NUTRI AI

ABSTRACT

With increasing awareness of health and nutrition, both individuals and professionals require reliable tools to assess the caloric and nutritional content of food. This paper presents an innovative application that automates the process of measuring calories in fruits and vegetables using advanced machine learning techniques. The system captures images of the produce, identifies them accurately, and estimates their calorie content based on mass and nutritional databases. After evaluating five machine learning classifiers, the Random Forest algorithm was chosen for deployment due to its superior performance. The application integrates seamlessly with a smart camera and mat, providing a user-friendly interface. Experimental results demonstrate an average accuracy of 88%, making the system a valuable tool for health-conscious individuals, dieticians, and wellness professionals. This solution offers a practical and efficient method to track caloric intake, contributing to healthier lifestyle choices.

Keywords—Calorie measurement, feature extraction, classification

I. INTRODUCTION

As global awareness of health and nutrition continues to rise, the need for accurate and efficient systems to measure the caloric and nutritional content of fruits and vegetables has become increasingly important. This paper introduces a machine learning-based model designed to assist individuals, patients, and dieticians in tracking daily caloric intake. The system leverages five different machine learning models to evaluate classification performance, achieving an impressive accuracy of 88% using cross-validation on various training and testing datasets. The system incorporates a smart camera and mat to capture images of fruits and vegetables, allowing for automatic calorie calculation.

This tool is intended to support healthier eating habits at a time when obesity and poor nutrition are major global health concerns. Since 1980, obesity rates have doubled, making it one of the top five causes of death worldwide. By 2012, one in six people globally were obese, a significant increase from one in ten in 2008. The main contributors to this growing crisis are high-calorie diets and a lack of physical activity. The World Health Organization (WHO) highlights that inadequate fruit and vegetable consumption is responsible for a large proportion of deaths from ischemic heart disease, gastrointestinal cancers, and strokes.

Nutritional deficiencies among children and young adults are especially concerning. A study by the Health Behaviour in School-aged Children (HBSC), conducted across 33 countries, found that fewer than half of adolescents aged 13 to 15 regularly consumed fruits and vegetables. This deficiency increases the risk of developing obesity, cardiovascular diseases, and various psychological conditions. To combat these risks, the WHO recommends a daily intake of at least 600 grams of fruits and vegetables. Addressing these nutritional gaps requires the development of an intelligent, automated system that not only tracks daily caloric intake but also provides insights into potential nutritional deficiencies.

The system proposed in this paper aims to meet this need. By using machine learning algorithms for image processing and classification, it identifies various fruits and vegetables and calculates their caloric value. The application provides real-time feedback on daily, weekly, and monthly nutritional intake, empowering users to make more informed dietary choices.

This paper is organized as follows: Section 2 offers a review of the relevant literature, Section 3 details the materials and methods used, including feature extraction techniques, while Section 4 explains the classification algorithms employed. Section 5 outlines the proposed method for calorie measurement, followed by the presentation of experimental results in Section 6. The conclusion and discussion are provided in Section 7.

Several existing methods for calorie estimation involve steps such as image acquisition, enhancement, feature extraction, and classification. The effectiveness of these methods depends heavily on selecting the right image processing techniques and classification strategies. Our proposed system integrates these components into a unified framework, offering an accurate and efficient solution for automated calorie and nutritional assessment of fruits and vegetables, making it a valuable resource for health-conscious users and dieticians alike.

II. LITERATURE SURVEY

The paper discusses the use of machine learning methods for measuring the calories in fruits and vegetables. The authors developed a system that uses a camera and an intelligent mat to capture images of the food items and then apply various classification algorithms to predict the calorie content [1].

Several existing techniques have been explored for calorie and nutrition measurement from food images. For example, Pouladzadeh et al. [7] used color, size, shape, and texture features along with SVM to measure calories and nutrition from food images, achieving an accuracy of 90-92%. Ponrani et al. [8] also analyzed the performance of SVM for calorie and nutrition measurement from food images. Almaghrabi et al. [9] proposed a novel method for measuring nutrition intake based on food images [1].

Other related works include the use of image processing techniques for food recognition and volume estimation. Savakar [11] used artificial neural networks for identification and classification of bulk fruits. Bonilla et al. [12] estimated the mass and volume of passion fruit using digital images. Pouladzadeh et al. [13] and Kuhad et al. [14] explored the use of distance estimation and deep learning to simplify the calibration process in food calorie measurement [1].

