RESEARCH PAPER

Analysis of Daily Food Consumption and Nutrient Patterns Using Machine Learning Models

Jaishree Jain[1]*, Mani Dublish[2], Santosh Kumar Upadhyay[1], Dheeraj Kumar[2], Shashank Sahu[1] and Ajay Kant Dubey[3]

[1]Department of Computer Science & Engineering, Ajay Kumar Garg Engineering College Ghaziabad, U.P., India
[2]Department of Computer Applications, Ajay Kumar Garg Engineering College Ghaziabad, U.P., India
[3]Department of Mechanical Engineering, Ajay Kumar Garg Engineering College Ghaziabad, U.P., India
jaishree3112@gmail.com[1], manidublish2@gmail.com[2], ersk2006@gmail.com[1], dheerajsingh.phd@gmail.com[2], sahushashank@akgec.ac.in[1] and dubeyajaykant@akgec.ac.in[3]

Received: 24th April, 2026; Revised: 20th May 2026; Accepted: 30th June, 2026; Available Online: 10th July, 2026

ABSTRACT

To remain healthy and to prevent problems tied to one's diet, it is necessary to assess one's diets and the nutrients ingested on a daily basis. In this paper, the authors recommend employing a Machine Learning Model so as to identify the eating habits of individuals and analyse the caloric content of the food they ingest. We tested out multiple methodologies, such as through a CNN-LSTM + XGBoost, in order to establish which model exhibited the highest performance levels for predicting caloric intake, macronutrient intake, and micronutrient intake. Based on these tests, we have found that our recommended methodology outperformed the previous statistical models at categorizing an individual's caloric intake, and that it can provide a means of scientifically and easily establishing reliable eating patterns on a population-wide level. The outcomes of this learning are indicative of the likely value of Machine Learning as a means of assessing the dietary patterns of individuals. In addition, these results could be utilized in the development of personalized meal plans, and in the design of Public Health Programmes. Experimental results demonstrate that this technology is able to very accurately categorize foods automatically. As such, this technology represents a valuable resource both to the general population and to professionals such as Dietitians for making better food selections and supporting healthy lifestyle choices. In conclusion, our findings indicate that the results produced by our Machine Learning methodology represent a complementary analysis to traditional dietary recommendations by providing researchers and practitioners with accurate, repeatable data.

Keywords: Machine learning, energy estimation, CNN, Calorie Intake, Dietary Habit, XG-Boost.

How to cite this article: Jain J, Dublish M, Upadhyay SK, Kumar D, Sahu S, Dubey AK, Analysis of Daily Food Consumption and Nutrient Patterns Using Machine Learning Models. Int J Drug Deliv Technol. 2026;16(6): 795-810. DOI: 10.25258/ijddt.16.6.120
Source of support: Nil.
Conflict of interest: None

INTRODUCTION

Food is not only responsible for sustaining life and promoting health; but also capable of sustaining daily emotional and mental function. Food affects people nutritionally, psychologically, emotionally and socially, at times having both positive and negative impacts on us as human beings.

Each meal's nutritional value varies considerably because the type and quality of ingredients used when preparing the meal can vary. Today's health-conscious consumer is interested in finding out what is included in the food they eat. Interest in learning more about nutrition is growing among individuals as the number of people with metabolic issues increases as a result of obesity among all age groups. There are significant safety implications for those with obesity (particularly developing nations) from the increased risks of cardiovascular events due to the rising rates of obesity and its associated health consequences. Obesity is a serious issue in developing countries where there has been a dramatic increase in the prevalence of both cardiac disease and the related risk factors for developing these diseases, including obesity [2].

Diet is one significant factor associated with those problems; therefore, scientists are interested in investigating the relationship between various types of food and health, as well as the possible hazards associated with the consumption of different types of food. While tracking the number of calories consumed is challenging, it is essential for maintaining healthy weight, meeting nutritional needs and balancing total nutrient intake over the long-term [3]. Despite having access to established nutritional guidelines that help to measure the caloric content of food, calorie counting has historically been extremely challenging to do consistently and those Associated with the same have always involved long, tedious procedures in a laboratory setting. Machine learning models are able to analyze large datasets of

**Author for Correspondence: jaishree3112@gmail.com*

information and identify patterns that will allow them to better estimate calorie counts than traditional laboratory type methods and rudimentary calculations such as the old-fashioned way of measuring calories in foods. Through our work, we have attempted to use various machine learning algorithms to estimate the caloric values of food items and develop nutrition-based models for assessing health risks based on these estimates. Other researchers are looking at similar data to determine how different types of food contribute to an individual's overall health with respect to their caloric values and composition. A notable tool, Food Prox, uses Random Forest algorithms to classify food items based on information from the DFND dataset obtained from nutritional information about these foods. Running multiple types of food through a Nutritional Analysis application can be very beneficial in determining how Different Extremities of Fats, Carbs, and Glycaemic Index Values interact with each other individually and as a collective whole to influence energy levels and long-term health. Based on data compiled by the World Health Organization (WHO, 2020) excessive amounts of saturated fats consumed results in elevated levels of LDL cholesterol which contributes to the development of cardiovascular disease. Additionally, the overconsumption of sodium has contributed to the development of cardiovascular disease as well through elevated blood pressure, which subsequently increases the likelihood of stroke and cardiovascular disease (WHO, 2020). The use of machine learning can help us better understand how nutrients affect our health and potentially identify which foods could be detrimental to our health. Because we can use these models in the diet tracking devices to help identify foods that are high in unhealthy things and offer alternatives, our research aims to build a Random Forest regression model for predicting meal calorie loads. This model will be able to use data on carbohydrates, proteins, total fats and total sugars in order to make a prediction. Poor dietary habits rise the menace of growing chronic diseases such as obesity, hypertension and cardiovascular disease. Collectively, the NCDs account for 74% of death globally.

In Australia, the situation is even worse. Young adults' diets in Australia have gone completely off track -- in 2022, only 2.1% of 18-to-24-year-olds were able to meet the Australian guidelines for fruit and vegetables. The period between ages 18-24 is also a critical time for nutritional intervention because many of the poor dietary behaviours that are associated with poor nutrition originate from this time frame. To develop effective strategies, we need an understanding of what factors influence eating behaviours among young adults. Examination of EOs considers the timing and the manner in which we consume food, beverages, and main meals; snacking; etc.; additionally, any type of significant beverage consumed needs to be evaluated also, along with the frequency of alcohol consumption relative to eating. The relationship between our regular activities and how frequently we consume food; the location of our consumption; and the rationale for eating will all affect our food choices. The above factors relate not only to the physical attributes of an individual but also to situational attributes that allow an individual to make the selection of what they consume. Therefore, many of the food products that ultimately enter your mouth are significantly influenced by both your personal tastes and the environment.

Most people will be influenced at both levels in terms of their identity as an eater, simply by their age, gender, income level and preferences/likes of specific types of food. The physical attributes will include where you consume food, with whom you consume food and what activities you participate in as you consume food; situational attributes will affect the physical characteristics of your food selections. For example, since eating is highly individualised and can vary from one day to the next, you need to take into consideration both personal characteristics and situational characteristics.

