**Nutritional Intelligence: Classifying Food Ingredients Using Machine Learning**

Bipin Kumar Rai  Chandan N S Divya Neelappa Marangappanavar

Computer Science and Engineering Computer Science and Engineering Computer Science and Engineering  
Dayananda Sagar University Dayananda Sagar University Dayananda Sagar UniversityBangalore, India Bangalore, India Bangalore, India

[bipinkrai@gmail.com](mailto:bipinkrai@gmail.com) [nschandan31@gamil.com](mailto:nschandan31@gamil.com) [divyamarangappanavar@gmail.com](mailto:divyamarangappanavar@gmail.com)

<https://orcid.org/0000-0002-9834-8093>

Indira S

Computer Science and Engineering  
Dayananda Sagar University

Bangalore, India  
[indirasm2004@gmail.com](mailto:indirasm2004@gmail.com)

**ABSTRACT**

In the age when diet defines health, knowing about the nutrition of food ingredients becomes very essential. In this work, a mechanism using machine learning is presented for classifying food ingredients in two categories, and the healthiness nature depends on composition and flavor types. Classification methods were trained using an Indian food dataset with ingredient percentage and categorical flavor data. The features studied are average ingredient content and main taste types in different correlations with health indicators.

To check the predictability of performance, six machine-learning algorithms, namely Decision Tree, Random Forest, SVM, Logistic Regression, KNN, and XGBoost, were implemented and compared. Among all these methods, XGBoost had consistently proved its superiority with higher accuracy and a well-balanced performance in all evaluation metrics. Since it was able to take care of complex patterns and intricate interactions between the compositions of ingredients, it becomes particularly well suited for this classification problem. The results give a way to demonstrate that XGBoost can be used ideally as the model for nutrition-driven decisions and recommendations to public health.

***Keywords***: Machine Learning, Food Ingredient Classification, Predictive Modelling.

**1.Introduction**

Food plays a chief role in sustaining life and quality living with health. This food gets digested, helping someone with energy to move. But this is only half of what goes out-the effect, either good or bad, his nourishment has on the physique and psyche in the long run. The nutrient value we give to our food depends largely on the individual ingredients we use in cooking: nutrients-rich ingredients, which keep the muscles strong, provide hormonal balance, assist brain functions, and facilitate metabolism: premium-proteins, unsaturated fats, dietary fiber, vitamins (A, B-complex, C, D, and E), and minerals (iron, calcium, and zinc) [25],[27]. On the other side, common ingredients are the sniper of promising nutrition; at best, they could be loaded with sodium, trans-fats, refined sugars, or some kind of chemical additives [26]. The body receives all the nutrients it requires to operate at its best when a nutritionally balanced diet is maintained. It has been demonstrated that a balanced intake of macro- and micronutrients reduces the risk of non-communicable diseases like obesity, type 2 diabetes [28], heart disease[29], and even some cancers.

We are seeing a boom in the use of intelligent technologies like machine learning in a variety of fields, including nutrition, at the same time that health science is making strides [30] [31]. Machine learning stand out as a significant tool when it comes to analyzing large and complicated datasets, identifying patterns, and making highly accurate data-driven decisions][32]. Classifying food ingredients based on nutritional and chemical properties constitutes an application of critical significance in the food and nutrition sciences. Thus, classification of ingredients through algorithms offers computer-aided classification rather than the traditional food analysis methods [10].

We have proposed a binary classification method that uses machine learning classifiers to classify food ingredients as either "healthy" or "unhealthy." The classification procedure is constructed with cleaning, normalization, and One-Hot Encoding of categorical variables being key preprocessing stages. To solve the class imbalance, SMOTE is applied to provide equity in prediction [18][19]. Subsequently, the various classification algorithms of Logistic Regression(LR), K-Nearest Neighbors (KNN), Decision Tree(DT), Random Forest(RF),Support Vector Machine(SVM), and XGBoost [9][11] are trained and assessed using metrics such as accuracy, precision, recall, and F1-score. These metrics provide a fair comparison among methods and thus aid in the selection of the best approach.

**2.Literature Review**

With the gradual increase in food safety and nutritional information relevance, there has grown the need for an intelligent system that can classify food ingredients as either healthy or harmful. ML therefore constitutes a big force in our arena with its automated, data-driven solutions for ingredient analysis based on several attributes. Classification algorithms help the ingredients listings with health monitoring, food quality checking, and healthy eating. In order to find the best method with respect to accuracy and generalization performance, the present work looks at the application of multiple machine learning models for classifying food ingredients. The following section intends to review all the work related to approaches and methodologies used in this field.