The proposed system in the paper aims to address the challenges of obesity and malnutrition by providing a reliable and efficient method for calorie measurement in fruits and vegetables. The authors have used five different machine learning models, including Decision Tree, SVM, Random Forest, K-Nearest Neighbors, and Generalized Linear Model, to predict the classification accuracy [1].

Overall, this paper provides a comprehensive literature survey on the use of machine learning techniques for calorie and nutrition measurement from food images, which can be useful for researchers working in this domain.

III . METHODOLOGY

The proposed model for flight price prediction is given in Fig 1. This model-building stage describes the various stages in building an entire model. For the implementation of ML model, requirement of previous historical data is important. The dataset collection, understanding & preprocessing techniques like feature engineering, data encoding & feature selection are the equally important to improve the quality of dataset. Thirdly the data splitting into testing, training, and validation is important to analyze the results and get the accuracy of the implemented models [16]. Lastly, the model deployment is done using the Flask framework on the web application using the techniques like HTML, CSS, and JavaScript.

A. Source Data

The primary data source for the NutrIAI project is the "What We Eat in America" (WWEIA) dataset, a part of the National Health and Nutrition Examination Survey (NHANES). This dataset provides comprehensive insights into the dietary habits of individuals in the U.S., including detailed information on food composition, nutrient levels, serving sizes, and demographic variables. Initially, the dataset contained 1,849,843 records, capturing a wide range of food items and their nutritional content. After thorough data preprocessing, cleaning, and feature engineering to focus on relevant nutritional attributes, the dataset was refined to 142,580 records. This enriched dataset serves as the foundation for building accurate models aimed at predicting calorie intake and analyzing nutrient deficiencies, ensuring high precision and reliability in our predictions.

B. Model Evaluation

1) Decision Tree Regressor: The two primary decision tree formats are regression and classification trees, where regression is utilized for continuous data and classification is utilized for categorical values. An independent variable is selected as the decision node by the decision tree from the dataset.

$$\mathbf{G} = 1 - \sum_{i=1}^{k} P_i^2$$

The Gini index has a range of 0 to 1, where 0 denotes absolutely pure examples (all instances belong to the same class) and 1 denotes evenly scattered instances (highest impurity) across all classes.

2) SVM Regressor: By selecting the hyperplane with the least amount of tolerance that best matches the data, Support Vector Machine (SVM) as a regressor forecasts the target variable . The anticipated value for a fresh data point is computed mathematically as follows:

$$\textbf{SVM=}\ f(x) = \left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b \right)$$

3) KNN Regressor: Using the average of its k nearest neighbors' values, K-Nearest Neighbors (KNN) as a regressor forecasts the output variable. The expected value in mathematics. The target values of a new data point's k nearest neighbors are averaged to determine y.

$$\textbf{k-NN:}\ \hat{y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

4) Bagging Regressor: Bagging, also known as Bootstrap Aggregating, is an ensemble learning method that averages the predictions of several models trained on various subsets of the training data to increase the stability and accuracy of machine learning models.

$$\textbf{Bagging=}\ \hat{f}_{\textbf{bag}}(x) = \frac{1}{B} \sum_{b=1}^{B} f_b(x)$$

5) XGBoost Regressor: Extreme Gradient Boosting, or XGBoost, is a sophisticated gradient boosting algorithm implementation that prioritizes speed and efficiency over more conventional gradient boosting approaches .

XGBoost= $\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}$

6) Random Forest Regressor: An ensemble learning technique called the Random Forest Regressor makes predictions by utilizing several decision trees [24]. To increase accuracy and decrease overfitting, it functions by averaging the predictions made by each individual tree.For each new data point in a Random Forest Regressor, the projected value Y-Pred can be computed as follows:

$$\textbf{Random Forest: } \hat{y} = \frac{1}{B} \sum_{b=1}^{B} h_b(x)$$

where: • N is the number of trees in the forest, • Ti (x) is the prediction of the i-th tree for the new data point x.

C. Performance metrices

Performance metrics, or statistical models, will be used to compare the accuracy of machine learning models trained by different algorithms. The sklearn.metrics module will be used to create regression metrics for each model's error measurement routines [12]. 1) The Mean Absolute Error (MAE): The mean absolute error is basically calculated by adding the mean of the absolute difference between the expected and actual quantities.