Since tracking eating behaviour is so complex, the application of machine learning will aid in identifying trends and developing personalised plans for healthier eating. There is also Ecological Momentary Assessment (EMA), which is a methodology of data collection for eating behaviour by collecting time-sensitive data through diaries/applications on mobiles and other devices. This will allow us to track eating behaviours in real-time as opposed to relying on people’s memories later. An example of this would be what times of the day or season fruits are commonly reported as being consumed more for snacks versus meals. Understanding this type of detail helps us to better understand how different contexts can influence our food choices and overall diet quality. ML can identify complex relationships and connections in large datasets, such as non-linear relationships and the interaction between different variables, without being explicitly instructed as to what to look for [13].

It might find that, for example, we eat more vegetables when eating with others but less when eating alone. ML models can produce personalized dietary suggestions by using details about an individual's personal situation and also details about an individual's event situation. These personal approaches can help to improve eating habits in the long term because the suggested changes are relevant to people's unique circumstances [14].

There have been interesting advances in ML to improve prediction of eating behaviours and enable personalized nutrition advice. However, most existing studies do not predict well food choices in real-life settings, especially for young adults. In addition, focusing on specific food groups restricts our understanding of the health effects of overall diet. Neural networks have been used to predict how well a person will stick to a diet, but these systems can be incredibly complicated, meaning they can be difficult to interpret and implement in real-world environments. There is growing interest in AI-based food substitution systems, but most of the research focuses on how acceptable such changes are rather than on actual improvements to people’s diets. One strategy to enhance

machine learning-based advice is to embed context and details about all food groups in the Eating Occasions and the overall daily diet quality [15-17]. This study analyzes the dietary behaviors of free-living young adults using ML to decompose their consumption during various contexts. They try different models to see whether context around meals impacts the food groups Aussie young adults eat. The aim of the study is to improve precision nutrition and better tailor recommendations by examining personal characteristics, eating contexts and overall diet quality.

LITERATURE REVIEW

Food safety and nutrition are receiving increased attention and there is a growing need for intelligent systems to detect whether food ingredients are safe or unsafe. Machine learning is therefore seeing a boom in this domain with automatic, evidence-based resolutions used for analysing food ingredients built on numerous criteria. Classification algorithms assist ingredient lists to track health, ensure food quality and encourage healthy eating [18-20]. The goal of the present contribution is to view the application of different machine learning models for the identification of food components, in order to find the optimal strategy in terms of accuracy and generalization performance. The complete Literature review of our proposed system is given in Table 1.

Data for this secondary analysis were obtained from the Measuring Eating in Everyday Life Study (MEALS), a cross-section study showed from April 2015 to April 2016 to examine eating forms in young adults aged 18-30 years. Design of the MEALS study and methods of dietetic valuation are defined in detail elsewhere [20].

Applicants were enlisted through online and offline recruitment strategies. Eligible participants were required to be a resident of Victoria, Australia, own a smartphone purchased in Australia, speak English as their primary household language and not be pregnant or breastfeeding. Applicants delivered advised written agreement and then finished an online survey via Qualtrics, and then recorded their food intake over four non-consecutive days using the “Food Now” smartphone application [21-23].

Written informed consent was obtained from all applicants before participation. Ethics approval was obtained from the Deakin University Human Ethics Advisory Group, faculty of health (HEAG-H 11_2015). Participants were given a AUD $25 voucher upon successful completion of the study. To comply with reporting guidelines, they included the STROBE-nut checklist of Strengthening the Reporting of Experimental Studies in Epidemiology-Nutritional Epidemiology group as supplementary material. We had to throw out people who didn’t give us data for fewer than three days. That gave us 675 young adults for the analysis. This study used the Food Now app to track diet. The people recorded what they ate immediately after eating it, for four days over two weeks, and the days had to include at least one weekend day. They would take pictures of their food and drinks and type in details of portion sizes and stuff.

The app pestered users to report leftovers, and pinged them every three hours they were awake, to make sure they got it all. And at bedtime it always reminded them if they hadn’t done an entry that day. This app used to work well, with a strong association between reported diets and measured energy use (ICC=0.75).

Trained nutritionists coded dietary records from food images and accompanying text and voice descriptions and matched the reported foods to entries in the Australian Food, Supplement and Nutrient Database 2011-2013 (AUSNUT 2011-13).

A duplicate review procedure was used to confirm data precision. Participants also recorded the smart time, EO type, eating location, social setting, concurrent activities, and food source through the app. For this analysis, foods and beverages consumed simultaneously and providing at least 210KK of energy were considered a single EO.

CLASSIFICATION OF FOOD TYPES AND COMPUTATION OF PORTIONS

All informed nutriment and beverages items were categorized as optional or required based on the Australian Bureau of statistics discretionary food list. Discretionary food consumption was expressed in servings, where one serving was equivalent to 600KJ of energy derived from discretionary foods.

Based on the Australian Dietary Guidelines (ADG), required foods were categorized into five foods groups: vegetables, fruits, grains, meats, and substitutions, and dairy and substitutes. Food group portions were assessed using the ADG food group record. For predictive modelling, the outcome variables were the servings consumed per eating occasion for each food group, which were transformed using the natural logarithm [ln(x +1)] to reduce skewness and normalize the distribution.

Quality of diet

The DGI is an assessment of the quality of the diet based on how well people follow the ADGs as determined by DGI using 2013 guidelines (ADGs) as a basis. The DGI contains 12 different components, all of which have been adjusted to fit 24-hour recall data for scoring purposes. Each of these components has been assigned a score between 0 and 10, where 0 indicates no adherence to the recommendations and 10 indicates maximum adherence. The total score possible is 0 to 120; therefore, the higher a person's total score is the better quality their diet is. A DGI score was calculated for a single day for each applicant.

Contextual Elements

Guided by existing evidence on determinants of food consumption, contextual factors were categorized as either person-level or eating occasion–level variables. Detailed definitions and descriptions of these features are presented. Person-level contextual factors were further classified into intrapersonal, socio-environmental, and physical-environmental domains and were assessed using an online questionnaire comprising validated item from previously published studies.

Intrapersonal factors were found to be sociodemographic and behavioral features such as sex/gender, age, country of birth, weekly gross income, educational level, smoking status, and physical activity level. For income, the categories were as follows: <$120, $120-$499, $500-$999, and >$1,000 per week, and the categories for education were low (i.e., high school or lower), medium (TAFE/VET), and high (i.e., university). Smoking status was broken down into three categories of never, ex, or current smoker; and the IPAQ was used for the measurement of physical activity, which classified individuals as either within the recommended amount of physical activity or not based on the guidelines for Australian physical activity and sedentary behavior. Intrapersonal contextual factors also included food preparation behavior, food shopping behavior, self-efficacy, and cooking confidence. Higher scores on these measures indicated greater in food preparation activities (12-84), greater involvement in food shopping tasks (0-21), stronger confidence in maintaining healthy eating behaviors (17-85), and higher cooking confidence (0-21), respectively.

The questionnaires from the Project EAT study were used to assess perceived duration shortages (score ranged from 4 to 16, with scores that were higher suggesting a stronger feeling of duration stress) and dietary constraints (score range from 0 to 11, with more high levels suggesting more apparent challenges to choosing nutritious meal options) [18].

