THE EUROPEAN SOCIETY OF MEDICINE
Medical Research Archives, Volume 12 Issue 7

RESEARCH ARTICLE

Machine Learning Techniques for Modelling and Predicting the Influence of Kefir in a Low-Protein Diet on Kidney Function

Vesna Knights [1], Elena Damjanovska Gavriloska[1]

[1] University "St. Kliment Ohridski" Bitola, Faculty of Technology and Technical Science, Republic of North Macedonia

OPEN ACCESS

PUBLISHED
31 July 2024

CITATION
Knights, V., Gavriloska, ED., et al., 2024. Machine Learning Techniques for Modelling and Predicting the Influence of Kefir in a Low-Protein Diet on Kidney Function. Medical Research Archives, [online] 12(7). https://doi.org/10.18103/mra.v12i7.5631

DOI
https://doi.org/10.18103/mra.v12i7.5631

ISSN
2375-1924

ABSTRACT

Introduction: This study aims to investigate the effects of a low-protein diet supplemented with kefir on protein catabolism and kidney function in stable chronic kidney disease patients. By employing advanced machine learning techniques, this research will explore the potential impact of kefir as a probiotic fermented dairy product on kidney function within the recommended intake. The study seeks to understand whether kefir supplementation in a low-protein diet can help maintain kidney function and identify any potential benefits or limitations.

Methods: The study employed a dataset comprising kidney health indicators and kefir intake records collected from 150 randomly selected patients in stage G1 to G5 during one year. Data preprocessing was performed to ensure data quality and feature relevance. Subsequently, a range of machine learning algorithms, including decision trees, random forests, and neural networks, but also and Stochastic Gradient Boosting and XgBoost model were implemented to model and predict the impact of kefir on kidney function.

The clinical data: age, sex, blood pressure (systolic and diastolic), BMI, albumin, bacteria, level of blood glucose, hemoglobin, creatinine, urea, UNAPCR, protein intake and kefir, GFR, MDRD, proturija, the existence of hypertension, diabetes mellitus, coronary artery disease, stage of CKD at the beginning and CKD stage after 12 months, binary output (patient stay in a same stage or is reduced kidney function).

Results: The analysis results revealed promising predictive capabilities of the machine learning models, demonstrating associations between kefir consumption and kidney function. Binary output indicates the patient stayed in the same CKD stage using low-protein diet where source of protein is kefir.

Conclusion: This research underscores the value of machine learning techniques in modeling and predicting the impact of kefir on kidney function. By shedding light on potential associations, this study paves the way for further investigations into the role of kefir in kidney health and sets a precedent for future studies in this area.

Keywords: Kefir, low protein diet, chronic kidney disease, machine learning technique, Modeling and the Prediction

1. Introduction

Dietary interventions play a crucial role in medical nutrition therapy, particularly in the management of various diseases. The significance of dietary regimens in preventing and treating conditions such as cardiovascular diseases, diabetes, cancer, infectious diseases, and kidney disorders is increasingly recognized in clinical practice[1-2].

Among these conditions, chronic kidney disease (CKD) stands out due to the profound impact of diet on disease progression and patient outcomes. CKD is characterized by a gradual decline in kidney function, typically progressing through five stages based on glomerular filtration rate (GFR). As the disease advances, dietary modifications become essential in managing symptoms and delaying progression to end-stage renal disease (ESRD), where renal replacement therapy like hemodialysis or transplantation becomes necessary [3-4].

One of the well-established dietary intervention in CKD management is a low-protein diet, which has shown promise in slowing disease progression and preserving kidney function[5-6]. Research has demonstrated beneficial responses to modified diets in treating CKD patients, highlighting the importance of dietary protein restriction in preserving kidney function[7-9].

Kefir, a probiotic fermented dairy product, has gained attention for its potential health benefits, including its impact on kidney function.

While a low-protein diet is recommended for CKD management, kefir, despite being a source of high-quality protein, vitamins, minerals, and beneficial microorganisms is used in controlled amounts to supplement essential nutrients[10-11]. Studies have demonstrated the efficacy of a low-protein diet supplemented with kefir in stabilizing kidney function and improving patient outcomes[12].

To further explore the potential benefits of kefir in CKD management, advanced analytical techniques are needed[13]. Machine learning techniques offer a powerful tool for modeling and predicting the influence of nutrients and body function [14], but also kidney function in CKD patients[15-16] . By leveraging large-scale multidimensional databases and advanced algorithms, machine learning enables the development of predictive models that can identify patterns and associations between dietary interventions and clinical outcomes[15-17].

In this paper several machine learning models were trained and evaluated for their performance in predicting CKD, employing both techniques, regression[18-21] and classification[22-26].

Compared to Ghosh and Khandoker's[27], study, which focused on developing a machine learning-driven nomogram specifically for predicting CKD stages 3–5, this study provides a broader evaluation of multiple machine learning models for predicting CKD stages, including the novel consideration of dietary factors like kefir intake. While Ghosh and Khandoker[27], emphasize prediction accuracy within a specific CKD range, this research highlights the comprehensive performance of various models and underscores the significance of dietary interventions in managing CKD.

Through interdisciplinary collaboration and innovative research methodologies, we aim to contribute to the growing body of knowledge on personalized nutrition and precision medicine in renal health.

2. Material and Methods

This study explores the application of machine learning techniques for modeling and predicting the influence of kefir consumption in a low-protein diet on kidney function. The database used in this analysis consists of relevant information on documents, and relevant information about patients, including demographics, medical history, lab test results, and the stage of the presence of CKD, and intake or not kefir as part of a low-protein diet.
Patients with CKD stages G1 to G5 were selected randomly, and data were collected over a 12-month period (from 2021 to 2022) at regular intervals. Protein intake was recommended using tables for protein composition derived from various products as per nutritional recommendations. As part of the study, patients were advised to consume kefir three times per week within the recommended protein intake. The recommended daily permissible protein intake ranged from 0.8 g/kg body weight to 1 g/kg body weight of protein. Notably, a significant source of protein in the diet was kefir, consumed at least three times per week. Knowing, the total protein intake (Total protein intake or DPI (Daily Protein Intake)) is calculated using the Maroni formula [28]:

$$PCR\ [g/24h] = 6.25 \times (UNU\ [g/24h] + NUNU\ 0.03 \times body\ weight\ [kg]) \quad [1]$$

Equation [1] reflects the relationship between the nitrogen excreted in urine and the total protein catabolized in the body, assuming stable metabolic conditions. Where PCR (Protein Catabolic Rate) under stable conditions is equals the daily protein intake. The constant 6.25 indicates 1 gram of excreted nitrogen corresponds to 6.25 g of processed proteins.

