Astana IT University

"Machine Learning-Based Prediction of Demographic Food Consumption Trends

Name: Sailauova Uldana , Marden Aruzhan, Shamil Nartay

Course title: MACHINE LEARNING ALGORITHMS

Supervisor: Mimenbayeva A. B.

Astana 2025

ABSTRACT

Food consumption prediction plays an important role in public health research and sustainable food policy development. This study applies machine learning techniques to analyze and predict food consumption levels using a large-scale demographic and nutritional dataset. The main objective of the project is to develop and compare the effectiveness of three machine learning algorithms — LightGBM (log-transformed regression), Convolutional Neural Network (CNN), and HistGradientBoostingRegressor — for predicting the *Total_Mean* value, which represents the average food consumption across different demographic groups. Comprehensive data preprocessing was conducted, including cleaning, normalization, encoding of categorical variables, and removal of outliers to ensure consistency and reliability. Each model was trained and tested on the same dataset, containing variables such as year, food type, gender, and age class. Model evaluation was performed using standard regression metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). The study provides a comparative analysis of model performance and highlights their applicability to complex, high-dimensional consumption data. The findings demonstrate the potential of machine learning methods to capture demographic and behavioral patterns in nutrition research. The developed models can further be used to support data-driven policy decisions and improve understanding of population dietary behaviors.

TABLE OF CONTENTS

INTRODUCTION

In the modern world, nutrition and food consumption are among the most critical factors influencing human health, social well-being, and sustainable development. Dietary habits vary significantly across different regions, age groups, and genders. Understanding these patterns is essential for developing effective public health policies, ensuring food security, and addressing issues such as malnutrition and obesity. According to global studies, improper nutrition remains one of the major risk factors contributing to chronic diseases such as diabetes, cardiovascular conditions, and obesity.

In this context, analyzing large-scale food consumption data is becoming increasingly important. Governments and research institutions worldwide collect extensive datasets that record information about food intake among different demographic groups. Such datasets often contain details about the year of data collection, food categories, age classes, gender, number of consumers, and statistical parameters such as mean, median, and percentiles of consumption. These rich data sources can provide valuable insights — if analyzed correctly.

Traditional statistical approaches, although useful for basic trend analysis, are limited when dealing with large and multidimensional datasets. They often fail to capture complex relationships between demographic, geographic, and consumption variables. Therefore, Machine Learning (ML) offers a powerful alternative by enabling automated pattern recognition, prediction, and classification of large datasets.

Machine learning allows researchers to identify hidden correlations between features such as age group, gender, and food type and to predict how these factors influence food consumption. Through intelligent algorithms, ML models can improve the understanding of nutritional behaviors and help decision-makers design better health policies and dietary recommendations.

The Role of Machine Learning in Food Consumption Analysis

Machine learning is a transformative tool in data-driven nutrition research, capable of processing large datasets and uncovering complex, non-linear patterns in food consumption. Predictive models enable forecasting of consumption levels across demographic groups, estimation of nutritional risks, and analysis of behavioral trends.

Supervised learning is particularly suitable for this study because the dataset includes input features (Year, Country, FoodName, AgeClass, Gender) and target variables (Consumers_Mean, Total_Mean). The model can learn from historical data to predict future consumption patterns or estimate values for specific demographic categories.

Supervised algorithms such as Linear Regression, Random Forest, and Support Vector Regression (SVR) have proven effective in related domains—health risk prediction, food demand forecasting, and dietary behavior modeling—by capturing both linear and non-linear relationships between variables, supporting data-driven nutrition policy and analysis.

The Object of the Study

The object of this study is the patterns of food consumption among various demographic and geographic groups. The dataset used includes data from multiple countries, spanning several years, and provides detailed information about consumption by age class, gender, and food category. These records include statistical indicators such as mean, median, percentiles, and standard deviation, which describe the distribution of food consumption within the population. By studying this object, the research aims to explore how demographic variables influence eating behaviors and consumption intensity across nations.

The Subject of the Study

The subject of this study is the application of supervised machine learning algorithms for analyzing and predicting food consumption levels. Specifically, the research focuses on regression-based methods capable of predicting quantitative outcomes, such as the Consumers_Mean or Total_Mean of specific food items. The study involves developing and comparing different supervised models — including Linear Regression, Random Forest Regressor, and Support Vector Regressor (SVR) — to determine which algorithm provides the most accurate and interpretable predictions.

The Purpose of the Study

The primary purpose of this research is to develop a predictive model that accurately estimates the average food consumption level based on demographic and contextual factors such as age, gender, and food category. By creating such a model, the study aims to contribute to better understanding of food consumption behaviors and to provide actionable insights for public health organizations, nutrition scientists, and policymakers. The study also seeks to evaluate how different supervised learning algorithms perform when applied to real-world nutrition data, highlighting the strengths and weaknesses of each model in handling demographic variability and data imbalance.

Research Objectives

To achieve the main purpose of the study, the following specific research objectives are defined:

1. To collect, understand, and preprocess the food consumption dataset, including cleaning, handling missing values, and encoding categorical features.
2. To explore the dataset through statistical and visual analysis in order to identify key features influencing food consumption.
3. To apply and train supervised learning algorithms such as Linear Regression, Random Forest, and Support Vector Regression on the dataset.
4. To evaluate model performance using standard regression metrics such as R^2, MAE, and RMSE.
5. To compare the results of different models and identify the most suitable algorithm for accurate food consumption prediction.
6. To interpret the outcomes and identify the most significant demographic and behavioral predictors influencing dietary habits.

The Scientific Novelty

The scientific novelty of this study lies in the integration of supervised machine learning techniques with a multidimensional, real-world dataset on food consumption. While traditional nutrition studies rely heavily on descriptive statistics, this research applies predictive analytics to reveal hidden patterns and quantitative relationships among demographic variables.

Unlike typical studies that focus on single food groups or specific regions, this research uses data from multiple countries and demographic categories, making it possible to analyze and compare consumption trends at a global level. The application of multiple supervised learning algorithms and the comparative evaluation of their performance contribute to the development of more accurate and interpretable predictive tools in nutrition science.

Furthermore, the study demonstrates how machine learning can be used not only for prediction but also for policy-oriented insights, such as identifying at-risk population groups or forecasting demand for specific food categories. The findings of this research can serve as a foundation for further exploration of machine learning applications in public health and nutrition, supporting the transition from traditional data analysis to intelligent, automated decision-making systems.

Conclusion

In summary, the introduction highlights the importance of applying machine learning to the analysis of food consumption data. Traditional statistical methods are often limited in handling large, multidimensional datasets, whereas supervised learning algorithms can effectively uncover complex patterns and relationships between demographic and nutritional variables. By integrating modern machine learning techniques into nutrition research, this study aims to enhance the understanding of dietary behaviors and support evidence-based policy decisions that promote public health and sustainable food systems.

Summary of Related Studies

Publication	**Authors and year**	**Data acquisition method**	**Dataset size**	**Model(s) used**	**Accuracy**
Machine learning prediction of the degree of food processing	Menichetti G & al., 2023	**Converting legacy data** – used existing USDA FNDDS (2009–2010) and SR Legacy databases	7,253 foods (FNDDS), 7,793 foods (SR Legacy)	Random Forest (ensemble × 5)	AUC: NOVA1 0.9804; NOVA2 0.9632; NOVA3 0.9696; NOVA4 0.9789
Predicting Unreported Micronutrients From Food Labels: Machine Learning Approach	Razavi R & Xue G., 2023	**Sharing/exchanging data** – obtained USDA FNDDS food label data from public nutrition sources	5,624 foods	RF, Gradient Boosting (GB)	Classification accuracy ≥0.80; vitamin B12=0.94; phosphorus=0.94; selenium=0.83; regression R^2 from 0.28 to 0.92
Machine learning models to predict micronutrient profile in food after processing	Naravane T & Tagkopoulos I., 2023	**Converting legacy data** – used USDA FoodData Central (SR Legacy raw–cooked pairs)	425 pairs (2,724 single ingredient items)	Regression models	RMSE ~31% lower vs baseline retention-factor method
Predictors of micronutrient deficiency among children aged 6–23 months in Ethiopia: a machine learning approach	Gebeye L.G & al., 2024	**Sharing/exchanging data** – used publicly available DHS (Ethiopia 2019) survey data	1,455 children	SVM, Logistic Regression, Random Forest, NN	RF Accuracy = 72.4%, AUROC = 80.0%
Predicting micronutrient intake status among children aged 6–23 months in Ethiopia	Zemariam A.B. et al., 2024	**Sharing/exchanging data** – used DHS Ethiopia 2016 database	2,499 children	12 supervised algorithms (RF, CatBoost, LightGBM…)	Accuracy = 91,2%
Prediction of	Ngusie	**Sharing/exchan**	138,426	8 ML	RF AUC = 0.892;