Food shopping behavior was calculated using questions related to participant's food and grocery shopping practices, including the use of shopping lists. Self-efficacy and cooking confidence were assessed using adapted versions of validated questionnaires, with cooking confidence measured using items derived from the United Kingdom's 1993 health and lifestyle surveys. The socio-environmental domain comprised family and friend social support, measured using a composite score ranging from 3 to 15. Higher grades reflected stronger perceived support for healthy eating from family and friends.

Table 1. **Table 1.** Reviews all of the work on techniques and strategies utilized in this sector

Author name with reference	Model/technique used	Key findings	Pros	Cons
Menichetti et al. [14]	FoodProx(random forest classifier), NOVA Classification, FPro(Food Processing Score)	1)Ultra-processed foods form over 73% of total foods in the United States. FoodProX outperforms NOVA with better coverage and continuous scoring. In other words, in some way, the capabilities and ability to allow execution differ between the systems. NOVA is a system for classifying processed foods based on their nutritional value.	Classifies complex foods and identifies healthier foods	Doesn't directly measure process-ing techniques
Qasrawi et al. [15]	Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), and Gradient Boosting (GB). Evaluation: confusion matrix, accuracy, precision, sensitivity, F1-score, and AUC	Through Random Forest and Gradient Boosting, food could be classified.	Offers insights for tar- geted interventions, Detailed validation strategy with robust performance metrics	A small ini Initial sample size and potential imbalances

Gupta et al. [16]	Uses Linear Regression for predicting nutritional val- ues, Decision Tree (DT) for categorical data, RT for im- proving accuracy, Support Vector Machine (SVM) for classification. Evaluation by mean squared error (MSE) and accuracy	RT and DT balanced accuracy and in- treatability. LR struggles with complexity. SVM was effective but required parameter tuning	DT offers explain-ability, Regression models provide stable and interpre-table. SVM handles high dimensional data well	Regression models experience difficulty identifying the non-linear relationship and therefore are not very flexible. Decision trees will overfit unless they are pruned adequately. Random trees are too complex and slow for large datasets. Support Vector Machines experience difficulty dealing with large datasets as they grow exponentially.
Tachie et al. [17]	DT, RF, & LightGBM were used for predicting Nutri Score & micronutrient content. Pre-processed data was needed to clean data & manage collinearity (by using VIF). Pre-processing steps also involved transforming categorical variables.	Random Forest (RF) and Light Gradient Boosting Model (LightGBM) achieved the most accurate performance, while Decision Tree learning (DT) had the least accurate performance based on the results of predictive analysis. The Random Forest model provided the most accurate test prediction results.	RF provides high levels of accuracy while being able to deal with missing values, and DT has lower levels of required data preparation and can handle both binary (i.e., categorical) and continuous (i.e., numeric) data. LR is used to assist in generating probability values.	Decision trees exhibit a high degree of sensitivity to noise, resulting in a failure to generalize appropriately. Random forests can be very computationally intensive and take longer to train due to the ensemble nature and complexity. Logistic regression has difficulty dealing with non-linearity between dependent and independent variables.
Kirk et al. [18]	The use of interpretable machine learning techniques to classify, predict and cluster data pertaining to nutrition included Random Forest, k-means clustering, logistic regression, SVM, XGBoost with SHAP, and Lasso regression.	Machine Learning (ML) methods predicted the severity of symptoms of obesity/malnutrition and glucose levels on an individual basis by using phenotype data from individuals.	High levels of accuracy were achieved, allowing for low-cost, non-invasive predictions with interpretable results. XAI demonstrated high proficiency in identifying biomarker-based health insights and tailoring interventions.	Some models, without xAI, were not easily interpretable due to their complexity. Population-specific data can hinder model generalizability.
Adjuik et al. [19]	used a combination of conventional and	FoodProX RF identified highly	While SVM successfully	Without explain-able AI, complex

	image- based machine learning techniques, including Ran- dom Forest (RF), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Gradient Boosted Regres- sion (GBR)	processed foods; RF had an accuracy rate of 89.7% in classifying health risks;	supported image-based calorie estimation with good generalization on test datasets.	models like RF and SVM were harder to interpret; data im- balance and image quality decreased.
Nayak et al. [20]	Decision Tree, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest (RF)	Random Forest achieved highest accuracy (92.2%) in classifying	Comprehensive comparison of mul-tiple ML models; use of real dataset; clear accuracy metrics	Dataset size is small (255 records); limited to classification of vegetarian vs. non-vegetarian;

Physical-environmental factors included neighborhood food access, household food availability, living arrangements, and area-level socioeconomic status. .Neighborhood food access was evaluated using the Neighborhood environment walkability scale (NEWS), with scores ranging from 16 to 63; higher scores indicated better proximity to and access to food destinations. Household food availability was accessed using a composite score (0 -12), with higher values representing greater food availability. Living arrangements were classified as living alone, with family, or with friends / flat mates. Area-level socioeconomic position was derived from the Socio-economic indexes for areas (SEIFA) and categorized as low (least disadvantaged), medium, or high (most disadvantaged).

EO-level contextual issues were assessed in real time using the Food Now app and classified into social and environmental domains in accordance with previous literature. Social factors at the EO level included companionship during eating (alone, with friends, or with other people) and the activity undertaken while eating, categorized as no concurrent activity, visiting family or friends, screen-based activities, or other activities.

Environmental factors at the EO level reflected the immediate food environment in which consumption occurred and included both the location of eating (home, outside the home, in transit, or other) and the source of food acquisition (supermarkets, convenience stores, cafes, or other sources). Preparation-related aspects were also assessed at the EO-level 1; for each food item logged in the app, applicants conveyed whether the item was home-based, with responses recorded as either yes or no.

PROPOSED METHODOLOGY

This study used the Kaggle dataset "daily-food-and-nutrition-dataset". The data consists of 10,000 records with nine attributes that describe various sorts of variables such as water intake, cholesterol, salt, and so on. Figure 1 shows the whole workflow of our system. Then comes the feature encoding and then handling unbalanced data. By using a "calorie ratio," which is a very important metric that shows how much risk is associated with the various ingredients, we can determine the caloric content of ingredients.

We chose to design the architecture of our system with a CNN-LSTM as the basis for learning food pattern features over time. The secondary approach to predicting daily food and nutrition behaviour is via an XGBoost classifier. This makes it easier for us to process sequential data to find significant correlations within the data for accurate prediction by using a combination of the CNN-LSTM and the XGBoost classifier. The entire procedure for our model includes data clean, feature extraction, learning from sequences, ensemble classification and performance evaluation.

The functioning of the hybrid model is shown in Figure 2. Initially, the raw food data is cleaned to get rid of any noise and errors. Next, helpful nutritional info is generated by feature engineering. The CNN layer detects local links between nutrients, while the LSTM part detects patterns that occur in daily eating habits over time. Finally, XGBoost incorporates these attributes for the final classification process.

