Computers in Biology and Medicine 135 (2021) 104606

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

HANA: A Healthy Artificial Nutrition Analysis model during COVID-19 pandemic

Mahmoud Y. Shams [a], Omar M. Elzeki [b,*], Lobna M. Abouelmagd [c], Aboul Ella Hassanien [d,f], Mohamed Abd Elfattah [c], Hanaa Salem [e]

[a] Faculty of Artificial Intelligence, Kafrelsheikh University, 33511, Egypt
[b] Faculty of Computers and Information, Mansoura University, 35516, Mansoura, Egypt
[c] Misr Higher Institute for Commerce and Computers, Mansoura, Egypt
[d] Faculty of Computers and Artificial Intelligence, Cairo University, Egypt
[e] Faculty of Engineering, Delta University for Science and Technology, Gamasa, Egypt
[f] Scientific Research Group in Egypt (SRGE), Cairo, Egypt

ARTICLE INFO

Keywords:
COVID-19
Healthy food
Regression
Artificial intelligence
Machine learning
Nutrition analysis

ABSTRACT

Background and objective: The impact of diet on COVID-19 patients has been a global concern since the pandemic began. Choosing different types of food affects peoples' mental and physical health and, with persistent consumption of certain types of food and frequent eating, there may be an increased likelihood of death. In this paper, a regression system is employed to evaluate the prediction of death status based on food categories.

Methods: A Healthy Artificial Nutrition Analysis (HANA) model is proposed. The proposed model is used to generate a food recommendation system and track individual habits during the COVID-19 pandemic to ensure healthy foods are recommended. To collect information about the different types of foods that most of the world's population eat, the COVID-19 Healthy Diet Dataset was used. This dataset includes different types of foods from 170 countries around the world as well as obesity, undernutrition, death, and COVID-19 data as percentages of the total population. The dataset was used to predict the status of death using different machine learning regression models, i.e., linear regression (ridge regression, simple linear regularization, and elastic net regression), and AdaBoost models.

Results: The death status was predicted with high accuracy, and the food categories related to death were identified with promising accuracy. The Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R^2 metrics and 20-fold cross-validation were used to evaluate the accuracy of the prediction models for the COVID-19 Healthy Diet Dataset. The evaluations demonstrated that elastic net regression was the most efficient prediction model. Based on an in-depth analysis of recent nutrition recommendations by WHO, we confirm the same advice already introduced in the WHO report[1]. Overall, the outcomes also indicate that the remedying effects of COVID-19 patients are most important to people which eat more vegetal products, oilcrops grains, beverages, and cereals - excluding beer. Moreover, people consuming more animal products, animal fats, meat, milk, sugar and sweetened foods, sugar crops, were associated with a higher number of deaths and fewer patient recoveries. The outcome of sugar consumption was important and the rates of death and recovery were influenced by obesity.

Conclusions: Based on evaluation metrics, the proposed HANA model may outperform other algorithms used to predict death status. The results of this study may direct patients to eat particular types of food to reduce the possibility of becoming infected with the COVID-19 virus.

1. Introduction

In many countries throughout the world, the current coronavirus (COVID-19) pandemic has led to general lockdowns that have resulted in the closure of all but essential services, such as grocery stores and pharmacies. Such closures have had an immediate and predictable effect

* Corresponding author.
E-mail address: omar_m_elzeki@mans.edu.eg (O.M. Elzeki).

https://doi.org/10.1016/j.compbiomed.2021.104606
Received 24 April 2021; Received in revised form 17 June 2021; Accepted 21 June 2021
Available online 30 June 2021

on food obtainability and selection. The pandemic has restricted food selections, which may impact mealtimes and diet, as well as having a general effect on both physical and psychological health [1].

Food systems have a direct and indirect impact on human health, and now it is more important than ever that they should become sustainable. In 2015, the United Nations' 2030 Plan for Sustainable Development issued an immediate call for action, including 17 sustainable development goals, by developed and emerging countries in global collaboration [2].

Consequently, diagnostic tools for food prediction and saving food in surrounding environments after the lockdown are needed. In addition, food and industry supply chains need to be monitored to determine if they have contributed to the spread of COVID-19. This is performed by examining how COVID-19 spreads through surfaces, the food supply chain, and surrounding environments [3]. For example, Mishra and Rampal [4] presented a study of the effect of the COVID-19 pandemic on food insecurity in India. They started by tracking the general status of food insecurity and hunger everywhere in the world, focusing on lower- and middle-income countries. They found that there are significant relationships between economic growth, joblessness, and starvation resulting from food shortages during pandemic lockdowns.

More precisely, as reported by Laborde et al. [5] that there are no lack of main food has appeared recently. Nevertheless, food markets and agriculture are in front of disruptions because of labor lacks outcomes from restrictions on activities of individuals and changes in food requests during the lockdown of restaurants and schools in addition to the shortfalls in people's income.

Nowadays, ML plays a vital role in diagnosis and prognosis issues especially tracking diseases in a medical application using image recognition systems, preventing and treat the spread of specific diseases especially in dealing with imbalanced data using ML approaches [6–8]. The diagnosis and classification of COVID-19 chest X-ray images as CT images-based ML approaches is the key feature to fight the spread of the COVID-19 virus. Moreover, the prediction system-based ML is used to forecast the effect of the current pandemic in different areas specifically in diagnosing diseases and healthcare systems [9–13].

When purchasing food, the number of choices is excessive to be capable of considering them all [14]. Individuals have dissimilar dietary needs, habits, and distinguish flavor in different habits. Consequently, the only choice is to realize their requirements by discussing the person. In some cases, the recommendation system is performed to assist a modest starving consumer, cooking supporter, concerning health, dieter, or somebody hostile seeking to enhance his/her medical prominence, which will enhance the impact of the final selection. Furthermore, the food's existence in time is required to make the customer more stratified and happier. A significant feature when building these systems is the data collection sources and customer habits. They can be collected based on user's feedback to certain posts such as tracking the likes or dislikes of the customers. As well as the recorded ratings by the public watched videos and/or images in the social media. The achievement of a food recommendation system is associated with its capability to track user favorites, maximize the number of fresh and healthy food and in contrast minimize and avoid the unhealthy.

With so many people currently getting sick from the COVID-19, unhealthy diets contribute to pre-existing medical conditions that make them more vulnerable to the virus [15]. In many parts of the world, getting sick means losing income. Hence, the pandemic has increased the risks faced by consumers, producers, and policymakers around the world [16]. What is required to get a portion of proper healthy food? The answer to this question is more urgent and necessary than ever. There is a great deal of ambiguity regarding the components of a healthy diet and appropriate policy interventions. However, there is a growing body of evidence and analysis that points to actions that will save lives-or at least a little-improve the well-being of billions of people.

Excessive metabolic risk (cholesterol, blood pressure, body mass index, blood sugar) is responsible for the most important risk factors for infection and death. There are more than two billion people infected with death or dying out of 70% of them [17]. Non-common-price foods caused 600 million illnesses, and 420,000 cases each beginning of 2010, the case for every global condition, undermining human health and food security. Emerging evidence suggests that people with pre-existing medical conditions related to diets, such as obesity, heart disease, and diabetes, suffer more serious consequences from infection with the Coronavirus, such as the severity of illness and an increased need for intensive health care, such as respirators. Therefore, a good nutrition system during the recent pandemic is recommended to provide a suitable decision for the individuals to avoid the side effects of wrong diet habits [18]. Artificial intelligence tools can present effective and promising methods to predict, plan, and provide a suitable decision for decision-makers in the field of diet and nutrition [19].

In this paper, a Healthy Artificial Nutrition Analysis (HANA) approach is proposed. The HANA used ML algorithms on available public data to generate a food recommendation system as well as tracking individual habits during the COVID-19 pandemic to ensure healthy foods. The primary contributions of this paper are as follows.

- Proposing the HANA model emphasize nutrition styles with higher mortality from Covid-19.
- We use ML-based analysis to seek food compatible nutrition styles during COVID-19 incubation.
- We predict the number of deaths resulting from poor habits related to food during the COVID-19 pandemic.
- Designing light and fast learning model as a healthy food recommendation system.
- Using ReliefF and Stochastic Gradient Descent (SGD) for optimal feature reduction based on PCA.
- HANA model outcomes confirm WHO nutrition advice for COVID-19 and nutrition studies.

The remainder of this paper is organized as follows. In Section II, focusing on models that are relevant to the recent COVID-19 pandemic, studies related to food prediction models are discussed. Section III describes the system architecture of the proposed food recommendation system. Experimental and evaluation results are discussed in Section IV. Conclusions and suggestions for future work are presented in Section V.

2. Related work

In general, three food trends depend on Artificial Intelligence (AI) that are considered when dealing with food problems. Industrial food is commercially regulated according to the stages of manufacture to improve and facilitate the consumption process. Moreover, it was introduced to provide most of the food consumed by the world's population. In agriculture, an important issue related to AI is to help farmers eliminate diseases and pests that affect plants, which in turn affects the quantity and quality of the crop, and consequently affects the volume of food. We identified many studies that help to identify plant diseases. Food is being used in the fight against poverty by developing a recommended AI-based diet to track and monitor the nutritional level in developing countries.

Patients with certain diseases, such as diabetes, heart disease, high blood pressure, and insulin resistance, are most vulnerable to COVID-19. Therefore, to avoid that, the patients should monitor and decrease the bad habits of eating especially foods with high insulin. Low carbohydrate, moderate proteins, and moderate fat are mainly required to maintain the normal insulin in the patient's body. Food containing Zinc is an efficient way to increase the human immune system performance. Oysters, shellfish, red meat, and cheese are rich in Zinc. Vitamin D also required and existed in Cod liver oil, and salmon. Vitamin C additionally is very important to decrease the percentage of COVID-19 virus existence. The food rich in vitamin C for instance leafy greens, sauerkraut, and berries [20–23].

For industrial food, Shen et al. [24] proposed an application to measure the food attributes to help people balancing their diet, as it detects food items in an image and recognizes them. The application uses the Convolution Neural Network for food recognition. The system can evaluate food properties by transferring data from the internet. They used Inception-v3 and Inception-v4 models. These models are based on Convolutional Neural Networks (CNN) and the results obtained to tackle the problem are more reliable.

Furthermore, Onu et al. [25] utilized AI models to expect low moisture content in drying potatoes. They used three different models; the Response Surface Methodology (RSM), Adaptive Neuro-Fuzzy Inference Systems (ANFIS), and Artificial Neural Network (ANN). They founded that the three models gave good prediction with the experimental data yet, RSM and ANFIS gave better results than ANN. In food processing, three cases are selected and studies gathering the machine learning and expert interaction as presented in Ref. [26]:

- In the first one, they hired experts to design the structure of the Bayes dynamic network for constructing a camembert maturing model, including variables from micro-scale (presence of bacteria and chemical components) to macro-scale (perceptual assessments).
- In the second one, they built a model to assist winemakers in assessing when to harvest grapes, depending on weather conditions, the model is also a Bayesian network model.
- Third, they used a graphical model based on symbolic regression to assist specialists making a model for bacterial production and stabilization.