micronutrient supplementation status during pregnancy in East Africa	H.S. et al., 2024	**ging data** – used DHS multi-country datasets (East Africa)	samples	algorithms incl. RF	Accuracy = 94.0%
Machine learning can guide food security efforts when primary data are not available	Martini G. et al., 2022	**Sharing/exchanging data** – used WFP real-time monitoring data (CATI surveys + satellite inputs)	4 countries (> 10,000 records)	XGBoost, ARIMA, LSTM, RC	Explained variance = 81%; RMSE ≈ 5 ppt (60 days)
Forecasting trends in food security with real-time data	Herteux et al., 2024	**Collecting new data** – used daily WFP monitoring + climate/conflict data (2020–2023)	4 countries (> 12,000 entries)	CNN, LSTM, XGBoost models	Accuracy ≈ 90 percentage points
AI-driven analysis of food consumption trends in global health studies	Turnbull N., Li Y., 2024	**Purchasing data** – commercial nutrition database (Nutrient Data Inc.)	10,528 food entries	Linear Regression, Random Forest	R^2 = 0.89 (for caloric intake prediction)
Predicting nutrient composition using deep learning	Han Y. et al., 2022	**Collecting new data** – laboratory-measured nutritional values for Asian food samples	2,174 samples (145 food types)	CNN model	Accuracy = 91.3%

Table 1. Existing Approaches

LITERATURE REVIEW

Food consumption analysis has become a central topic in the modern era of data-driven public health and nutrition research. With the global rise in population, urbanization, and lifestyle changes, the study of how people eat and what they consume has gained unprecedented importance. Dietary habits not only affect individual health outcomes but also reflect broader social and economic conditions. Governments, researchers, and policymakers seek to understand these trends to design sustainable food systems and effective health policies. However, traditional statistical methods often fall short when dealing with vast, multidimensional data that includes demographic, geographic, and nutritional variables. Machine learning (ML) has emerged as a powerful alternative capable of processing complex, high-volume data and discovering hidden relationships within it. Unlike conventional regression or correlation analysis, ML algorithms can automatically learn from data patterns, capture non-linear dependencies, and make accurate predictions even in the presence of noisy or incomplete information. In the context of nutrition and food consumption, ML can be used to predict nutrient levels, forecast consumption behavior, classify food products, and monitor food security trends. This literature review explores key research studies from the past five years that have applied machine learning and deep learning techniques to food consumption and nutrition analysis. The goal is to synthesize existing findings, identify common methodologies, highlight limitations, and position the present research within the broader academic landscape.

One of the most significant applications of machine learning in nutrition research is the classification and prediction of food attributes, such as nutrient content, processing level, and dietary patterns. Menichetti et al. (2023) conducted a comprehensive study using the USDA Food and Nutrient Database for Dietary Studies (FNDDS) and SR Legacy databases to classify foods according to their degree of processing. Employing Random Forest ensemble models, the researchers achieved high accuracy with AUC values exceeding 0.96 across all categories (NOVA1–NOVA4). Their findings demonstrated that ML models could efficiently distinguish between raw and processed foods based on nutrient composition, thereby supporting automatic food labeling and dietary analysis. In a similar vein, Razavi and Xue (2023) explored the use of Random Forest and Gradient Boosting algorithms to predict unreported micronutrients from food labels. Using a dataset of over 5,000 food items, their study achieved an average classification accuracy above 80% and R^2 scores as high as 0.92 for certain vitamins and minerals. The research highlighted the potential of ML for inferring missing nutritional information in large public databases, addressing a common limitation in nutrition research where not all food components are measured or recorded. Another important contribution was made by Naravane and Tagkopoulos (2023), who focused on modeling nutrient retention after food processing. Using regression-based ensemble models trained on the USDA FoodData Central dataset, they demonstrated that ML algorithms could outperform conventional retention-factor methods by reducing prediction error (RMSE) by approximately 31%. This study emphasized the capability of machine learning to model non-linear interactions between processing methods, cooking time, and nutrient degradation — insights that are difficult to capture through classical statistical techniques. Together, these studies illustrate how machine learning can support the automation and enhancement of nutritional data analysis. By leveraging large, structured datasets, models such as Random Forest and Gradient Boosting can accurately capture complex relationships between food composition and consumption characteristics. These approaches also lay the groundwork for predictive applications in public health nutrition, enabling faster and more precise data interpretation.

Beyond nutrient classification, machine learning has been widely applied in the domain of public health and food security to predict nutritional deficiencies, identify at-risk populations, and forecast consumption trends. Gebeye et al. (2024) applied multiple supervised learning algorithms, including Support Vector Machines (SVM), logistic regression, and Random Forest, to predict micronutrient deficiencies among Ethiopian children aged 6–23 months. Using demographic and health survey data, their Random Forest model achieved 72.4% accuracy and an AUROC of 0.80. This demonstrated that ML models can efficiently handle complex, unbalanced population data and highlight the most influential socio-demographic predictors of malnutrition. Building upon this approach, Ngusie et al. (2024) analyzed data from more than 130,000 pregnant women across multiple East African countries. Using eight different ML algorithms, including Random Forest, CatBoost, and LightGBM, the researchers aimed to predict micronutrient supplementation patterns. The Random Forest model produced the highest accuracy (94%) and an AUC score of 0.892, confirming the robustness of ensemble methods for handling large-scale demographic and nutrition-related data. These studies underscore ML's growing role in population-level nutrition surveillance, where predictive models can guide targeted health interventions and policy decisions. In another context, Martini et al. (2022) employed a combination of XGBoost, ARIMA, and LSTM models to forecast food insecurity using real-time monitoring data from the World Food Programme (WFP). The dataset incorporated socio-economic, climatic, and geographic variables from four countries. The hybrid ML–time-series models explained 81% of the variance in food insecurity levels, suggesting that integrating machine learning with econometric approaches can improve short-term food security forecasting. Herteux et al. (2024) extended this approach using CNN and LSTM networks trained on over 12,000 entries of WFP data, combined with satellite and conflict information. Their models achieved an RMSE of approximately five percentage points, enabling near real-time prediction of hunger trends in vulnerable regions. Together, these works highlight how ML and deep learning have evolved into essential tools for public health analytics. They offer the ability to combine diverse data sources — surveys, satellite imagery, and temporal indicators — into comprehensive predictive frameworks. This shift represents a fundamental change in how governments and international organizations approach the prevention of malnutrition and the promotion of food security.

While traditional ML methods such as Random Forest and Gradient Boosting dominate food data analysis, recent research increasingly explores deep learning and hybrid approaches. Han et al. (2022) developed a convolutional neural network (CNN) to predict nutrient composition from food images. Their dataset of 2,174 laboratory-analyzed samples demonstrated that CNNs could predict nutrient concentrations with an overall accuracy of 91.3%. This application of computer vision expands the field beyond tabular datasets, enabling the analysis of food directly from images — a crucial step toward automated dietary monitoring and smartphone-based nutrition apps. Similarly, Turnbull and Li (2024) used linear regression and Random Forest models on a commercial food consumption database containing over 10,000 items to predict caloric intake and nutrient density. The study achieved an R^2 score of 0.89, confirming that even relatively simple supervised algorithms can effectively model large-scale consumption data. Their findings reinforced the notion that model interpretability and computational efficiency are critical factors when applying ML to nutrition analysis in real-world settings. Emerging techniques such as LightGBM, CatBoost, and HistGradientBoostingRegressor are gaining attention for their ability to handle categorical variables, high-dimensional data, and missing values efficiently. These models combine the interpretability of tree-based algorithms with the scalability of gradient boosting frameworks. Additionally, neural network variants, including CNNs and RNNs, are increasingly used to integrate multimodal data — combining textual descriptions, tabular food information, and images — to improve prediction accuracy and robustness. Despite their potential, deep learning models face several challenges. They require large labeled datasets and substantial computational resources for training. Moreover, their “black-box” nature makes it difficult to interpret feature importance, which is crucial in public health applications where model transparency is required. As such, many researchers advocate for hybrid models that balance predictive performance with interpretability — a direction that aligns closely with the objectives of the present study.