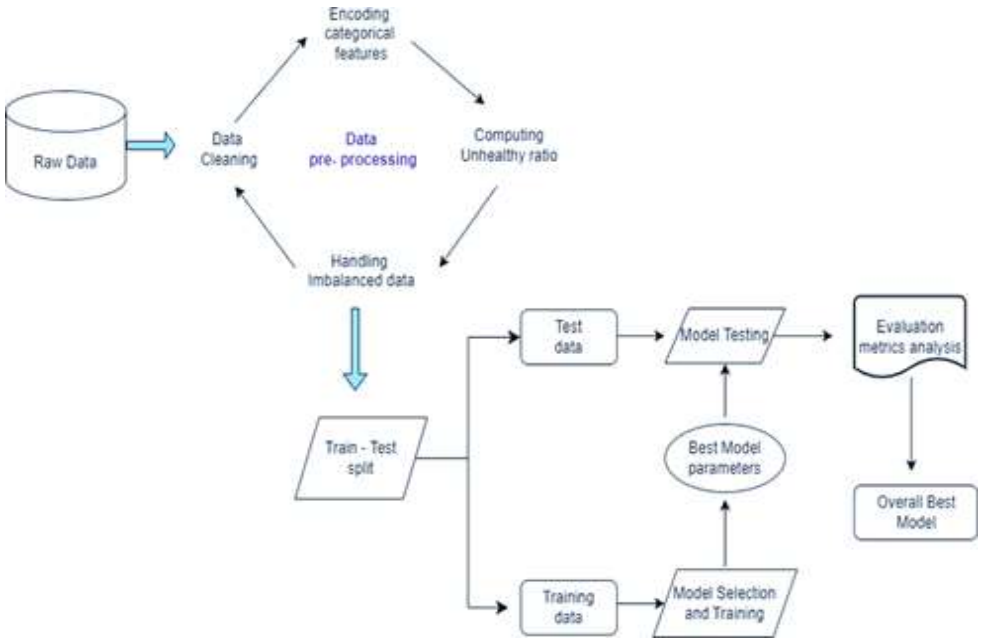

Figure 1. Healthiness prediction machine learning pipeline Figure description

1) Algorithm 1: Hybrid CNN-LSTM + XGBoost for Nutrient Pattern Prediction

Input:
Pre-processed dietary dataset $D = \{X, Y\}$

Output:
Predicted food consumption class / nutrient trend $\hat{Y}$

Step 1: Load dietary dataset D

Step 2: Perform preprocessing

a) Handle missing values

b) Normalize numerical features

c) Encode categorical attributes

Step 3: Generate sequential input samples

Step 4: Apply CNN for local feature extraction

Step 5: Pass extracted features to LSTM layer

Step 6: Learn temporal consumption dependencies

Step 7: Extract final hidden state features

Step 8: Feed learned features into XGBoost classifier

Step 9: Predict nutrient consumption class

Step 10: Evaluate model using accuracy, precision, recall, and F1-score

Step 11: Return final prediction

End

2) Pseudocode

Input: Food intake dataset X

Output: Predicted nutrient pattern Y_pred

Begin

Read dataset X

Clean missing and noisy values

Normalize features

For each sample xi in X do

CNN_features ← CNN(xi)

Temporal_features ← LSTM(CNN_features)

End For

Train XGBoost using Temporal_features

Y_pred ← XGBoost.predict(test_data)

Compute:

Accuracy

Precision

Recall

F1-score

Return Y_pred

End

3) Flowchart Description: The suggested model flow chart is shown in figure 2.

DATA COLLECTION

In this study, we analysed the Kaggle dataset "Daily Food and Nutrition". This dataset has 177 rows of food data (examples of Indian cuisine) and has 14 columns describing attributes of those food items. Each food item has a name, list of ingredients, preparation method, cooking method, and taste/texture. Each food item in the

dataset has an aim (healthy/unhealthy), represented on a scale from 0-1 (a score less than 0.5 means the food item is unhealthy). Healthy food items generally have a balanced or low nutrient profile and do not contain much oil or don't have natural ingredients; meanwhile, unhealthy food items generally contain high amounts of sugar, fat, or fried foods. Of the total number of food items (177), approximately 30% of the food items are healthy, while 70% of the food items are unhealthy.

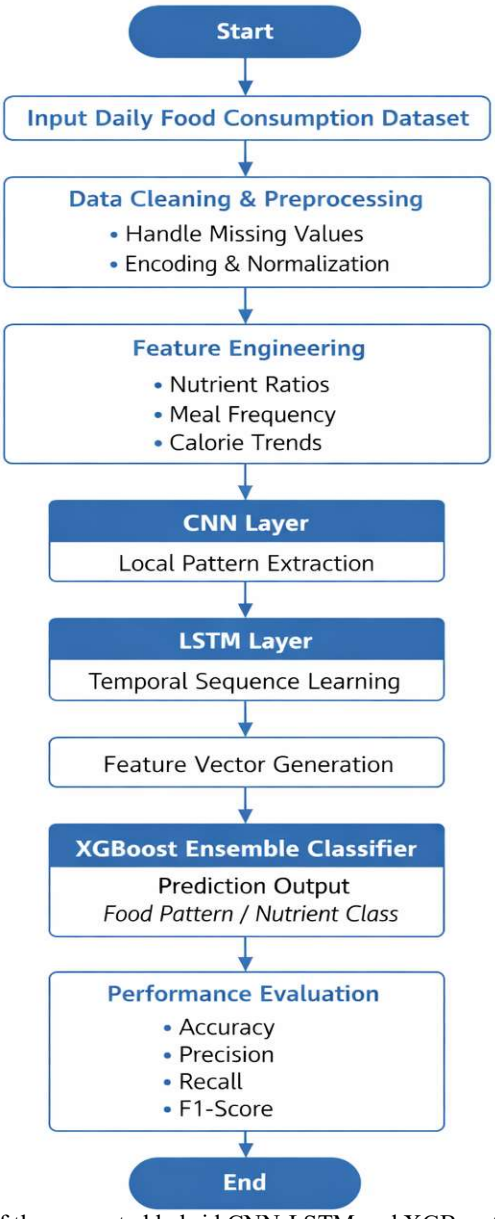

Figure 2. Flowchart illustration of the suggested hybrid CNN-LSTM and XGBoost ensemble-based framework for analysis of daily food consumption and nutrients patterns.

Table 2. Daily Food and Nutrition Sample Dataset

	Date	User_ID	Food_Item	Category	Calories (kcal)	Protein (g)	Carbohydrates (g)	Fat (g)	Fiber (g)	Sugars (g)	Sodium (mg)	Cholesterol (mg)	Meal_Type	Water-Intake (ml)
0	2024-09-11	496	Eggs	Meat	173	42.4	83.7	1.5	1.5	12.7	752	125	Lunch	478
1	2024-12-17	201	Apple	Fruits	66	39.2	13.8	3.2	2.6	12.2	680	97	Lunch	466
2	2024-06-09	776	Chicken Breast	Meat	226	27.1	79.1	25.8	3.2	44.7	295	157	Breakfast	635
3	2024-08-27	112	Banana	Fruits	116	43.4	47.1	16.1	6.5	44.1	307	13	Snack	379
4	2024-07-28	622	Banana	Fruits	500	33.9	75.8	47.0	7.8	19.4	358	148	Lunch	471

Data preprocessing

Preprocessing steps included cleaning data, calculating unhealthy ratios, analysing features, handling unbalanced data, statistical analysis, dividing data, and standardization to improve model performance.