	Attribute	Data Type	Description/Unit
0	id	Numeric	Identifier (0, 1, 2, ...)
1	age	Numeric	Age in years
2	sex	Numeric	Gender (0 for female, 1 for male)
3	blood_pressure_systolic	Numeric	Systolic blood pressure in mm/Hg
4	blood_pressure_diastolic	Numeric	Diastolic blood pressure in mm/Hg
5	BMI	Numeric	Body Mass Index in kg/m^2
6	albumin	Numeric	Albumin level in g/dL
7	bacteria	Categorical	Presence of bacteria (values: 'no', 'yes')
8	blood_glucose	Numeric	Blood glucose level in mg/dL
9	Hb	Numeric	Hemoglobin level in g/dL
10	Creat	Numeric	Creatinine level in umol/L
11	Urea	Numeric	Urea level in mg/dL
12	UNU	Numeric	Urinary Nitrogen as Urea level in g/24h
13	Protein Catabolic Rate (PCR)	Numeric	PCR level g/24h
14	Daily Protein Intake (DPI)	Numeric	DPI in grams per kilogram of body weight
15	GFR	Numeric	Glomerular Filtration Rate in ml/min/1.73m^2
16	MDRD	Numeric	Estimated Glomerular Filtration Rate in mL/min
17	Proteinuria	Numeric	Proteinuria in g/mmol
18	Hypertension	Categorical	Hypertension (values: 'no', 'yes')
19	diabetes_mellitus	Categorical	Diabetes Mellitus (values: 'no', 'yes')
20	coronary_disease	Categorical	Coronary Disease (values: 'no', 'yes')
21	kefir	Numeric	Consumption of kefir (0 for yes, 1 for no)
22	CKD_1m (at 1st month)	Categorical	('GFR1', 'GFR2', 'GFR3', 'GFR4', 'GFR5')
23	CKD_12m (after 12 months)	Categorical	('GFR1', 'GFR2', 'GFR3', 'GFR4', 'GFR5')
24	output after 12 months	Numeric	Outcome (0 for same stage, 1 for different stage)

Table 1. Medical Attributes Dataset

NUNU It is known as Non-Urinary Nitrogen and amounts 0.03 g per kg of body weight (70 kg individual resulting in 2.1 g NUNU in 24-hour urine). For example, if we have received 7 g of nitrogen as urinary urea, then the total protein intake will be obtained by multiplying 7 by 6.25, i.e., 45 grams for 24 hours. To this value, 13 grams that do not belong to the urea cycle (2.1 × 6.25) are added, resulting in the total amount of ingested proteins of 58 grams. This value is then divided by body weight, and it amounts to 58/70 kg = 0.8 g/kg body weight.

Data Collection: The dataset utilized in this study was obtained from at the University Clinic of Nephrology in Skopje. It consisted of 150 entries, each representing an individual patient. The dataset (Table 1), included demographic information such as age and sex, clinical measurements including blood pressure readings (systolic and diastolic), body mass index (BMI), and various laboratory parameters such as albumin levels, blood glucose levels, hemoglobin levels, creatinine levels, urea levels, uric acid levels, and protein creatinine ratio.

The details of the process of implementing various machine learning models to predict chronic kidney disease (CKD) stages based on patient data are presented at Figure 1.

Dataset: it is data in a CSV file format.
Reading the data file: dataset is read with function 'data = pd.read_csv'.

Data Preprocessing: was conducted to ensure the quality and reliability of the dataset. This involved:
-Cleaning the data ('cleaned_data = df.drop_duplicates().reset_index(drop=True)'),
- Checking for null values ('null_values = df.isnull().sum()'),
-Normalization ('from sklearn.preprocessing import StandardScaler')

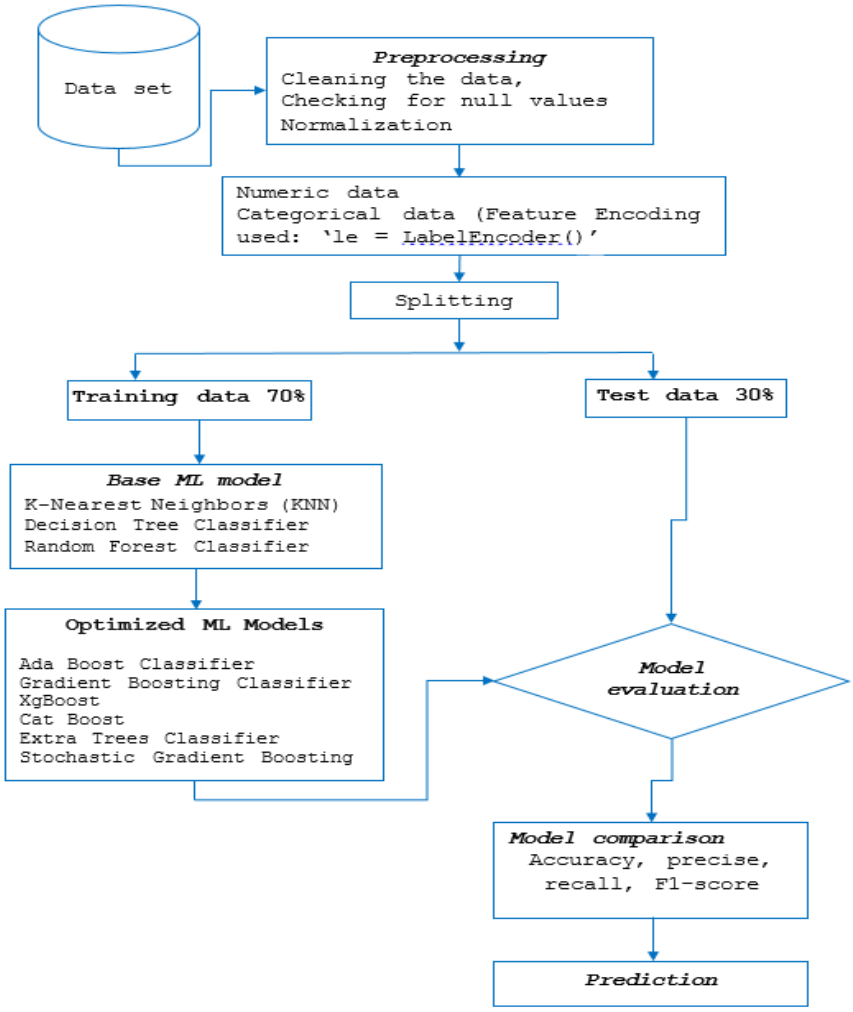

Figure 1 The Algorithm ML prediction in Python.

The next step is to do categorical features or encoding using label encoding. As all of the categorical columns have 2 categories we can use label encoder. In this process, the class LabelEncoder from the sklearn.preprocessing library is used 'le = LabelEncoder()'to transform categorical data into numerical labels. Each unique category within a column is assigned a unique integer label, 0 or 1. This encoding enables machine learning algorithms to interpret categorical data as numerical inputs. Categorical data are attributes (bacteria, hypertension, diabetes_mellitus, coronary disease, CKD_1m and CKD_12m)

Data splitting: The dataset was divided into two subsets: 70% for training and 30% for testing subsets. This split ensures that the model's performance can be evaluated on unseen data, providing a realistic estimate of its predictive capabilities.

Model Building: Several machine learning models were trained and evaluated for their performance in predicting CKD. These models include K-Nearest Neighbors (KNN), Decision Tree Classifier, Random Forest Classifier, AdaBoost Classifier, Gradient Boosting Classifier, Stochastic Gradient Boosting, XGBoost, CatBoost, Extra Trees Classifier, and LightGBM Classifier.

Base models:

Nearest Neighbors (KNN): is a simple and intuitive algorithm used for classification and regression tasks. It works based on the principle that data points with similar features tend to belong to the same class. Given a new data point, KNN calculates the distance to all other data points in the training set and classifies the new point based on the majority class of its k nearest neighbors. KNN begins by storing all available data points and their corresponding class labels or target values. When a new data point is presented for classification or prediction, KNN calculates the distance between this point and every other point in the training set. Common distance metrics include Euclidean distance and Manhattan distance.