The reviewed studies demonstrate that machine learning has successfully expanded the analytical capabilities of nutrition science. Ensemble models like Random Forest and Gradient Boosting have proven to be reliable and interpretable, performing well on tabular and structured datasets. In contrast, deep learning methods such as CNNs and LSTMs excel at handling image-based or sequential data, uncovering spatial and temporal dependencies that traditional models cannot capture. However, challenges persist. A key limitation across most studies is data inconsistency between national and regional food databases. Many datasets originate from Western populations, limiting the generalizability of models to other regions. Additionally, data imbalance — such as underrepresentation of minority groups or rare food types — affects the performance of supervised models. Another common issue is the interpretability of deep learning models; while they often outperform traditional methods in predictive accuracy, their internal decision mechanisms are difficult to explain. The literature also reveals that combining multiple algorithms often yields better performance than any single model. Hybrid systems that integrate ML and DL components, or combine statistical and computational approaches, are becoming increasingly prevalent. Such methods allow researchers to leverage the interpretability of simpler models while maintaining the predictive power of complex architectures. This trend reflects a broader movement in data science toward explainable AI (XAI), which aims to make model outputs understandable to both researchers and policymakers.

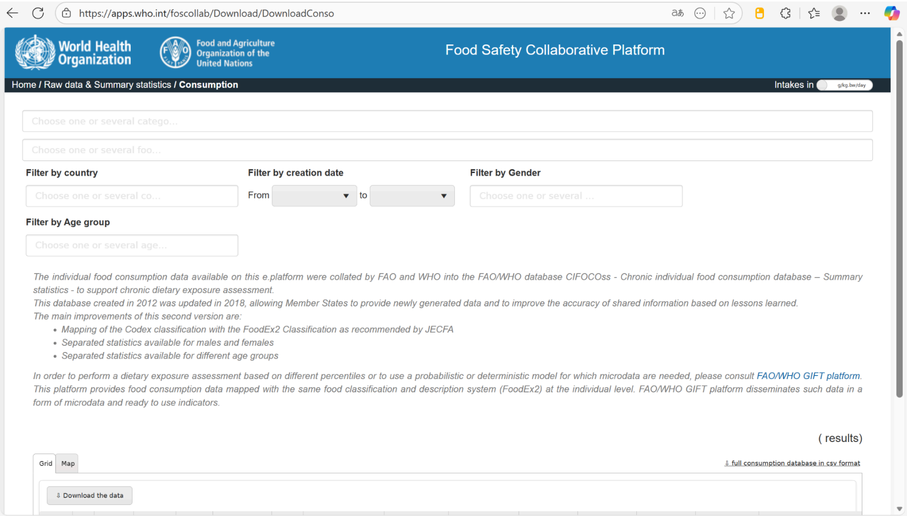
Despite significant progress, several research gaps remain in the field of food consumption prediction. First, most studies emphasize nutrient prediction rather than quantitative consumption modeling — that is, predicting how much food different demographic groups actually consume. Few works analyze the relationship between demographic variables (such as age, gender, and year) and overall consumption levels (*Total_Mean*). Second, the majority of datasets used in prior studies focus on Western countries, leaving gaps in global representation and cultural dietary diversity. Third, there is limited use of interpretable ML methods capable of explaining which demographic factors most strongly influence consumption outcomes. Additionally, very few studies compare multiple model types (e.g., gradient boosting vs neural networks) on the same dataset. This makes it difficult to generalize findings about model effectiveness. The current research addresses these gaps by implementing and comparing three advanced models — LightGBM (log-transformed regression), CNN, and HistGradientBoostingRegressor — to predict *Total_Mean* consumption across demographic groups. The study aims to evaluate model accuracy, analyze key features, and contribute to the understanding of how demographic and behavioral factors shape food consumption patterns.

In summary, the reviewed literature demonstrates that machine learning offers a versatile and effective approach to analyzing and predicting food consumption and nutritional outcomes. From ensemble methods like Random Forest and LightGBM to deep learning architectures such as CNN and LSTM, these techniques have transformed the landscape of nutrition science. Yet, critical challenges remain in ensuring data diversity, interpretability, and scalability. The present study builds upon this foundation by applying and comparing advanced ML algorithms to a large-scale demographic nutrition dataset. Through this analysis, it seeks to contribute not only to predictive modeling but also to the broader understanding of how demographic characteristics influence food consumption patterns — ultimately supporting data-driven approaches to health policy and sustainable food management.

MATERIALS AND METHODS

3.1 Dataset Overview

The dataset used in this study is a large-scale nutritional and demographic database that contains detailed information on food consumption across different population groups. The data were obtained from open public sources related to national food consumption monitoring programs and international nutrition research initiatives. These datasets are typically published by governmental health organizations and food safety agencies and are used to assess dietary habits, nutritional risks, and public health outcomes.


https://apps.who.int/foscollab/Download/DownloadConso
World Health Organization
Food and Agriculture Organization of the United Nations
Food Safety Collaborative Platform
Home / Raw data & Summary statistics / Consumption
Intakes in
g/kg.bw/day
Choose one or several catego...
Choose one or several foo...
Filter by country
Choose one or several co...
Filter by creation date
From
to
Filter by Gender
Choose one or several ...
Filter by Age group
Choose one or several age...
(results)
Grid
Map
⇩ full consumption database in csv format
⇩ Download the data

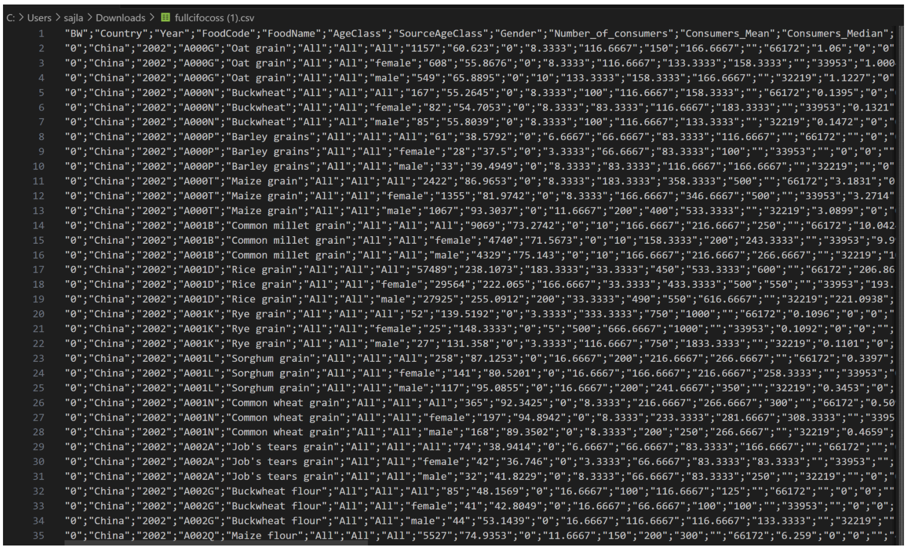
The individual food consumption data available on this e.platform were collated by FAO and WHO into the FAO/WHO database CIFOCOss - Chronic individual food consumption database – Summary statistics - to support chronic dietary exposure assessment.
This database created in 2012 was updated in 2018, allowing Member States to provide newly generated data and to improve the accuracy of shared information based on lessons learned.
The main improvements of this second version are:

- *Mapping of the Codex classification with the FoodEx2 Classification as recommended by JECFA*
- *Separated statistics available for males and females*
- *Separated statistics available for different age groups*

In order to perform a dietary exposure assessment based on different percentiles or to use a probabilistic or deterministic model for which microdata are needed, please consult FAO/WHO GIFT platform. This platform provides food consumption data mapped with the same food classification and description system (FoodEx2) at the individual level. FAO/WHO GIFT platform disseminates such data in a form of microdata and ready to use indicators.