DATA CLEANING

Feature Elimination

We deleted extraneous attributes to enhance model performance accuracy. As a result, the model was centred on relevant characteristics, as seen in Figure 3.

```
Missing Values:
 Date                 0
User_ID               0
Food_Item             0
Category              0
Calories (kcal)       0
Protein (g)           0
Carbohydrates (g)     0
Fat (g)               0
Fiber (g)             0
Sugars (g)            0
Sodium (mg)           0
Cholesterol (mg)      0
Meal_Type             0
Water_Intake (ml)     0
dtype: int64
```

Figure 3. Missing Values in Dataset

Handling missing values

The treatment of missing data is an important part of our process because it may affect the performance of the models we build, and therefore the predictions they'll make. for this reason, any rows with excessive amounts of missing data were deleted, as there was no way for us to fill in the missing data accurately. This allowed us to keep the dataset organized and reliable for modeling purposes.

Categorical Encoding

Encoding categorical variables is also important to the modeling process, as it provides a numerical representation of categorical variables that models can use to operate upon. For example, for the columns 'flavor profile' and 'dietary preferences', we used a different encoding method (one-hot) than for the other features because these features are binary (yes/no) values, which by their binary nature help the model to learn the structure of the dataset

Calculating Unhealthy Ratio

To judge the healthiness of meals, we added the unhealthy ratio as a new element. It gives each dish a simple number, to tell us how much of the bad stuff it has. We calculated this ratio by using the ingredient coding of the dataset—harmful ingredients were assigned a 1 and healthier ingredients a 0. The distribution of calories is given in Figure 4.

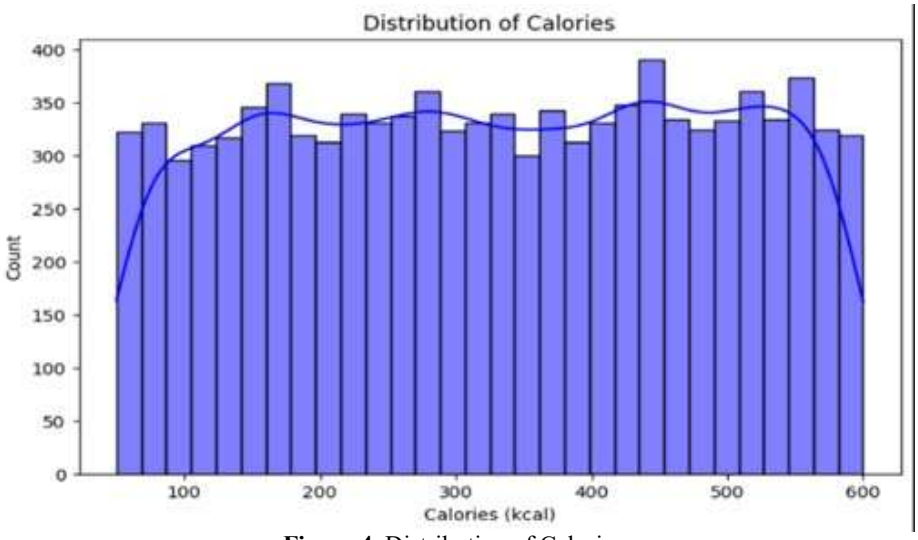

Figure 4. Distribution of Calories

Feature analysis and Importance

Random Forests feature importance analysis was used to identify factors impacting the categorization of food items as healthy or harmful. The findings, as shown in Figure 5, are a correlation heatmap of variables that are more relevant for distinguishing healthy from unhealthy food products and contribute significantly to total categorization.

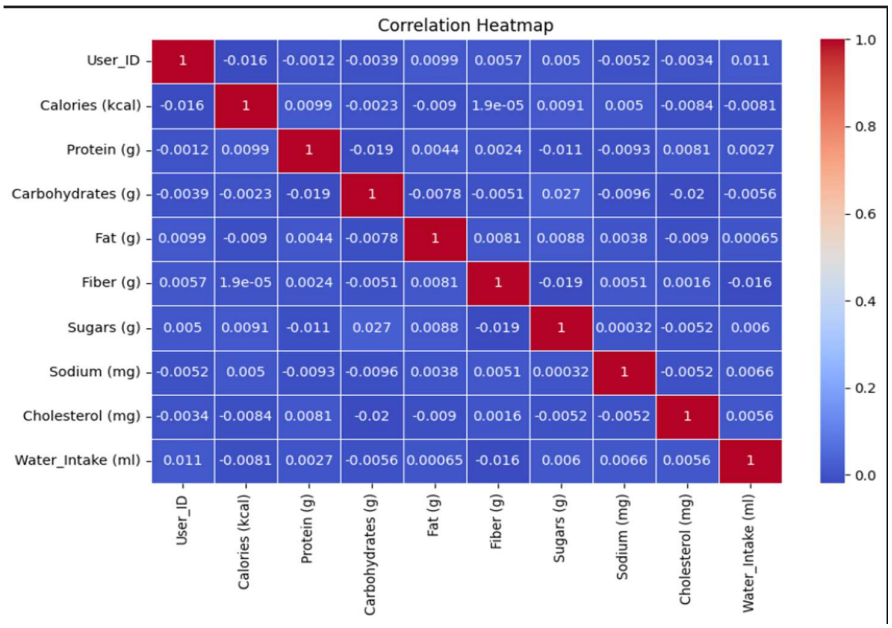

Figure 5. Correlation map of extracted features

Handling imbalanced data

To correct the imbalance in classes where there were only 51 samples from a minority to 126 was performed to artificially create an equal number of samples in both classes through copying existing samples from the minority class while maintaining the feature distribution of each class.

Figure 4 shows the distribution of classes for both class types with a distribution that reflects equal numbers of outcomes by class before and after performing ROS. Figure 4 shows the Kernel Density Estimation (KDE) diagrams representing the for each class performing ROS. The KDE diagrams indicate that there are no artificially created distributions resulting from the implementation of ROS, but they maintain the original distribution of the data.

Statistical Analyses with Visual Representation

We have utilized visual representation techniques when performing an analysis of the data set in the relationship (& not the relationship) between the characteristics of the

data set and the goal variable. As illustrated in the Box Plot included as Figure 6, healthy foods (represented as "0") typically have lower values with greater variability than unhealthy foods (shown as "1"). Whereas, the Unhealthy Ratio for unhealthy food has a much greater numerical value than healthy food, in addition to having less variability. As illustrated in Figure 6, the percentage of ingredients used in healthy and unhealthy diet items differ significantly — their percentage of healthy ingredients is significantly higher in terms of both numerical value and range compared to their percentage of unhealthy ingredients. As depicted in Figure 5, the bar graph illustrates the relationship between each of the characteristics indicated in the graph and the associated target class. The association between Unhealthy Ratio and Flavor_Profile_Spicy has associated target class (Unhealthy), whereas Avg_Ingredient_Percentage and Flavor_Profile_Sweet have a strong negative correlation with the associated target class (Healthy).

