

# IngriGen - A Tool to pick Ingredients

Karan Navin Javali

**Abstract**—IngriGen, a tool for ingredient selection, integrates nutrition data analysis and machine learning. This paper synthesizes insights from the food nutrition data, mapping different relationships between the macros selected and using them to gain new insights and predictions. By utilizing the USDA National Nutrient Database, IngriGen aims to inspire new recipes tailored to specific dietary needs, exploring the relationships between ingredients and nutritional goals, one of the ways being by predicting the fat content of food. Fat is an essential nutrient required by the human body, but in limited amounts. While trying to achieve body goals, it is important to note the fat content intake to prevent any health issues.

**Index Terms**—new recipes, fat content, ideal ingredients

Template by the IEEE Computer Society

## 1 INTRODUCTION

In the vast culinary landscape, creating new recipes often involves navigating through a multitude of ingredients. IngriGen addresses this challenge by providing a systematic approach to ingredient selection. This project taps into nutrition data to predict ideal ingredient combinations based on diverse requirements, such as calorie content, weight goals, and body goals. One of the ways it has been achieved is by predicting the estimated fat content of the food resulting by using various ingredients. The key elements used for calculation are the estimated carbohydrates, protein, and calorie of the meal. I have tried using a few models to get the result, however, I found out that Random Forest works best in this scenario. The error is the least and the fat content predicted is relatively accurate.

Fat content estimation plays a crucial role in the creation of a diet plan because dietary fat is a key macronutrient that significantly influences overall health and well-being. Here are several ways in which fat content estimation contributes to the development of an effective diet plan: [2] Dietary fats play a crucial role in human nutrition, serving multiple functions. On one hand, fats serve as a vital energy source, facilitate the absorption of fat-soluble vitamins, and act as essential structural components of cell membranes. Conversely, an elevated intake of fats is linked to health concerns such as obesity, type 2 diabetes, cancer, and coronary heart disease. Attention is frequently directed towards animal fats, characterized by a significant proportion of saturated fatty acids, when addressing strategies to reduce overall fat consumption.

To further emphasize the significance of fat in the diet, it is mentioned how it promotes general health. Good fats, found in foods like nuts and avocados, support heart health, improve mental clarity, and facilitate the absorption of important fat-soluble vitamins, such as A, D, E, and K. Furthermore, a diet that includes healthy fats in a balanced manner is linked to better metabolic health and weight management. Understanding the subtle differences between different forms of fat enables people to make well-informed dietary decisions that support their long-term health goals as well as their nutritional

needs.

To summarize, fat content prediction is a valuable tool for nutritionists and individuals alike when creating a diet plan. This aligns well with the required intake of energy, carbohydrate, fiber, fat, fatty acids, cholesterol, protein and amino acids, which are other important macros required by the human body [10]. It allows for the customization of dietary recommendations based on individual health goals, preferences, and specific nutritional requirements. Understanding the culinary domain is essential for the success of IngriGen. By incorporating nutrition data from sources such as the USDA National Nutrient Database, which has a comprehensive dataset with the latest information of the ingredients, we gain valuable insights into the nutrient composition of foods, especially American food items.

We can pose an important question here - Which foods are healthy sources of fat? [1] While the general public is well-informed about overall dietary fat intake, there is a lack of awareness regarding the significance of fat quality and the diverse origins of dietary fats. Notably, foods like pizza, grain-based desserts, and various chicken dishes rank prominently among the primary sources of fats in the U.S. diet.

The following are the research objectives -

- To perform exploratory visualization work on the nutrition data to describe trends within ingredients, including calories per unit weight, fat content, and vitamin content, organized by categories such as meat, vegetables, oils, and flavorings.
- To apply a machine learning technique, either classification or regression, to predict the ideal ingredients and quantities for different requirements, such as calorie content, weight goals, and body goals, based on the nutrition data.
- To defend the model used for ingredient selection and quantity prediction, ensuring its reliability and accuracy.

To assess the ability of the model to provide insights into the relationships between different ingredients and their

impact on different weight groups and goals, thereby generating knowledge to assist in creating new recipes for various target audiences.