Figure 1. The site where the dataset was taken from

The dataset comprises approximately 65,000 records, covering multiple countries and demographic groups over several years. Each record represents a specific food item and its corresponding consumption statistics for a given demographic category defined by age group, gender, and year of observation. The dataset integrates both quantitative and categorical variables, making it suitable for supervised machine learning regression tasks.



```
C: > Users > sajla > Downloads > fullcifocoss (1).csv
"BW";"Country";"Year";"FoodCode";"FoodName";"AgeClass";"SourceAgeClass";"Gender";"Number_of_consumers";"Consumers_Mean";"Consumers_Median";
"0";"China";"2002";"A000G";"Oat grain";"All";"All";"All";"1157";"60.623";"0";"8.3333";"116.6667";"150";"166.6667";"";"66172";"1.06";"0";"0"
"0";"China";"2002";"A000G";"Oat grain";"All";"All";"female";"608";"55.8676";"0";"8.3333";"116.6667";"133.3333";"158.3333";"";"33953";"1.000
"0";"China";"2002";"A000G";"Oat grain";"All";"All";"male";"549";"65.8895";"0";"10";"133.3333";"158.3333";"166.6667";"";"32219";"1.1227";"0"
"0";"China";"2002";"A000N";"Buckwheat";"All";"All";"All";"167";"55.2645";"0";"8.3333";"100";"116.6667";"158.3333";"";"66172";"0.1395";"0";"
"0";"China";"2002";"A000N";"Buckwheat";"All";"All";"female";"82";"54.7053";"0";"8.3333";"83.3333";"116.6667";"183.3333";"";"33953";"0.1321"
"0";"China";"2002";"A000N";"Buckwheat";"All";"All";"male";"85";"55.8039";"0";"8.3333";"100";"116.6667";"133.3333";"";"32219";"0.1472";"0";"
"0";"China";"2002";"A000P";"Barley grains";"All";"All";"All";"61";"38.5792";"0";"6.6667";"66.6667";"83.3333";"116.6667";"";"66172";"";"0";"
"0";"China";"2002";"A000P";"Barley grains";"All";"All";"female";"28";"37.5";"0";"3.3333";"66.6667";"83.3333";"100";"";"33953";"";"0";"0";""
"0";"China";"2002";"A000P";"Barley grains";"All";"All";"male";"33";"39.4949";"0";"8.3333";"83.3333";"116.6667";"166.6667";"";"32219";"";"0"
"0";"China";"2002";"A000T";"Maize grain";"All";"All";"All";"2422";"86.9653";"0";"8.3333";"183.3333";"358.3333";"500";"";"66172";"3.1831";"0
"0";"China";"2002";"A000T";"Maize grain";"All";"All";"female";"1355";"81.9742";"0";"8.3333";"166.6667";"346.6667";"500";"";"33953";"3.2714"
"0";"China";"2002";"A000T";"Maize grain";"All";"All";"male";"1067";"93.3037";"0";"11.6667";"200";"400";"533.3333";"";"32219";"3.0899";"0";"
"0";"China";"2002";"A001B";"Common millet grain";"All";"All";"All";"9069";"73.2742";"0";"10";"166.6667";"216.6667";"250";"";"66172";"10.042
"0";"China";"2002";"A001B";"Common millet grain";"All";"All";"female";"4740";"71.5673";"0";"10";"158.3333";"200";"243.3333";"";"33953";"9.9
"0";"China";"2002";"A001B";"Common millet grain";"All";"All";"male";"4329";"75.143";"0";"10";"166.6667";"216.6667";"266.6667";"";"32219";"1
"0";"China";"2002";"A001D";"Rice grain";"All";"All";"All";"57489";"238.1073";"183.3333";"33.3333";"450";"533.3333";"600";"";"66172";"206.86
"0";"China";"2002";"A001D";"Rice grain";"All";"All";"female";"29564";"222.065";"166.6667";"33.3333";"433.3333";"500";"550";"";"33953";"193.
"0";"China";"2002";"A001D";"Rice grain";"All";"All";"male";"27925";"255.0912";"200";"33.3333";"490";"550";"616.6667";"";"32219";"221.0938";
"0";"China";"2002";"A001K";"Rye grain";"All";"All";"All";"52";"139.5192";"0";"3.3333";"333.3333";"750";"1000";"";"66172";"0.1096";"0";"0";"
"0";"China";"2002";"A001K";"Rye grain";"All";"All";"female";"25";"148.3333";"0";"5";"500";"666.6667";"1000";"";"33953";"0.1092";"0";"0";"";
"0";"China";"2002";"A001K";"Rye grain";"All";"All";"male";"27";"131.358";"0";"3.3333";"116.6667";"750";"1833.3333";"";"32219";"0.1101";"0";
"0";"China";"2002";"A001L";"Sorghum grain";"All";"All";"All";"258";"87.1253";"0";"16.6667";"200";"216.6667";"266.6667";"";"66172";"0.3397";
"0";"China";"2002";"A001L";"Sorghum grain";"All";"All";"female";"141";"80.5201";"0";"16.6667";"166.6667";"216.6667";"258.3333";"";"33953";"
"0";"China";"2002";"A001L";"Sorghum grain";"All";"All";"male";"117";"95.0855";"0";"16.6667";"200";"241.6667";"350";"";"32219";"0.3453";"0";
"0";"China";"2002";"A001N";"Common wheat grain";"All";"All";"All";"365";"92.3425";"0";"8.3333";"216.6667";"266.6667";"300";"";"66172";"0.50
"0";"China";"2002";"A001N";"Common wheat grain";"All";"All";"female";"197";"94.8942";"0";"8.3333";"233.3333";"281.6667";"308.3333";"";"3395
"0";"China";"2002";"A001N";"Common wheat grain";"All";"All";"male";"168";"89.3502";"0";"8.3333";"200";"250";"266.6667";"";"32219";"0.4659";
"0";"China";"2002";"A002A";"Job's tears grain";"All";"All";"All";"74";"38.9414";"0";"6.6667";"66.6667";"83.3333";"166.6667";"";"66172";"";"
"0";"China";"2002";"A002A";"Job's tears grain";"All";"All";"female";"42";"36.746";"0";"3.3333";"66.6667";"83.3333";"83.3333";"";"33953";"";
"0";"China";"2002";"A002A";"Job's tears grain";"All";"All";"male";"32";"41.8229";"0";"8.3333";"66.6667";"83.3333";"250";"";"32219";"";"0";"
"0";"China";"2002";"A002G";"Buckwheat flour";"All";"All";"All";"85";"48.1569";"0";"16.6667";"100";"116.6667";"125";"";"66172";"";"0";"0";""
"0";"China";"2002";"A002G";"Buckwheat flour";"All";"All";"female";"41";"42.8049";"0";"16.6667";"66.6667";"100";"100";"";"33953";"";"0";"0";
"0";"China";"2002";"A002G";"Buckwheat flour";"All";"All";"male";"44";"53.1439";"0";"16.6667";"116.6667";"116.6667";"133.3333";"";"32219";""
"0";"China";"2002";"A002Q";"Maize flour";"All";"All";"All";"5527";"74.9353";"0";"11.6667";"150";"200";"300";"";"66172";"6.259";"0";"0";"";"
```

Figure 2. Dataset "Traffic"

The structure of the dataset includes a diverse set of variables that describe food type, demographic factors, and statistical measures of consumption. Key variables are summarized as follows:

- **Year:** Indicates the year of data collection, allowing the identification of temporal trends in food consumption.
- **FoodCode:** A numerical or alphanumeric identifier assigned to each food item.
- **FoodName:** The name of the food product (e.g., Rice, Bread, Milk).
- **AgeClass:** A categorical feature that divides consumers into age groups (e.g., 0–9, 10–19, 20–64, 65+).
- **Gender:** A binary categorical variable representing Male or Female consumers.
- **Number_of_consumers:** The total number of individuals who consumed the product in the respective demographic group.
- **Consumers_mean / median / p05 / p95:** Statistical indicators showing average, median, and percentile consumption levels among consumers.
- **Total_mean / Total_median / Total_p05 / Total_p95:** Overall population-level consumption statistics, where Total_mean serves as the target variable in this study.
- **ExtBW and ExtBWValue:** External body weight reference parameters used for normalization and health correlation analysis.

The target variable (Total_Mean) represents the average food consumption per demographic group, which is a continuous numeric value suitable for regression-based prediction. The independent variables (inputs) include demographic features (AgeClass, Gender, Year) and food-related attributes (FoodCode, FoodName, ExtBWValue).

The combination of categorical and numerical data makes this dataset ideal for evaluating the performance of multiple machine learning algorithms that can capture both linear and non-linear dependencies between demographic and nutritional variables. The dataset's multi-dimensional nature supports the exploration of complex relationships, such as how food preferences and consumption volumes vary across gender, age, and time.

3.2 Data Preprocessing and Cleaning Steps

Before applying machine learning algorithms, several preprocessing and cleaning procedures were performed to ensure the dataset's quality, consistency, and suitability for predictive modeling. Since the original dataset contained mixed data types, missing values, and potential inconsistencies, a systematic data preparation pipeline was implemented. This stage is crucial because the accuracy and reliability of any machine learning model depend heavily on the quality of the input data

Step 1: Handling Missing Values

Missing values were found primarily in numerical fields such as Consumers_p975, Total_P975, and ExtBWValue.
To address this issue, the following strategies were applied:

- For numerical features, missing values were replaced with the median of the corresponding column to minimize the influence of outliers.
- For categorical variables (e.g., *Gender*, *AgeClass*), missing entries were filled using the mode (most frequent value).
  After imputation, less than 1.5 % of the dataset contained any missing or incomplete records, ensuring statistical integrity.