1) **The Mean Absolute Error (MAE):** The mean absolute error is basically calculated by adding the mean of the absolute difference between the expected and actual quantities.

$$\textbf{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

The real output numbers are y, while the intended values are y'. A total of n data points are present. The lower the MAE number, the better your model will work .

2) **Mean square error (MSE):** Rather than utilizing an absolute number to sum the true and anticipated output values, the root mean square error exponentiates the difference between them.

$$\textbf{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

3) **Root mean square error (RMSE):** To calculate the root mean square error (RMSE), one must take the square root of the predicted and actual values.

$$\textbf{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

The actual output numbers are y, while the predicted values are y'. The total number of data points is n. A model's RMSE value is greater than its MAE and smaller than its RMSE value when comparing models with higher performance. The coefficient of determination, or R2, will show you how well the independent variable changed in your model when there was a deviation.

4) **R2 Score**: You can use it to gain insight into how well your model's independent variable changed when there was a deviation.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

The value of R-squared falls between 0 and 1. Your model performs better when compared to the values of other models if its value for one is closer

D. Data Analysis and visualisation

Data analysis and visualization play a critical role in understanding and interpreting the dataset, providing insights into the relationships

between various food nutrients. Through visual representations like bar charts, scatter plots, and heatmaps, trends, patterns, and outliers in the dataset are easily identifiable. Such insights help in the preprocessing and feature selection process for building more accurate predictive models. Furthermore, visualization simplifies communication of complex findings to non-technical audiences, enhancing data-driven decision-making.

In this section, we explore multiple visualizations created from the dataset that contain nutrient information such as energy content, protein levels, and calcium concentrations. These visualizations provide an overview of the highest contributing foods for each nutrient and their relationships to other variables. Each figure illustrates important aspects of the dataset, which is instrumental in feature selection and model development.

1. **Top 10 Foods with Highest Energy Content** (Fig. 1) – The bar plot highlights the top 10 food items with the highest energy content. By visualizing energy (in kcal), we can see which foods provide the most caloric content per serving. This is useful in selecting energy-dense foods for certain dietary applications.
2. **Top 10 Foods with Highest Protein Content** (Fig. 2) – The bar chart depicts the top 10 foods with the highest protein content. Protein-rich foods are essential for various health-related recommendations, and visualizing them helps in understanding the distribution of protein content across food items.
3. **Comparison of Calcium, Iron, and Fiber in Top 5 Foods** (Fig. 3) – A grouped bar plot is used to compare the calcium, iron, and fiber content in the top 5 foods with the highest calcium concentration. The comparison aids in selecting foods that provide balanced levels of these essential nutrients.
4. **Correlation Heatmap of Nutrients** (Fig. 4) – The heatmap displays the correlations between various nutrients such as calcium, carbohydrates, cholesterol, energy, protein, sodium, sugars, and fat. The darker colors represent stronger correlations, which highlight the key relationships between different nutrients. This insight is crucial for selecting features for predictive modeling.
5. **Top 10 Foods with Highest Cholesterol Content** (Fig. 5) – Another bar plot illustrates the top 10 foods with the highest cholesterol content. This visualization helps identify which foods contribute the most cholesterol, which is particularly relevant for dietary guidelines focusing on heart health.
6. **Scatter Plot of Carbohydrates vs Energy** (Fig. 6) – A scatter plot is used to show the relationship between carbohydrates and energy. The color and size of the points represent fat and protein content, respectively, adding a multi-dimensional view of the data. This helps in understanding how the combination of macronutrients affects energy content.

Through these visualizations, key insights into nutrient distribution, correlations, and outliers are uncovered, which guide the subsequent steps of model building and evaluation. The visualizations serve as a foundation for feature selection and facilitate more robust decision-making processes.

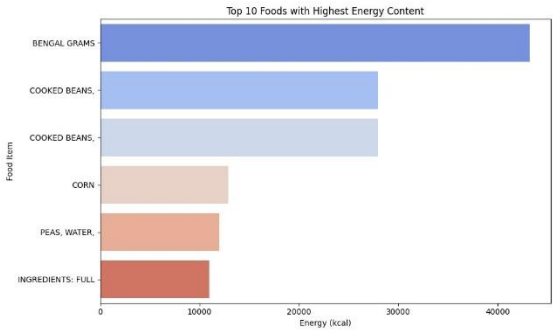

Fig. 1: Bar chart showing the top 10 foods with the highest energy content (kcal).