If point one is $P = (x_1, x_2, \ldots, x_n)$ and point two is $Q = (y_1, y_2, \ldots, y_n)$ in n-dimensional space distance calculations are:

Euclidean Distance $\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ [2]

Manhattan Distance $\sum_{i=1}^{n}|x_i - y_i|$ [3]

Decision Tree Classifier: is also algorithms for classification and regression tasks. They partition the feature space into regions and make predictions based on the majority class (for classification) or the average value (for regression) within each region. Each internal node of the tree represents a decision based on a specific feature, leading to a split in the data. Decision trees are easy to interpret and visualize, making them useful for understanding the decision-making process of the model. In the context of decision trees, Gini impurity is used as a criterion to evaluate the quality of a split.

It quantifies the probability of incorrectly classifying a randomly chosen element if it were randomly labeled according to the distribution of classes in the set.

A lower Gini impurity indicates a purer node with more homogeneity in class labels. The formula for Gini Impurity is

$$G = 1 - \sum_{i=1}^{C}(p_i)^2 \quad [4]$$

Entropy is another measure of impurity or disorder in a set of data points. In the context of decision trees, entropy is used as an alternative criterion for evaluating splits.

It measures the average amount of information needed to predict the class label of a randomly chosen element in the set.

Like Gini impurity, lower entropy values indicate greater homogeneity in class labels.

$$\text{Entropy} \quad E = -\sum_{i=1}^{C} p_i log_2(p_i) \quad [5]$$

Information gain is a concept used to quantify the effectiveness of a split in a decision tree.
It measures the reduction in entropy or Gini impurity achieved by splitting the data on a particular feature. Information Gain

$$IG = Entropy\ (parent) - \sum_{i=1}^{k}(\frac{N}{N_i} \times\ Entropy\ (child_i)) \quad [6]$$

Random Forest Classifier: is ensemble learning methods that build multiple decision trees during training and output the mode of the classes (for classification) or the mean prediction (for regression) of the individual trees.

They are robust against overfitting and tend to produce high-quality predictions by aggregating the outputs of multiple trees.

For a Random Forest, the final prediction is an aggregate of the predictions from individual decision trees. The aggregation method differs between classification and regression tasks:

Classification: The final prediction is the mode (majority vote) of the predictions from all individual trees.

$$\hat{y} = mode\ \{h_1(x), h_2(x), \dots, h_B(x)\} \quad [7]$$

where $h_i(x)$ is the prediction from the *i-th* tree, and B is the total number of trees in the forest.

Regression: The final prediction is the average of the predictions from all individual trees.

$$\hat{y} = \frac{1}{B}\sum_{i=1}^{B} h_i(x) \quad [8]$$

where $h_i(x)$ is the prediction from the *i-th* tree, B is the number of trees in the forest.
$\hat{y}$ is the final predicted value; x represents the input feature vector; $h_i(x)$ is the prediction of the *i-th* tree; B is the number of trees in the random forest ensemble.

Optimized Models:
AdaBoost (Adaptive Boosting) Classifier is an ensemble learning method that combines multiple weak learners to create a strong classifier. It iteratively trains a sequence of weak learners, giving higher weights to the misclassified data points at each iteration. The updated weight of the i-th data point at the i-th iteration is given by:

$$\omega_i^{(t+1)} = \omega_i^{(t)} \cdot e^{-\alpha_t \cdot y_i \cdot h_i(x)} \quad [9]$$

Here, y_i is the true label of the i-th data point, $h_i(x)$is the prediction of the weak learner, and $\alpha_t$is the weight of the weak learner.

Gradient Boosting Classifier: Gradient boosting is another ensemble learning technique that builds a strong classifier by sequentially adding weak learners, each correcting the errors of its predecessor. It minimizes a loss function by gradient descent in the function space of weak learners. Gradient boosting models are known for their predictive power and flexibility but may require careful tuning of hyperparameters. Gradient Boosting minimizes a loss function L by adding weak learners iteratively. Let F(x) be the additive model representing the ensemble of weak learners. The objective is to find F(x) that minimizes the loss function L over the training data.

$$F_o(x) = \arg min \sum_{i=1}^{n} L(y_i, \gamma) \quad [10]$$

Here, $F_o(x)$ is the initial model (a constant value that minimizes the loss function), y_i are the true values, and n is the number of training samples. The parameter γ represents the optimal constant value that minimizes the loss function over all training data.

XGBoost Extreme Gradient Boosting is an additive model in a similar way to traditional gradient boosting, but with optimizations for speed and performance. The objective function L for XGBoost combines the training loss and a regularization term to penalize the complexity of the model.

The objective function $L(\Theta)$ that XGBoost aims to minimize is defined as:

$$L(\Theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{i=1}^{K} \Omega(f_k) \quad [11]$$

Where, $l(y_i, \hat{y}_i)$ is the loss function (e.g., mean squared error) measuring the difference between the true label y_i and the predicted label $\hat{y}_i$. n is the number of training samples. K is the number of trees in the model. Θ represents the parameters of all the trees. $\Omega(f_k)$ is the regularization term for the $k-th$ tree, which helps to control the complexity of the model and prevent overfitting. XGBoost employs both L1 (lasso) and L2 (ridge) regularization techniques to control model complexity and prevent overfitting. This combination helps XGBoost to handle large-scale datasets effectively while maintaining model simplicity and interpretability.

The regularization term $\Omega(f_k)$ is typically defined as:

$$\Omega(f_k) = \gamma\ T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j{}^2 + \alpha \sum_{j=1}^{T} |\omega_j| \quad [12]$$

Where: γ controls the complexity cost by penalizing the number of leaves T in the three; λ is the L2 regularization term (ridge regularization) that penalizes the sum of the

squared leaf weights ω_j; α is the L1 regularization term (lasso regularization) that penalizes the sum of the absolute values of the leaf weights ω_j. T is the number of leaves in the tree; ω_j is the weight of the $j-th$ leaf.

Uses a combination of L1 and L2 regularization to control model complexity, which helps in both feature selection (L1 regularization) and weight shrinkage (L2 regularization), thus providing a more flexible approach to prevent overfitting.

CatBoost Classifier: is a gradient boosting library developed by Yandex that is designed to handle categorical features efficiently. It automatically converts categorical features into numerical values and uses a specialized algorithm to deal with categorical data during training. The objective function $L(\Theta)$ that CatBoost t aims to minimize is defined as:

$$L(\Theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{i=1}^{K} \Omega(f_k) \quad [13]$$

Where, $l(y_i, \hat{y}_i)$ is the loss function (e.g., mean squared error) measuring the difference between the true label y_i and the predicted label $\hat{y}_i$. n is the number of training samples. K is the number of trees in the model. Θ represents the parameters of all the trees. $\Omega(f_k)$ is the regularization term for the $k-th$ tree, which helps to control the complexity of the model and prevent overfitting.

CatBoost primarily uses L2 regularization (ridge regularization) to control the complexity of the model. The regularization term $\Omega(f_k)$ for CatBoost is defined as:

$$\Omega(f_k) = \gamma\, \mathrm{T} + \frac{1}{2}\lambda \sum_{j=1}^{T} {\omega_j}^2 \quad [14]$$

Where: γ controls the complexity cost by penalizing the number of leaves T. λ is the L2 regularization term on leaf weights ω_j.

Focuses on controlling model complexity using L2 regularization, which tends to shrink weights and reduce variance, making the model more robust to overfitting.

Extra Trees Classifier: (Extremely Randomized Trees) is an ensemble learning method similar to random forests. It builds multiple decision trees using random subsets of features and random thresholds for each feature, leading to faster training times and potentially higher predictive accuracy.