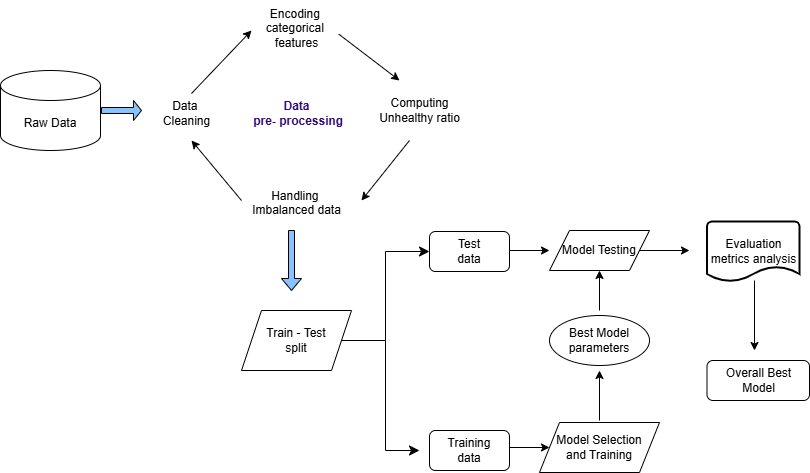
**Table 1: Comparative Analysis of Related Works**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Author name with reference | Model / technique used | Key findings | Pros | Cons |
| [1]G. Menichetti, B. Ravandi, D. Mozaffarian, and A.-L. Barabási | FoodProX(random forest classifier), NOVA Classification, FPro (Food Processing Score) | Over 73% of US food is ultra-processed..FoodProX outperforms NOVA with better coverage and continuous scoring. | Classifies complex foods and identifies healthier foods. | doesn’t directly measure processing techniques |
| [2] R. Qasrawi, S. Sgahir, M. Nemer, M. Halaikah, M. Badrasawi, M. Amro, S. Vicuna Polo, D. Abu Al-Halawa, D. Mujahed, L. Nasreddine, I. Elmadfa, S. Atari, and A. Al-Jawaldeh | Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), and Gradient Boosting (GB).  Evaluation: confusion matrix, accuracy, precision, sensitivity, F1-score, and AUC. | Random Forest and Gradient Boosting, you can classify foods as healthy or unhealthy based on **nutritional features, ingredients** | Offers insights for targeted interventions, Detailed validation strategy with robust performance metrics. | a small initial sample size and potential data imbalances |
| [3] Ruchika Sharma, Sushil Kumari, Navneet Gupta 2012. | Uses Linear Regression for predicting nutritional values, Decision Tree(DT) for categorical data, RT for improving accuracy, Support Vector Machine(SVM) for classification. Evaluation by mean squared error(MSE) and accuracy. | RT and DT balanced accuracy and interpretability. LR struggles with complexity. SVM was effective but required parameter tuning. | DT offers explainability, Regression models provide stable and interpretable. SVM handles high dimensional data well. | Regression models struggle with non-linear relationship lacking flexibility. DT overfit if not pruned effectively. RT is complex and slow. SVM struggles with large datasets. |
| [4] Christabel Tachie, Nii Adjetey Tawiah, Alberta N.A. Aryee 2023. | DT, RF, Light Gradient Boosting Machine(LightGBM) employed to predict Nutri-Score and micronutrient content. Preprocessing includes cleaning, handling multicollinearity(VIF), and transforming categorical variables. | RF and LightGBM achieved the highest accuracy. DT performed least effectively with lower accuracy. Rf showed the best overall testing performance. | RF robust to missing data and high accuracy. DT requires less data cleaning and can handle both numerical and categorical data. LR provides probability outputs. | DT sensitive to noise which leads to poor generalization. RF computationally intensive slower to train due to ensemble nature and high complexity. LR struggles with non-linearity. |
| [5] Daniel Kirk, Esther Kok, Michele Tufano,Bedir Tekinerdogan, Edith JM Feskens,and Guido Camps.(2022). | For classification, prediction, and clustering in nutritional contexts, interpretable machine learning techniques such as Random Forest, k-means clustering, logistic regression, SVM, XGBoost with SHAP, and Lasso regression were employed. | ML methods forecasted the degree of symptoms, obesity, malnutrition, and glucose response. Personalization was made easier by phenotyping. Explainable AI provided new information. Additionally, ML highlighted individual variations and optimized diagnostic resources. | High accuracy was attained, allowing for affordable, non-invasive forecasts with comprehensible outcomes. demonstrated a high level of proficiency in finding biomarker-based health insights and tailoring interventions. beneficial for a range of applications pertaining to nutrition. | Without xAI, some models were difficult to understand due to their complexity. Data that is specific to a population can lead to problems with generalizability. high reliance on clustering assumptions and data quality. Some lacked information about how the model was implemented. |
| [6] Adjuik, T. A., Boi-Dsane, N. A. A., & Kehinde, B. A. (2024). | used a combination of conventional and image-based machine learning techniques, including Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Gradient Boosted Regression (GBR). | FoodProX RF identified highly processed foods; RF had an accuracy rate of 89.7% in classifying health risks; Strong accuracy in estimating calories from food photos was attained by SVM and CNN. | While SVM successfully supported image-based calorie estimation with good generalization on test datasets, RF offered high accuracy in classifying food and health risks. | Without explainable AI, complex models like RF and SVM were harder to interpret; data imbalance and image quality decreased reliability; and some models had lower testing accuracy. |
| [7]Sasmita Nayak, Mamata Beura, Mohammed Siddique, Siba Prasad Mishra | Decision Tree, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest (RF) | Random Forest achieved highest accuracy (92.2%) in classifying Indian food as vegetarian or non-vegetarian. | Comprehensive comparison of multiple ML models; use of real dataset; clear accuracy metrics | Dataset size is small (255 records); limited to classification of vegetarian vs non-vegetarian; no hybrid models tested yet |

**3. Proposed system and Methodology**

**3.1 Workflow**

The detailed workflow of our proposed system is shown in figure 1. The process starts with collecting raw data and applying data cleaning to remove noise and irrelevant data. This is further followed by feature encoding and handling imbalanced data. An important step in preprocessing involves calculating a domain specific feature “unhealthy\_ratio” which signifies the risk posed by each ingredient. Post train – test split we applied six machine learning models. Evaluation metrics like accuracy, precision, recall, f1 – score and confusion matrix are used to assess the model’s performance. The best parameters are retained to improve generalization of the model. Through all these steps the main goal of our work evaluating and classifying food ingredients as healthy or not for human being is achieved.

****

**Figure 1. Machine Learning Pipeline for Healthiness Prediction**

**3.2 Data collection**

The dataset called "Indian Food Classification with TensorFlow" is obtained from Kaggle. In order to facilitate food classification tasks using machine learning and deep learning models, this dataset includes detailed information on a broad range of Indian food items. The dataset comprises 177 records (rows) and 9 distinct features (columns). Each food item is described by a number of features in the dataset, including its nutritional value, method of preparation, and classification according to health. Among the dataset's salient characteristics are: Name, ingredients, preparation time, cooking time, flavor profile, course, state, location, percentage of ingredients, units, and target. The dataset’s target column is a binary classification label that indicates the healthiness of a food item. A nutritious food item is indicated by a value of 0 in the target column and a food item with a value of 1 is considered unhealthy.

**3.3 Data preprocessing**

The preprocessing involved data cleaning, computing unhealthy ratio, feature analysis and importance, handling imbalanced data, Statistical Analysis, data splitting, and standardization to optimize model performance.