An approach based on k-cluster segmentation and color detection is presented by Ref. [27] for grading, sorting fruits and vegetables, and the extracted features are calculated such as entropy, mean, and standard deviation.

In [28] the researchers produce a system where they used image processing with the help of SVM classifier to classify healthy rice plants and diseased rice plants. The system got a resolution of over 90%. Furthermore, in Ref. [29] the researchers present a proposed network structure for classifying potato leaf diseases based on CNN. The suggested architecture is consisting of 14 layers, and the average overall test accuracy is 98%. In Ref. [30] also identify leaf diseases of the apple, they use CNN based on the pre-train network AlexNet, the experiments of the proposed disease identification based on CNN give accuracy about 97.62%. To increase crop production, the researchers in Ref. [31] suggested a framework for fruit harvesting robots. The framework includes three classification models that are used to classify images of fruits in real-time date according to their type, ripeness, and harvest decision, as traditional methods may delay the production cycle of dates and represent more than 45% of the cost of production date, they used CNN with fine-tuning and transfer learning on pre-trained models. The proposed models achieve 99.01%, 97.25%, and 98.59% accuracy.

The researchers in Ref. [32] also work on date fruit; they offer a new and more accurate way to distinguish between healthy and damaged date fruits. They used deep CNN; this method can predict the maturity stage of healthy dates. The CNN model managed to achieve an overall rating accuracy of 96.98%. Furthermore, researchers were attempting to detect the food defects especially for fruits such as apple in Ref. [33] they used the modified AlexNet model with an eleven-layer structure, along with a comparison study was performed to boost classification results obtained. They use three well-known algorithms back-propagation neural networks (BPNN), Particle Swarm Optimization (PSO), and SVM. The proposed CNN model for apple detection achieves a recognition rate of 92.50%, which is higher than other algorithms commonly used, such as BPNN, SVM, and PSO algorithm.

Another issue in AI is fighting the poverty presented in Refs. [34,35], where the researchers in Ref. [29] study the data collected from five African countries: Tanzania, Nigeria, Uganda, Malawi, and Rwanda, it demonstrated that CNN can be prepared to distinguish image features up to 75% of the variance in the local economic level Results.

Their method could change efforts to track and target poverty in developing nations. In Ref. [35] the researchers present a more accurate approach for predicting the essential dimensions of poverty, health, education, and standard of living (Pearson correlation 0.84–0.86). They used Gaussian Process regression, a Bayesian learning technique, providing uncertainty associated with predictions. The model is built with an elastic net regularization to prevent overfitting. The results show maximum accuracy when using disparate data such as the resulting Pearson Correlation reached 0.91.

O'Hara and Toussaint [36] observe the insecurity of food in Washington, DC. They discovered the new chances in urban agriculture and the production of food with sustainable simultaneous food access to tackle the insecurity of local food and the required infrastructure.

Ordás et al. [37] present the habits of individuals in eating. The study of 170 countries were performed to discover the relationships between these habits and death rates caused by COVID-19 based on ML approaches taking the distribution of energy, fat, and protein through twenty-three different sorts of diets into consideration. The results indicate that 95% predicted correctly using a regression model based on Principal Component Analysis (PCA). Moreover, a course of treatment is performed for SARS-COV-2 patients to estimate the death cases using ML and Deep Learning (DL) approaches as investigated by Kivrak et al. [38].

Shams et al. [39] proposed a regression model Based on Support Vector Machine (SVM) and Deep Learning (DL) approaches given a dataset contains both confirmed deaths and recovered cases. The results achieved indicate that the RMSE using SVM's with the Radial Basis Function (RBF) kernel is 0.27, while the SVM with linear Kernel achieves 0.18 RMSE, and the deep regression model achieves 0.29 RMSE.

In this work, we can conclude that the general structure of food systems in the period during the COVID-19 pandemic is illustrated in food directions based on AI. We observed that there are four important variables or parameters that are used in the food systems applied during the COVID-19 pandemic. These criteria include food security outcomes from individual closures, food safety affected by the recent pandemic, individual public health support system, and food sustainability [2].

Patients with severe pneumonia have been identified as vulnerable to the protein-energy deficiency that significantly damages respiratory muscle contractility and the immune defense system [40,41]. According to Ref. [41], SARS-CoV-2 infected individuals are most seriously and critically unwell and at nutritional danger. For assessing the dietary risks and treatment of severe and critical COVID-19 patients a study [42] was presented in 2021. A total of 523 people were enrolled in Wuhan, China from four hospitals. The window for inclusion was between January 2, 2020 and 15 February. The computerized medical records, nursing records and associated exams were used for clinical features and laboratory data. So, the power of data science was shown in this study. It concluded to the following the high risk of malnutrition in critical and serious patients with COVID-19. The low concentration of BMI and protein was substantially related to adverse outcomes. In individuals with COVID-19, early nutrition assessment and treatment are required.

In some cases, the COVID-19 pandemic has been observed to have significant impacts on food systems around the world, through both the vulnerabilities it has revealed within food supply chains, food demand, and the purchasing power of consumers [43]. The death cases are reduced in some countries compared with other countries in Europe, the major reason is the food habits as reported by Bousquet et al. [44]. Moreover, some types of food like Cabbage and fermented vegetables are taken into consideration from the mortality heterogeneity of countries with mitigation candidates [45].

3. The proposed approach

The USDA center for Nutrition Policy and Promotion recommends a balanced diet comprising 10% fruits, 20% protein, 30% grains, and 40% vegetables. However, most people do not follow these

recommendations. The impact of an unbalanced diet is more significant during a global pandemic. In this paper, we merge the world's overweight population, starvation, and types of food as regression features and the death rate due to COVID-19 as expected values to learn more about how healthy eating can assist in fighting the disease.

To learn more about how a healthy eating style could help combat the coronavirus, we propose an enhanced and optimized feature reduction algorithm and regression model. Based on the ReliefF algorithm, the feature reduction algorithm selects the top relevant features, to predict the probability of death due to the followed diet style as COVID-19 infected candidate.

As shown in Fig. 1, the proposed HANA model consists of four stages. The first stage is data pre-processing, which includes all processes involved in collecting and managing the data. The second stage consists of feature enhancing, selection, and dimension reduction steps. The third stage utilizes different regression and prediction models. Finally, the evaluation matrices stage is used to evaluate the applied regression models as shown in Figure. The primary contributions of this proposed approach are as follows (see Fig. 2).

- First, we proposed a hybrid feature reduction algorithm based on SGD and the ReliefF algorithm. The proposed hybrid algorithm

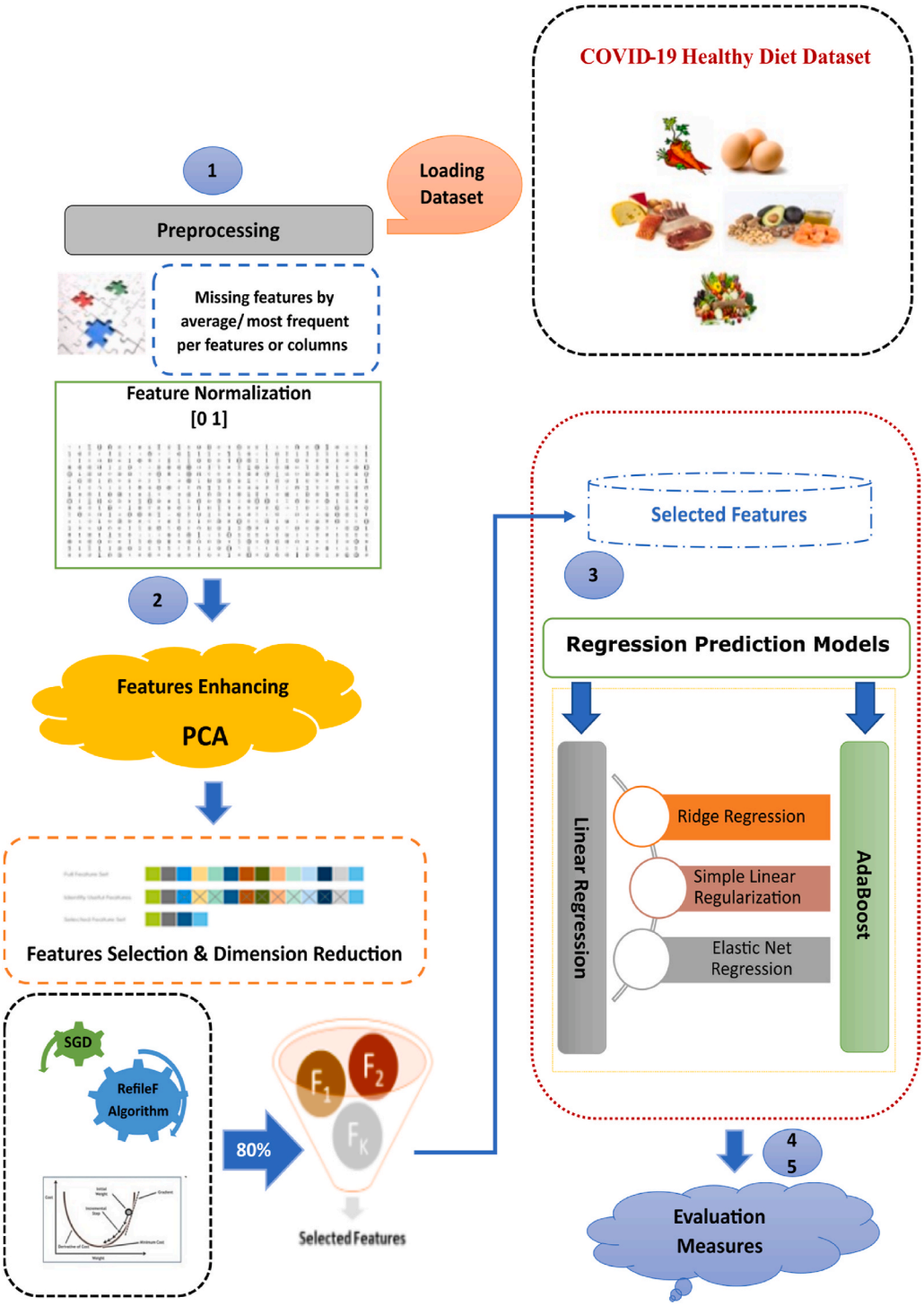

Fig. 1. Proposed HANA model.