The final prediction for an input sample x is made by aggregating the predictions from all the individual trees in the ensemble, as mode (majority vote) of the predictions from all the individual trees.Both Random Forest and Extra Trees Classifier use ensemble methods that build multiple decision trees, but they differ primarily in how they select splits for these trees.

Stochastic Gradient Boosting is a variant of gradient boosting that introduces randomness into the training process. This randomness can improve the model's generalization ability and reduce overfitting by making the model less sensitive to individual training examples.

Model evaluation Models were evaluated using various metrics such as accuracy, precision, recall, and F1-score. This comprehensive evaluation ensured the selection of the most effective model for predicting CKD stages based on the given dataset.

Model comparison and *prediction* Choose an appropriate machine learning algorithm or model for CKD prediction based on the nature of the data and the problem at hand.

3. Results

3.1. DESCRIPTIVE STATISTICS

The average age of patients in the dataset is approximately 54.75 years, with a gender distribution of 79 men and 71 women, making about 52.7% of the patients male (Figure 2). The average systolic blood pressure is approximately 138.3 mmHg, with a standard deviation of 16.75 mmHg, while the average diastolic blood pressure is approximately 86.72 mmHg, with a standard deviation of 9.60 mmHg.

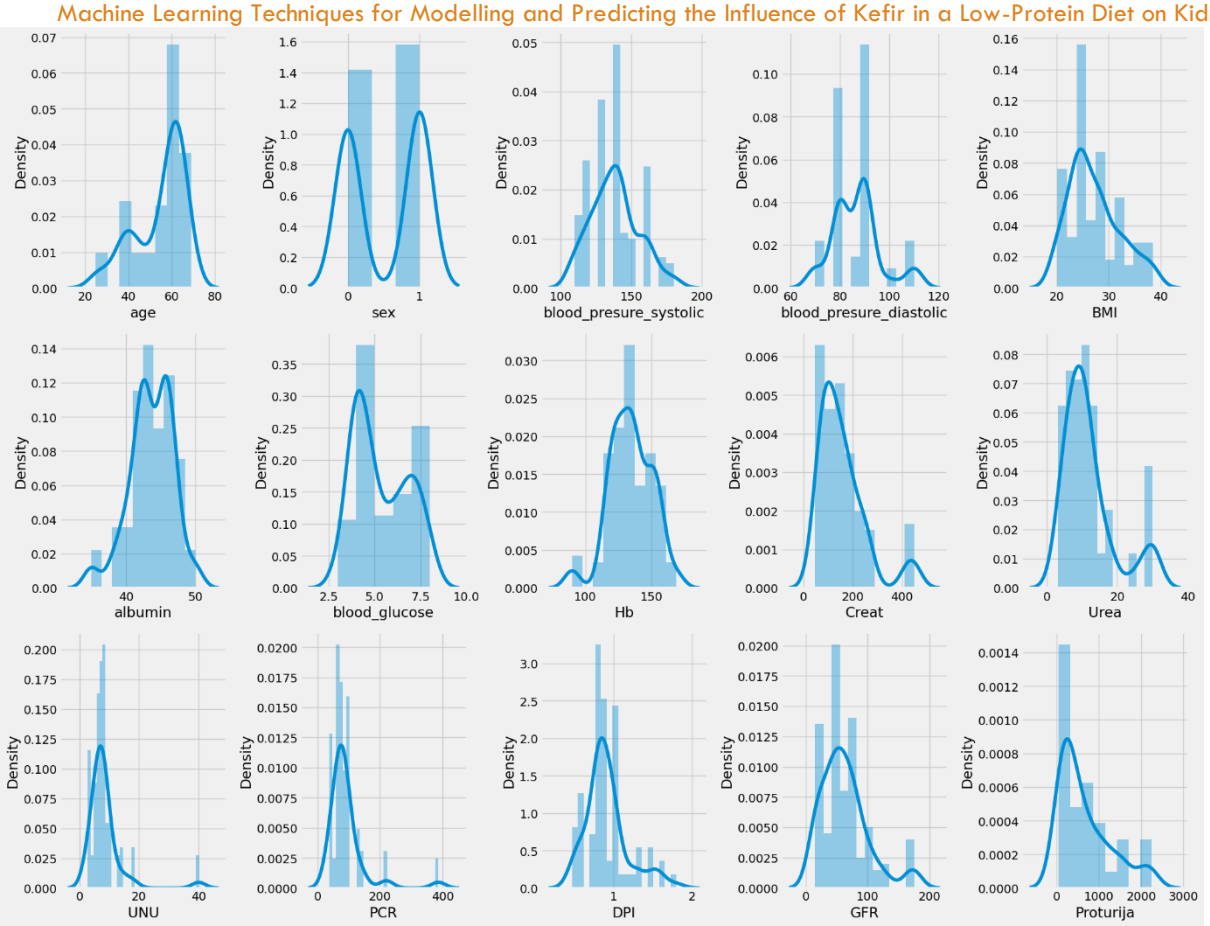

Figure 2 Numerical values of dataset

At the beginning the mean Body Mass Index (BMI) for patients in the study is calculated to be approximately 27.52, with a standard deviation of 4.99 (Table 2). The minimum BMI recorded is 20.1, while the maximum BMI is 38.5, indicating a diverse range of body weight among the patients. The BMI distribution shows that 25% of patients have a BMI below 24.1, 50% fall below 26.1, and 75% fall below 29.4. After 12 months average BMI for patients is lest then beginning (27.46), but is not statistical significant (Table 2). Values of variables at the beginning are marked with I, the average values level after 12 months is marked with II. It was found average lower values for albumin, Urea, UNU, PCR and Proteinuria but not statistical significant difference. Statistical significant differences are found for creatinine (Creat), GFR and MDRD (Table 2).

The parameters and their respective thresholds used in this study can be further understood by referring to the sources [29-34].

Attribute	Mean (I)	Sta. Dev (I)	Mean (II)	Sta. Dev (II)	Pearson Correlation	t-Stat	P(T<=t) two-tail	Signif. (p<0.05)
BMI	27.52	4.99	27.46	5.03	0.998	0.996	0.328	No
Albumin	43.66	3.18	43.64	2.63	0.746	0.044	0.965	No
Hb	135.26	17.79	135.97	16.58	0.968	-0.848	0.403	No
Creat	156.48	94.59	147.10	83.56	0.963	1.892	0.034	Yes
Urea	11.70	7.56	10.81	6.01	0.858	1.230	0.229	No
UNU	8.52	6.90	8.22	4.80	0.909	0.503	0.619	No
PCR	92.00	67.30	87.78	42.79	0.889	0.646	0.523	No
DPI	1.18	0.67	1.10	0.56	0.913	1.309	0.201	No
GFR	62.07	33.57	63.63	34.45	0.990	-1.733	0.047	Yes
MDRD	57.40	33.71	59.97	32.84	0.983	-2.059	0.024	Yes
Proteinuria	813.11	904.70	668.82	609.96	0.801	1.379	0.179	No

Table 2. Statistical Analysis of Attributes Changes Over 12 Months

The average creatinine levels observed in this study were consistent with known ranges for different stages of CKD. For example, the average creatinine levels for patients in Stage 1 CKD (GFR > 90 mL/min) typically range from 0.6 to 1.2 mg/dL, which aligns with our findings (American Kidney Fund, 2022). Similarly, the proteinuria levels across CKD stages also corresponded with established medical data. For instance, Stage 1 CKD patients generally exhibit normal to mildly increased protein levels, typically less than 30 mg/g creatinine (ACR < 30 mg/g) (National Kidney Foundation, 2022), while patients in Stage 4 CKD (GFR = 15-29 mL/min) usually have severe proteinuria, often greater than 2000 mg/g creatinine (ACR > 2000 mg/g) (Kidney Research UK, 2022).