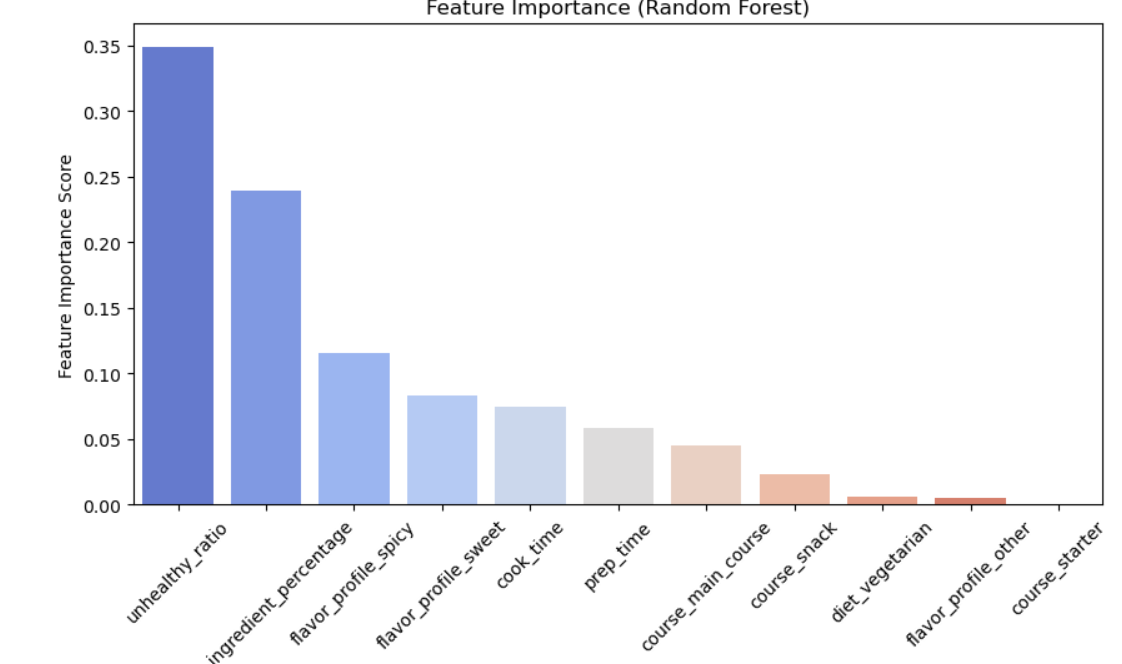
**Data cleaning**

Food\_name, state, and region were among the features that were eliminated because they would not have contributed to the binary classification of food items as either healthy or unhealthy. Rows with too many missing values were removed because they lacked sufficient information for meaningful imputation, preserving the high quality of the data and preventing bias. In order to transform categorical features like course, flavor\_profile, and diet into binary vectors for learning algorithms, one-hot encoding was further used.

**Computing unhealthy ratio**

We have included a new feature-the unhealthy ratio-to enhance food-healthiness classification at a finer resolution. The unhealthy ratio measures the share of unhealthy ingredients among the ingredients of each dish, thus providing a numerical value that supports the overall classification exercise. The calculation of the ratio was based on values in an existing column containing encoded ingredient information. The unhealthy ingredients were first highlighted, after which their proportion in relation to all ingredients was calculated. Having created this new feature, it was added to the dataset, and the original columns were removed to avoid redundancy.

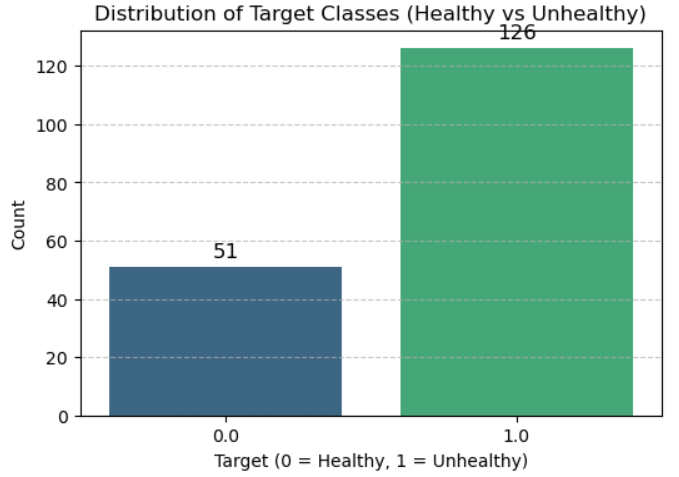
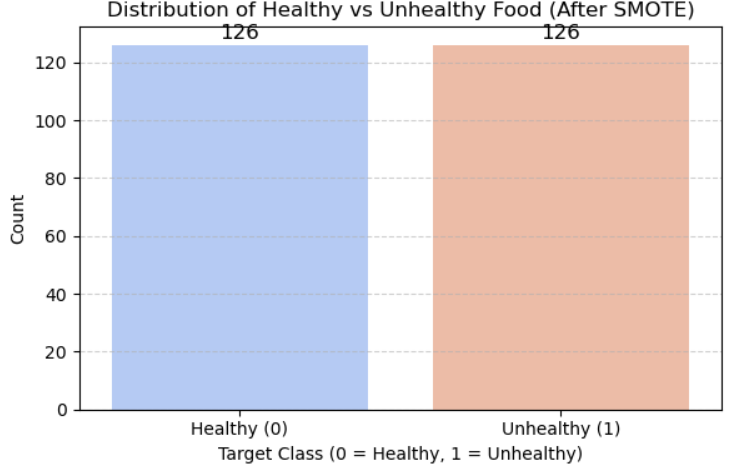
**Feature analysis and importance**

To get an understanding of factors that lead to classification of food ingredients as healthy or not, we performed a feature importance analysis using the Random Forest. Based on the output depicted in the feature importance plot, some two criteria are in the spotlight: unhealthy\_ratio and ingredient\_percentage. These two characterizations represent the most important factors in distinguishing between foods that are classified as healthy and those classified as unhealthy.

**Figure 2. Feature Importance using Random Forest**

**Handling imbalanced data**

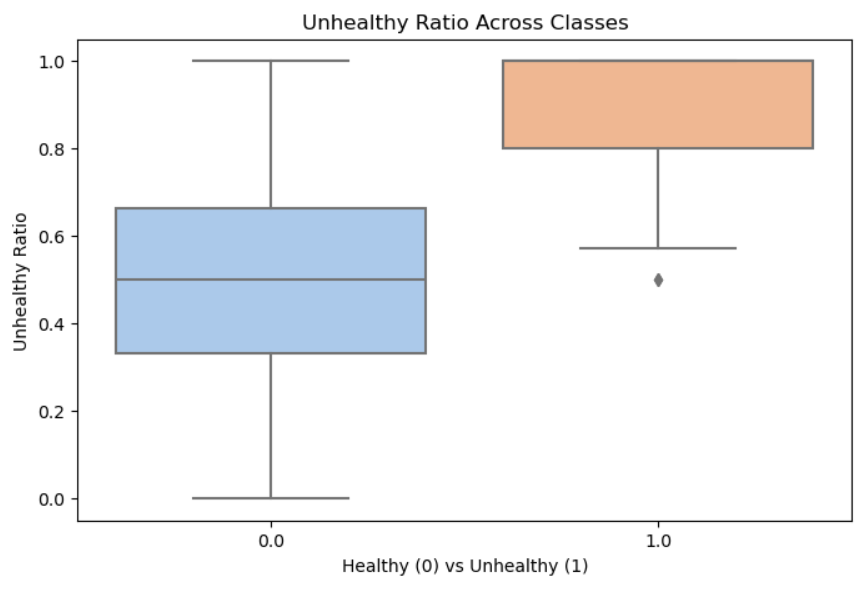
In classification tasks, Unbalanced datasets, in which one class is substantially underrepresented, can result in biased models that underperform the minority class. The initial distribution in our dataset was skewed, with only 51 samples classified as healthy (0) and 126 samples classified as unhealthy (1). To address this, We used **SMOTE** (Synthetic Minority Over-sampling Technique) [19], a popular method that interpolates between existing samples to create new instances of the minority class [18]. After applying SMOTE, the class distribution was successfully balanced, with 126 samples in both classes.