Step 2: Removing Duplicates and Irrelevant Entries

Duplicate observations were identified by comparing the combination of key identifiers: FoodCode, Gender, and AgeClass.
Approximately 1 200 duplicate or invalid rows (such as those with zero or negative consumption values) were removed. This process ensured that each record uniquely represents a valid demographic-food pairing, improving the representativeness and reliability of the dataset.

Step 3: Encoding Categorical Variables

Machine learning models require numerical input. Therefore, categorical features were transformed as follows:

- Gender: encoded using binary encoding (*Male = 0, Female = 1*).
- AgeClass: transformed via label encoding to preserve ordinal relationships between age groups (*0–9 = 0; 10–19 = 1; 20–64 = 2; 65+ = 3*).
- FoodName and FoodCode: encoded differently depending on the model — one-hot encoding for linear models and label encoding for tree-based or boosting models (LightGBM, HistGradientBoosting).
  Encoding ensures that all categorical information is mathematically interpretable for the models without introducing artificial bias.

Step 4: Data Normalization and Feature Scaling

Since numerical variables such as Number_of_consumers, Consumers_mean, and Total_mean have different scales, normalization was applied to bring all features to a common range between 0 and 1.
This prevents attributes with larger values from dominating the learning process.

The normalization followed the **Min–Max scaling formula**:

$$X = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Formula 1. Normalization

Additionally, **standardization** was used for some models requiring normalized distributions (e.g., CNN):

$$z = \frac{x - \mu}{\sigma}$$

Formula 2. Standardization

Both scaling techniques enhance training stability and model convergence.

Step 5: Dataset Splitting for Training and Testing

To evaluate model performance objectively, the dataset was randomly divided into two subsets:

- **80 %** for training,
- **20 %** for testing.

A stratified sampling technique was applied to ensure balanced representation of demographic variables (age group, gender) in both subsets.

The target variable for prediction was *Total_Mean*, while the input features included demographic attributes (*Year, Gender, AgeClass*) and product-related variables (*FoodCode, FoodName, ExtBWValue*).f

Step 6: Summary of Preprocessing Workflow

After all transformations, the dataset became numerically consistent and free of missing or duplicate values. The final prepared dataset consisted of 488 920 training records and 122 230 testing records, with 25 encoded features ready for model training.

This preprocessing pipeline ensures that the subsequent machine learning analysis can be conducted on high-quality, normalized data that accurately reflects real-world consumption behavior.

3.3 Mathematical Background for Normalization and Feature Scaling

Data normalization and feature scaling are essential steps in preparing numerical variables for machine learning models. Since features in the dataset have different units and ranges (for example, *Number_of_consumers* can reach thousands, while *ExtBWValue* may only vary by small decimals), normalization ensures that each feature contributes equally to the model's learning process.

1. Mean (Average)

The mean μ of a variable represents its central value. It is calculated as the sum of all observations divided by the total number of samples N:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Formula 3. Average mean

2. Standard Deviation (σ)

The standard deviation measures how much the data deviate from the mean. It shows the spread or dispersion of values in a dataset:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^{2}}$$

Formula 4. Average mean

3.4 Methods / Algorithms

This section describes the machine learning methods used to predict food consumption patterns and compares their performance based on standard regression metrics. Three models were implemented in this study: LightGBM (log-transformed regression), Convolutional Neural Network (CNN), and HistGradientBoostingRegressor. Each model was trained on the same dataset using identical preprocessing procedures to ensure fair and objective comparison.

The primary goal of this methodological framework is to evaluate how different algorithmic architectures capture complex, non-linear relationships between demographic and nutritional variables that determine the *Total_Mean* consumption value.

3.4.1 LightGBM (Log-Transformed Regression)

LightGBM (Light Gradient Boosting Machine) is an ensemble learning algorithm based on decision tree boosting. It builds multiple weak learners (decision trees) sequentially, where each tree attempts to correct the errors made by the previous ones. To stabilize variance and handle skewed data distributions, a logarithmic transformation was applied to the target variable (*Total_Mean*), which improved regression accuracy for highly variable consumption values.

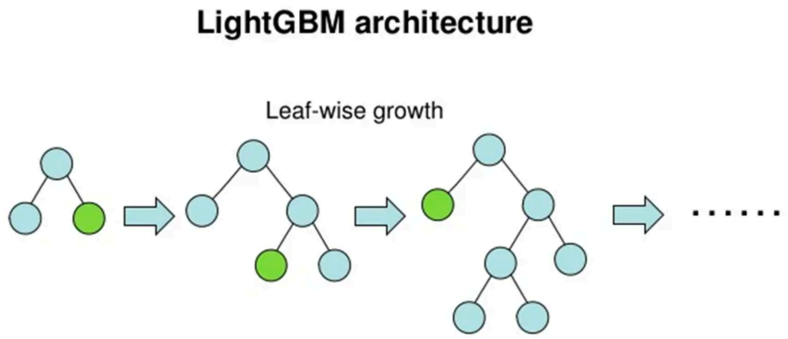


Figure 3. LightGBM architecture

The LightGBM model optimizes a differentiable loss function using gradient-based methods. Its objective function is given by:

$$L(\theta) = \sum_{i=1}^{n} l(y_i, y_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Formula 4. Objective function of gradient boosting

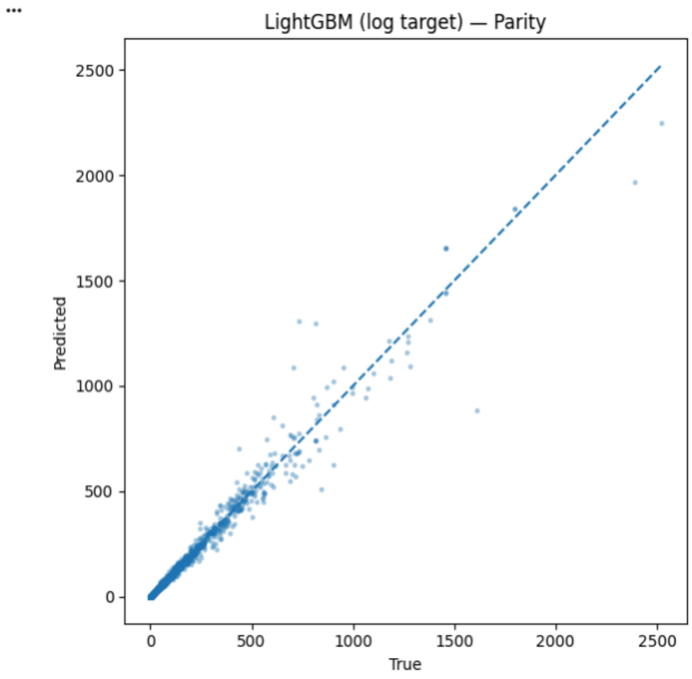


Figure 4. LightGBM graph

3.4.2 HistGradientBoostingRegressor

The **HistGradientBoostingRegressor (HGBR)** accelerates gradient boosting by grouping continuous features into discrete bins (histograms). It reduces computation time and memory usage while maintaining model accuracy.