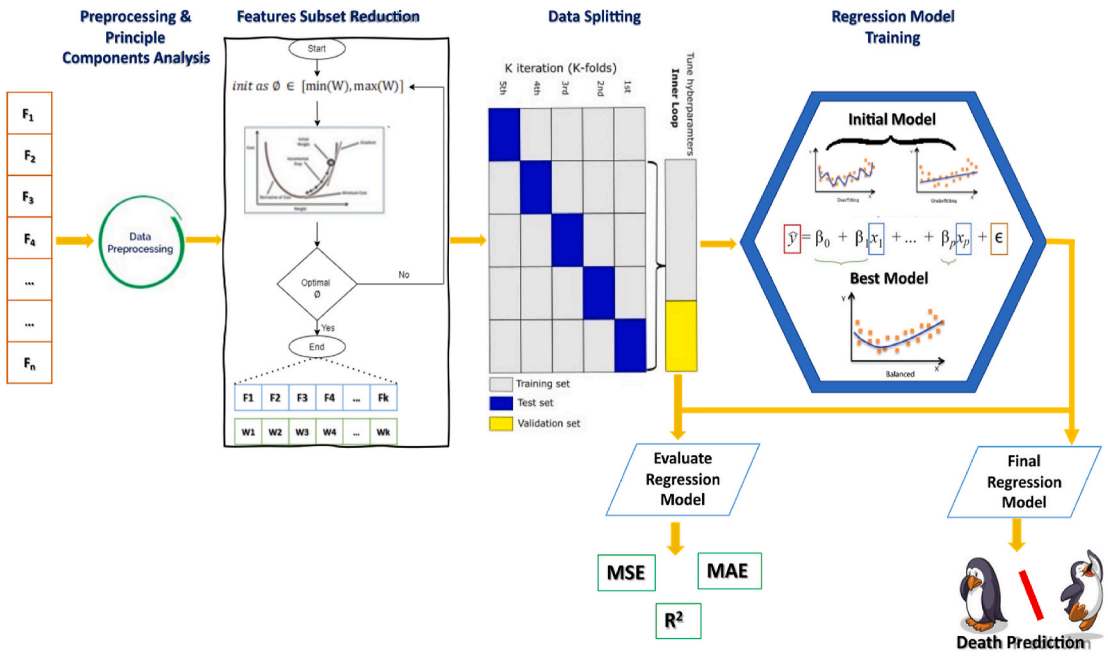

Fig. 2. Representative stages of the proposed HANA model pipeline.

should obtain the optimal threshold values used in the reduction process. The selected threshold is used by the ReliefF algorithm to select the most relevant features from the relevant vector.

- Second, we built the regression ML model using linear regression, i.e., elastic net regression). The regression MSE, RMSE, MAE, and R^2 are determined and evaluated. Furthermore, the experiment proves that the features transformation using PCA increases the efficiency of regression metrics.
- Third, to the best of our knowledge, this study is the first to report hybrid feature reduction based on PCA, ReliefF, and SGD algorithms to predict the death rate of COVID-19 cases correlated to diet.

3.1. Stage 1: data preprocessing

Dealing with data correctly is crucial to achieving a highly accurate prediction model. Problems related to the dataset, e.g., missing values, imbalance, etc. Must be addressed. Here, missing values are completed based on the following benchmarks.

3.1.1. Missing value

The missing values of this applied dataset are imputed based on an average and most frequent property. This means that the imputation technique used is to substitute any missing value with the mean or most frequent of the variable values for all other cases. The major advantage of this approach is that no further change in the sample mean is required for this variable. The value for the recorded $\bar{y}_h$ is the average sample of respondent data within some class h.

In addition, the average imputation can be determined within the classes and can be expressed as $\hat{y}_i = \bar{y}_h$ where $\hat{y}_i$is the imputed.

3.1.2. Feature normalization

Feature normalization an important step to is ensure a high accuracy model. The features are normalized to the interval [0, 1] to produce a feature vector template to be applied in the next feature processing stage.

3.2. Stage 2: feature processing

Generally, the proposed approach depends on using PCA to extract the most significant features through a hybrid algorithm based on Relief features and the SGD optimizer. Then, the AdaBoost and other regression models are presented to forecast death cases resulting from the consumption of unhealthy food during the COVID-19 pandemic. The process to output an ML regression model and the expected death rate is given in Algorithm 1, an Intelligent Healthy Food Regression System that describes the steps of the COVID-19 Pandemic Roadmap.

Algorithm 1. Healthy Artificial Nutrition Analysis (HANA): A COVID-19 Pandemic Roadmap

3.2.1. PCA

PCA is among the most commonly used unsupervised dimensional reduction techniques. The purpose of PCA is to locate the space that represents the high variance path of the data. PCA space consists of major orthogonal elements, i.e., axes or vectors. The PCs are determined either by resolving the covariance matrix or by using singular value decomposition [46].

In this work, we would like to show that the PCA-enhancement dataset (features) approach is not only limited to a specific subspace learning method but also represents the main PCA concept, i.e., mapping $\varphi(y)$ on the data. If the input is uploaded to a higher-dimensional space, the subspace can be approximated more easily. Although the high dimension space becomes very significant, the main point behind PCA is to try to prevent explicit computing of φ and to work with $\varphi(y_j) \geq \varphi(y_i)^T \varphi(y_j)$. The C matrix of the covariance is determined as in Equation (1) [47].

$$C = \frac{1}{n}\sum_{i=1}^{n}(y_i)(y_i)^T \tag{1}$$

where n is the number of instances for all enrolled features i, and T is the transformed function inside the orthogonal domain. The main objective of using PCA in this paper is to modify and enhance the features that help achieve a higher accuracy prediction model.

Algorithm 1: Healthy Artificial Nutrition Analysis (HANA): A COVID-19 Pandemic Roadmap

Input: A Feature Vector acquired from a health food COVID-19 dataset.
Initialization: Parameters setting for ML regression models and experiments
Procedure:
1 Setting the PCA for the input feature vectors.
2 Determine the PCs of the features.
3 For loop using SGD optimizer:
3.1 Randomize threshold.
3.2 Apply ML regression
3.3 Evaluate using MSE, RMSE, MAE, and R^2
3.4 Update the best threshold value.
4 Setting the threshold value for selecting the features subset using Relief method.
5 Transform the selected features using PCA.
6 Applying ML regression prediction models based on:
I. linear regression model
II. Ada boost regression model
7 Using Kfold cross validation method to determine the evaluation metricises.
Output:
ML regression model,
Expected Death Rate in the population giving the food style.

3.2.2. Feature selection and dimension reduction

Selecting the most important features from a whole dataset is considered an open challenge [48]. In this paper, to extract the most important features two algorithms, i.e., the ReliefF algorithm and SGD, are combined in this step. The steps of the hybrid algorithm are summarized in Algorithm (1).

3.2.2.1. ReliefF algorithm: the greatest recognized variant. In practice, the original Relief algorithm is no longer used [49] and has now been replaced by ReliefF [50] because it is one of the best and most commonly used RBA algorithms. The "F" was added to indicate that the algorithm differs from the sixth variation (A to F) of the algorithm proposed by Kononenko [51]. The ReliefF adopts a filter method approach to feature selection, and classification, accuracy is not used as an evaluation metric directly.

Assume that x is the training dataset and that the ith sample of a training dataset denoted by l is $y_i^{(l)}(i = 1, 2, 3, \ldots\ldots, N_l)$. Therefore, to measure the similarity between two samples, Euclidean distance is applied to determine the Nearest Neighbour (NN) training dataset $M(y_i^{(l)})$ and $\overline{M}(y_i^{(l)})$.

The Given a training dataset $y_i^{(l)}, M_j(y_i^{(l)})(j = 1, 2, 3, \ldots., k)$ shows the jth NN sample of the training dataset denoted l, whereas $\overline{M_j}(y_i^{(l)}).(j = 1, 2, 3, \ldots., k)$ shows jth NN sample of the other training dataset. The difference of nth features between $y_i^{(l)}$and $M_j(y_i^{(l)})$ are determined by Equation (2) [46]:

$$D_n\left(y_i^{(l)}, M_j\left(y_i^{(l)}\right)\right) = \frac{1}{k}\sum_{j}^{k}\left[\frac{y_i^{(l)}(n) - M_j\left(y_i^{(l)}(\mathrm{n})\right)}{max(y(n)) - \min(y(n))}\right] \quad (2)$$

Further, the difference of nth features between $y_i^{(l)}$and $\overline{M_j}(y_i^{(l)})$ are determined by Equation (3):

$$D_n\left(y_i^{(l)}, \overline{M_j}\left(y_i^{(l)}\right)\right) = \sum_{v\neq 1}\sum_{j=1}^{k}\frac{p(v)\left|y_i^{(l)}(n) - \overline{M_j}\left(y_i^{(l)}(\mathrm{n})\right)\right|}{k(1 - p(l)[\ \max(y(n)) - \min(y(n))]} \quad (3)$$

In this case, the score of nth features are determined using Equation (4),

$$\mathrm{S(n)} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{N}}\left[\mathrm{D_n}\left(\mathrm{y_i^{(l)}}, \overline{\mathrm{M_j}}\left(\mathrm{y_i^{(l)}}\right)\right) - \mathrm{D_n}\left(\mathrm{y_i^{(l)}}, \mathrm{M_j}(\mathrm{y_i^{(l)}})\right)\right] \quad (4)$$

where S(n) is the score of the nth feature, and N represents the total training dataset. There is always competition between higher S(n) and more discriminatory features. Thus, different differing from the contribution rate in the PCA, S(n) is an intuitive value for assessing used to assess the performance of the classification capability of the features. The ReliefF approach will identify many more discriminatory features than PCA. Even these larger, However, even this greater number of significant features are kept and shaped as a low dimensional subspace.

3.2.2.2. SGD. SGD is an effective approach to scale down as it achieves enhanced consequences for scarce data. SGD means that the gradient curve descends to the lowest point. SGD is applied iteratively until the minimum set of points has been achieved. There are three forms of SGD, SGD, full-batch stochastic, and mini-batch gradient approaches. The three variables are derived depending on the characteristics taken for each iteration. Further SGD operates on a random probability basis. Rather than using the entire dataset, random features are selected as samples from the specified dataset for each iteration. These samples are called batch samples. For a precise prediction, conventional gradient descent is applied to the entire dataset simultaneously, which is difficult when the dataset is extremely large [52].

As a result, the SGD algorithm has been developed. The SGD algorithm attempts to address the problem associated with large datasets by selecting subcategories of the dataset at random for every iterative process. SGD is a discriminatory learning approach that prototypes and classifies detected data. SGD considers a batch to be a single sample in an iterative process. Thus, from each iterative process, the cost function gradient for each sample is calculated. However, SGD is faster than traditional gradient descent approaches because the information is introduced instantly after every sample has been trained. The general steps of the SGD algorithm are as follows [53].

1. Take the derivatives of the loss function for the enrolled features such that the resulting loss function is given by $L(\varphi) = (\dot{y} - y)^2(x)$where $\dot{y}$is the resulting predicted value and y is the actual value relative to x.
2. Calculate the gradient ∇ of the loss function results from step 1.

3. Select the initial random value of the enrolled features to start f_0.
4. Update L(f) such that the feature values are included.
5. Calculate the value of size stage (SS) as SS = Evaluation_Measures*Gradient.
6. Calculate the new feature value (NFV) in terms of the Old Feature Value (OFV) such that as NFV = OFV – SS. Therefore, the NFV is updated in the reverse direction of the gradient.

$$f_1 = f_0 - (stage\_size * \nabla L(\varphi))$$

7. Repeat steps 3 to 5 after each iteration until the gradient is closer to zero.

At every step, the features are selected and randomized. Evaluation_Measures has a more significant effect on the algorithm. It is better if the Evaluation_Measures value is greater at the beginning of the iteration process as it tends to make the algorithm take big stages. The Evaluation_Measure should be reduced when it approaches the minimum value, to prevent losing the minimum point.