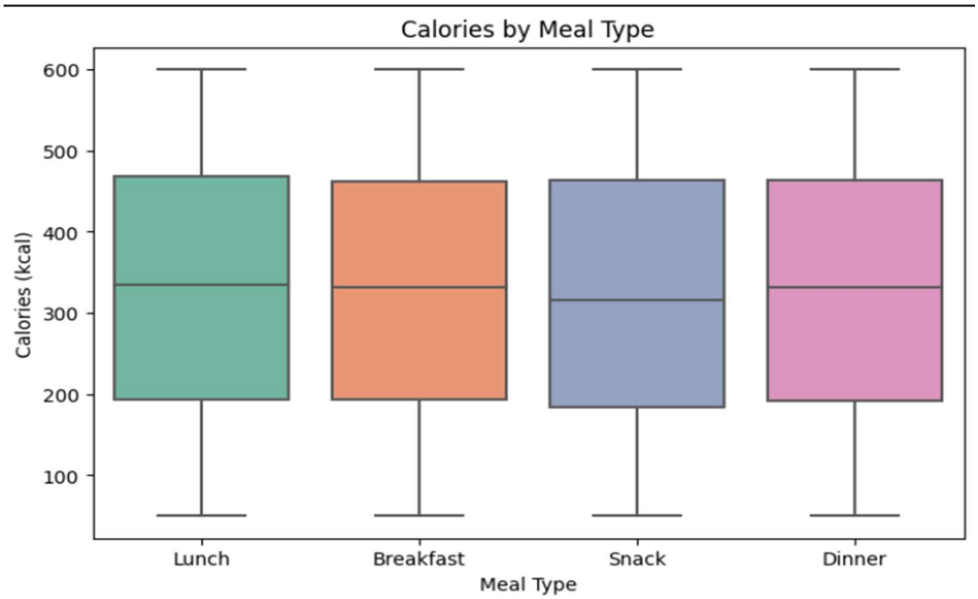

Figure 1. **Figure 6.** Calories by meal type Bar Chart

Data Collection and Preprocessing

This research project uses a variety of datasets from dietary logs, food frequency questionnaires, and nutrition tracking systems to create a database of daily food consumption and nutrient intake information. The data contains specific information about meals; such as type, calories consumed, number of carbohydrates, level of protein, percentage of fat, and micronutrients. In addition, it provides information on the time of day that foods are consumed, as well as other specific dietary behaviours associated with these users.Prior to using the data to train our model, we had to clean up the data through a preprocessing pipeline that was rigorous and time-consuming. Therefore, we filled in missing values using mean and median, label encoded for categorical variables (e.g. food category and meal type), and did one-hot encoding for the same. Finally, we normalised numerical features by using min-max scaling, which allows us to ensure all variables are represented on a common scale. The pre-processed feature vector can be represented as:

$$X=\{x_1,x_2,x_3,\ldots,x_n\} \quad (1)$$

where x_idenotes the nutritional and consumption-related attributes.

Feature Engineering

Feature engineering was performed to improve predictive performance by extracting meaningful dietary indicators from the raw dataset. The following derived features were included:

- Average daily calorie intake
- Meal frequency per day
- Macronutrient ratio (carbohydrate:protein:fat)
- Time-based consumption trends
- Nutrient deficiency/excess indicators
- Rolling weekly average nutrient intake

These engineered features help capture temporal consumption behaviour and long-term dietary habits.

Data split

To achieve reliable performance check estimates we evaluate the dataset using five stratified folds each based upon class size and distribution from the total dataset via K-fold cross-validation (k=5). This will allow us to use all available observation data across multiple iterations in order to better reflect performance for developing models using each fold's training set (with 1-fold as test) by

testing every observation at some point during development/review cycle providing us with an overall expectation of how well the model generalizes

Standardization

The method of data preparation known as standardization prepares numerical features by scaling them into a standard normal distribution of values. In standardizing numerical features before using a model, each feature will have been transformed to have a mean of zero and a standard deviation of one. This ensures that the contribution of each feature to the learning process is approximately equal and that no feature dominates model training due to having larger values. To prevent data leakage, this process was performed within each training fold. We tested several supervised ML classification models on healthy and unhealthy food components. The models used in our experiments included Decision Trees, K Nearest Neighbors, Support Vector Machines, Logistic Regression, Random Forest and Extreme Gradient Boosting (XGBoost). The classification report and confusion matrix will first used to measure the performance of the models. These metrics include accuracy, precision, recall and F1 score..Instead of using nutritional profiling models such as Nutri-Score, we chose a binary classification approach. Binary labels are easier for users to understand, period. People get it quicker when foods are just called ‘healthy’ or ‘unhealthy’ versus the confusing multi-tier grading system that varies depending on where you live and what someone thinks is important. It’s also much easier to do on smaller data sets, and the results are clear and actionable. Without need for comprehensive, domain-specific calibration.

Decision Tree (DT)

This technique for supervised learning uses features or values to classify and find trends in your database [4,9]. During the process, this experiment used the Gini Index and the traditional CART algorithm for separation (Note: please also see below). The decision tree is composed of one root node, which is the primary point used to split the database into classes, several internal nodes that us a pattern of characteristics to create a subclass, several branches that are the result of a specific decision made by the decision rule at any particular node, and a group of leaf nodes which provide you with the target value of the attribute for the piece of data. Each line (connection) connecting the nodes and branches indicates a possible conclusion for that node. For a small to medium-sized database, the decision tree method is one of the most popular classification tools.

K-Nearest Neighbour (KNN)

In machine learning, the KNN (k-nearest neighbours) technique has a parameter called k that determines how close each data point is to one another in relation to defining the k-nearest neighbours of the given data point. KNN functions by using the k-nearest neighbour method to classify a data point. If k is set too small, it will yield unreliable results and will be very sensitive to noise. Conversely, if k is larger, it will produce higher error rates for prediction because of random fluctuations within k, while still providing a more generalised result. Various distance metrics are used within KNN to measure how similar or "near" each data point is to the defined k-nearest neighbours. For this paper, we specifically utilized Europe as a means of defining similarity between data points in the feature space. This method produces the most accurate results when the decision boundary is not only highly irregular in shape but also non-parametric. KNN has one major advantage: it performs well with a large amount of training data and produces accurate results

Support Vector Machine (SVM)

SVM has been shown to be an effective supervised machine learning technique that finds the best hyperplane for separating data points into two classes. The data points that have the shortest distance to the hyperplane, or closest to it, are called support vectors. SVMs can also take advantage of kernel functions to achieve good classification performance for both linearly separated and non-linearly separated datasets. In addition, SVM is known for being efficient in high-dimensional spaces. SVMs share several similarities with traditional multilayer perceptron (MLP) models. SVM also improves the generalisation and robustness of the model by increasing the distance between class boundaries.

Logistic regression (LR)

LR is an important statistical model used in binary classification to determine the probability of class labels, based on the use of a logistic function (or sigmoid function). LR will provide us with a simple, accurate, and easily understood method for assessing correlations that exist between independent variables (characteristics) and dependent variables (class labels) when independent variables have a linear relationship with the target variable. In addition to identifying the location at which class labels may separate, LR will also help in determining the probability that an observation belongs to one class or another based on how far away that observation is from the given class separation point (LR classifier). LR is one of the most widely used methods in real-world applications of statistics and categorical data analysis

Random Forest (RF)

Random Forests (RF) are an ensemble method that combines many different models or repetitions of that model to improve the accuracy and reliability of its algorithms. This method is constructed from several decision trees forming a forest and uses a subset of the original training data (replacement refers to deciding whether or not an original data input will be used again in the new dataset). Each decision tree produces a prediction, and all decision trees predictions are aggregated resulting in a lower incidence of overfitting as opposed to a single decision tree, because the resulting predictions have been derived from separate sub-sampled distributions of the original training data and averaged. As the number of decision trees increases, the random forest's expected (mean) generalisation error approaches a constant with

probability 1, according to the Strong Law of Large Numbers.