**Figure 3. Target Class Distribution Before and After SMOTE**

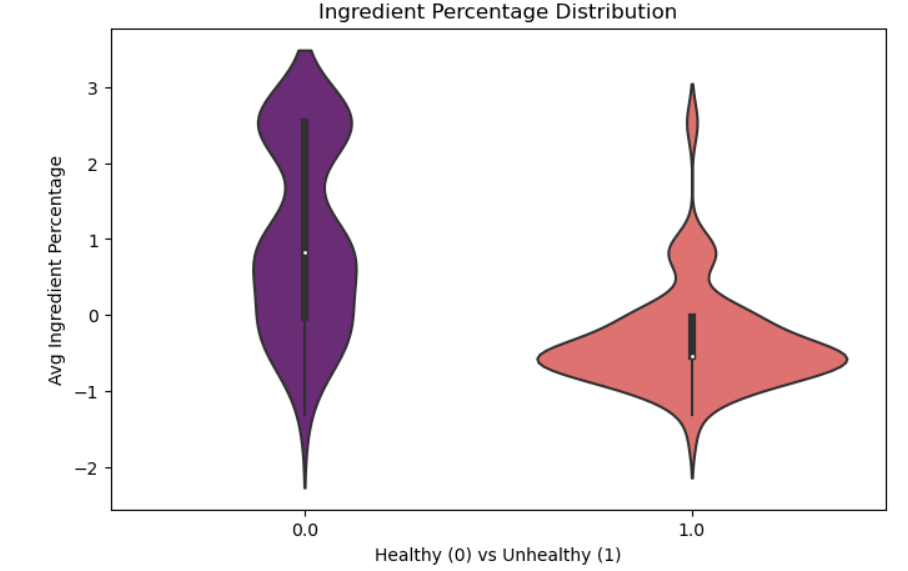
**Statistical Analysis**

Visual methods were used for statistical analysis in order to better understand the dataset and the relationship between its features and the target variable.



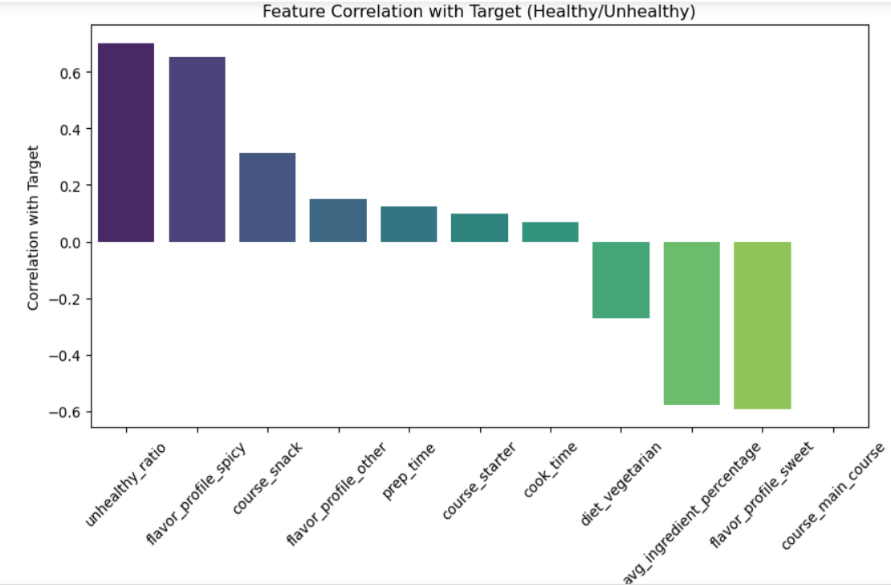
**Figure 4. The distribution of the unhealthy ratio across classes**

The box plot unequivocally demonstrates that healthy foods (0) have lower and more dispersed values, whereas unhealthy foods (1) have significantly higher unhealthy\_ratio values with less variance.



**Figure 5. The ingredient percentage distribution across classes.**

The distribution of average ingredient percentages for nutritious and unhealthy foods is contrasted in this graph. Healthy food items are typically made up of a higher percentage of healthy ingredients, as evidenced by their higher and more varied ingredient percentage. Conversely, the ingredient percentages of unhealthy foods exhibit a narrower and smaller range, indicating a comparatively small amount of healthy ingredients.



**Figure 6. The feature correlation with the target variable (Healthy/Unhealthy)**

The bar graph shows the correlation between each feature and the target class. Unhealthy\_ratio and flavor\_profile\_spicy are good indicators of unhealthiness, inasmuch as they exhibit a strong positive correlation with the "Unhealthy" target class. Contrariwise, avg\_ingredient\_percentage and flavor\_profile\_sweet show a negative correlation and so are more in tandem with the "Healthy" target class.

**Data Split**

The dataset was split into training and testing, with an 80-20 split, with a more significant purpose of testing the methods of model learning for their performance and generalizability. Training consists of using 80 percent of the data to follow the underlying trends, while the remaining 20 percent is set aside to test the model's ability in actually predicting data in the real world.

**Standardization**

Standardization is a method used in data preprocessing where scales of numerical features are changed to the scales of a standard normal distribution. Thus, standardization was done so that each numerical feature could have an equal say in the process of model learning. This method simply puts a scale to all the features by making the data have zero mean and unit standard deviation.

**3.4 Model Selection and Evaluation**

Various supervised machine learning algorithms were analyzed regarding their performance and efficacy in classifying food ingredients into healthy or unhealthy ingredients. DT [8],[10], KNN [8],[14], SVM [8],[14], LR [13],[14],[20], RF [14], and XGB [15],[16],[17] were employed as the models. The performance of each model was evaluated using classification reports, and confusion matrices focusing on precision, recall, F1-score, and accuracy.