3.3. Stage 3: regression using ML

The proposed HANA model approach was used to predict the status of death using two prediction models, i.e., linear regression (standard linear regression, ridge regression, and elastic net regression), and AdaBoost models.

3.3.1. Linear regression models

3.3.1.1. Simple linear regression method. Simple linear regression models the relationship between two variables such that one variable is utilized to define the value of another variable. This relationship can be expressed as a complex mathematical equation that is applied to numerous values of the two variables given certain assumptions to describe the data. Simple linear regression can be expressed as follows:

$$Y = \beta_1 + \beta_2 X + \epsilon \tag{5}$$

where X is the independent variable, Y is the dependent variable, and ε is the error. β1 and β2 are the interception and the slope of the regression model, respectively.

For a given parameter β1, β2 the cost function (ϑ) can be determined as follows.

$$\vartheta(\beta_1, \beta_2) = \frac{1}{2n}\sum_{k=0}^{n}\left(h_\beta\left(x^k\right) - y^k\right)^2 \tag{6}$$

The main objective of simple regression is to minimize the cost function such that: $\overset{minimize}{\beta_1}{}_{,\beta_2}\vartheta(\beta_1, \beta_2)$ [54].

3.3.1.2. Ridge regression method. Regularization and feature selection tasks are done using the ridge regression method. Even when the variables are highly correlated, this regression method determines the feature subset that is significant to the classification and prediction problems. Ridge regression also helps to handle missing values in the variables. A specific type of estimator for coefficient shrinkage, i.e., a ridge estimator, is used in the ridge regression method [55].

This type of regression corresponds to the regularization form of L2 in which a penalty called the L2 penalty is applied. The L2 penalty is determined as the square of the coefficient's magnitude. By simply summing the penalty values, the cost function in the ridge regression method is adjusted. The cost function used in ridge regression is presented in Equation (7).

$$\sum_{i=1}^{n}\left(y_i - \check{y}_i\right)^2 = \sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{m} w_i * \ x_{ij}\right)^2 + \lambda\sum_{j=0}^{n} w_j^2 \tag{7}$$

where n is the feature numbers, m is the total number of features n the dataset, and λ is the penalty value. Note that ridge regression applies constraints on *w* coefficients.

3.3.2. Elastic net regression method

An elastic net regression model is a standard type of regularized linear regression that associates two penalties, L1 and L2 penalty functions. Further, during the training process, elastic net regression can regularize and develop a linear regression model by adding penalties to the loss function. The new instance data can be evaluated automatically to make the final prediction based on a grid search strategy. To avoid the shortages initiated in lasso regularization, the elastic net includes a quadratic expression in the penalty. Moreover, isolation is performed based on ridge regression. The elastic net method can regularize variable selection instantaneously. In addition, it fits the dimensional data to be more than the number of samples used. The assemblages and selection of variables are crucial processes in the elastic net method. Assume that the given data are (y, X) such that the penalty parameters are (l1, l2). For any static non-negative λ1 and λ2, the naïve elastic net used to solve the lasso problem is determined as follows [56].

$$L(l_1, l_2, \beta) = |y - X\beta|^2 + l_2|\beta|^2 + l_2|\beta|_1 \tag{8}$$

where $|\beta|^2 = \sum_{j=1}^{p}\beta_j^2$, and $|\beta|_1 = \sum_{j=1}^{p}|\beta_j|$

The estimator of the elastic net denoted $\dot{\tilde{\beta}}$, is determined as the minimizer of Equation (6) and is calculated as follows:

$$\dot{\tilde{\beta}} = argmin_\beta\{L(l_1, l_2, \beta)\} \tag{9}$$

3.3.3. AdaBoost method

In the proposed approach, the AdaBoost model [57] was used for regression. Normally, two main parameters have a direct effect on AdaBoost, i.e., the number of iterations (T) and the weights of the training patterns (w), which are initialized to be equal [58].

Given training sets $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ such that each instance xi belongs to a domain X, and each label $x_i \in \{-1, +1\}$. For regression and classification issues, Equation (10) determines the expected values to match and minimize the sign of $f_\lambda(x_i)$ to y_i, such that

$$\sum_{i=1}^{m}[y_i f\lambda(y_i) \le 0] \tag{10}$$

where $y_i f\lambda(y_i) \le 0$ is 1 if $y_i f\lambda(y_i) \le 0$ is validated, otherwise 0.

Classification and regression can reduce the number of errors; however, the most significant contribution is to minimize some other non-negative loss function. For example, the Ada-boosting algorithm is presented as a loss function, as in Equation (11).

$$Adaboost_{loss} = \sum_{i=1}^{m} e^{(-(y_i f\lambda(y_i)))} \tag{11}$$

3.4. Stage 4: validation analysis

3.4.1. Evaluation measures

To evaluate the proposed prediction system, four well-known measures metrics are used, i.e., MSE, RMSE, MAE, and R^2.

3.4.1.1. MSE. MSE is one of the most commonly used metrics for regression tasks. It is essentially an estimate of the square of the difference between the target value and the regression model's expected value. It penalizes small differences when it squares the discrepancies, leading to over-estimating how poor the model is as shown in Equation (12).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2 \quad (12)$$

where y_i is the actual expected output, $\widehat{y}_i$ is the model's prediction, and n is the number of samples.

3.4.1.2. RMSE. RMSE measures the difference between values predicted by a model and the actual values. RMSE is calculated as follows:

$$RMSE = \sqrt{\frac{1}{\mathrm{ns}}\sum_{\mathrm{i}=1}^{\mathrm{ns}}\left(\mathrm{y_i} - \widehat{\mathrm{y}}_\mathrm{i}\right)^2} \quad (13)$$

where n and s denote the number of data and the forecast value, respectively.

3.4.1.3. MAE. MAE is used to measure prediction model accuracy, and it is computed by Equation (14).

$$MAE = \frac{1}{ns}\sum_{\mathrm{i}=1}^{\mathrm{ns}}\left|\mathrm{y_i} - \widehat{\mathrm{y}}_\mathrm{i}\right| \quad (14)$$

where y_i is the actual expected output, $\widehat{y}_i$ is the model's prediction, and n is the number of samples.

3.4.1.4. R^2. The main objective of R^2 is to measure the correlation between forecast and measured data. A dataset has n values denoted y_1, $y_2, \ldots n$ (communally identified as y_i or as a vector $y = [y_1, y_2, \ldots n]^T$), respectively related to a predicted value $f_1, \ldots, f_n$. To determine the total sum of squares and the sum of residual squares Equations (15) and (16), are used as follows.

Total sum of squares:

$$S_{tot} = \sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)^2 \quad (15)$$

The sum of residual squares is also referred to as the residual sum of squares:

$$S_{res} = \sum_{i=1}^{n}\left(y_i - f_i\right)^2 \quad (16)$$

The most common expression of the coefficient of determination is given in Equation (17).

$$R^2 = 1 - \frac{S_{res}}{S_{tot}} \quad (17)$$

Low MSE, RMSA, and MAE values indicate the best result, and high R^2 values indicate high accuracy.

4. Experimental and results

This section describes experiments conducted to determine the efficiency of the proposed HANA model. All experiments were performed using the Python 3.0 software package running on a machine with a Core i7 processor, 16 RAM, and NVIDIA 4G- GT 740 m GPU environment. We present a dataset description, dataset statistical analysis, PCA to enhance features, feature selection, dimension reduction, and performance validation of the proposed methodology in the following subsections.

4.1. Dataset description

In the proposed HANA model, the COVID-19 Healthy Diet Dataset[2] was used. The dataset contains percentages of fats in different foods in 170 countries around the world. For comparison, the dataset also includes obesity, undernourished, and COVID-19 cases as percentages of the total population. The COVID-19 Healthy Diet Dataset includes information about various categories of food, alcoholic beverages, animal products, animal fats, aquatic products, cereals excluding beer, oil crops, eggs, seafood, sugar and sweeteners, fruits, meat, miscellaneous, milk excluding butter, offal, spices, starchy roots, pulses, stimulants, sugar crops, tree nuts, vegetable oils, vegetal products, and vegetables. It consists of five files in comma-separated values format containing the following.

1. Fats in each category of food in the dataset are represented as a percentage.
2. The percentage of food consumed (kg) in 170 countries.
3. The percentage of energy consumption (calories) from different categories of food.
4. The percentage of protein in different categories of food.
5. Detailed subcategories of food in each category.

4.2. Dataset statistical analysis

4.2.1. Dataset visualization using scatter diagrams

As mentioned previously, the database covers 170 countries and 25 types of food. In the proposed HANA, scatter graphs are used to observe and represent the relationship between food types (features). The points in a scatter plot reveal patterns in the entire dataset as well as individual data values. This information can be used to determine correlation relationships [56]. In other words, a scatterplot diagram is used to visualize the data in the dataset. Scatter plots represent a group of dispersion plots showing each pair of features. The most significant and weak relationships can be easily deduced from scatter plot diagrams. In turn, this enables us to explain the connection between each pair of features. Example scatterplot diagrams for the COVID-19 Healthy Diet Dataset are shown in Fig. 3.

Scatter graphs were used to make the scatter plots to correlate food types and death. Death is a dependent variable and different types of food are independent variables. As shown an examples graphs illustrate a triple relationship between two different types of food in the COVID-19 Healthy Diet Dataset and death.

4.2.2. Dataset distribution analysis

Pearson correlation is one of the most common methods used with numeric variables, and its value range from 1 to −1 where 1 indicates a positive correlation, −1 denotes a negative correlation, and 0 indicates no correlation [59,60]. The proposed framework used the Pearson correlation coefficient to find the correlation between the different types of food in the COVID-19 Healthy Diet Dataset and the extent of their positive and negative correlation with death. The top six categories of food that are correlated with death are listed in Table 1.

The proposed HANA model also analyzes the distribution of the COVID-19 Healthy Diet Dataset and the results show that the food categories have a logarithmic relationship to death. The mean μ and variance σ2 for each food category are shown in Fig. 4.