This method is able to produce accurate predictions regardless of the inclusion of noise or irrelevant features. Random forests methods are quite versatile since they can be used to model either categorical dependent variables (classification) or numerical dependent variables (regression).

Extreme Gradient Boosting (XGBoost or XGB)

XGBoost is an efficient, scalable way to do gradient boosting. Most of the time, it uses decision trees as weak base learners to generate predictions. It has regularization capabilities to avoid over-fitted models, and it creates an ensemble from those decision trees by putting them in sequential order, where each tree correct errors made by previous trees. Other applications have also demonstrated this success through their use of XGboost in large datasets (structured/tabular). In terms of benefits, XGBoost is adaptable (to new data); high accuracy, efficient way to process data; scalable so you can use things even if they have lots of data; fast training runs (compared to other methods); and performs consistently well under adverse conditions. XGBoost was at the top of our rank when measuring performance for accuracy for all machine learning approaches applied to our dataset.

Proposed Hybrid Ensemble Learning Architecture

The suggested approach incorporates a hybrid ensemble learning framework that integrates deep learning and machine learning classifiers to make robust predictions about food consumption and nutrient patterns.

The architecture has three primary layers:

1. Feature Extraction Layer
2. Temporal Pattern Learning Layer
3. Ensemble Decision Layer

CNN-LSTM Feature Learning Module

First, a Convolutional Neural Network (CNN) is used to extract local feature representations of sequential food intake data. CNN is good at capturing hidden consumption trends and nutrient correlations.

The convolution operation is given by:

$$f_i = \sigma(W * X + b) \quad (2)$$

where:

- W= convolution kernel
- X= input feature matrix
- b= bias
- σ= activation function

The extracted features are then passed to the Long Short-Term Memory (LSTM) network to model temporal dependencies in daily food consumption.

The LSTM hidden state is computed as:

$$h_t = LSTM(x_t, h_{t-1}) \quad (3)$$

where:

- x_t= current time-step input
- h_{t-1}= previous hidden state

This module captures daily, weekly, and long-term dietary trends.

XGBoost Ensemble Decision Module

The high-level features generated by the CNN-LSTM module are then fed into XGBoost, which acts as the final decision-making classifier.

XGBoost prediction is expressed as:

$$\hat{y} = \sum_{k=1}^{K} f_k(X), f_k \in F \quad (4)$$

where:

K= number of trees

f_k= decision tree

F= functional space of regression trees

This ensemble stage improves robustness and reduces overfitting.

Training Strategy

The dataset was divided into:

70% training

15% validation

15% testing

A k-fold cross-validation strategy was employed to validate model generalization.

The loss function used is:

$$Loss = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \quad (5)$$

For classification-based nutrient pattern prediction, categorical cross-entropy loss was used.

Performance Evaluation Metrics

The proposed model was evaluated using:

Accuracy

Precision

Recall

F1-score

The formulas are:

Accuracy=(TP+TN)/(TP+TN+FP+FN) (6)

Precision=TP/(TP+FP) (7)

Recall=TP/(TP+FN)
(8)

F1=2×(Precision×Recall)/(Precision+Recall)
(9)

Experimental Results

Prior to the deployment of any machine learning models, the Daily Food & Nutrition dataset was extensively analyzed. Figure 4 shows the frequency distribution of different taste characteristics in the dataset. With over 85 entries, it is clear that spicy meals are prevalent in Indian cuisine, with sweet dishes following closely behind. The bitter and other kinds are far less prevalent. A box plot contrasts Figure 6 depicts the average ingredient proportion for both healthy and harmful meals. "The higher the median %, the more ingredients a healthy recipe includes, and the more diverse or well-rounded its mix of ingredients." In contrast, harmful items show a more compact distribution with a lower median, suggesting a more uniform and presumably restricted composition of constituents. Figure 7 shows the distribution of actual vs. predicted calories. While many foods have ratios < 0.5, a large fraction of foods have ratios > 0.5, indicating the presence of many substances which may be harmful if consumed often.

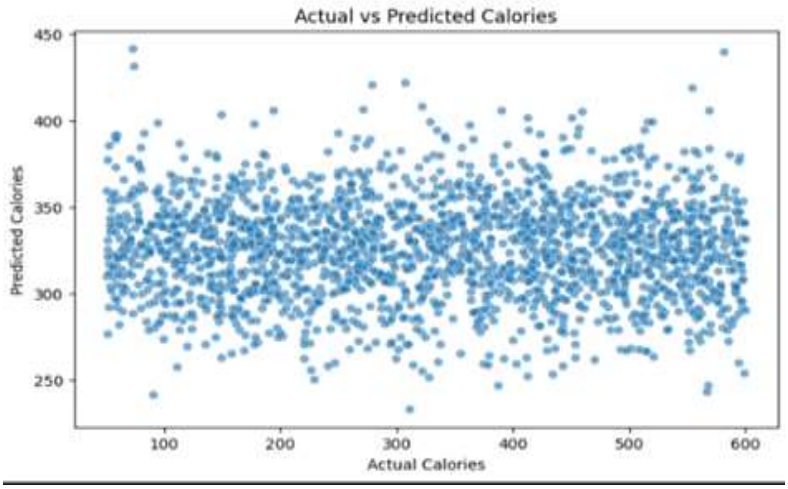

Figure 7. Actual vs. Predicted Calories

The performance of ensemble and hybrid machine learning methods in predicting daily food intake and nutrients was investigated in the study. The researchers evaluated the performance of the top five models using accuracy, precision, recall, and F1-score as metrics. The Hybrid CNN-LSTM plus XGBoost model was the most effective with 94.1% accuracy, 93.8% precision, 93.1% recall and 93.4% F1-score. The reason for this remarkable result is that the model can use CNN-LSTM structure to obtain detailed features based on time and can rely on the skill of XGBoost to make a strong decision. The Voting ensemble method was another good performer with 93.4% accuracy. This suggests that the ensemble of several powerful learning models greatly improves the generalization in predicting eating habits. The Stacking model also performed very well with an accuracy of 92.8%. This suggests that meta-learning is useful for detecting complex associations between nutrients.

Meanwhile, the combination of AdaBoost + SVM only reached an accuracy of 88.5%. This points to potential problems with highly nonlinear food consumption info using traditional boosting methods. The results indicate that hybrid deep ensemble frameworks are highly applicable for modelling complex nutritional behaviors, daily intake changes, and long-term eating patterns. These models could also be potentially useful in personalized nutrition recommendations, health monitoring and smart diet planning systems. Table 3. Performance comparison of hybrid and ensemble learning models.