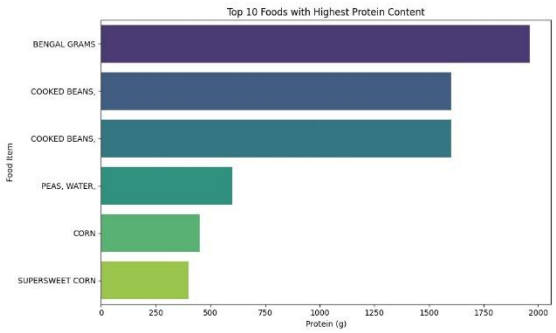

Fig. 2: Bar chart showing the top 10 foods with the highest protein content (g).

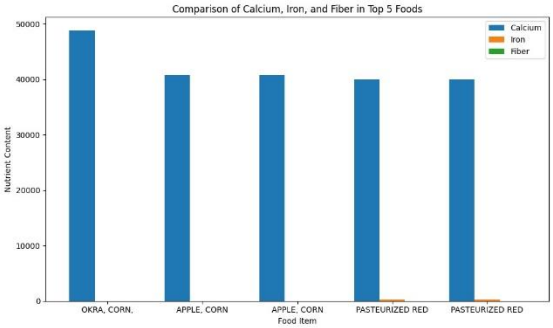

Fig. 3: Comparison of calcium, iron, and fiber content in the top 5 foods with the highest calcium concentration.

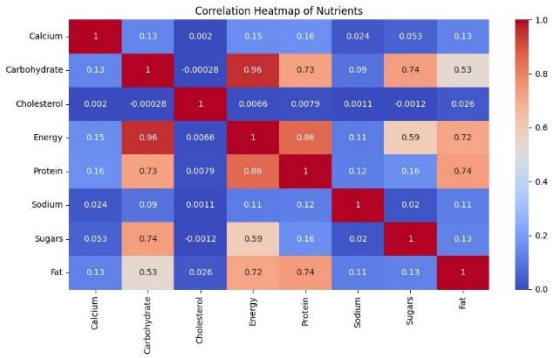

Fig. 4: Heatmap showing the correlation between various nutrients in the dataset.

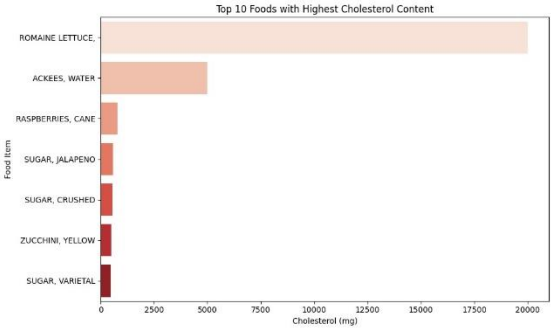

Fig. 5: Bar chart showing the top 10 foods with the highest cholesterol content (mg).

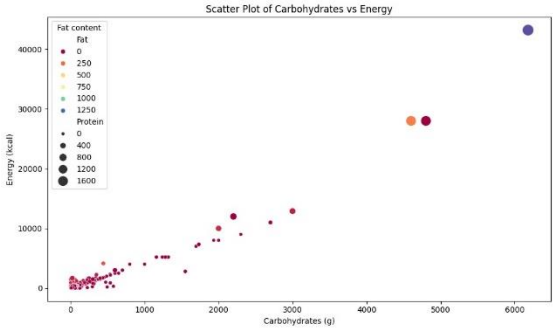

Fig. 6: Scatter plot of carbohydrates vs energy content, with fat represented by color and protein by size.

REFERENCES

[1] D. R. Whiting, L. Guariguata, C. Weil, and J. Shaw, "IDF diabetes atlas: global estimates of the prevalence of diabetes for 2011 and 2030," Diabetes research and clinical practice, vol. 94, no. 3, pp. 311-321, 2011.

[2] C. Hedinger, Histological typing of thyroid tumours. Springer Science & Business Media (2012).

[3] Q. Zhang and Y. Wang, "Socioeconomic inequality of obesity in the United States: do gender, age, and ethnicity matter?," Social science & medicine, vol.58, no.6, pp. 1171-1180, 2004.

[4] W. Jia, R. Zhao, N. Yao, J.D. Fernstrom, M. H. Fernstrom, R. J. Sclabassi, and M. Sun, "A food portion size measurement system for image-based dietary assessment," IEEE Proc. of 35th Annual Northeast Conference on Bioengineering, USA, pp. 1-2, 2009.