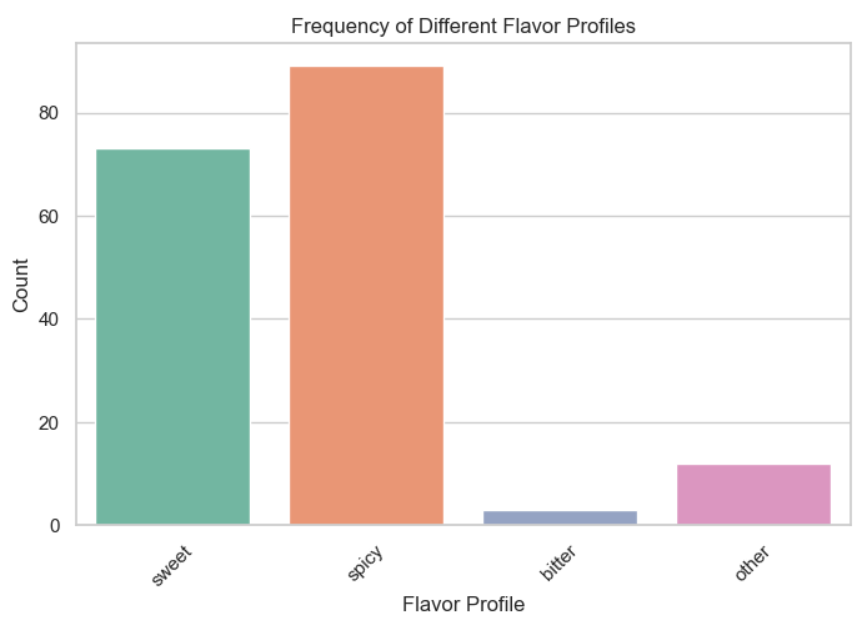
DT is a supervised learning algorithm that uses feature values to classify data and identify patterns in the dataset [7], [9]. The DT employed the Gini Index as the splitting criterion and the default CART algorithm for this investigation. In a decision tree, each node corresponds to an attribute that is used to classify a piece of data, and each connecting line represents a potential outcome the attribute could achieve [10]. KNN is a simple, instance-based learning algorithm where data points are categorized using the majority class of their k nearest neighbors in the feature space [9]. For our work, we chose Euclidean distance to measure how similar the data points are [8]. SVMis a potent supervised learning algorithm [7],[10], used to identify the best hyperplane for dividing data points into distinct classes [9],[13].The points lying closest to the hyperplane are called support vectors [9]. SVM improves generalization and robustness by concentrating on maximizing the margin between classes [9],[10],[13]. LRis a popularly used statistical model for binary classification that estimates the probability of a class label using a logistic (sigmoid) function [10],[12].

RF is a kind of ensemble approach [7],[15] that combines various models or iterations of the same model [7]. It can be mathematically shown, by the Strong Law of Large Numbers, that the generalization error of a random forest converges with probability 1 to a constant as the number of trees increases [15],[22]. Even when noise and irrelevant features are included, this method performs fairly well [22],[23]. This method is capable of dealing with response variables from both sides: it acts as regression if the response is continuous and classification if it's categorical [24]. XGBoost is a scalable and extremely effective gradient boosting algorithm [15],[21]. It employs regularization to avoid overfitting and builds an ensemble of decision trees in a sequential fashion, with each tree fixing the mistakes of the ones before it [15],[16]. XGBoost offers good adaptability, fast learning convergence and higher accuracy in large range datasets [11]. Among all the machine learning algorithms we applied to our dataset, XGBoost achieved the highest accuracy.

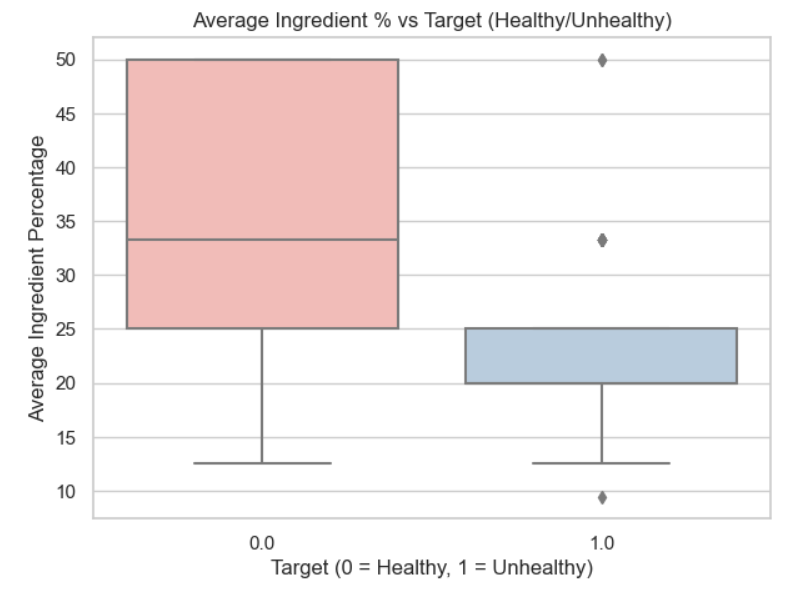
**4. Results and Comparison**

The Indian food dataset was thoroughly examined based on a number of significant nutritional and preparation-related characteristics, including ingredient composition, cooking and preparation times, diet type, course, and flavor profile, prior to the application of any machine learning models. The frequency distribution of the dataset's various flavor profiles is shown in Figure 7. With more than 85 entries, it is evident that spicy foods predominate in Indian cuisine, with sweet dishes following closely behind. The categories of bitter and others are much less common. Figure 9 shows the distribution of the ratio of unhealthy ingredients. Although many foods have low unhealthy ratios, a sizable percentage of foods have ratios greater than 0.5, suggesting the presence of multiple ingredients that may be hazardous if consumed frequently.

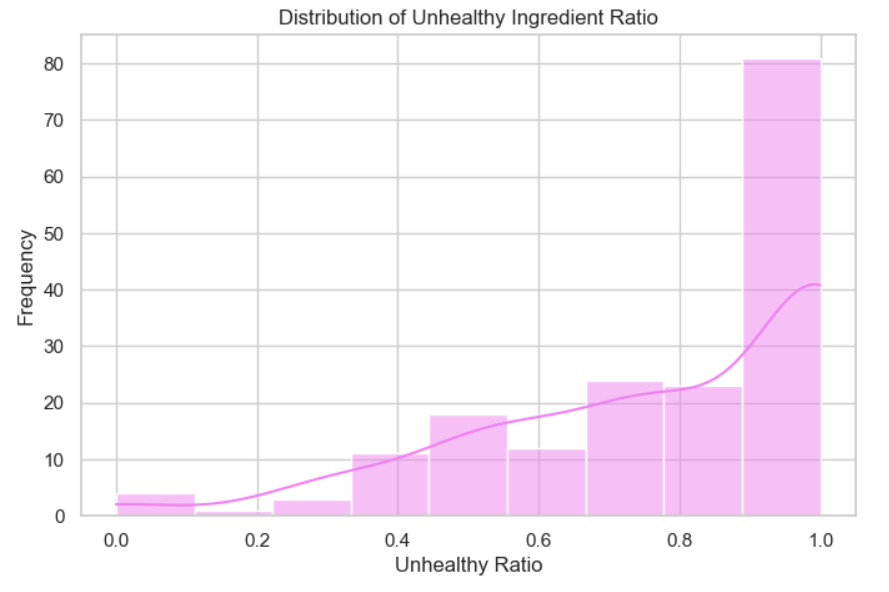
The percentage of course types in the dataset is displayed in Figure 12. Main courses are found to make up the majority of the meal, making up roughly 46.9% of all the dishes. Snacks account for a smaller portion at 12.4%, while desserts come in second at about 39.5%. With only 1.1% of the total, starters are the least represented. The log-transformed distribution of cooking time is shown in Figure 13. The data displays a near-normal (bell-shaped) distribution following the logarithmic transformation, with a log value of roughly 3 to 3.5 at its center. This suggests that there are fewer cases of extremely short or extremely long cooking times and that most dishes have moderate cooking times.