4.2.3. PCA for enhancing features

In this study, the goal of PCA was to extract the main data representative of the typical features from the COVID-19 Healthy Diet Dataset and present it as a new set of independent parameters of the principal component. The function of the PCA is to transform correlated variables into non-correlated variables called the principal components by applying the orthogonal transformation. The transformation is designed such that the main component represents the greatest amount of data variance, and the orthogonal axes are sorted in descending order depending on the amount of variance [61,62]. PCA is calculated by decomposition of a covariance matrix produced from a dataset or decomposition of individual values of a dataset matrix [63]. PCA can be applied to extract and classify features to identify target samples in a dataset. Thus, in this paper, PCA is used to enhance and improve the features in the dataset. As shown in Fig. 5 and Fig. 6, the values analyzed

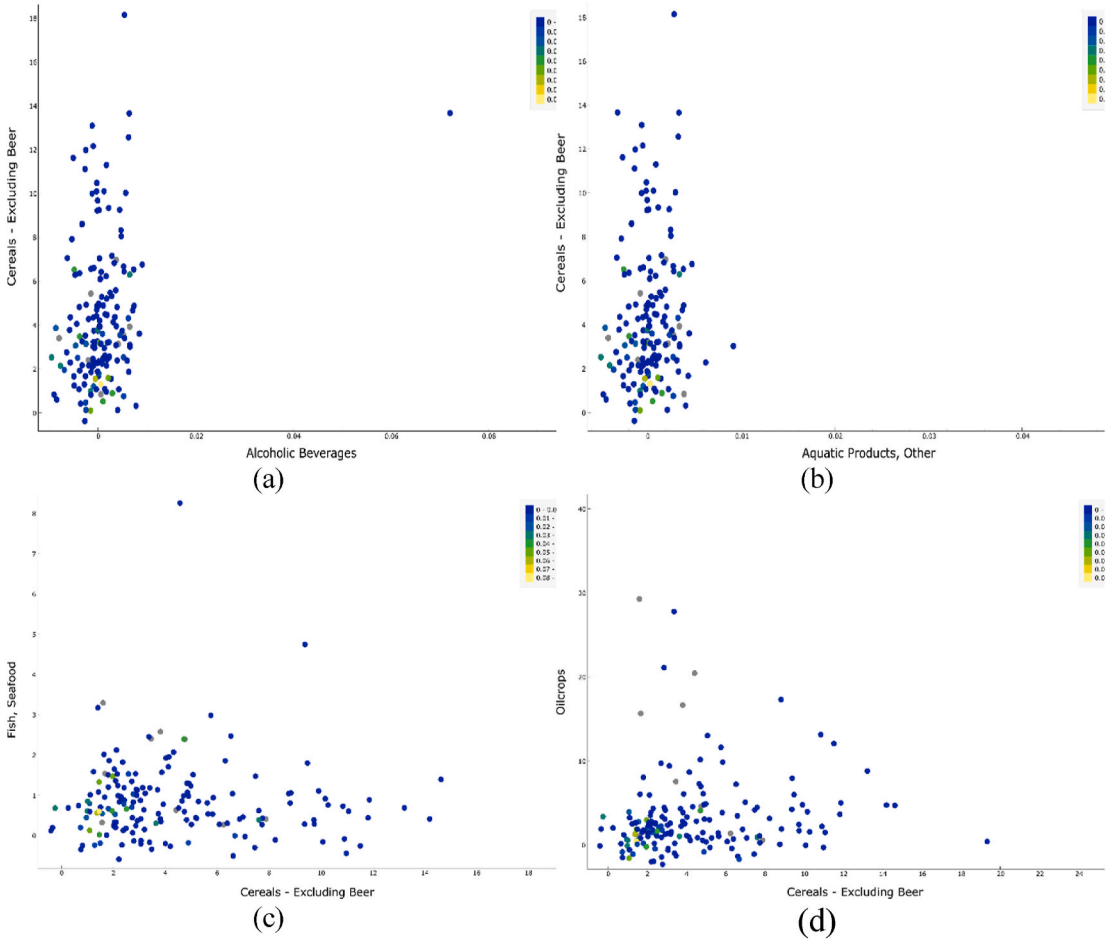

Fig. 3. Examples of a scatter graph for COVID-19 Healthy Diet Dataset.

Table 1
Top six food category (features) more correlated to death.

Features	μ	σ^2
Animal Fats	4.138451	3.277778
Cereals Excluding Beer	4.376548	3.174437
Vegetal Products	29.304396	7.978798
Animal Products	20.695714	7.979141
Eggs	0.953890	0.642060
Milk Excluding Butter	5.109061	3.321855
Spices	0.281251	0.447500

at PC_{20} increased as the component variance decreased from 0.007 to 0.04 and the cumulative variance increased to 0.8 from 0.22, thus increasing the value of the features analyzed, which, in turn, improved the features.

4.3. Features selection and dimension reduction

4.3.1. Relief F feature selection

PCA is also used to control the correlation among features. Next in the new space, the ReliefF algorithm is used to discover more discriminatory features. Table 2 shows the result for each dimension of the features after the PCA transformation. Compared with the scores obtained by the ReliefF method, it is evident that higher scores can be achieved using the feature dimensions.

4.3.2. SGD parameter optimizer

Feature reduction is considered an optimization problem that can be by SGD. The SGD threshold will determine the dimensionality of the subspace. The main purpose of SGD is to limit the features used in prediction methods. If there is a prepared dataset (the values of the features' PC's resulting from using ReliefF), SGD is a computationally exceptionally sumptuous system. We executed linear regression using SGD, as shown in Equation (18) [64,65].

$$M = m - \eta \nabla \mathrm{E}_{\mathrm{j}}(\mathrm{y}) \tag{18}$$

Here, $\mathrm{E}_{j}(y)$ represents the estimated features, and E_{j} is the present value of the PC's features resulting from using ReliefF. Commonly, Ɇ is an error function; then, by tracking the gradient direction in the value space of (y), we move in the (y), which reduces the error. SGD calculates the best (y) by minimizing Ɇ simultaneously. More importantly, with either perceptron segmentation or linear regression, (y) requires the model's weight parameters, and Ɇ(y) is the model's error. Regular gradient descent is expressed as follows:

$$\mathrm{M} \leftarrow \eta \nabla \mathrm{E}_{\xi}(y) \tag{19}$$

where the objective of the error is calculated as follows.

$$\mathrm{E}_{\xi}(y) = \ln \sum j \mathrm{E}_{j}(y) \Rightarrow \nabla \mathrm{E}_{\xi}(my) = \ln \sum j \nabla \mathrm{E}_{j}(y) \tag{20}$$

As a result of the evolution matrix measures used to determine the exact prediction method, SGD adjusts the selection of the selected features due to the threshold. The resulting features amounted to 80% of the output of the ReliefF feature, as shown in Table 3.

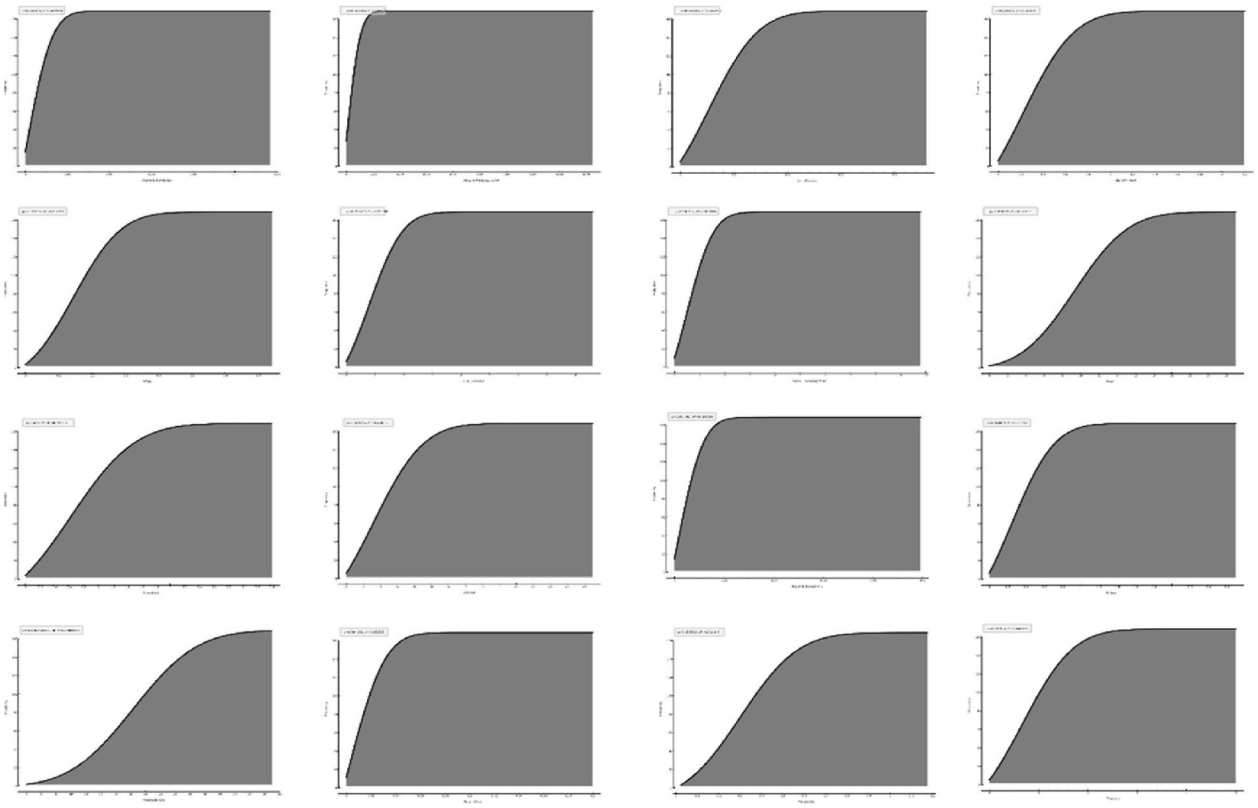

Fig. 4. The distribution curves for different food categories and death.

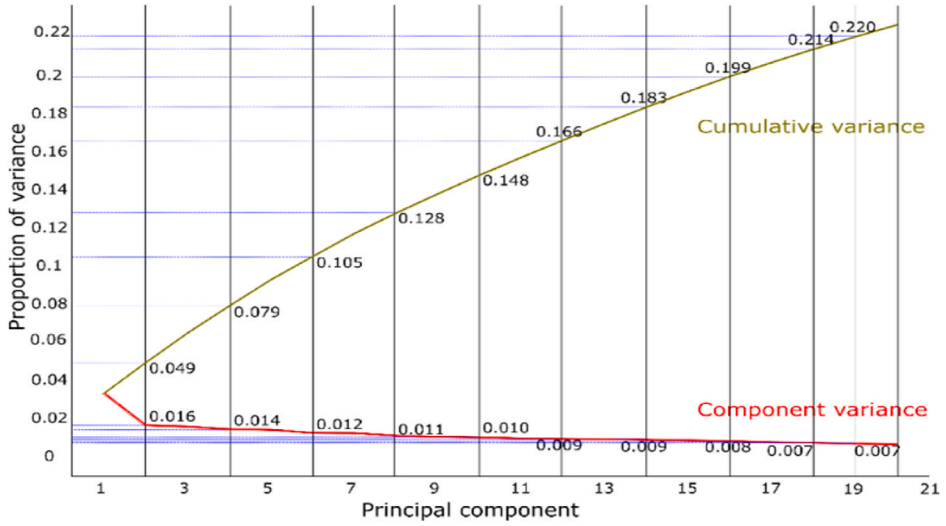

Fig. 5. The transformation of features by using PCA.

4.3.3. Regression prediction model results

The MSE, RMSE, MAE, and R^2 values for the proposed framework's regression prediction models are compared in Table 4.

As can be seen, death was predicted with high accuracy, and the values were predicted with acceptable accuracy. The MSE, RMSE, MAE, and R^2 metrics serve to evaluate the accuracy of the models for the COVID-19 Healthy Diet Dataset, as shown in Table 4. The most efficient regression prediction model was elastic net regression. As shown in Table 4, the MSE, MAE, and RMSE for the elastic net regression model were significantly lower than ridge regression, simple linear regularization, and AdaBoost models MSE, MAE, and RMSE metrics are frequently used to determine model accuracy. In this study, they provided the best fit when predicting death. Because of R^2 compares the fit of the chosen model with a horizontal straight line e.g null hypothesis, R^2 is negative for the chosen model which fits worse horizontal line. R^2 is a performance metrics that is not always the square of anything, hence it can be negative value without violating any rules of math; i.e. R^2 is negative denoting the chosen model does not follow the trend of the data. Conversely, the R^2 value for the elastic net regression model was significantly higher than other models.