### 1.1 Related Work

A study was conducted to predict the body fat content of a human being [3]. Both high and low fat content in the body could lead to several abnormalities. This study employs feature extraction methods - Factor Analysis, Principal Component Analysis, and Independent Component Analysis - coupled with machine learning models to predict body fat percentage. Evaluating real-world datasets, the aim is to assess and compare the efficacy of different feature extraction methods for body fat prediction, providing insights and a baseline for future research. The study also explores optimal feature selection while balancing accuracy and efficiency. We can use this study as an extension to this study. It may be possible to create a more comprehensive diet plan based on the body fat percentage goals, such as high body fat percent to average body fat percent, and map the exact macros required for this. This both studies could be combined sequentially (calculate body fat, then calculate the required amount of fat content in the food to reach the goal), or create a whole new methodology which implements parts of both the studies

Another study focuses on the metabolizable energy content and the digestibility of fiber in the diet. [4] While the primary focus of the study is on energy content and the influence of fiber, the findings have implications for predicting fat content in diets. The observed decrease in fat digestibility with increased fiber intake suggests that dietary fiber can be a factor affecting fat metabolism. This information may contribute to the development of models or algorithms for predicting the fat content of diets based on their fiber composition. Additionally, the use of an empirical formula for ME (Metabolizable Energy) prediction provides a practical approach that could be adapted or extended in the context of fat content prediction models.

Another study focuses on the application of deep learning in food. [5] While the study doesn't directly address fat content prediction, it provides a foundational understanding of deep learning in the food domain. Researchers in the field of nutrition and food science can use the knowledge and methodologies presented in this paper as a starting point for developing models specific to predicting fat content in foods, leveraging the success and insights gained from deep learning applications in related areas.

One of the studies emphasizes the process of the inclusion of sustainability in the Italian Dietary Guidelines (IDGs). [6] It underscores the interconnectedness of dietary choices, sustainability, and health. The emphasis on plant-based diets, reduction of animal food, and strategies to minimize food waste provides a foundation for considering the implications of these recommendations on fat content in the broader context of sustainable and healthy eating patterns. The study encourages a holistic approach to dietary guidance that can inspire further research into predictive models encompassing both health and sustainability aspects.

Another study systematically reviewed literature on

dietary patterns, examining their relationship with nutrient adequacy, lifestyle factors, demographics, and health outcomes [7]. Two main methods were identified in previous reports: diet indexes assessing compliance with dietary guidance, and data-driven methods using factor or cluster analysis. Regardless of the approach, patterns emphasizing fruit, vegetable, whole grain, fish, and poultry consumption were associated with positive outcomes, including micronutrient intake and biomarkers. Positive predictors of healthier dietary patterns included age, income, and education. While healthful patterns were inversely linked to all-cause mortality and cardiovascular disease risk, the risk reduction was modest and attenuated after controlling for confounders. Limited studies found associations between incident cancers and dietary patterns. Both approaches have limitations, are prone to measurement errors, and haven't introduced new diet-disease hypotheses. By leveraging insights from this study, fat content prediction models can be enriched, providing a more comprehensive and health-conscious perspective on food choices. This integration can contribute to the development of dietary recommendations that not only consider fat content but also align with broader patterns associated with positive health outcomes.

The Food Guide Pyramid, created by the USDA, is a visual tool aligned with the Dietary Guidelines for Americans. [8] Unlike earlier guides, it addresses both adequacy and moderation in dietary choices. It promotes higher intake of vegetables, fruits, and grains, with a focus on dark-green leafy vegetables, legumes, and whole grains. These foods offer essential nutrients, complex carbohydrates, and dietary fiber while being generally low in fat. Analysis of nutrient levels in the recommended diet patterns underscores the significance of whole-grain products, especially in lower-calorie diets, where it's advised that half the grain servings be whole grains. Regardless of calorie levels, it's recommended to include several servings of whole grains daily. Current intakes of vegetables, fruits, and grains fall below recommended levels. The pyramid graphic effectively communicates the need for increased consumption of these food groups for a healthful diet. This study provides foundational insights into dietary guidelines, moderation, and the importance of specific food groups. This information can be integrated into predictive models to enhance their relevance and effectiveness in promoting healthful dietary patterns, potentially influencing fat consumption in the broader context of balanced nutrition.

