines to be personalized based on new data for improved health. Further research could be carried out to investigate, for example the interaction of nutrients at a micronutrient level and food processing on health and disease to add more comprehensive knowledge in this critical nutrition-related public health area as pointed by [15].

REFERENCES

[1] L. M. Bodnar, S. I. Kirkpatrick and A. I. Naimi, "Machine learning can improve the development of evidence-based dietary guidelines," Public Health Nutrition, vol. 25, (9), pp. 2566-2569, 2022. Available: https://www.cambridge.org/core/product/identifier/S1368980022001392/type/journal article. DOI: 10.1017/S1368980022001392.

[2] S. Russo and S. Bonassi, "Prospects and Pitfalls of Machine Learning in Nutritional Epidemiology," Nutrients, vol. 14, (9), pp. 1705, 2022. Available: https://www.ncbi.nlm.nih.gov/pubmed/35565673. DOI: 10.3390/nu14091705.

[3] D. Kirk, C. Catal and B. Tekinerdogan, "Precision nutrition: A systematic literature review," Computers in Biology and Medicine, vol. 133, pp. 104365, 2021. Available: https://dx.doi.org/10.1016/j.compbiomed.2021.104365. DOI: 10.1016/j.compbiomed.2021.104365.

[4] L. Oliveira Chaves et al, "Applicability of machine learning techniques in food intake assessment: A systematic review," Critical Reviews in Food Science and Nutrition, vol. 63, (7), pp. 902-919, 2023. Available: https://www.tandfonline.com/doi/abs/10.1080/10408398.2021.1956425. DOI: 10.1080/10408398.2021.1956425.

[5] L. Khorraminezhad et al, "Statistical and Machine-Learning Analyses in Nutritional Genomics Studies," Nutrients, vol. 12, (10), pp. 3140, 2020. Available: https://www.ncbi.nlm.nih.gov/pubmed/33066636. DOI: 10.3390/nu12103140.

[6] J. D. Morgenstern et al, "Perspective: Big Data and Machine Learning Could Help Advance Nutritional Epidemiology," Advances in Nutrition (Bethesda, Md.), vol. 12, (3), pp. 621-631, 2021. Available: https://dx.doi.org/10.1093/advances/nmaa183. DOI: 10.1093/advances/nmaa183.

[7] H. Gao et al, "Food Nutrient Extraction Based on Image Recognition and Entity Extraction," WiMob, pp. 13-19, 2023. Available: https://ieeexplore.ieee.org/document/10187783. DOI: 10.1109/WiMob58348.2023.10187783.

[8] V. C. Silva et al, "Clustering analysis and machine learning algorithms in the prediction of dietary patterns: Cross‐sectional results of the Brazilian Longitudinal Study of Adult Health (ELSA‐Brasil)," Journal of Human Nutrition and Dietetics, vol. 35, (5), pp. 883-894, 2022. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/jhn.12992. DOI: 10.1111/jhn.12992.

[9] A. Young, M. J. Johnson and R. M. Beattie, "The use of machine learning in paediatric nutrition," Current Opinion in Clinical Nutrition and Metabolic Care, vol. 27, (3), pp. 290-296, 2024. Available: https://www.ncbi.nlm.nih.gov/pubmed/38294876. DOI: 10.1097/MCO.0000000000001018.

[10] G. Menichetti et al, "Machine Learning Prediction of Food Processing," MedRxiv, 2022. Available: https://search.proquest.com/docview/2674724981. DOI: 10.1101/2021.05.22.21257615.

[11] T. Naravane and I. Tagkopoulos, "Machine learning models to predict micronutrient profile in food after processing," Current Research in Food Science, vol. 6, pp. 100500, 2023. Available: https://dx.doi.org/10.1016/j.crfs.2023.100500. DOI: 10.1016/j.crfs.2023.100500.

[12] I. Parkar et al, "Personalized Health Recommendation System using Machine Learning," International Journal for Research in Applied Science and Engineering Technology, vol. 12, (4), pp. 5938-5943, 2024. Available: https://doi.org/10.22214/ijraset.2024.61379. DOI: 10.22214/ijraset.2024.61379.

[13] USDA National Nutrient DB. Available: https://data.world/craigkelly/usda-national-nutrient-db.

[14] R. Sombutkaew and O. Chitsobhuk, “Image-based Thai Food Recognition and Calorie Estimation using Machine Learning Techniques,” pp. 1–4, May 2023, doi: 10.1109/ECTI-CON58255.2023.10153183. [Online]. Available: https://ieeexplore.ieee.org/document/10153183

[15] P. G. Ferrario and K. Gedrich, “Machine learning and personalized nutrition: a promising liaison?,” vol. 78, no. 1, pp. 74–76, Jan. 2024, doi: 10.1038/s41430-023-01350-3. Available: https://www.ncbi.nlm.nih.gov/pubmed/37833568.