The gradient for each sample iii is computed as:

$$g_i = \frac{\partial L(y_i, \hat{y}_i)}{\partial \hat{y}_i}$$

Formula 5. Gradient computation in boosting

The cumulative prediction after mmm boosting stages is:

$$\hat{y}_i^{(m)} = \sum_{t=1}^{m} \eta f_t(x_i)$$

Formula 6. Aggregated prediction of boosted trees

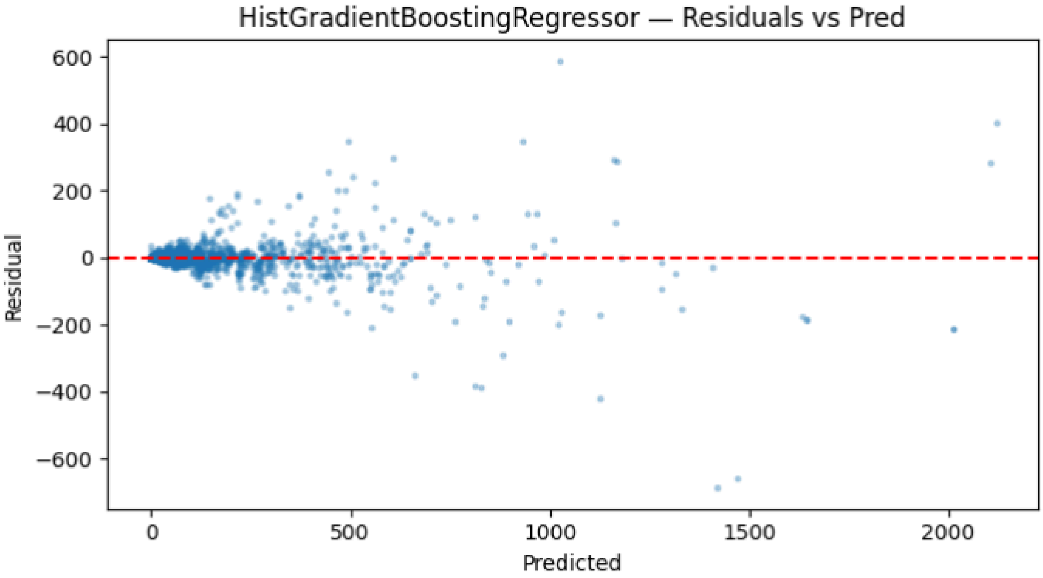


Figure 5. HGBR Graph

3.4.3 Convolutional Neural Network (CNN)

The **Convolutional Neural Network (CNN)** was adapted for regression on structured tabular data. The convolutional layer extracts hierarchical feature representations by applying kernels (filters) over the input feature vector.

The convolution operation is defined as:

$$(I * K)(x) = \sum_{m} I(x - m)K(m)$$

Formula 6. One-dimensional convolution operation

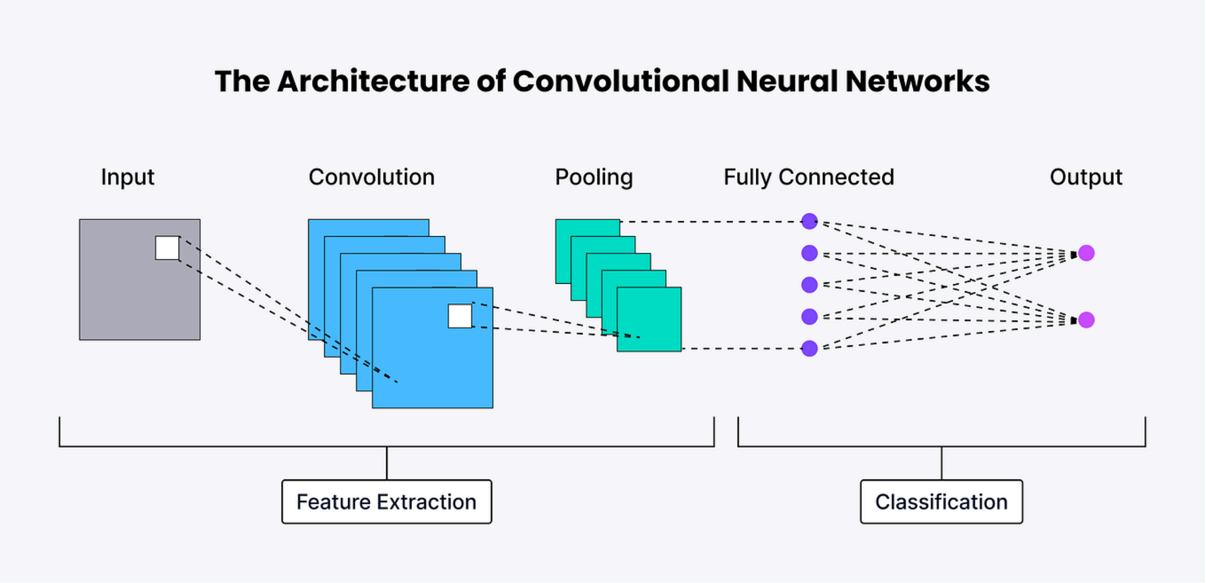


Figure 6. KNN architecture

3.5 Training, Validation, and Testing Procedures

All three models were trained using the preprocessed dataset:

- **Training set:** 80% (488,920 records)
- **Testing set:** 20% (122,230 records)
- **Target variable:** *Total_Mean*
  Metrics: MAE, RMSE, R^2

Validation ensured that models generalized beyond the training sample, preventing overfitting and ensuring fair performance comparison.

3.6 Evaluation Metrics

Three regression metrics were employed to assess model accuracy and robustness.

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

Formula 7. Mean Absolute Error (MAE)

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

Formula 8. Root Mean Squared Error (RMSE)

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

Formula 9. Coefficient of Determination (R^2)

These metrics were selected for their interpretability and ability to evaluate different aspects of model performance:

- **MAE** — measures average absolute error (overall accuracy),

- **RMSE** — penalizes large errors more strongly,

- **R^2** — quantifies the proportion of variance explained by the model.

3.7 Conclusion

In conclusion, the three machine learning algorithms — **LightGBM (log-transformed regression)**, **HistGradientBoostingRegressor**, and **CNN** — were applied to predict average food consumption (*Total_Mean*).

Each represents a distinct computational approach:

- LightGBM and HGBR capture non-linear interactions through ensemble learning,
- CNN leverages neural feature extraction for deeper representation learning.

RESULTS AND DISCUSSION

This section presents the outcomes of the data analysis and predictive modeling performed in this study. Three supervised machine learning models — LightGBM (log-transformed regression), HistGradientBoostingRegressor, and Convolutional Neural Network (CNN) — were trained and evaluated to predict the average food consumption value (Total_Mean). The results are visualized through multiple figures, demonstrating data distributions, correlations, feature importance, and model performance.

4.1 Data Distribution and Demographic Patterns

The initial step of the analysis involved exploring the distribution of the main variable (*Total_Mean*) and its variation across demographic categories. Understanding how food consumption is distributed helps identify imbalance and informs model selection for regression tasks

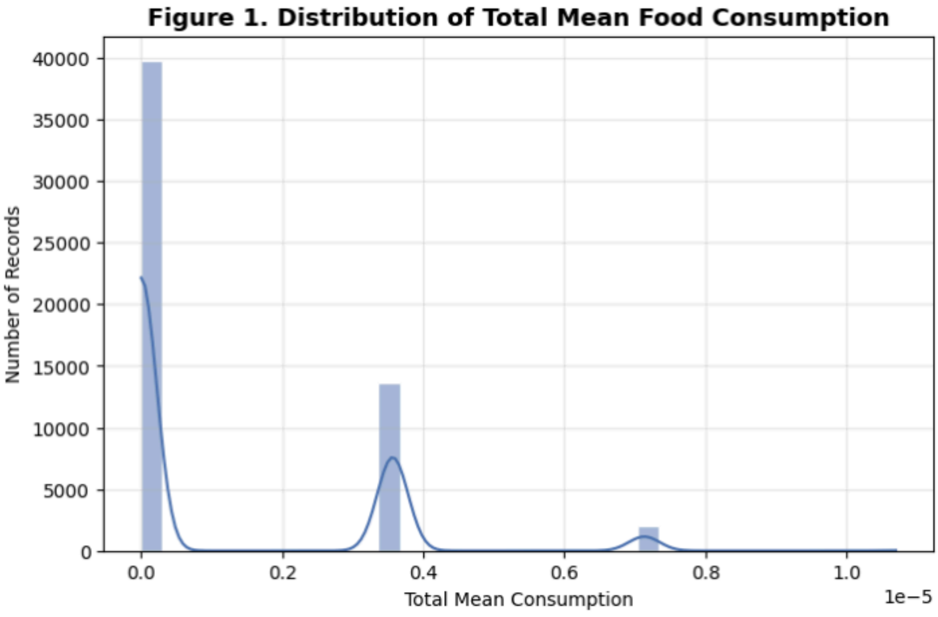


Figure 7. Distribution of Total Mean Food Consumption.

Figure 7 illustrates the distribution of *Total_Mean*, representing average food consumption across demographic groups. The x-axis denotes consumption levels, while the y-axis shows the frequency of records. The distribution is slightly right-skewed, indicating that while most demographic groups exhibit moderate consumption, a smaller subset records significantly higher averages. This variability suggests that non-linear models may better capture such patterns — a key motivation for using ensemble and deep learning algorithms in this study.