## 2 EXPLORATORY DATA ANALYSIS

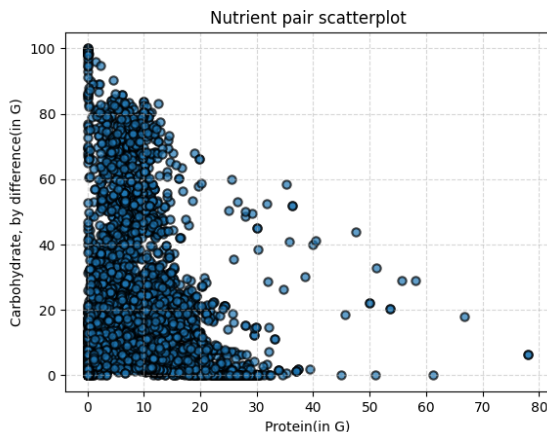
The dataset was used from the website of United States Department of Agriculture [9]. It has a comprehensive dataset of the basic ingredients, as well as the various products sold in the USA. I mainly focused on the retrieval of the basic ingredients. This was done by referring to the FNDDS dataset - FNDDS is a database that provides the nutrient values for foods and beverages reported in What We Eat in America, the dietary intake component of the National Health and Nutrition Examination Survey. It contains nutritional information of the basic ingredients. Dataset field meanings -

- `fdc_id` - ID of the food in the food table
- `food_code` - A unique ID identifying the food within FNDDS
- `wweia_category_number` - Unique Identification number for WWEIA food category to which this food is assigned.
- `start_date` - Start date indicates time period corresponding to WWEIA data
- `end_date` - End date indicates time period corresponding to WWEIA data

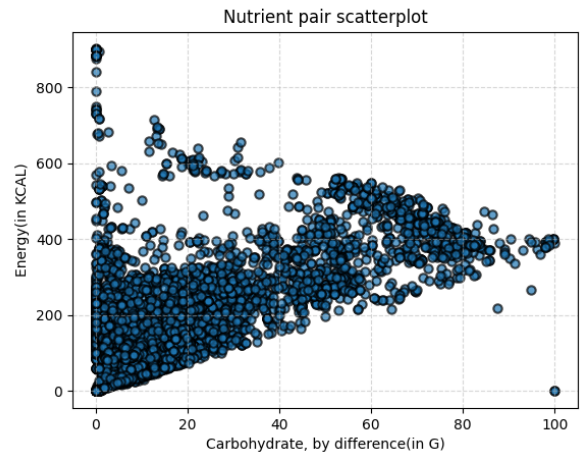
Example of raw data and it's interpretation -  
 "2340775","11121100","1002","2019-01-01","2020-12-31"

The row represents an entry of `fdc_id`, `food_code`, `wweia_category_number`, `start_date`, and `end_date`. The main identifier here is the `fdc_id`, which is the id of the ingredient. The other data is related to other databases - `food_code` (A unique ID identifying the food within FNDDS), `wweia_category_code` (Unique Identification number for WWEIA food category to which this food is assigned). The `start_date` and `end_date` indicates time period corresponding to WWEIA data. For the use of this project, we need the `fdc_ids` of the ingredients, after which we are calling an API from the USDA to get our relevant data. Here, "2340775" is the `fdc_id` we need. Upon calling the API multiple times, the data was processed to store the nutritional information of the food items, as well as the reference to the actual food items and nutrient codes.

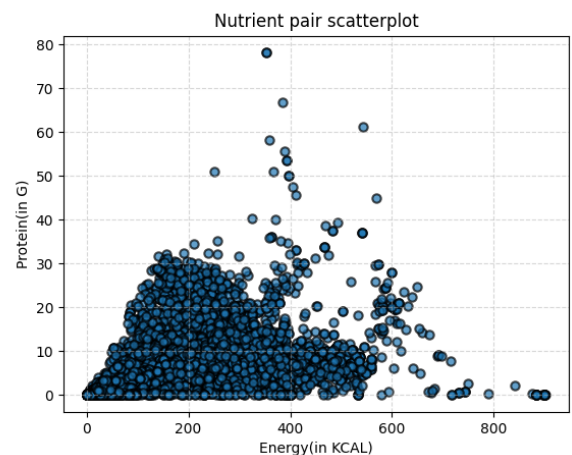
A few plots were created using the macros such as protein, carbohydrates, fat and calorie content of food. The trends of the macros in different foods were also noted based on the plot.