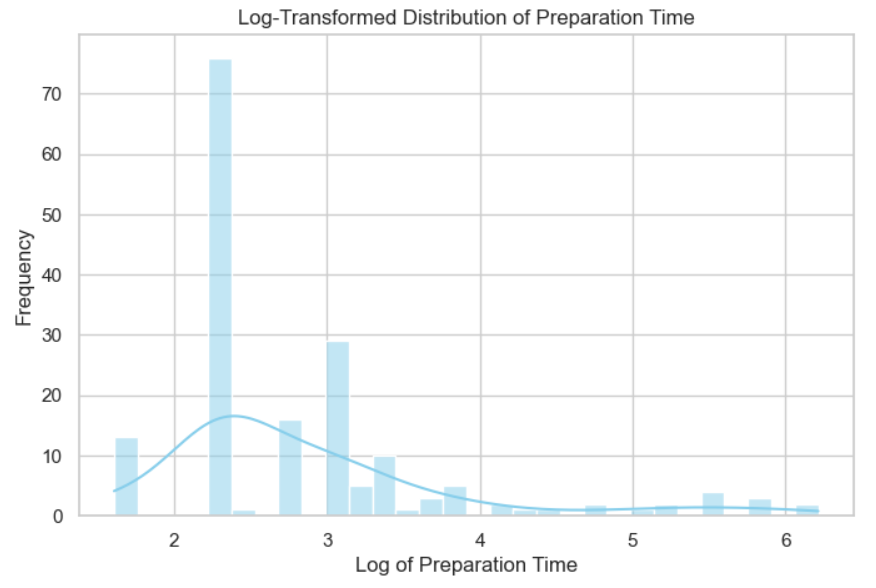
**Figure 7. shows the frequency of different flavor profiles in the dataset.**



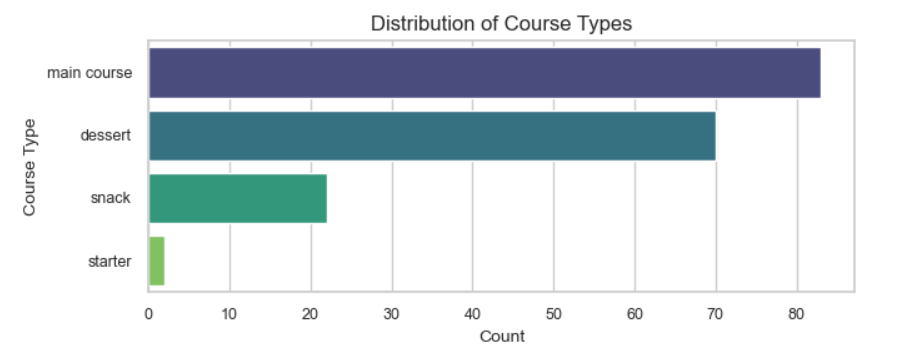
**Figure 8. Box Plot of Average Ingredient Percentage by Health Classification**



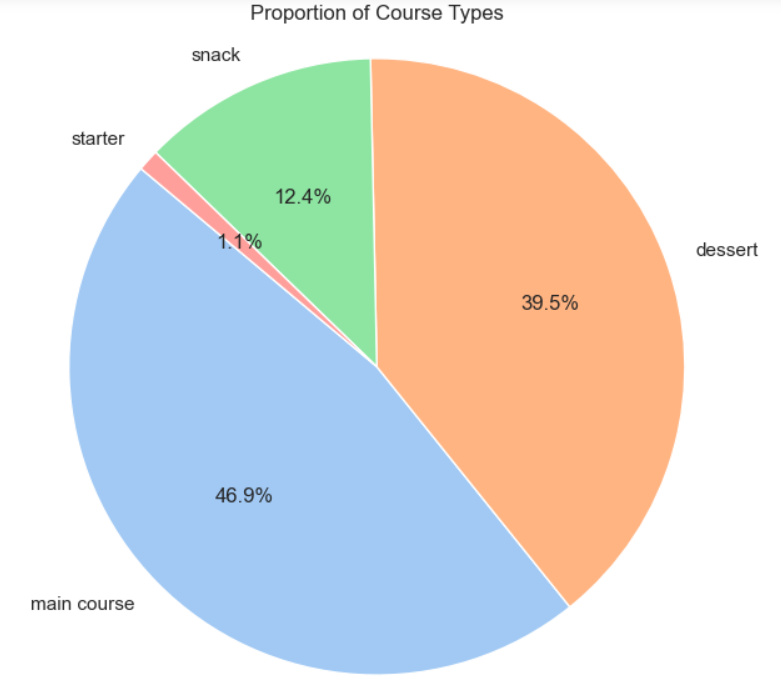
**Figure 9. Distribution of Unhealthy Ingredient Ratios in Food Items**



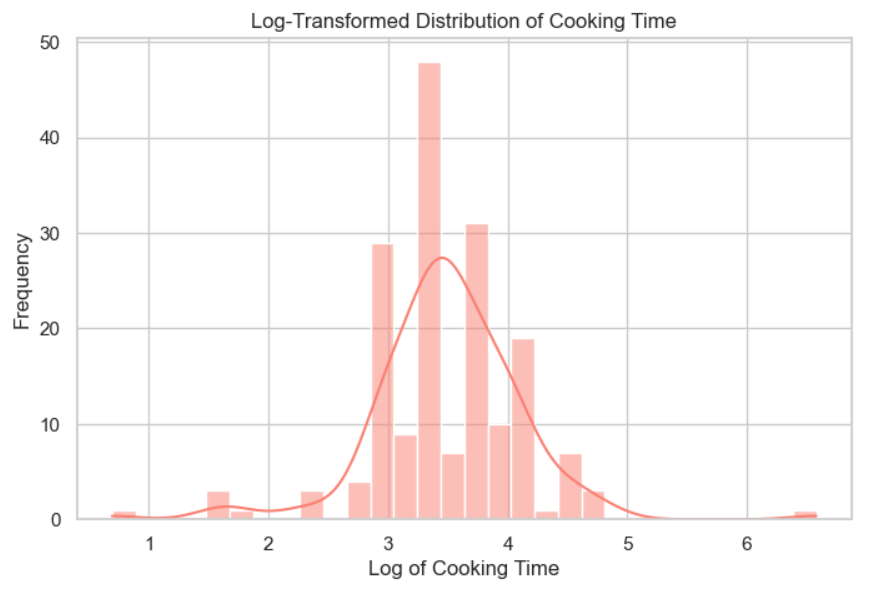
**Figure 10. Log-Transformed Distribution of Food Preparation Time**