The features are now prepared to be close-fitting to a model, but which one? We selected four models (linear regression (ridge regression, simple linear regularization, and elastic net regression), and AdaBoost models) and performed K-fold cross-validation to determine which one

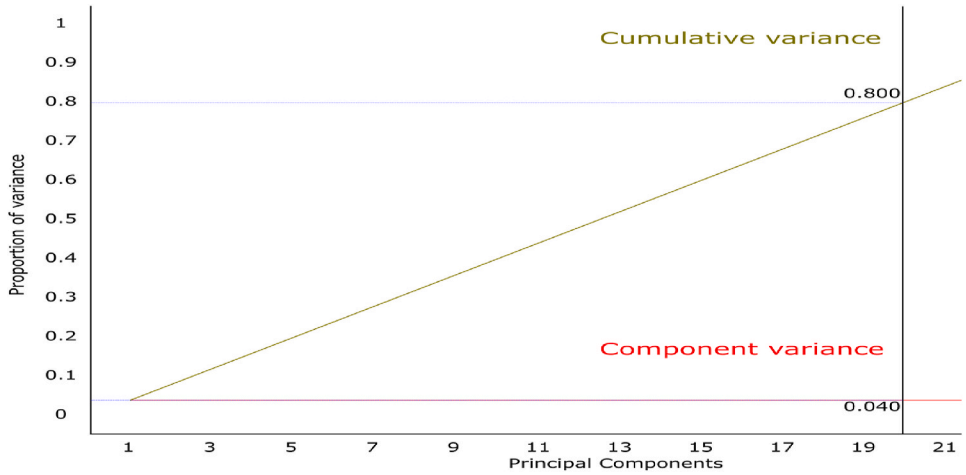

Fig. 6. Enhancing features by using PCA.

Table 2
Selected features based on PCA.

PC #	Value	PC #	Value	PC #	Value	PC #	Value	PC #	Value
PC1	0.185161	PC6	0.130819	**PC11**	0.106391	**PC16**	0.091944	**PC21**	0.026076
PC2	0.145267	PC7	0.059881	**PC12**	0.074000	**PC17**	0.052747	**PC22**	0.085176
PC3	0.095224	PC8	0.070417	**PC13**	0.119977	**PC18**	0.060325	**PC23**	0.140997
PC4	0.069354	PC9	0.084768	**PC14**	0.049968	**PC19**	0.103454	**PC24**	0.104525
PC5	0.123344	PC10	0.102241	**PC15**	0.038403	**PC20**	0.073350	**PC25**	0.080452

Table 3
PCA for reduction features.

PC #	Value	PC #	Value	PC #	Value	PC #	Value
PC1	0.185161	**PC13**	0.119977	**PC3**	0.095224	**PC12**	0.074000
PC2	0.145267	**PC11**	0.106391	**PC16**	0.091944	**PC20**	0.073350
PC23	0.140997	**PC24**	0.104525	**PC22**	0.085176	**PC8**	0.070417
PC6	0.130819	**PC19**	0.103454	**PC9**	0.084768	**PC4**	0.069354
PC5	0.123344	**PC10**	0.102241	**PC25**	0.080452	**PC18**	0.060325

Table 4
Comparison of the proposed framework regression prediction models based on evaluation metrics (The best results are represented in bold).

Model		Evaluation Metric			
		MSE	RMSE	MAE	R^2
Linear Regression	Ridge Regression	0.00023083	0.01519314	0.01023939	−0.15965093
	Simple Linear Regularization	0.00023091	0.01519604	0.01024034	−0.16009405
	Elastic Net Regression	0.00018113	0.01345867	0.00873109	**0.09001016**
	AdaBoost	0.00020749	0.01440446	0.00761746	−0.04237952

is best. We used 20-fold cross-validation to compare the output of the four models and conclude that elastic net regression has a slightly higher probability of providing better prediction accuracy.

4.4. Performance validation of the proposed methodology

To evaluate the performance of the SGD method, we compared using and not using SGD for feature reduction. Table 5 shows the probabilities before and after feature reduction. In Table 5, the score for the model in the row is higher than the score for the model in the column. Small numbers indicate the probability that the difference is negligible. Table 5 represents the different enhancement of results before and after feature transformation. The different enhancements are compared concerning the MSE, RMSE, MAE, and R^2 metrics. The proposed method obtains better results for all metrics using simple linear regression compared to regression using regularization with the ridge function. The MSE, RMSE, and MAE values improve with simple linear regression and regression using elastic net compared with standard linear regression, regression using the elastic net, and AdaBoost. AdaBoost increases the R^2 value compared to regression using ridge and elastic net regularization. Overall, the results support the following points regarding hybrid PCA-Relief feature selection with SGD parameter optimization.

a. The reduced and enhanced COVID-19 Healthy Diet Dataset has higher predictive accuracy than the original COVID-19 Healthy Diet Dataset due to a lower number of features.
b. The elastic net regression model has higher accuracy (lower MSE, RMSE, MAE, and higher R^2) than other models with the reduced COVID-19 Healthy Diet Dataset.

Table 5
Regression models comparison by evaluation metrics (MSE, RMSE, MAE, and R^2) before and after reduction.

			Linear Regression						AdaBoost	
			Ridge Regression		Simple linear Regularization		Elastic Net Regression			
			Before feature reduce.	After feature reduce.	Before feature reduce	After feature reduce	Before feature reduce	After feature reduce	Before feature reduction	After feature reduce
Linear Regression	Ridge Regression	**MSE**			0.147	0.020	0.931	0.918	0.758	0.502
		RMSE			0.135	0.020	0.960	0.942	0.890	0.853
		MAE			0.110	0.013	0.986	0.974	0.998	0.994
		R^2			0.956	0.854	0.045	0.089	0.110	0.162
	Simple Linear Regularization	**MSE**	0.853	0.980			0.930	0.918	0.758	0.502
		RMSE	0.865	0.980			0.960	0.942	0.890	0.853
		MAE	0.890	0.987			0.986	0.974	0.998	0.994
		R^2	0.044	0.146			0.045	0.089	0.110	0.162
	Elastic Net Regression	**MSE**	0.069	0.082	0.070	0.082			0.071	0.291
		RMSE	0.040	0.058	0.040	0.058			0.313	0.724
		MAE	0.014	0.026	0.014	0.026			0.949	0.980
		R^2	0.955	0.911	0.955	0.911			0.228	0.213
	AdaBoost	**MSE**	0.242	0.498	0.242	0.498	0.929	0.709		
		RMSE	0.110	0.147	0.110	0.147	0.687	0.276		
		MAE	0.002	0.006	0.002	0.006	0.051	0.020		
		R^2	0.890	0.838	0.890	0.838	0.772	0.787		

c. Results show that deaths increase with increasing consumption of animal fat, animal products, eggs, and milk - including butter.
d. Results show that deaths decrease with increased consumption of vegetal products, spices, pulses, oil crops, cereals - excluding beer, and starchy roots.

4.5. Comparative study

In this section, we introduce the comparative study of the proposed regression model with recent regression models related to the study of the COVID-19 effects on human food habits during a lockdown. To the best of our knowledge, investigations into the direct relationship between food and COVID-19 and the prediction of death cases resulting from bad dietary habits during the recent COVID-19 pandemic are limited. Therefore, Table 6 introduces a comparative study of the proposed regression model compared with studies performed by Ordás et al. [37], and Shams et al. [39]. The comparative study demonstrated the superiority of the proposed regression model compared with other models in terms of accuracy and MSE.

5. Results analysis and discussion

The problem is refined to be more concise and clearer for assisting the nutrition experts and infected COVID-19 subjects. The results discuss the intercorrelation between different types of foods also the intercorrelation of food categories and mortality of COVID-19. The analysis is supported by recent literature discussions and research in both nutrition analysis and artificial intelligence directions. The proposed HANA model works on predicting the mortality of COVID-19 depending on the nutritional system style of a given population and determines the surviving infected cases from others using machine learning. In the subsequent items, we have illustrated these insights in more detail. *First*, Table 1 emphasizes more the foods associated with higher correlated to death from Covid-19. Based on an in-depth analysis of recent nutrition recommendations by WHO, we concise the same advice already introduced in the WHO report. During the COVID-19 pandemic, WHO report recommends avoiding eating out, using less salt and sugar, and moderate eating amounts of i) oil as avocado, nuts, sunflower, olive oil, soy, canola, and corn oils; ii) fat including animal fats iii) animal products such as fish, fatty meat and iv) the saturated fats used in manufacturing vegetal products like, cream, coconut oil, and cheese. *Second*, the nutritional experts emphasize that a nutritious diet can assist overall enhancement of the immune system and becomes less vulnerable to COVID-19 diseases, and they recommended that foods rich in vitamins and fresh vegetables can assist in strengthening the immunity system [66]. One of the current nutrition approaches is the Mediterranean diet for COVID-19 [67] in which discusses the impact of diabetes (Type II) and cardiovascular disease's direct effects on the infected COVID-19 subjects. It is noteworthy their contributions confirm a highly positive outcome of the HANA model representing Eat less salt and sugar and eat fresh and unprocessed foods every day.

Finally, Matching the evaluation result of the proposed HANA model in determining the success rate of infected COVID-19 cases verse the presented review by Mechanick et al. [68], we confirm and concise there is a high gap between clinical nutrition applied to COVID-19 and unsafe food infrastructure. They put forward many questions including the use of nutritional supplements and nutrients in COVID-19 patients with diabetes, and nutritional interventions in patients especially older than 65 years, or who are frail, to prevent severe COVID-19 disease.

The analysis of the Root Cause (RC) is the procedure for determining root causes of problems to find suitable solutions. RC assumes it is far more harmful than helpful ad hoc symptomatology and fire removal to systemically avoid and focus on solving problems. As shown in Fig. 7 the results were analyzed for the proposed model HANA and the work of RC. As shown in the root cause, it is recommended to eat less salt and sugar as low potential impact with low actionability. While it is the highest

Table 6
The comparative study between the proposed ML regression model and [37,39] using the same database.

Author	Methodology	Dataset used	Problem Type	Metrics	Comment
Ordás et al. [37]	(PCA, K-means Algorithm)	https://www.kaggle.com/mariaren/covid19-healthy-diet-dataset	Classification	Accuracy = 95%	The correlations between the eating habits and death cases of 170 countries during the COVID-19 pandemic were assessed to find the relationship between these habits and death rates-based ML.
Shams et al. [39]	SVM Model based on RBF SVM Model with Linear SVM Model with Linear Kernel Deep Learning	https://www.kaggle.com/mariaren/covid19-healthy-diet-dataset	Classification	Accuracy = 99.73% Accuracy = 99.83% Accuracy = 79.30% Accuracy = 99.72%	This architecture can forecast the human cases affected by the COVID-19 pandemic due to each patient's diet habits and system.
Our Proposed (HANA Model)	Elastic Net Regression (PCA, Backpropogation Neural Netwroks)	https://www.kaggle.com/mariaren/covid19-healthy-diet-dataset https://www.kaggle.com/mariaren/covid19-healthy-diet-dataset	Regression Classification	MSE = 0.00018113 Accuracy = 98.76%	This proposed regression model able to forecast the human cases affected by the COVID-19 pandemic due to each patient's diet habits and system using MSE.