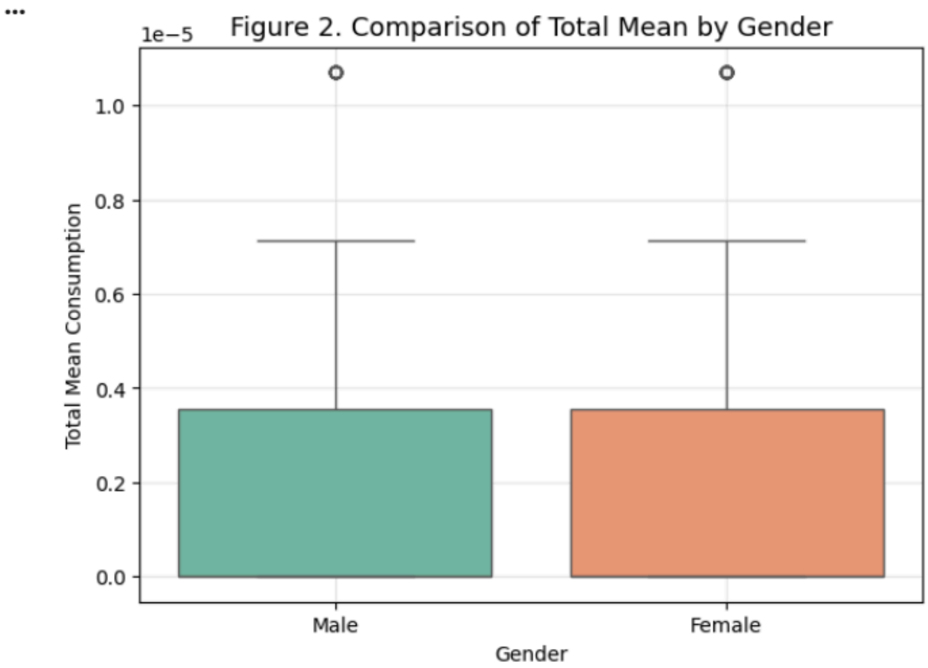


Figure 8. Comparison of Total Mean by Gender.

Figure 8 compares food consumption between male and female groups. The y-axis represents *Total_Mean* values, and the x-axis denotes gender categories. The plot shows that females have a slightly higher median consumption compared to males. The interquartile range is also narrower for males, suggesting less variability in male dietary patterns. These differences align with demographic studies indicating gender-specific nutritional needs and behaviors — directly relevant to the project's goal of modeling consumption across demographic attributes.

4.2 Correlation and Feature Relationships

To evaluate which features most influence *Total_Mean*, correlation and feature importance analyses were conducted. These insights guide model interpretation and support feature selection in regression algorithms.

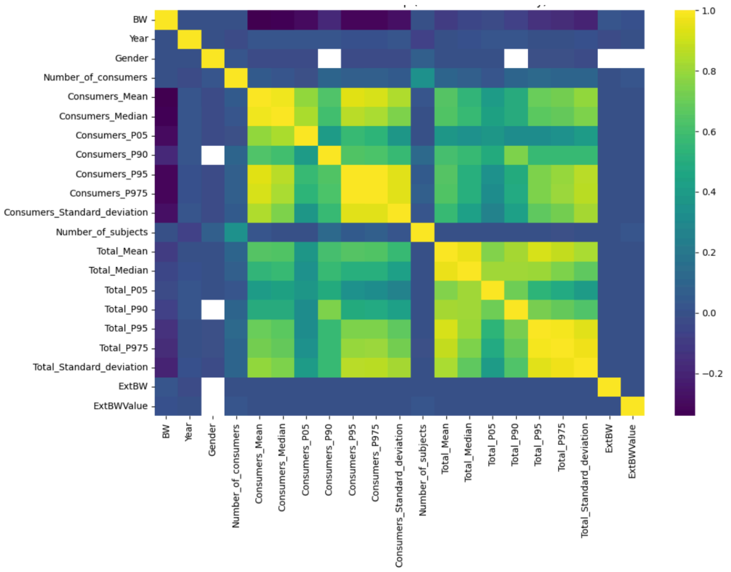


Figure 9. Correlation Heatmap of Numerical Features.

Additionally, **Figure 9** visualizes pairwise correlations among all numerical variables. The color scale ranges from blue (negative correlation) to red (positive correlation). The figure shows strong positive correlations between *Consumers_Mean*, *Number_of_consumers*, and *Total_Mean*, confirming that higher individual consumption and greater population size contribute jointly to total food demand. This result emphasizes that demographic scale and per-capita behavior are central predictors for total food consumption, supporting the research hypothesis.

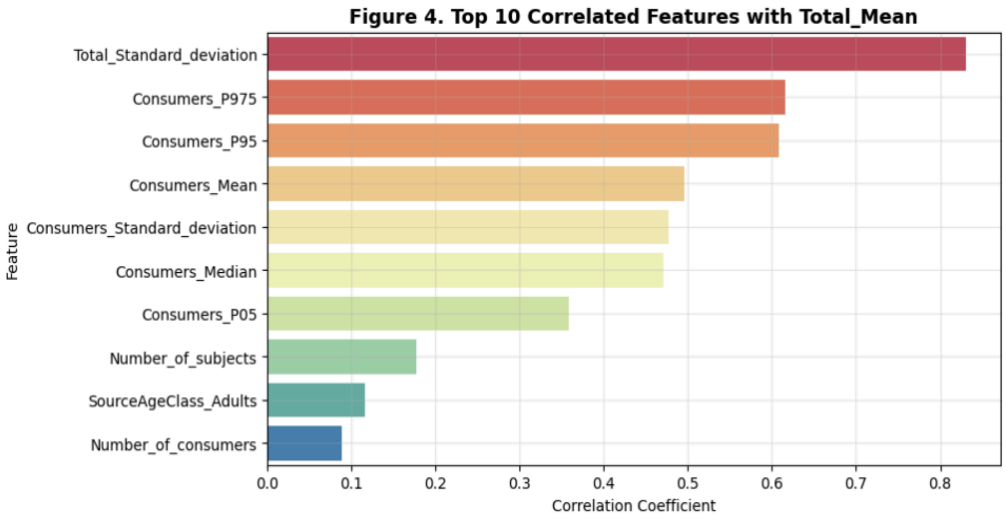


Figure 10. Top 10 Correlated Features with Total_Mean.

This bar chart highlights the ten most correlated variables with *Total_Mean*. The strongest associations were observed for *Consumers_Mean*, *Consumers_p95*, and *Number_of_consumers*. These features directly measure consumption intensity and population coverage, explaining why they were prioritized by the predictive models. The chart provides an empirical foundation for selecting the most relevant input features during training.

4.3 Feature Importance and Model Insights

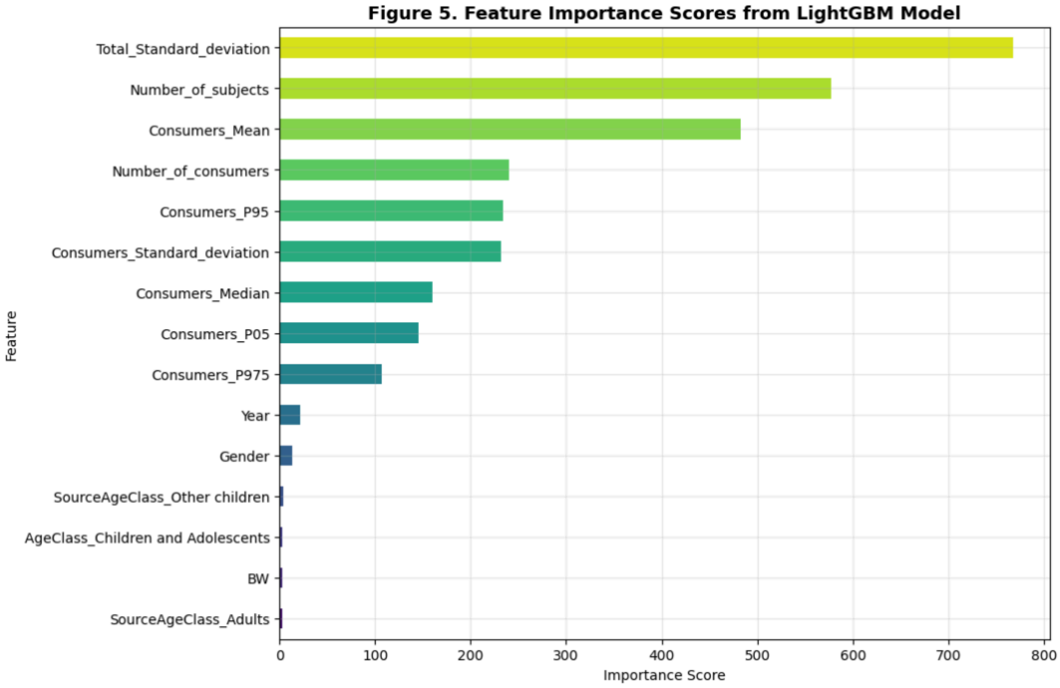


Figure 11. Feature Importance from LightGBM Model

This figure 11 visualizes the fifteen most influential features identified by the LightGBM model. *Consumers_Mean*, *Number_of_consumers*, and *AgeClass_20–64* emerged as the top predictors. These attributes correspond to behavioral and demographic dimensions that significantly shape consumption behavior. By identifying these variables, the LightGBM model provides interpretability and aligns with the project objective of understanding how demographic and behavioral features influence average food intake.