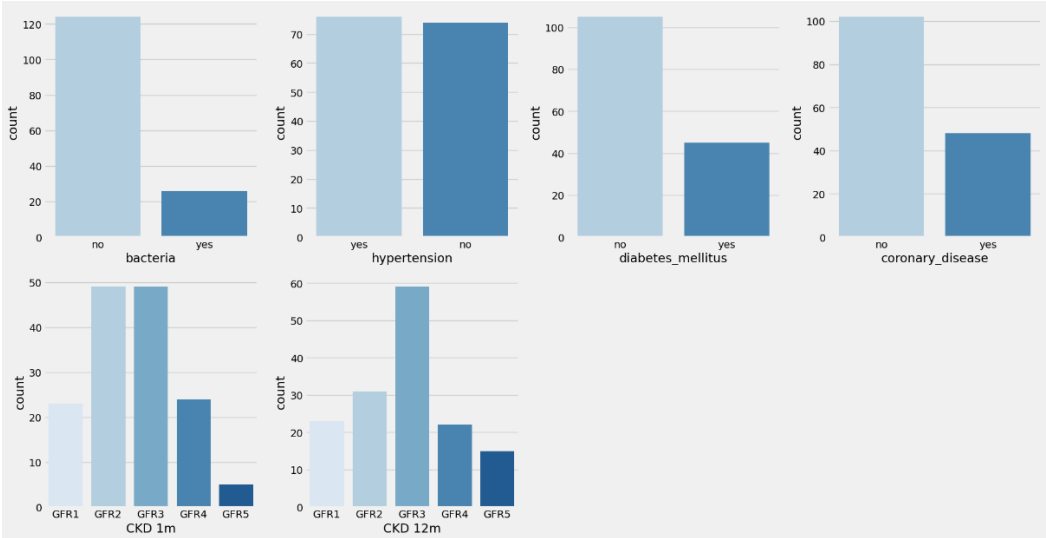

Figure 3 Categorical variables of data set

Figure 3, presents the categorical variables, including the 50% presence of hypertension (76 yes, 74 no), 30% presence of diabetes mellitus (105 no, 45 yes), 32% presence of coronary artery disease (102 no, 48 yes), and 68% consumption of kefir (98 yes, 52 no). Furthermore, the dataset contains information on the staging of Chronic Kidney Disease (CKD from stage 1 to stage 5) taken at the 1st month and after 12 months, as well as the outcome variable indicating whether the patient remained in the same CKD stage (output 0) or progressed to a different stage within a 12-month period (output 1).

The provided table (Table 2), displays the mean and standard deviation values for DPI and GFR across different stages of CKD. The table provides insights into how the mean and variability of GFR values change across different stages of CKD after 12 months regulated DPI and dietary modifications and kefir intake as a source of proteins. The data of DPI and GFR at the beginning is marked with I, and after 12 months marked with II, when DPI is < =1 grams per kilogram of body weight, across different stages of CKD. The GFR values show a declining trend with advancing CKD stages, which aligns with the typical progression of the disease. Intake values (DPI) appear to be relatively stable across stages but show slight variations, possibly indicating changes in dietary habits or compliance with medical advice over the 12 months.

CKD Stage	DPI _I **(Mean ± SD)**	DPI _II **(Mean ± SD)**	**GFR_I (Mean ± SD)**	**GFR_II (Mean ± SD)**
GFR > 90	1.32 ± 1.14	1.00 ± 0.38	116.75 ± 28.44	120.95 ± 32.46
GFR 60-90	1.00 ± 0.30	0.94 ± 0.31	70.35 ± 10.12	73.75 ± 8.50
GFR 45-59	1.05 ± 0.53	0.92 ± 0.34	47.80 ± 5.32	50.50 ± 4.65
GFR 30-44	1.47 ± 0.64	0.95 ± 0.37	33.70 ± 6.95	35.40 ± 6.30
GFR 15-29	1.44 ± 0.90	1.00 ± 0.42	21.20 ± 4.19	22.80 ± 4.52
GFR < 15	1.32 ± 1.14	1.00 ± 0.50	11.15 ± 2.50	12.45 ± 2.85

Table Average and Standard Deviation of DPI and GFR by CKD Stage

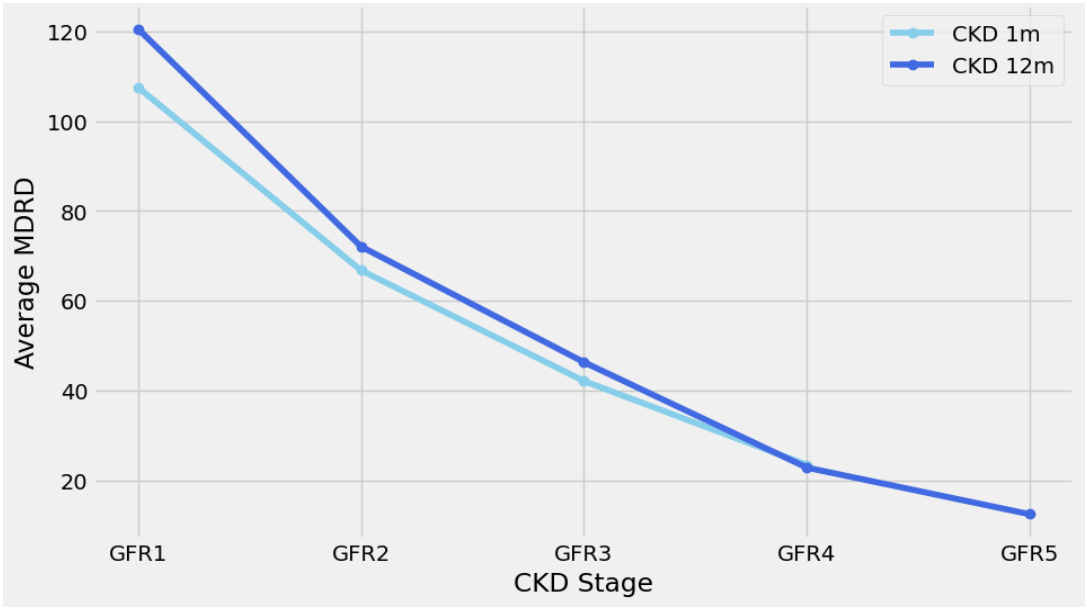

Figure 4 Average MDRD at beginning (first month –CKD 1m) and after 12 months (CKD 12m) by CKD Stage

At Figure 4 is presented graphically the statistical significant changes in MDRD value between the beginning and after 12 months. These changes reflect to the CKD stage and the impact of interventions such as dietary modifications of controlled DPI and kefir intake.

The dataset includes additional clinical attributes such as the presence of proteinuria, kefir intake, and an output parameter, which may be relevant for further analysis and correlation with other clinical outcomes.