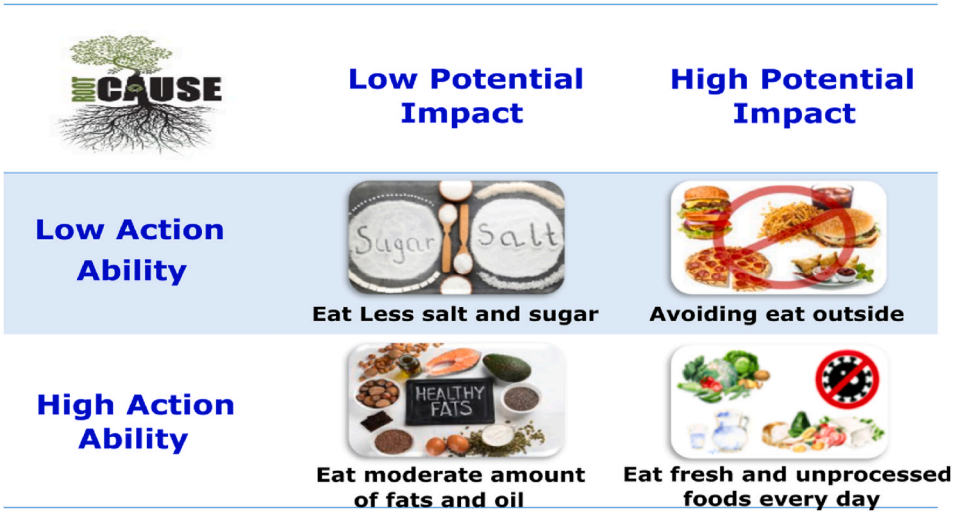

Fig. 7. The Root cause analysis of the proposed HANA model

potential impact to avoid eating out with a low actionability. The high actionability recommended, is to eat a moderate amount of fats and oil as well as eating fresh and unprocessed foods every day that are presented in root cause analysis as low, and high potential impact, respectively.

6. Conclusion and future work

Among this COVID-19 Healthy Diet dataset, consisting of 170 countries around the world, have different types of food, and the increase in eating some food categories may increase the death status since the start of the COVID-19 virus. The world had to change its eating habits to improve health and reduce the death rate. In this work, a Healthy Artificial Nutrition Analysis (HANA) was conducted on the death status of people infected with the COVID-19 virus taking into account the type of food. For this purpose, 25 features related to different food groups were used. HANA model is presented as an algorithm based on ML and data analysis to provide an effective decision tool for the nutrition experts to predict and analyze the suitable diet and nutrition during the current pandemic. Furthermore, we determine the root cause analysis of the food to recommend which diet habits are recommended based on potential impact and actionability. A statistical analysis of the COVID-19 Healthy Diet dataset was carried out to clarify the correlation of different types of food with the death rate and also their distribution, and to calculate the mean and standard deviations. Furthermore, the HANA model is presented to ensure predictable results with enhanced results compared with the well-known methods MSE, MAE, and RMSE.

To extract from the COVID-19 Healthy Diet Dataset the main data representative of the typical features and present it as a new set of independent parameters of the principal component the PCA was used. Hence the idea of using PCA in this paper to enhance and improve the features in the dataset. Then use Relief as a feature selection, after that the reduction of features is conceived as an optimization problem solved by SGD.

Death was forecasted using two regression prediction models, namely linear regression (ridge regression, No regularization, and elastic net regression), and AdaBoost models. The most efficient regression prediction model is the elastic net regression. The MSE, MAE, and RMSE for the Elastic Net Regression model were significantly lower than ridge regression, simple linear regularization, and AdaBoost models. On the other hand, the R^2 value was for the Elastic Net Regression model significantly higher than other models. In future work, a complete diet will be developed, either for specific diseases such as obesity or as a model for a healthy life. During the future development, smart health care is increasing to adapt smart city principles. The HANA model is an IoT-based appreciation component to such cases in post actions of the COVID-19 pandemic.

Declaration of competing interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

Acknowledgement

The authors thank Smart Science Lab, Mansoura, Egypt (UPID Number: 260201-2021) for technical support.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2021.104606.

Data availability

A data availability is found in https://www.kaggle.com/mariaren/covid19-healthy-diet-dataset.

Funding statement

N/A.

References

[1] S. Snuggs, S. McGregor, Food & meal decision making in lockdown: how and who has Covid-19 affected? Food Qual. Prefer. 89 (2021) 104145.
[2] C.M. Galanakis, The food systems in the era of the coronavirus (COVID-19) pandemic crisis, Foods 9 (4) (2020) 523.
[3] M. Rizou, I.M. Galanakis, T.M. Aldawoud, C.M. Galanakis, Safety of foods, food supply chain and environment within the COVID-19 pandemic, Trends Food Sci. Technol. 102 (2020) 293–299.
[4] K. Mishra, J. Rampal, The COVID-19 Pandemic and Food Insecurity: A Viewpoint on India, vol. 135, *World Development*, 2020, p. 105068.
[5] D. Laborde, W. Martin, J. Swinnen, R. Vos, COVID-19 risks to global food security, Science 369 (6503) (2020). Art. no. 6503.
[6] Y. Zhou, Y. Lu, Z. Pei, "Intelligent diagnosis of Alzheimer's disease based on internet of things monitoring system and deep learning classification method, Microprocess. Microsyst. 83 (2021) 104007.
[7] M.H. Mahmoud, S. Alamery, H. Fouad, A. Altinawi, A.E. Youssef, An automatic detection system of diabetic retinopathy using a hybrid inductive machine learning algorithm. ", " *Personal and Ubiquitous Computing*, 2021, pp. 1–15.
[8] O. M. Elzeki, M. Abd Elfattah, H. Salem, A. E. Hassanien, and M. Shams, "A novel perceptual two layer image fusion using deep learning for imbalanced COVID-19 dataset," PeerJ Computer Science, vol. 7, 2021.
[9] O.M. Elzeki, M. Shams, S. Sarhan, M. Abd Elfattah, A.E. Hassanien, COVID-19: a new deep learning computer-aided model for classification, PeerJ Computer Science 7 (2021) e358.
[10] S.M. Abir, S.N. Islam, A. Anwar, A.N. Mahmood, A.M.T. Oo, Building resilience against COVID-19 pandemic using artificial intelligence, Machine Learn. IoT: A Survey of Recent Progress," *IoT* 1 (2020), 2, Art. no. 2.
[11] S. Bhattacharya, et al., Deep learning and medical image processing for coronavirus (COVID-19) pandemic: a survey, Sustain. Cities Soc. 65 (2021) 102589.
[12] M.Y. Shams, O.M. Elzeki, M. Abd Elfattah, T. Medhat, A.E. Hassanien, Why are generative adversarial networks vital for deep neural networks? A case study on COVID-19 chest X-ray images. In *Big Data Analytics and Artificial Intelligence against COVID-19: Innovation Vision and Approach*, Springer, 2020, pp. 147–162.
[13] T. Alafif, A.M. Tehame, S. Bajaba, A. Barnawi, S. Zia, Machine and deep learning towards COVID-19 diagnosis and treatment: survey, challenges, and future directions, Int. J. Environ. Res. Publ. Health 18 (3) (2021). Art. no. 3.
[14] A.K. Gopalakrishnan, "A Food Recommendation System Based on BMI, BMR, K-NN Algorithm, and a BPNN," in *Machine Learning for Predictive Analysis*, Springer, 2021, pp. 107–118.
[15] I. Zabetakis, R. Lordan, C. Norton, A. Tsoupras, COVID-19: the inflammation link and the role of nutrition in potential mitigation, Nutrients 12 (5) (2020) 1466.
[16] H.B. Sharma, et al., Challenges, opportunities, and innovations for effective solid waste management during and post COVID-19 pandemic, Resour. Conserv. Recycl. 162 (2020) 105052.
[17] M. Marazuela, A. Giustina, M. Puig-Domingo, Endocrine and metabolic aspects of the COVID-19 pandemic, Rev. Endocr. Metab. Disord. 21 (4) (2020) 495–507.
[18] R. De Amicis, et al., Patients with severe obesity during the COVID-19 pandemic: how to maintain an adequate multidisciplinary nutritional rehabilitation program? Obes. Facts (2021) 1–9.
[19] S. Camaréna, Artificial intelligence in the design of transition to sustainable food systems, J. Clean. Prod. (2020) 122574.
[20] C. Pérez-Rodrigo, et al., Patterns of Change in dietary habits and physical activity during lockdown in Spain due to the COVID-19 pandemic, Nutrients 13 (2) (2021). Art. no. 2.
[21] L. Marty, B. de Lauzon-Guillain, M. Labesse, S. Nicklaus, "Food choice motives and the nutritional quality of diet during the COVID-19 lockdown in France, Appetite vol. 157 (2021) 105005.
[22] J.M. Soon, I. Vanany, I.R.A. Wahab, R.H. Hamdan, M.H. Jamaludin, Food safety and evaluation of intention to practice safe eating out measures during COVID-19: cross sectional study in Indonesia and Malaysia, Food Contr. 125 (2021) 107920.
[23] R.M. Fanelli, Changes in the food-related behaviour of Italian consumers during the COVID-19 pandemic, Foods 10 (1) (2021). Art. no. 1.
[24] Z. Shen, A. Shehzad, S. Chen, H. Sun, J. Liu, Machine learning based approach on food recognition and nutrition estimation, Procedia Comput. Sci. 174 (2020) 448–453.
[25] C.E. Onu, P.K. Igbokwe, J.T. Nwabanne, C.O. Nwajinka, P.E. Ohale, Evaluation of optimization techniques in predicting optimum moisture content reduction in drying potato slices, Artif. Intell. Agric. 4 (2020) 39–47.
[26] A. Tonda, et al., "Interactive Machine Learning for Applications in Food Science," in *Human and Machine Learning*, Springer, 2018, pp. 459–477.