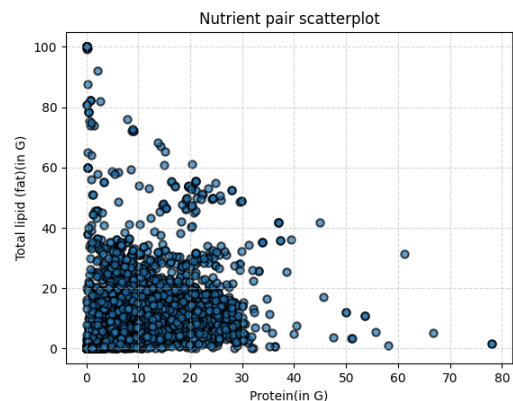
Protein - Carbohydrate plot: Most of the ingredients with high carbohydrates have low protein. There are very few ingredients with low carbohydrate content and high protein content.



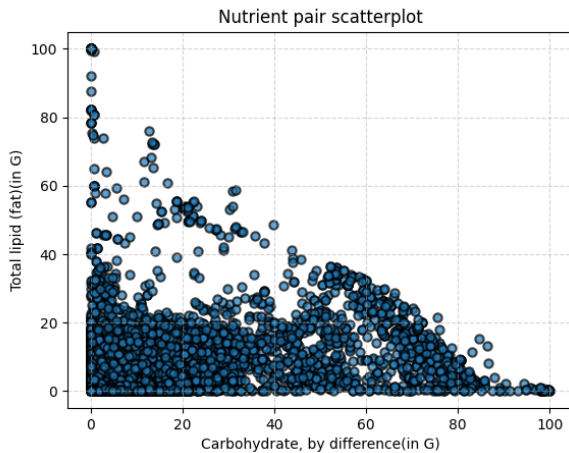
Energy - Carbohydrate plot: There seems to be a threshold for amount of energy present with respect to carbohydrates (a clear line is seen above which most values are concentrated). There are very few outliers which do not obey this threshold (high carbohydrate with low energy).



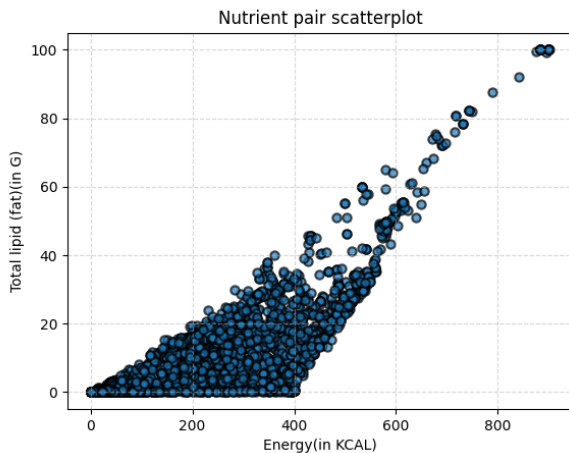
Protein - Energy plot: There seems to be no relationship of protein and energy amounts except that there is an upper limit for the amount of protein with respect to energy (can be seen as the top slope of the plot).



Protein - Fat plot: There is no relationship between the amount of protein and fat in the ingredient. The values seem to be random. It can be seen that there is no ingredient with high amount of fat and protein.



Carbohydrate - Fat plot: Most ingredients are concentrated in 0-40 gm (fat) range. It can be seen that there is no ingredient with high amount of fat and carbohydrates.



Fat - Energy plot: With increase in fat, there is an increase in energy. The values are restricted between 2 slopes.

### 3 METHODOLOGY

Few models were used to predict fat content. Techniques such as regression (linear, LASSO, Ridge), decision trees, SVM, neural networks, random forest were used. After experimenting with the data and after selecting the calorie, carbohydrate, and protein content of the foods, the prediction of fat content was possible. The evaluation metrics used were Mean Squared Error and R-Squared. Due to the nature of the data (non-linear and feature interactions), random forest was the best model which yielded the least error. Because Random Forest is an ensemble technique that combines several Decision Trees trained on arbitrary subsets of features and data, it is especially useful for managing non-linear data. The model can now capture intricate linkages and interactions between predictors and the target variable thanks to this method. Because Random Forest reduces overfitting through randomization and Decision Trees' expressiveness and flexibility, it is a good choice for situations where non-linear patterns are common. Its efficacy in capturing complex non-linear correlations in

datasets is attributed to its implicit feature engineering, high-dimensional spatial handling capabilities, and resilience to noisy data. Nonetheless, it's critical to take into account features unique to the dataset and use the right metrics to assess the model's performance.