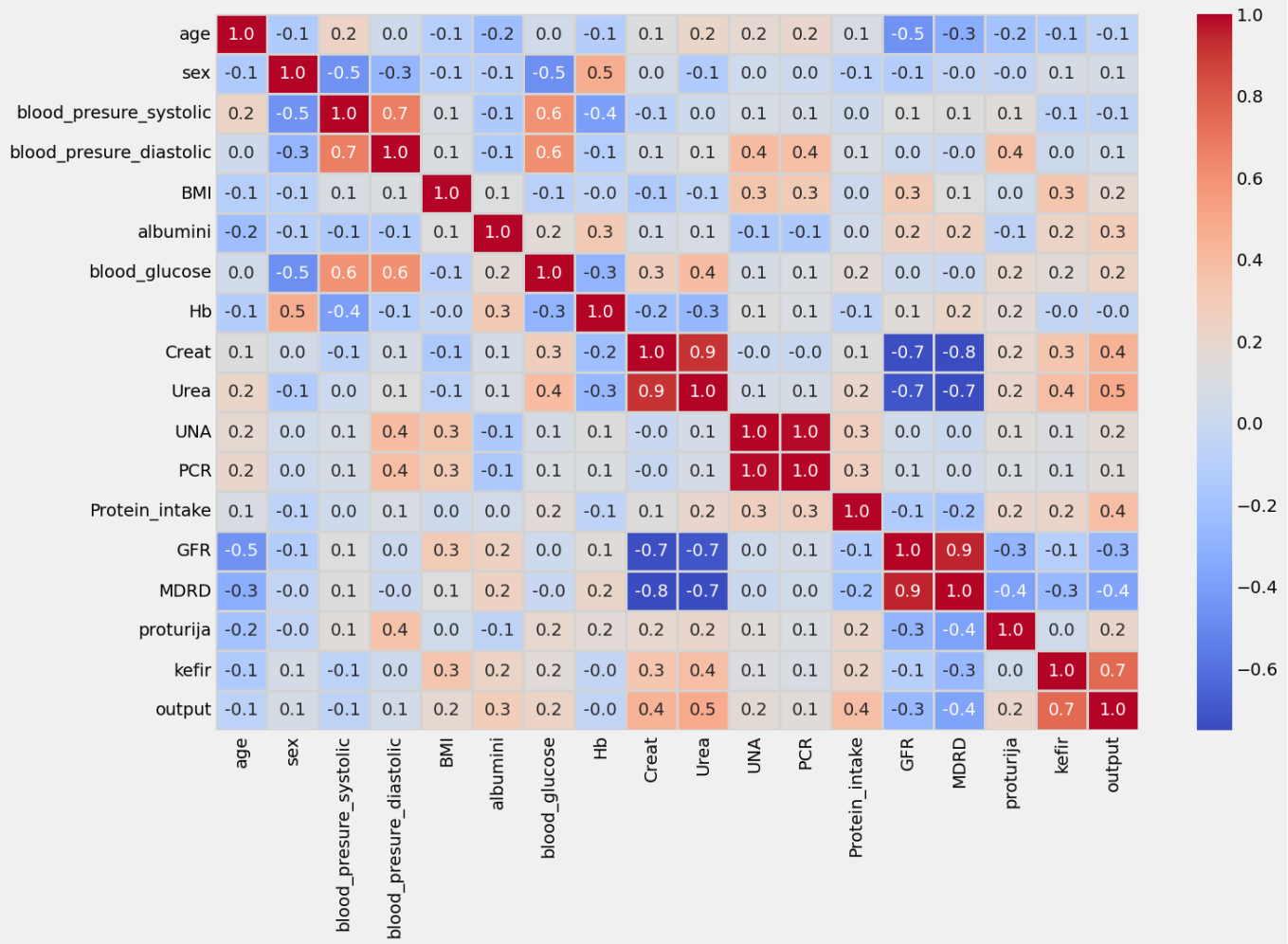

Figure 5 Heatmap of data

The positive correlation between regulated daily protein intake and kefir intake and improved clinical outcomes (0.7) underscores the potential of kefir as a dietary intervention in CKD management. By providing a fermented high-quality protein source, kefir supports an optimal protein catabolic rate, which is crucial for maintaining kidney function in CKD patients. The integration of kefir into dietary plans for CKD patients could enhance nutritional status and slow disease progression, contributing to better patient outcomes.

The relatively weak negative correlation between kefir and GFR (-0.09) and MDRD (-0.29) is in the context of protein, which likely represents the progression of CKD, where a higher value indicates more advanced disease stages.

These findings suggest that kefir as a fermented dairy product, may have a beneficial nutritional effect on kidney function, helping to manage CKD progression more effectively. Further studies could explore the specific mechanisms through which kefir exerts its positive effects, as well as its potential role in personalized dietary interventions for CKD patients.

3.2. MODEL PERFORMANCE

The models were trained using the training dataset and evaluated on the test dataset. The performance metrics used include accuracy score, confusion matrix, and classification report. These metrics provide insights into the models' ability to correctly classify CKD stages and handle class imbalances. The models were evaluated using metrics such as accuracy, precision, recall, and F1-score to determine their performance and suitability for predicting CKD stages. The metrics are defined as follows:

- Accuracy: The proportion of correctly classified instances among the total instances.
- Precision: The proportion of true positive results in relation to the total predicted positives.
- Recall: The proportion of true positive results in relation to the actual positives.
- F1-Score: The harmonic mean of precision and recall, providing a single metric that balances both concerns.

The performances of the machine learning models in predicting CKD stages, evaluated using the above-mentioned metrics, are summarized in Table 4.

Model	Accuracy	Precision	Recall	F1-Score
Random Forest Classifier	0.96	0.97	0.96	0.96
AdaBoost Classifier	0.96	0.97	0.96	0.96
CatBoost	0.96	0.97	0.96	0.96
Extra Trees Classifier	0.96	0.97	0.96	0.96
Gradient Boosting Classifier	0.93	0.94	0.93	0.93
Stochastic Gradient Boosting	0.93	0.94	0.93	0.93
XGBoost	0.93	0.94	0.93	0.93
Decision Tree Classifier	0.91	0.94	0.93	0.93
K-Nearest Neighbors (KNN)	0.88	0.90	0.89	0.89

Table 4. Performance Metrics for Machine Learning Models

Figure 6 provides a graphical representation of the accuracy scores for each model, making it easy to compare their performance visually.