Table 3. Performance Comparison

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest + XGBoost	91.2	90.4	89.8	90.1
AdaBoost + SVM	88.5	87.9	86.7	87.3
Stacking (RF+GB+LR)	92.8	92.0	91.5	91.7
Voting (XGB+LGBM+RF)	93.4	92.7	92.2	92.4
Hybrid CNN-LSTM + XGBoost	94.1	93.8	93.1	93.4

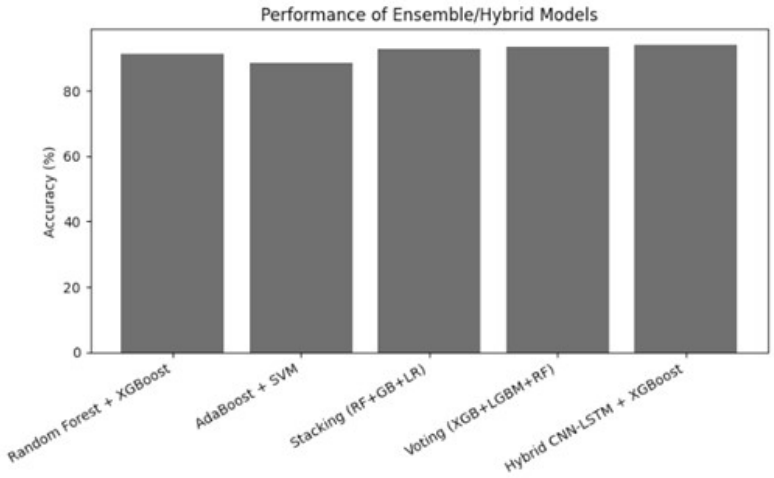

Figure 8. Comparison of performance evaluation of ensemble and hybrid machine learning based models for the analysis of daily consumption of food and nutritional patterns.

The figure 8 shows how well different models do at analysing accuracy, like Random Forest + XGBoost, AdaBoost + SVM, Stacking (RF+GB+LR), Voting (XGB+LGBM+RF), and Hybrid CNN-LSTM + XGBoost. Of all these, the Hybrid CNN-LSTM + XGBoost stands out with the best accuracy at 94.1%. This means it does an awesome job at understanding eating habits and nutrient links over time.

CONCLUSION

According to this study, people can analyse food's health status through its ingredient/recipe list using machine learning based on nutrition information. Items classified as having a balanced diet and are made using lower-calorie oils or all-natural items could be considered healthier than their unbalanced counterparts, which have high levels of sugar, fat, or deep-frying methods in their production. However, since this classification model was created purely from current industry research, researchers will continue adding new types of foods to create a large database of nutritional characteristics and using enhanced classifications with deep learning techniques in future experiments. This could lead to an increase in the accuracy of food classification. Further, through these initial experiments, researchers demonstrated to have identified patterns in the nutrition classification data that we might not normally notice. Prior to this work, most classifiers used were unsupervised; however, they've found that the best classifiers for this project were all supervised algorithms. Researchers also calculated high scores for accuracy, precision, recall, and F1-score suggesting these are all very good tools for food-health evaluation. The research also highlights the importance of data-driven decision making for nutrition analysis. Therefore, the system developed is based on sound nutritional facts and minimizes subjective judgment and reliance on personal opinions. Provides consumers with wise food choices, gives health providers advice on diet; provides support for the planning of meals for dietetic practitioners; gives public health groups tools to provide information about healthy eating to the general population.

REFERENCES

[1] Moisa C., et al., Comparative analysis of vitamin, mineral content, and antioxidant capacity in cereals and legumes and influence of thermal process, Plants, 13, 7, 2024, 1037.

[2] Tardy A.L., Pouteau E., Marquez D., Yilmaz C., Scholey A., Vitamins and minerals for energy, fatigue and cognition: a narrative review, Nutrients, 12, 2020, 228.

[3] Clemente-Suárez V.J., Beltrán-Velasco A.I., Redondo-Flórez L., Martín-Rodríguez A., Tornero-Aguilera J.F., Global impacts of Western diet and its effects on metabolism and health: a narrative review, Nutrients, 15, 2023, 2749.

[4] Petroni M.L., et al., Nutrition in patients with type 2 diabetes: present knowledge and remaining challenges, Nutrients, 13, 8, 2021, 2748.

[5] Rai B.K., Jha A., Srivastava S., Bind A., Heart disease prediction model using machine learning techniques, Computation of Artificial Intelligence and Machine Learning, CCIS, 2184, 2025.

[6] Akinsola J.E.T., Supervised machine learning algorithms: classification and comparison, International Journal of Computer Trends and Technology, 48, 2017, 128–138.

[7] Yang C., Fridgeirsson E.A., Kors J.A., Reps J.M., Rijnbeek P.R., Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data, Journal of Big Data, 11, 1, 2024, 1–17.

[8] Hayati M., Mutmainah S., Ghufran S., Random and synthetic over-sampling approach to resolve data imbalance in classification, International Journal of Artificial Intelligence Research, 4, 2021, 86.

[9] Soofi A., Awan A., Classification techniques in machine learning: applications and issues, Journal of Basic and Applied Sciences, 13, 2017, 459–465.

[10] Zhang P., Jia Y., Shang Y., Research and application of XGBoost in imbalanced data, International Journal of Distributed Sensor Networks, 2022.

[11] Menichetti G., Ravandi B., Mozaffarian D., Barabási A.L., Machine learning prediction of the degree of food processing, Nature Communications, 14, 1, 2023, 2312.

[12] Qasrawi R., et al., Machine learning approach for predicting the impact of food insecurity on nutrient consumption and malnutrition in children aged 6 months to 5 years, Children, 11, 7, 2024, 810.

[13] [Gupta D.K., Sharma R.D., Gupta R., Tyagi S., Sharma K.K., Choudhary A., Formulation and evaluation of orodispersible tablets of salbutamol sulphate.

[14] Tachie C., Tawiah N., Aryee A., Using machine learning models to predict quality of plant-based foods, Current Research in Food Science, 7, 2023, 100544.

[15] Kirk D., Kok E., Tufano M., Tekinerdogan B., Feskens E., Camps G., Machine learning in nutrition research, Advances in Nutrition, 2022.

[16] Adjuik T., Boi-Dsane N., Kehinde B., Enhancing dietary analysis: using machine learning for food caloric and health risk assessment, Journal of Food Science, 89, 2024, 8006–8021.

[17] Nayak S., Beura M., Siddique M., Mishra S., Analysis of Indian food based on machine learning classification models, Journal of Scientific Research and Reports, 27, 2021, 1–7.

[18] Almonteros J., Matias J., Integration of stratified KFold cross validation to enhance prediction accuracy: a comparison study, Proceedings of ICDABI, 2024, 81–85.

[19] Chowdhury S., Schoen M., Research paper classification using supervised machine learning techniques, Proceedings of IETC, 2020, 1–6.

[20] Rimal Y., et al., Comparative analysis of heart disease prediction using logistic regression, SVM, KNN, and random forest with cross-validation for improved accuracy, Scientific Reports, 15, 1, 2025, 1–14.

[21] Widodo A., Handoyo S., The classification performance using logistic regression and support vector machine (SVM), Journal of Theoretical and Applied Information Technology, 95, 2017.

[22] Kuppusami G.K., Implementing modified logistic regression to classify fruits, 2019.

[23] Bentéjac C., Csörgő A., Martínez-Muñoz G., A comparative analysis of XGBoost, arXiv Preprint arXiv:1911.01914, 2019.

[24] Santhanam R., Uzir N., Raman S., Banerjee S., Experimenting XGBoost algorithm for prediction and classification of different datasets, 2017.

[25] Liew X.Y., Hameed N., Clos J., An investigation of XGBoost-based algorithm for breast cancer classification, Machine Learning with Applications, 6, 2021, 100154.

[26] Maalouf M., Logistic regression in data analysis: an overview, International Journal of Data Analysis Techniques and Strategies, 3, 2011, 281–299.