### 4 RESULTS AND DISCUSSION

Following are the resulting metrics when using Random Forest

- Random Forest Model MSE: 1.055
- Random Forest Model R-squared value: 0.9908

It would seem that it is possible to get a close estimate of fat content in resulting food items. Most of the research objectives were achieved. I have displayed the trends of macros within ingredients. I tried applying classification and regression, however the results were very poor. Hence, I was forced to change my approach. I have tried to reduce the error as much as possible. Upon using the model, various trends can be seen in the macros. In general, increase in calorie content leads to an increase in fat content of food. An increase of carbohydrates causes a decrease of fat content. An increase of protein content also causes an increase in total fat. The decrease caused by carbohydrates is much less compared to the increase caused by the other 2 macros.

### 5 CONCLUSION

In conclusion, it is possible to predict fat content based on other macros such as calorie, carbohydrate, and protein content of food using Random Forest model. Initially, the attempt was to create an ingredient picker when considering all of the nutritional information of foods including vitamins, SFA, and other minerals, however, the irregularity of data was too great. Not all food items contained a consistent amount of any of the other nutrients. Hence the most common occurring macros were chosen. After trying different combinations of input features and target features, the resulting conclusion was reached.

### REFERENCES

- [1] Liu, A.G., Ford, N.A., Hu, F.B. et al. A healthy approach to dietary fats: understanding the science and taking action to reduce consumer confusion. *Nutr J* 16, 53 (2017). <https://doi.org/10.1186/s12937-017-0271-4>
- [2] Schmid A. The role of meat fat in the human diet. *Crit Rev Food Sci Nutr*. 2011 Jan;51(1):50-66. doi: 10.1080/10408390903044636. PMID: 21229418.
- [3] Fan Z, Chiong R, Hu Z, Keivanian F, Chiong F. Body fat prediction through feature extraction based on anthropometric and laboratory measurements. *PLoS One*. 2022 Feb 22;17(2):e0263333. doi: 10.1371/journal.pone.0263333. PMID: 35192644; PMCID: PMC8863283.
- [4] David J. Baer, William V. Rumpler, Carolyn W. Miles, George C. Fahey, Dietary Fiber Decreases the Metabolizable Energy Content and Nutrient Digestibility of Mixed Diets Fed to Humans, *The Journal of Nutrition*, Volume 127, Issue 4, 1997, Pages 579-586, ISSN 0022-3166, <https://doi.org/10.1093/jn/127.4.579>.
- [5] Zhou L, Zhang C, Liu F, Qiu Z, He Y. Application of Deep Learning in Food: A Review. *Compr Rev Food Sci Food Saf*. 2019 Nov;18(6):1793-1811. doi: 10.1111/1541-4337.12492. Epub

- 2019 Sep 16. PMID: 33336958.
- [6] Rossi L, Ferrari M, Ghiselli A. The Alignment of Recommendations of Dietary Guidelines with Sustainability Aspects: Lessons Learned from Italy's Example and Proposals for Future Development. *Nutrients*. 2023 Jan 20;15(3):542. doi: 10.3390/nu15030542. PMID: 36771249; PMCID: PMC9921064.
  - [7] Kant AK. Dietary patterns and health outcomes. *J Am Diet Assoc*. 2004 Apr;104(4):615-35. doi: 10.1016/j.jada.2004.01.010. PMID: 15054348.
  - [8] Welsh S, Shaw A, Davis C. Achieving dietary recommendations: whole-grain foods in the Food Guide Pyramid. *Crit Rev Food Sci Nutr*. 1994;34(5-6):441-51. doi: 10.1080/10408399409527674. PMID: 7811377.
  - [9] <https://fdc.nal.usda.gov/download-datasets.html>
  - [10] Trumbo P, Schlicker S, Yates AA, Poos M; Food and Nutrition Board of the Institute of Medicine, The National Academies. Dietary reference intakes for energy, carbohydrate, fiber, fat, fatty acids, cholesterol, protein and amino acids. *J Am Diet Assoc*. 2002 Nov;102(11):1621-30. doi: 10.1016/s0002-8223(02)90346-9. Erratum in: *J Am Diet Assoc*. 2003 May;103(5):563. PMID: 12449285.