[27] S. Usha, M. Karthik, R. Jenifer, P.G. Scholar, Automated sorting and grading of vegetables using image processing, Int. J. Eng. Res. Gen. Sci. 5 (6) (2017). Art. no. 6.
[28] T. Kodama, Y. Hata, "Development of classification system of rice disease using artificial intelligence," in *2018*, IEEE Int. Conf. Syst. Man Cybern. (2018) 3699–3702.
[29] N.E.M. Khalifa, M.H.N. Taha, L.M. Abou El-Maged, A.E. Hassanien, "Artificial Intelligence in Potato Leaf Disease Classification: A Deep Learning Approach," in *Machine Learning and Big Data Analytics Paradigms: Analysis, Applications and Challenges*, Springer, 2021, pp. 63–79.
[30] B. Liu, Y. Zhang, D. He, Y. Li, Identification of apple leaf diseases based on deep convolutional neural networks, Symmetry 10 (1) (2018). Art. no. 1.
[31] H. Altaheri, M. Alsulaiman, G. Muhammad, Date fruit classification for robotic harvesting in a natural environment using deep learning, IEEE Access 7 (2019) 117115–117133.
[32] A. Nasiri, A. Taheri-Garavand, Y.-D. Zhang, Image-based deep learning automated sorting of date fruit, Postharvest Biol. Technol. 153 (2019) 133–141.
[33] A. Wu, J. Zhu, T. Ren, Detection of apple defect using laser-induced light backscattering imaging and convolutional neural network, Comput. Electr. Eng. 81 (2020) 106454.
[34] N. Jean, M. Burke, M. Xie, W.M. Davis, D.B. Lobell, S. Ermon, Combining satellite imagery and machine learning to predict poverty, Science 353 (6301) (2016). Art. no. 6301.
[35] N. Pokhriyal, D.C. Jacques, Combining disparate data sources for improved poverty prediction and mapping, Proc. Natl. Acad. Sci. Unit. States Am. 114 (46) (2017). Art. no. 46.
[36] S. O'Hara, E.C. Toussaint, Food access in crisis: food security and COVID-19, Ecol. Econ. 180 (2021) 106859.
[37] M.T. García-Ordás, N. Arias, C. Benavides, O. García-Olalla, J.A. Benítez-Andrades, Evaluation of country dietary habits using machine learning techniques in relation to deaths from COVID-19, 4, in: Healthcare 8, 2020, p. 371.
[38] M. Kivrak, E. Guldogan, C. Colak, Prediction of death status on the course of treatment in SARS-COV-2 patients with deep learning and machine learning methods, Comput. Methods Progr. Biomed. 201 (2021) 105951.
[39] M.Y. Shams, O.M. Elzeki, M. Abd Elfattah, L.M. Abouelmagd, A. Darwish, A. E. Hassanien, "Impact of COVID-19 pandemic on diet prediction and patient health based on support vector machine," in *advanced machine learning Technologies and applications*, Proceedings of AMLTA 2021 (2021) 64–76.
[40] C. Uno, et al., Nutritional status change and activities of daily living in elderly pneumonia patients admitted to acute care hospital: a retrospective cohort study from the Japan Rehabilitation Nutrition Database, Nutrition 71 (2020) 110613.
[41] X. Zhao, et al., "Evaluation of nutrition risk and its association with mortality risk in severely and critically ill COVID-19 patients, J. Parenter. Enteral Nutr. 45 (1) (2021) 32–42.
[42] G. Li, et al., Nutritional risk and therapy for severe and critical COVID-19 patients: a multicenter retrospective observational study, Clin. Nutr. 40 (4) (2021) 2154–2161.
[43] S. Nordhagen, U. Igbeka, H. Rowlands, R.S. Shine, E. Heneghan, J. Tench, COVID-19 and small enterprises in the food supply chain: early impacts and implications for longer-term food system resilience in low-and middle-income countries, World Dev. 141 (2021) 105405.
[44] J. Bousquet, et al., "Is diet partly responsible for differences in COVID-19 death rates between and within countries? Clin. Transl. Allergy 10 (1) (2020) 16.
[45] J. Bousquet, et al., "Cabbage and fermented vegetables: from death rate heterogeneity in countries to candidates for mitigation strategies of severe COVID-19, Allergy 76 (3) (2021) 735–750.
[46] A. Tharwat, Principal component analysis-a tutorial, Int. J. Appl. Pattern Recogn. 3 (3) (2016) 197–240.
[47] A. Ehlers, F. Baumann, R. Spindler, B. Glasmacher, B. Rosenhahn, PCA enhanced training data for adaboost. In *International Conference on Computer Analysis of Images and Patterns*, 2011, pp. 410–419.
[48] R.J. Urbanowicz, M. Meeker, W. La Cava, R.S. Olson, J.H. Moore, Relief-based feature selection: introduction and review, J. Biomed. Inf. 85 (2018) 189–203.
[49] K. Kira, L.A. Rendell, The feature selection problem: traditional methods and a new algorithm, in: Aaai 2, 1992, pp. 129–134.
[50] K. Kira, L.A. Rendell, A practical approach to feature selection. In *Machine learning Proceedings*, 1992, pp. 249–256. Elsevier, 1992.
[51] I. Kononenko, "Estimating Attributes: Analysis and Extensions of RELIEF," in *European Conference on Machine Learning*, 1994, pp. 171–182.
[52] J. Flynn, et al., Deepview: view synthesis with learned gradient descent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2367–2376.
[53] R. Gemulla, E. Nijkamp, P.J. Haas, Y. Sismanis, Large-scale matrix factorization with distributed stochastic gradient descent. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge discovery and Data Mining*, 2011, pp. 69–77.
[54] N. Altman, M. Krzywinski, Simple Linear Regression, Nature Publishing Group, 2015.
[55] T. Sirimongkolkasem, R. Drikvandi, On regularisation methods for analysis of high dimensional data, Ann. Data Sci. 6 (4) (2019) 737–763.
[56] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. Roy. Stat. Soc. B 67 (2) (2005) 301–320.
[57] M. Collins, R.E. Schapire, Y. Singer, "Logistic regression, AdaBoost and bregman distances, Machine Learning 48 (1) (2002) 253–285.
[58] A. Vezhnevets, V. Vezhnevets, Modest AdaBoost-teaching AdaBoost to generalize better, 5, in: Graphicon 12, 2005, pp. 987–997.
[59] J. Benesty, J. Chen, Y. Huang, I. Cohen, Pearson correlation coefficient, in: Noise reduction in Speech Processing, Springer, 2009, pp. 1–4.
[60] J. Adler, I. Parmryd, "Quantifying colocalization by correlation: the Pearson correlation coefficient is superior to the Mander's overlap coefficient, Cytometry 77 (8) (2010) 733–742.
[61] C.L. Sabharwal, B. Anjum, Data reduction and regression using principal component analysis in qualitative spatial reasoning and health informatics, Polibits 53 (2016) 31–42.
[62] I.T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments, Phil. Trans. Math. Phys. Eng. Sci. 374 (2065) (2016) 20150202.
[63] H. Abdi, L.J. Williams, Principal component analysis, Wiley interdisciplinary reviews: Comput. Stat. 2 (4) (2010) 433–459.
[64] L. Yang, A. Shami, On hyperparameter optimization of machine learning algorithms: theory and practice, Neurocomputing 415 (2020) 295–316.
[65] M. Yaqub, et al., State-of-the-Art CNN optimizer for brain tumor segmentation in magnetic resonance images, Brain Sci. 10 (7) (2020) 427.
[66] A. Bhatta, Choice of food: a preventive measure during Covid-19 outbreak, Europasian J. Med. Sci. 2 (1) (2020) 88–92.
[67] A.M. Angelidi, A. Kokkinos, E. Katechaki, E. Ros, C.S. Mantzoros, Mediterranean diet as a nutritional approach for COVID-19, Metab., Clin. Exp. 114 (2021).
[68] J.I. Mechanick, et al., "Clinical nutrition research and the COVID-19 pandemic: a scoping review of the ASPEN COVID-19 task force on nutrition research, J. Parenter. Enteral Nutr. 45 (1) (2021) 13–31.

Mahmoud Y. Shams received the bachelor's degree in electronics and communication from the Faculty of Engineering, Mansoura University, in 2004, the master's degree in computer vision and pattern recognition from the Faculty of Computer and Information Sciences, Mansoura University, and the Ph.D. degree from the Computer Science Department, Mansoura University. He is currently an Assistant Professor with the Machine Learning and Information Retrieval Department, Faculty of Artificial Intelligence, Kafr Elsheikh University. He has published over ten articles in refereed international journals. His research interests include using deep learning approaches and cloud computing. ORCID:0000-0003-3021-5902

O. M. ELZEKI received his bachelor's degree in 2007 from Computer Science Department, Mansoura University, Egypt and received his Master's Degree from Mansoura University of Computer and Information Systems, Egypt in 2013. He received Ph.D in artificial intelligent on 2019 from Computer Science Department, Mansoura University, Egypt. He interests in Data science, Big-Data analysis, and machine learning in the different computational environment. My publications appear in different press houses including Elsevier, Springer, IEEE and international conferences.Now, he is assistant professor in Computer Science Department, Faculty of Computers & Information, Mansoura University, Egypt. ORCID:0000-0001-5409-1305

Lobna Mohamed AboEl-Magd received the B.Sc., M.Sc. degree and Ph.D. degrees in Computer Science from the Faculty of Computer and Information, University of Mansoura, in 2000, 2008 and 2014, respectively. Since January 2015, she working as an assistant professor at the Department of computer Sciences, Misr High Institute for Commerce and Computers (MET). Her research interests Artificial Intelligent, Internet of Things, computer networks, Operating system, and Language processing. ORCID:0000-0002-1247-8373

ABOUL ELLA HASSANIEN is currently the Founder and the Head of the Egyptian Scientific Research Group (SRGE) and a Professor of information technology with the Faculty of Computer and Information, Cairo University. He is an Ex-Dean of the Faculty of Computers and Information, Beni Suef University. He has more than 800 scientific research articles published in prestigious international journals and over 40 books covering such diverse topics as data mining, medical images, intelligent systems, social networks, and smart environment. He received several awards, including the Best Researcher of the Youth Award of Astronomy and Geophysics of the National Research Institute, Academy of Scientific Research, Egypt, in 1990, the Scientific Excellence Award in Humanities from the University of Kuwait for the 2004 Award, the Superiority of Scientific—University Award, Cairo University, in 2013, the Islamic Educational, Scientific and Cultural Organization (ISESCO) prize on Technology, in 2014, the State Award for excellence in engineering sciences 2015, and the Medal of Sciences and Arts of the first class by the President of the Arab Republic of Egypt, in 2017. Also he honored in Egypt as the Best Researcher in Cairo University, in 2013.

Mohamed Abd Elfattah is a Lecturer in Department of Computer Science at Misr Higher Institute of Commerce and Computers, Mansoura, Egypt. He Obtained Master's and Ph.D. Degrees in computer science from Mansoura University, Egypt in 2013, 2019, respectively, his research interests include data mining, intelligent computing, image processing, and

data science. Abd Elfattah is a reviewer to many international conferences, journals. Abd Elfattah published multiple publications including papers in international conference proceedings. His publications have appeared in international conferences and publishing houses such as IEEE, and Springer.

Hanaa Salem received a bachelor's degree in electronics engineering from the Faculty of Engineering, Mansoura University, in 2000, the master's degree in automatic control system engineering the Faculty of Engineering, Mansoura University, and the Ph.D. degree in artificial intelligence and image processing from the Computer Science and Engineering Department, Electronic Engineering Faculty, Minufia University. She interests in data science, big-data analysis, image processing and machine learning. She is currently an Assistant Professor with the Communications and Computers Engineering Department, Faculty of Engineering, Delta University for Science and Technology. ORCID:0000-0002-8714-567X