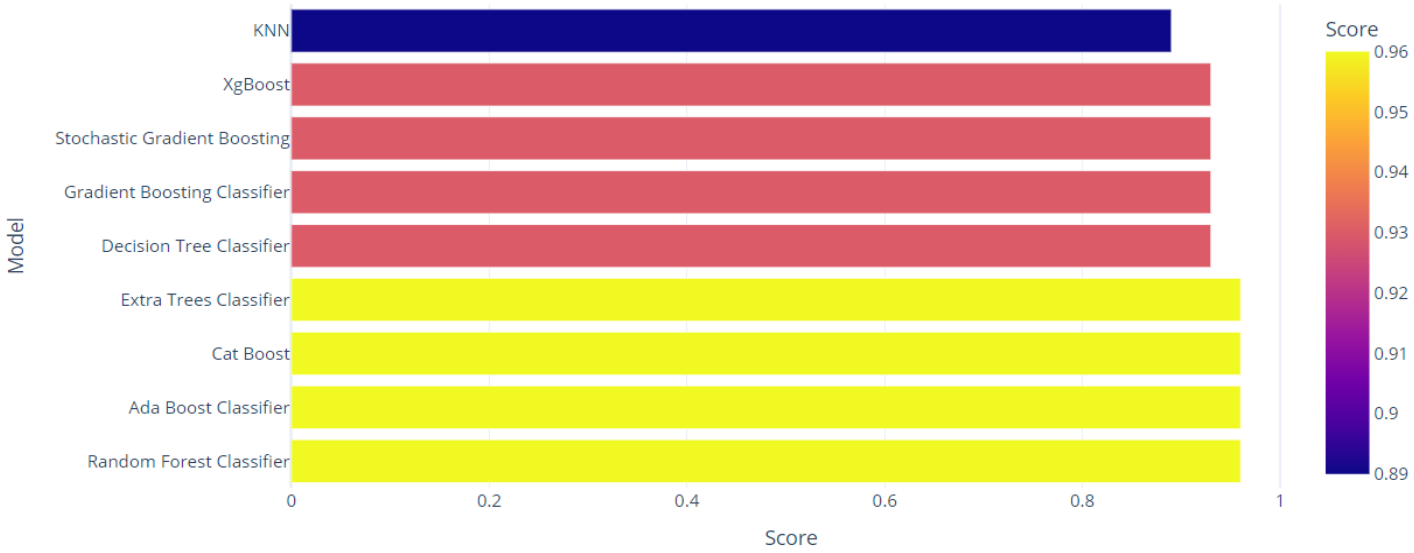

Figure 6 Models comparison score

The ensemble methods, particularly Random Forest, AdaBoost, CatBoost, and Extra Trees all achieved the accuracy of 96%, indicating their strong performance in predicting CKD stages, exhibited the highest accuracy scores, indicating their robustness and ability to generalize well to new data. These models effectively handle the complexities and nuances of the CKD dataset, providing reliable predictions for patient outcomes. The success of these models in predicting CKD stages demonstrates the potential of machine learning in transforming clinical practice and advancing personalized medicine.

3.3. MODEL PREDICTIONS

The ensemble models, particularly Random Forest Classifier, AdaBoost Classifier, CatBoost, and Extra Trees Classifier, exhibited the highest accuracy scores, indicating their robustness and ability to generalize well to new data. These models effectively handle the complexities and nuances of the CKD dataset, providing reliable predictions for patient outcomes.

Further analysis of clinical attributes, such as the presence of proteinuria and kefir intake, revealed significant correlations with CKD progression, highlighting the importance of dietary interventions in managing kidney health.

4. Discussion

The machine learning models exhibited strong predictive capabilities in CKD Risk Prediction based on patient data[18]. The comprehensive evaluation and comparison of these models underscore the potential of machine learning techniques in enhancing CKD management through precise and early detection[19].

In this study, the ensemble methods, particularly Random Forest Classifier, AdaBoost Classifier, CatBoost, and Extra Trees Classifier, demonstrated the highest accuracy scores of 96%. These models showed robustness and excellent generalization capabilities, which are crucial for clinical applications where patient data variability is high. The effectiveness of these models in handling the complexities of CKD data highlights their potential in providing reliable predictions for patient outcomes. Similar to our findings, Saha et al.[23], found that ensemble methods, particularly Random Forest, XGBoost, and CATBoost, performed exceptionally well. In the mention study [23], the dataset included 25 features, and the random forest algorithm achieved the highest accuracy of 99.08% This reinforces the effectiveness of these algorithms in CKD prediction. However, our accuracy rates (96%) are slightly lower, which could be attributed to differences in datasets and feature selection.

In the comparative study by Maria Youse[25], it was demonstrated that the number of features significantly impacts the accuracy of CKD predictions. The study showed that as the number of attributes decreases, the accuracy varies, with 6 attributes yielding an average accuracy of 97.4%, while using 23 attributes resulted in 80.5% accuracy. This reinforces the importance of feature selection in machine learning models for CKD prediction. Our findings align with this observation, as the

high accuracy of our ensemble models underscores their effectiveness in utilizing selected features to predict CKD stages accurately.

In "Comparative Study of Classifier for Chronic Kidney Disease prediction using Naive Bayes, KNN and Random Forest"[22], evaluated K-nearest neighbor, had the similar performance as our study. In the study by Saha et al.[23], K-NN performed with an accuracy of 0.7875, precision of 0.8571, and F-measure of 0.8090. Unlike this study, "Improving Prediction of Chronic Kidney Disease Using KNN Imputed SMOTE Features and TrioNet Model" proposed improvement of KNN, model performing with 98,98% accuracy. Our results showed KNN had the lowest performance (88% accuracy) among the models tested. This discrepancy might be due to sensitivity to noisy data (KNN) and overfitting (Decision Tree). However, these models still performed reasonably well, highlighting their utility as base models in the ensemble approaches.

Overall, the ensemble methods, with their ability to combine the strengths of multiple weak learners, proved to be the most effective for predicting CKD stages. These findings underscore the importance of using advanced machine learning techniques in healthcare decision-making, paving the way for personalized dietary interventions to improve kidney health.

By leveraging the predictive power of these models, healthcare practitioners can better understand the impact of dietary factors, such as kefir consumption, on kidney function and make informed decisions to enhance patient outcomes. The success of these models in predicting CKD stages demonstrates the potential of machine learning in transforming clinical practice and advancing personalized medicine.

Conclusion

NOVEL CONTRIBUTIONS OF THIS STUDY INCLUDE:
Comprehensive Model Evaluation: This study provides a detailed comparison of various machine learning models, highlighting the strengths and weaknesses of each in the context of CKD prediction.

Incorporation of Dietary Factors: Including dietary factors, especially kefir intake, and demonstrating its positive impact on kidney function, adds significant value to the research.

Robust Performance Metrics: The extensive use of multiple performance metrics ensures a thorough evaluation of model performance, enhancing the reliability and applicability of the findings in clinical settings.

Impact on Personalized Medicine: The findings underscore the potential of machine learning in advancing personalized dietary interventions, paving the way for tailored treatment plans based on individual dietary habits and clinical profiles.

Further longitudinal studies are needed to establish the long-term benefits of kefir consumption on kidney function and CKD progression. Investigating the specific bioactive components in kefir that contribute to its beneficial effects on kidney health could provide deeper insights into its role in CKD management.

In summary, these findings underscore the significance of advanced machine learning techniques in healthcare decision-making, paving the way for personalized dietary interventions to enhance kidney health. By leveraging the predictive power of these models, healthcare practitioners can better understand the impact of dietary factors, such as kefir consumption, on kidney function and make informed decisions to improve patient outcomes.