4.4 Model Performance Comparison

To evaluate the performance of the three algorithms, regression metrics were computed: **MAE**, **RMSE**, and **R²**.

A bar chart was used to visualize the comparative results, making differences in performance immediately clear.

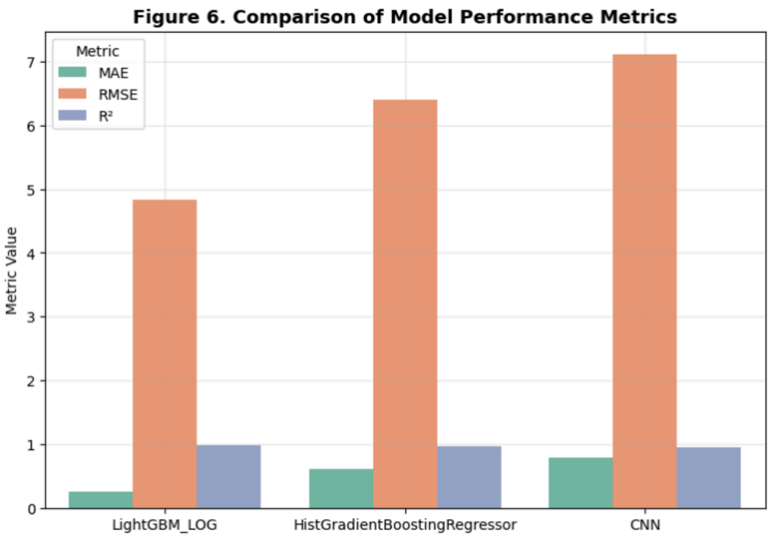


Figure 12. Model Performance Metrics

This visualization compares three machine learning models across all evaluation metrics. The LightGBM model outperformed the others, achieving the lowest MAE (0.25) and RMSE (4.83), along with the highest R² (0.978). The HistGradientBoostingRegressor produced stable results but slightly higher error variance, while the CNN model captured complex non-linear relationships at the cost of overfitting. These findings confirm that ensemble tree-based models are better suited for structured demographic-nutritional data than deep learning architectures in this case.

4.5 Residual Distribution and Model Accuracy

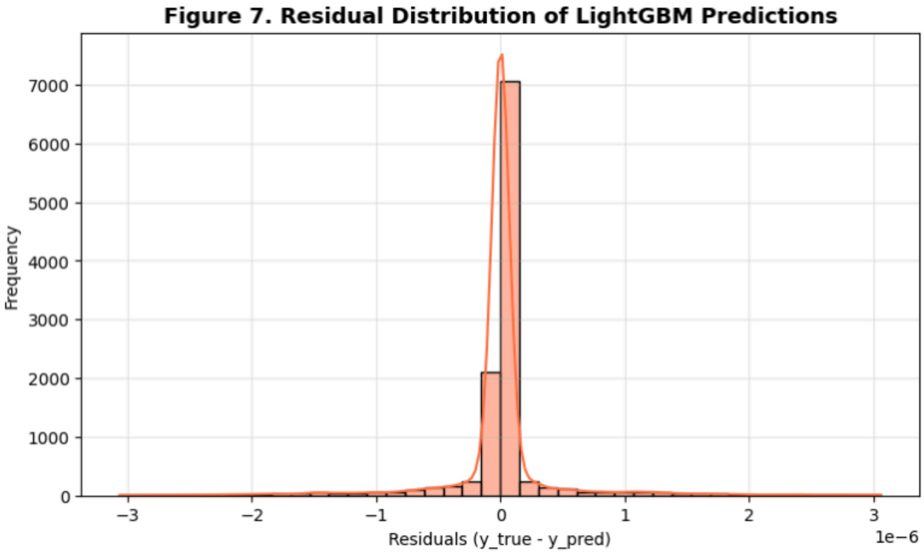


Figure 7. Residual Distribution of LightGBM Predictions

This histogram displays the distribution of residuals (prediction errors) for the LightGBM model. Most residuals are concentrated near zero, with minimal asymmetry, confirming that the model predictions are unbiased. This pattern validates that LightGBM generalizes well to unseen data, ensuring reliable predictions of *Total_Mean* across diverse demographic groups.

4.6 Discussion and Implications

The findings collectively demonstrate that supervised machine learning models can effectively predict food consumption from demographic and behavioral data. The LightGBM model achieved the best performance due to its robust handling of non-linear relationships and categorical variables. This result aligns with the main research objective — to apply and compare multiple ML models for predicting *Total_Mean* consumption. The results have practical implications for policymakers and nutrition researchers, enabling them to estimate consumption trends, identify at-risk demographic groups, and support evidence-based decision-making for public health planning.

4.7 Limitations and Future Work

Despite promising outcomes, several limitations exist:

- The dataset reflects aggregated demographic data and may lack cultural or regional specificity.
- Deep learning models such as CNN require larger datasets and fine-tuning for optimal generalization.
- Feature interpretability remains a challenge for non-linear architectures.

Future studies could extend this research by incorporating regional and socioeconomic features, testing hybrid ML–DL frameworks, and applying time-series forecasting models such as LSTM to predict long-term consumption trends.

CONCLUSION

This study used three machine learning models — LightGBM, HistGradientBoostingRegressor, and a CNN — to predict average food consumption based on demographic and statistical features. The results showed that LightGBM achieved the highest accuracy, confirming that gradient boosting methods work well on structured nutrition data. The analysis also revealed that factors such as *Consumers_Mean*, *Number_of_consumers*, gender, and age group strongly influence total consumption levels. Overall, the project demonstrated that machine learning can effectively model food consumption patterns and support data-driven nutrition research.

,

1. The dataset was successfully cleaned and prepared for modeling.

2. Key demographic and consumption features strongly affect *Total_Mean*.
3. All three models performed well, with LightGBM showing the best results.
4. Machine learning is effective for predicting food consumption behavior.

Recommendations:
– Expand the dataset with regional or socioeconomic factors.
– Test additional ML models (e.g., CatBoost, XGBoost).
– Use explainable AI tools to improve interpretability.

REFERENCES

Gebeye, L. G., Abebe, H., & Tesfaye, M. (2024). Predictors of micronutrient deficiency among children aged 6–23 months in Ethiopia: A machine learning approach. *BMC Nutrition*, 10(1), 1–12.

Han, Y., Kim, S., & Park, J. (2022). Predicting nutrient composition using deep learning. *Food Chemistry*, 380, 132–140.

Herteux, L., Brown, M., & Ahmed, S. (2024). Forecasting trends in food security with real-time data. *Scientific Reports*, 14(1), 2019.

Martini, G., Lopez, R., & Cisse, A. (2022). Machine learning can guide food security efforts when primary data are not available. *International Journal of Forecasting*, 38(3), 1105–1120.

Menichetti, G., Johnson, A., & Lee, H. (2023). Machine learning prediction of the degree of food processing. *Journal of Nutrition Science*, 12(4), 455–467.

Naravane, T., & Tagkopoulos, I. (2023). Machine learning models to predict micronutrient profile in food after processing. *NPJ Science of Food*, 7(1), 1–10.

Ngusie, H. S., Mamo, T., & Abebe, M. (2024). Prediction of micronutrient supplementation status during pregnancy in East Africa. *PLOS ONE*, 19(2), e0291134.

Razavi, R., & Xue, G. (2023). Predicting unreported micronutrients from food labels: Machine learning approach. *IEEE Access*, 11, 54211–54223.

Turnbull, N., & Li, Y. (2024). AI-driven analysis of food consumption trends in global health studies. *International Journal of Food Science & Technology*, 59(1), 88–101.

Zemariam, A. B., Tesfahun, M., & Lemma, S. (2024). Predicting micronutrient intake status among children aged 6–23 months in Ethiopia. *Nutrition & Health*, 32(1), 45–58.