**Figure 11. The distribution of course types among the dishes**

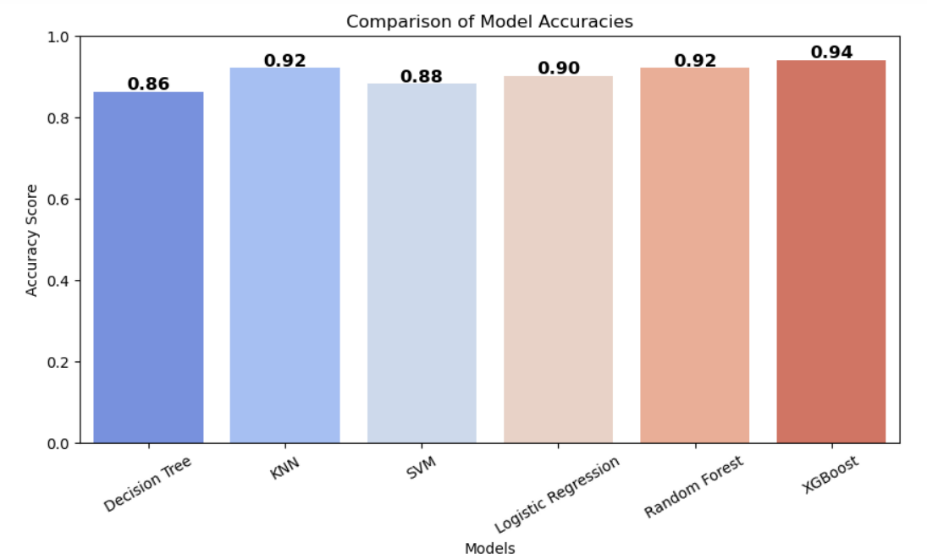


**Figure 12. Comparative Representation of Course Categories**



**Figure 13. Distribution of Cooking Time on a Logarithmic Scale**

The performance of various machine learning models such as DT, KNN, SVM, Logistic Regression, RF, and XGB was evaluated on the food ingredient classification task, and the results are summarized in the figure 14 below and in table 2. The findings show that XGB with an accuracy of 94% outperformed others in terms of generalization ability and predictive power. Its ensemble-based gradient boosting approach helped it perform better than other models in this binary classification task.



**Figure 14. Comparative Accuracy Analysis of ML Models**

**Table 2: Comparison of Algorithm Performance in Predicting Health Conditions**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Algorithm** | **Class – Healthy(0) Unhealthy(1)** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
| DT | 0.0 | 0.86 | 0.79 | 1.00 | 0.88 |
|  | 1.0 | 0.86 | 1.00 | 0.72 | 0.84 |
| KNN | 0.0 | 0.92 | 0.87 | 1.00 | 0.93 |
|  | 1.0 | 0.92 | 1.00 | 0.84 | 0.91 |
| SVM | 0.0 | 0.88 | 0.81 | 1.00 | 0.90 |
|  | 1.0 | 0.88 | 1.00 | 0.76 | 0.86 |
| LR | 0.0 | 0.90 | 0.84 | 1.00 | 0.91 |
|  | 1.0 | 0.90 | 1.00 | 0.80 | 0.89 |
| RF | 0.0 | 0.92 | 0.87 | 1.00 | 0.93 |
|  | 1.0 | 0.92 | 1.00 | 0.84 | 0.91 |
| XGB | 0.0 | 0.94 | 0.90 | 1.00 | 0.95 |
|  | 1.0 | 0.94 | 1.00 | 0.88 | 0.94 |

**Table 3: Confusion Matrix: Healthy vs. Unhealthy Prediction**

|  |  |  |
| --- | --- | --- |
|  | **Predicted: Unhealthy** | **Predicted: Healthy** |
| **Actual: Unhealthy** | 26 | 0 |
| **Actual: Healthy** | 3 | 22 |

**6. Conclusion**

In this study, we first attempt to separate food ingredients based on their nutritional characteristics into healthy and unhealthy. The classification algorithms with which the performance was tested are RF, XGB, KNN, DT, SVM, and LR, based on the measures of classification efficiency; accuracy, precision, recall, and f1-score. In all the experiments carried out, the ensemble method, XGB, has proven to be the most powerful, showing its robustness and competence in handling this classification problem.

Traditionally, machine learning models have shown good predictive power, especially for binary classification. However, these have room for potential further development. Adding deep learning models might enhance generalization and accuracy even more when we consider bigger and more complex datasets. To understand deeper patterns hidden in the data, future practitioners can explore networks and more sophisticated architectures, such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs).

**References**

[1] Menichetti G, Ravandi B, Mozaffarian D, Barabási AL. Machine learning prediction of the degree of food processing. Nat Commun. 2023 Apr 21;14(1):2312. doi: 10.1038/s41467-023-37457-1. PMID: 37085506; PMCID: PMC10121643.

[2] Qasrawi R, Sgahir S, Nemer M, Halaikah M, Badrasawi M, Amro M, Vicuna Polo S, Abu Al-Halawa D, Mujahed D, Nasreddine L, Elmadfa I, Atari S, Al-Jawaldeh A. Machine Learning Approach for Predicting the Impact of Food Insecurity on Nutrient Consumption and Malnutrition in Children Aged 6 Months to 5 Years. Children (Basel). 2024 Jul 2;11(7):810. doi: 10.3390/children11070810. PMID: 39062259; PMCID: PMC11274836.

[3] D.K. Gupta, R.D. Sharma, Ritu Gupta , S. Tyagi , K.K. Sharma and Anurag Choudhary, FORMULATION AND EVALUATION OF ORODISPERSIBLE TABLETS OF SALBUTAMOL SULPHATE

[4] Tachie, Christabel & Tawiah, Nii & Aryee, Alberta. (2023). Using machine learning models to predict quality of plant-based foods. Current Research in Food Science. 7. 100544. 10.1016/j.crfs.2023.100544.