[5] H. Vainio, and A. B. Miller, "Primary and secondary prevention in colorectal cancer," Actaoncologica, vol. 42, no.8, pp.809-815, 2003.

[6] P. Pouladzadeh, S. Shirmohammadi, and R. Almaghrabi, "Measuring calorie and nutrition from food image," IEEE Transactions on Instrumentation & Measurement, vol. 63, no. 8, pp. 1947-1956, 2014.

[7] D.S. Ponrani, S.N. Suveka, and S.K. Brabha, "Performance analysis of SVM to measure calorie and nutrition from food Images," International Journal of Advanced Research Trends in Engineering and Technology (IJARTET), vol. 1, no. 3, pp. 93-98, 2014.

[8] R. Almaghrabi, G. Villalobos, P. Pouladzadeh, and S. Shirmohammadi, "A novel method for measuring nutrition intake based on food image," IEEE International Conference on Instrumentation and Measurement Technology, Austria, pp. 366-370, 2012.

[9] S. Anushadevi, "Calorie measurement of food from food image, "International Journal on Applications in Information and Communication Engineering, vol. 1, no. 7, pp. 14-17, 2015.

[10] D. Savakar, "Identification and classification of bulk fruits images using artificial neural networks," International Journal of Engineering and Innovative Technology (IJEIT), vol. 1, no. 3, 2015.

[11] J. Bonilla, F. Prieto, and C. Pérez, "Mass and volume estimation of passion fruit using digital images," IEEE Latin America Transactions, vol. 15, no. 2, pp. 275-281, 2017.

[12] P. Pouladzadeh, P. Kuhad, S.V.B. Peddi, A. Yassine, and S. Shirmohammadi, "Mobile cloud based food calorie measurement," IEEE 4th International Conference on Multimedia and Expo Workshops (ICMEW), China, pp. 1-6, 2014.

[13] P. Kuhad, A. Yassine, and S. Shirmohammadi, "Using distance estimation and deep learning to simplify calibration in food calorie measurement," IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications [CIVEMSA], China, pp. 1-2, 2015.

[14] P. Pouladzadeh, G. Villalobos, R. Almaghrabi, and S. Shirmohammadi, "A novel SVM based food recognition method for calorie measurement applications," IEEE International Conference on Multimedia and Expo Workshops (ICMEW), Australia, pp. 495-498, 2012.

[15] G. Villalobos, R. Almaghrabi, P. Pouladzadeh, and S. Shirmohammadi, "An image processing approach for calorie intake measurement," IEEE Symposium on Medical Measurement and Applications, Budapest, Hungary, pp. 1-5, 2012.

[16] W. Gonzalez, and R. E. Woods, Digital image processing using MATLAB, Third New Jersey: Prentice Hall, 2004.

[17] D.G. Lowe, "Distinctive image features from scale-invariant key points," International journal of computer vision, vol.60, no.2, pp. 91-110, 2004.

[18] A. Ali, X. Jing, and N. Saleem, "GLCM-based fingerprint recognition algorithm," IEEE 4th International Conference on Broadband Network and Multimedia Technology (IC-BNMT), China, pp. 207-211, 2011.

[19] S. R. Kodituwakku, and S. Selvarajah, "Comparison of color features for image retrieval," Indian Journal of Computer Science and Engineering, vol.1, no.3, pp. 207-211, 2004.

[20] J. R. Quinlan, "Induction of decision trees," Machine learning, vol. 1, no.1, pp. 81-106, 1986.

[21] A. Ben-Hur, D. Horn, H. T. Siegelmann, and V. > Abhi: Vapnik, "Support vector clustering," Journal of Machine Learning Research, vol. 2, pp. 125-137, 2001.

[22] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5 32, 2001.

[23] T. Cover, and P. Hart, "Nearest-neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no.1, pp. 21-27, 1967.

[24] P. McCullagh, and J. A. Nelder, "Generalized Linear Models," 2nd edition, Chapman-Hall, London. Standard book on generalized linear models, 1989.

[25] Health Canada. (2011, November) Health Canada Nutrient Values.[Online].http://www.hcsc.gc.ca/fnan/nutrition/fiche-nutridata/nutrient_value-valeurs_nutritives-tc-tmeng.php.

[26] http://www.aqua-calc.com/page/density-table.

[27] K. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, "An introduction to kernel-based learning," IEEE Transactions on Neural Networks, vol. 12, no.2, pp. 181–201, 2001