References

1. Rees, K., Dyakova, M., Wilson, N., Ward, K., Thorogood, M., Brunner, E. (2013). Dietary advice for reducing cardiovascular risk. *Cochrane Database of Systematic Reviews, 2013*(12). https://doi.org/10.1002/14651858.CD002128.pub5
2. Santesso, N., Bianchi, M., Mente, M., Mustafa, R., Heels-Ansdell, D., Schünemann, H. J. (2012). Effects of higher versus lower protein diets on health outcomes: A systematic review and meta-analysis. *European Journal of Clinical Nutrition, 66*(12). https://doi.org/10.1038/ejcn.2012.37
3. Levin, A., Hemmelgarn, B., Culleton, B., Sheldon Tobe, S., McFarlane, P., Ruzicka, M., Tonelli, M. (2008). Guidelines for the management of chronic kidney disease. *Canadian Medical Association Journal, 179*(11). https://doi.org/10.1503/cmaj.080351
4. Carrero, J., Cozzolino, M. (2014). Nutritional therapy, phosphate control and renal protection. *Nephron Clinical Practice, 126*(1), 1-7.
5. Mitch, W. (2005). Beneficial responses to modified diets in treating patients with chronic kidney disease. *Kidney International, 67*(1), 133-135.
6. Kaysen, G., Odabaei, G. (2013). Dietary protein restriction and preservation of kidney function in chronic kidney disease. *Blood Purification, 35*(1-3), 22-25.
7. Fouque, D., Laville, M., Boissel, P. J. (2009). Low protein diets for chronic kidney disease in nondiabetic adults. *Cochrane Database of Systematic Reviews, 2009*(3). https://doi.org/10.1002/14651858.CD001892.pub2
8. Fouque, D., Wang, P., Laville, M., Boissel, J. P. (2000). Low protein diets delay end-stage renal disease in nondiabetic adults with chronic renal failure. *Nephrology, Dialysis, Transplantation, 15*(12), 1986-1992.
9. Gavrilovska, E., Knights, V., Simovska, V., Ivanovski, N. (2022). Statistical analysis concerning the importance of a low protein diet in the progression of chronic kidney disease. *Journal of Hygienic Engineering and Design, 39*(1), 116-121.
10. Damjanovska Gavriloska, E., Kalevska, T., Dimitrovska, G., Kljusurić, J. G. Antoska Knights, V. (2023). Mathematical modelling of determining the dynamics of the nutritional value content of classic kefir and three types of flavors of functional kefir. *Horizons - International Scientific Journal, 1*(1), 116-129. https://doi.org/10.20544/
11. Damjanovska Gavriloska, E., Kalevska, T., Dimitrovska, G., Gajdoš Kljusurić, J. (2023). Sensory and pH evaluations of novel varieties of kefir. *Horizons - International Scientific Journal, 1*(1), 77-90. https://doi.org/10.20544/
12. Gligorova Damjanovska, E., Severova, G., Cakalaroski, K., Antovska-Knight, V., Danilovska, I., Simovska, V., Ivanovski, N. (2018). Beneficial short term effect of low protein diet on chronic kidney disease progression in patients with chronic kidney disease stage G3a. A pilot study. *Hippokratia, 22*(4), 178-182. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6251848/
13. Chen, F., Kantagowit, P., Nopsopon, T., Chuklin, A., Pongpirul, K. (2023). Prediction and diagnosis of chronic kidney disease development and progression using machine-learning: Protocol for a systematic review and meta-analysis of reporting standards and model performance. *PLoS ONE, 18*(2), e0278729. https://doi.org/10.1371/journal.pone.0278729
14. Knights, V., Kolak, M., Markovikj, G., Gajdoš Kljusurić, J. (2023). Modeling and optimization with artificial intelligence in nutrition. *Applied Sciences, 13*(13), 7835.
15. Segal, Z., Kalifa, D., Radinsky, K., et al. (2020). Machine learning algorithm for early detection of end-stage renal disease. *BMC Nephrology, 21*(518). https://doi.org/10.1186/s12882-020-02093-0
16. Badrouchi, S., Bacha, M. M., Hedri, H., et al. (2023). Toward generalizing the use of artificial intelligence in nephrology and kidney transplantation. *Journal of Nephrology, 36*, 1087-1100. https://doi.org/10.1007/s40620-022-01529-0
17. Al-Lamki, R., Burlacu, A., Iftene, A., Jugrin, D., Popa, I. V., Lupu, P. M., ... Covic, A. (2020). Using artificial intelligence resources in dialysis and kidney transplant patients: A literature review. *BioMed Research International, 2020*, 9867872. https://doi.org/10.1155/2020/9867872
18. Dritsas, E., & Trigka, M. (2022). Machine learning techniques for chronic kidney disease risk prediction. *Big Data and Cognitive Computing, 6*(3), 98. https://doi.org/10.3390/bdcc6030098
19. Zhao, J., Zhang, Y., Qiu, J., Zhang, X., Wei, F., Feng, J., ... Li, W.-D. (2022). An early prediction model for chronic kidney disease. *Scientific Reports, 12*, 2765. https://doi.org/10.1038/s41598-022-06665-y
20. Debal, D. A., & Sitote, T. M. (2022). Chronic kidney disease prediction using machine learning techniques. *Journal of Big Data, 9*(109). https://doi.org/10.1186/s40537-022-00657-5
21. Knights, V., & Prchkovska, M. (2024). From equations to predictions: Understanding the mathematics and machine learning of multiple linear regression. *Journal of Mathematical & Computer Applications, 3*(2), 1-8.
22. Devika, R., Avilala, S. V., & Subramaniyaswamy, V. (2019). Comparative study of classifier for chronic kidney disease prediction using naive Bayes, KNN and random forest. In S. Tiwari, M. C. Trivedi, M. L. Kolhe, K. Mishra, B. K. Singh (Eds.), *Advances in Data and Information Sciences* (pp. 679-684). Springer. https://doi.org/10.1109/ICCMC.2019.8819654
23. Saha, I., Gourisaria, M. K., & Harshvardhan, G. M. (2022). Classification system for prediction of chronic kidney disease using data mining techniques. In S. Tiwari, M. C. Trivedi, M. L. Kolhe, K. Mishra, B. K. Singh (Eds.), *Advances in Data and Information Sciences* (pp. 423-428). Springer. https://doi.org/10.1007/978-981-16-5689-7_38
24. Sinha, P., & Sinha, P. (2015). Comparative study of chronic kidney disease prediction using KNN and

SVM. *International Journal of Engineering Research & Technology (IJERT), 4*(12). https://www.ijert.org/research/comparative-study-of-chronic-kidney-disease-prediction-using-knn-and-svm-IJERTV4IS120622.pdf

25. Youse, M. (2023). Prediction of chronic kidney disease using different classification algorithms: A comparative study. *Journal of Xi'an Shiyou University, Natural Science Edition, 17*(10), 453-462. http://xisdxjxsu.asia
26. Saif, D., Sarhan, A. M., & Elshennawy, N. M. (2024). Deep-Kidney: An effective deep learning framework for chronic kidney disease prediction. *Health Information Science and Systems, 12*(3). https://doi.org/10.1007/s13755-023-00261-8
27. Ghosh, S. K., & Khandoker, A. H. (2023). A machine learning driven nomogram for predicting chronic kidney disease stages 3–5. *Scientific Reports, 13*, 21613. https://doi.org/10.1038/s41598-023-48815-w
28. Maroni, B. J., Steinman, T. I., & Witch, W. E. (1985). A method for estimating nitrogen intake of patients with chronic renal failure. *Kidney International, 27*, 58-65.
29. American Heart Association. (n.d.). Understanding blood pressure readings. Retrieved May 31, 2024, from https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings
30. Mayo Clinic. (n.d.). Albumin test. Retrieved May 31, 2024, from https://www.mayoclinic.org/tests-procedures/albumin/about/pac-20384948
31. National Library of Medicine. (n.d.). Urine test. MedlinePlus. Retrieved May 31, 2024, from https://medlineplus.gov/ency/article/003772.htm
32. American Diabetes Association. (n.d.). Diagnosis. Retrieved May 31, 2024, from https://www.diabetes.org/a1c/diagnosis
33. National Kidney Foundation. (n.d.). Creatinine. Retrieved May 31, 2024, from https://www.kidney.org/atoz/content/creatinine
34. National Kidney Foundation. (n.d.). Proteinuria. Retrieved May 31, 2024, from https://www.kidney.org/atoz/content/proteinuria