[5] Kirk, Daniel & Kok, Esther & Tufano, Michele & Tekinerdogan, Bedir & Feskens, Edith & Camps, Guido. (2022). Machine Learning in Nutrition Research. Advances in Nutrition. 13. 10.1093/advances/nmac103.

[6] Adjuik, Toby & Boi-Dsane, Naa & Kehinde, Bababode. (2024). Enhancing dietary analysis: Using machine learning for food caloric and health risk assessment. Journal of Food Science. 89. 8006-8021. 10.1111/1750-3841.17421.

[7] Nayak, Sasmita & Beura, Mamata & Siddique, Mohammed & Mishra, Siba. (2021). Analysis of Indian Food Based on Machine learning Classification Models. Journal of Scientific Research and Reports. 27. 1-7. 10.9734/JSRR/2021/v27i730407.

[8] Chowdhury, Shovan & Schoen, Marco. (2020). Research Paper Classification using Supervised Machine Learning Techniques. 1-6. 10.1109/IETC47856.2020.9249211.

[9] Soofi, Aized & Awan, Arshad. (2017). Classification Techniques in Machine Learning: Applications and Issues. Journal of Basic & Applied Sciences. 13. 459-465. 10.6000/1927-5129.2017.13.76.

[10] Akinsola, J E T. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. International Journal of Computer Trends and Technology (IJCTT). 48. 128 - 138. 10.14445/22312803/IJCTT-V48P126.

[11] Zhang, P., Jia, Y., & Shang, Y. (2022). Research and application of XGBoost in imbalanced data. International Journal of Distributed Sensor Networks. https://doi.org/10.1177/15501329221106935.

[12] Maalouf, Maher. (2011). Logistic regression in data analysis: An overview. International Journal of Data Analysis Techniques and Strategies. 3. 281-299. 10.1504/IJDATS.2011.041335.

[13] Widodo, Agus & Handoyo, Samingun. (2017). The classification performance using logistic regression and support vector machine (SVM). Journal of Theoretical and Applied Information Technology. 95.

[14] Rimal, Y., Sharma, N., Paudel, S., Alsadoon, A., Koirala, M. P., & Gill, S. (2025). Comparative analysis of heart disease prediction using logistic regression, SVM, KNN, and random forest with cross-validation for improved accuracy. Scientific Reports, 15(1), 1-14. <https://doi.org/10.1038/s41598-025-93675-1>

[15] Bentéjac, Candice & Csörgő, Anna & Martínez-Muñoz, Gonzalo. (2019). A Comparative Analysis of XGBoost. 10.48550/arXiv.1911.01914.

[16] Santhanam, Ramraj & Uzir, Nishant & Raman, Sunil & Banerjee, Shatadeep. (2017). Experimenting XGBoost Algorithm for Prediction and Classification of Different Datasets.

[17] Liew, X. Y., Hameed, N., & Clos, J. (2021). An investigation of XGBoost-based algorithm for breast cancer classification. Machine Learning With Applications, 6, 100154. <https://doi.org/10.1016/j.mlwa.2021.100154>

[18] Hussein, Ahmed & Li, Tianrui & Chubato, Wondaferaw & Bashir, Kamal. (2019). A-SMOTE: A new preprocessing approach for highly imbalanced datasets by improving SMOTE. International Journal of Computational Intelligence Systems. 12. 10.2991/ijcis.d.191114.002.

[19] S. M. Imran and A. Geetha, "Evaluating the Effectiveness of Smote for Imbalanced Data Expansion and Its Impact on Classification Accuracy," 2024 First International Conference for Women in Computing (InCoWoCo), Pune, India, 2024, pp. 1-7, doi: 10.1109/InCoWoCo64194.2024.10863344.

[20] Kuppusami, Ganesh Kumar. (2019). Implementing modified Logistic Regression to classify Fruits. 10.13140/RG.2.2.35036.74882.

[21] Shaik, Subhani. (2025). Ensemble Learning Techniques for Rice Nutrient Disease Deficiency Detection and Prediction Analysis. Asian Journal of Research in Computer Science. 18. 129-139. 10.9734/ajrcos/2025/v18i5644.

[22] Leo Breiman. Random forests. Machine Learning, 45(1):5–32, 2001.

[23] Ali, Jehad & Khan, Rehanullah & Ahmad, Nasir & Maqsood, Imran. (2012). Random Forests and Decision Trees. International Journal of Computer Science Issues(IJCSI). 9.

[24] Cutler, Adele & Cutler, David & Stevens, John. (2011). Random Forests. 10.1007/978-1-4419-9326-7\_5.

[25] C. Moisa *et al.*, ‘Comparative Analysis of Vitamin, Mineral Content, and Antioxidant Capacity in Cereals and Legumes and Influence of Thermal Process’, *Plants*, vol. 13, no. 7, 2024.

[26] 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* **2023**, *15*, 2749.

[27] Tardy, A.L.; Pouteau, E.; Marquez, D.; Yilmaz, C.; Scholey, A. Vitamins and Minerals for Energy, Fatigue and Cognition: A Narrative Review of the Biochemical and Clinical Evidence. *Nutrients* **2020**, *12*, 228.

[28] Petroni ML, Brodosi L, Marchignoli F, Sasdelli AS, Caraceni P, Marchesini G, Ravaioli F. Nutrition in Patients with Type 2 Diabetes: Present Knowledge and Remaining Challenges. Nutrients. 2021 Aug 10;13(8):2748. doi: 10.3390/nu13082748. PMID: 34444908; PMCID: PMC8401663.

[29] Rai, B.K., Jha, A., Srivastava, S., Bind, A. (2025). Heart Disease Prediction Model Using Machine Learning Techniques. Computation of Artificial Intelligence and Machine Learning. ICCAIML 2024. Communications in Computer and Information Science, vol 2184. Springer, Cham. <https://doi.org/10.1007/978-3-031-71481-8_23> .

[30] Rai, B.K., Jain, I., Tiwari, B. *et al.* Multimodal mental state analysis. *Health Serv Outcomes Res Method* (2024). <https://doi.org/10.1007/s10742-024-00329-2>

[31] Rai BK, Kumar G, Balyan V (eds) AI and blockchain in healthcare. In: Advanced technologies and societal change. Springer, Singapore. <https://doi.org/10.1007/978-981-99-0377-1>

[32] Rai, B.K., Srivastava, A.K., Sharma, S., Kamboj, S. (2024). Proposed Model for Detection of Pneumonia Using Deep Learning. Lecture Notes in Electrical Engineering, vol 1096. Springer, Singapore. <https://doi.org/10.1007/978-981-99-7137-4_56>