Computer Modeling in
Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2023.045868

ARTICLE

Improving Prediction of Chronic Kidney Disease Using KNN Imputed SMOTE Features and TrioNet Model

Nazik Alturki[1], Abdulaziz Altamimi[2], Muhammad Umer[3,*], Oumaima Saidani[1], Amal Alshardan[1], Shtwai Alsubai[4], Marwan Omar[5] and Imran Ashraf[6,*]

[1]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh, 11671, Saudi Arabia

[2]Department College of Computer Science and Engineering, University of Hafr Al-Batin, Hafar, Al-Batin, 39524, Saudi Arabia

[3]Department of Computer Science & Information Technology, The Islamia University of Bahawalpur, P.O. Box 63100, Bahawalpur, Pakistan

[4]Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, P.O. Box 151, Al-Kharj, 11942, Saudi Arabia

[5]Information Technology and Management, Illinois Institute of Technology, Chicago, IL 60616-3793, USA

[6]Information and Communication Engineering, Yeungnam University, Gyeongsan, 38541, Korea

*Corresponding Authors: Muhammad Umer. Email: umer.sabir@iub.edu.pk; Imran Ashraf. Email: imranashraf@ynu.ac.kr

Received: 10 September 2023 Accepted: 29 November 2023 Published: 11 March 2024

ABSTRACT

Chronic kidney disease (CKD) is a major health concern today, requiring early and accurate diagnosis. Machine learning has emerged as a powerful tool for disease detection, and medical professionals are increasingly using ML classifier algorithms to identify CKD early. This study explores the application of advanced machine learning techniques on a CKD dataset obtained from the University of California, UC Irvine Machine Learning repository. The research introduces TrioNet, an ensemble model combining extreme gradient boosting, random forest, and extra tree classifier, which excels in providing highly accurate predictions for CKD. Furthermore, K nearest neighbor (KNN) imputer is utilized to deal with missing values while synthetic minority oversampling (SMOTE) is used for class-imbalance problems. To ascertain the efficacy of the proposed model, a comprehensive comparative analysis is conducted with various machine learning models. The proposed TrioNet using KNN imputer and SMOTE outperformed other models with 98.97% accuracy for detecting CKD. This in-depth analysis demonstrates the model's capabilities and underscores its potential as a valuable tool in the diagnosis of CKD.

KEYWORDS

Precision medicine; chronic kidney disease detection; SMOTE; missing values; healthcare; KNN imputer; ensemble learning

1 Introduction

Precision medicine emerged as a potential field concept that offers personalized medical care. Precision medicine is especially helpful for accurate diagnosis, offering effective treatment and timely intervention for special subgroups of patients. Precision medicine is particularly helpful for chronic kidney disease (CKD). The kidneys, which resemble fist-sized beans, are positioned just below the rib cage on both sides of the spine [1]. Each day, they filter as much as 150 quarts of blood, resulting in the production of nearly 2 quarts of urine. Primarily functioning within the excretory system, the kidneys carry out the vital task of expelling excess water and waste from the body through the process of urination [2]. This intricate process encompasses a series of excretion and re-absorption steps that are crucial for maintaining a stable equilibrium of body chemicals [3]. The kidneys are essential for maintaining the right levels of salt, potassium, and acidity in the body. They also produce hormones that affect other organs. For example, these hormones help control blood pressure, make red blood cells, and manage calcium levels, among other important tasks.

CKD earns its “chronic” classification due to its gradual onset and enduring impact on the functionality of the urinary system. Factors that elevate the risk of CKD encompass high blood pressure, diabetes, and cardiovascular disease [4]. Individuals affected by CKD, particularly during its advanced stages, may encounter unfavorable effects on their neurological and immunological systems. In less developed nations, certain patients may progress to a stage where the utilization of dialysis becomes imperative as they approach the end of their lives. The Glomerular Filtration Rate (GFR), a pivotal measure of kidney function, is derived from considerations such as results from blood tests, gender, the patient’s age, and other relevant criteria [5]. Several factors can elevate the probability of CKD, such as diabetes, high blood pressure, genetic conditions, autoimmune diseases, medication misuse, exposure to toxins, and family history, often progressing silently before symptoms appear [6].

The primary culprits behind CKD are diabetes and high blood pressure. Diabetes, marked by elevated blood sugar levels, can lead to kidney damage as well as affect blood vessels, the heart, and eyes. Additionally, inadequately managed hypertension can significantly contribute to heart attacks, strokes, and the development of chronic kidney disease. Other factors that can impact kidney health include dysplasia, kidney stones, metabolic diseases, genetic diseases, tumors, recurrent urinary tract infections (UTIs), glomerulonephritis, fatness, and aging [7]. CKD is often considered a silent disease due to being asymptomatic until normal functions of the kidneys are reduced by 80%–85% to awfully low levels of 15%–20% [8]. As CKD reaches its advanced stages, several prominent symptoms become evident, including fatigue, decreased appetite, reduced energy levels, swelling in the legs and ankles, difficulties with concentration, dry skin with intense itching, sleep disturbances, nocturnal muscle cramps, puffiness around the eyes, and frequent urination, particularly during the night [9].

Presently, CKD is rapidly becoming a major health issue, resulting in millions of deaths worldwide due to the lack of timely and affordable treatment. Particularly affected are individuals from low and middle-income countries [10]. As of 2013, approximately one million people lost their lives due to CKD [11]. Developing nations bear a more substantial burden of this disease, with a total of 387.5 million CKD patients in low to average-income countries, comprising 177.4 million males and 210.1 million females [12]. These figures highlight the growing prevalence of CKD in developing countries. However, efforts are being made to improve early detection of the disease to enable timely treatment during its initial stages.

Reshma et al. [13] employed a feature selection approach on the CKD Dataset. To perform this selection, they used the Ant Colony Optimization (ACO) method, a metaheuristic algorithm employed specifically for feature selection, categorized as a Wrapper method. Initially, the dataset comprised 24 attributes. However, through the application of the feature selection algorithm, they narrowed it down to 12 pertinent features, which formed the basis for constructing their model. Notably, the model was built using the Support Vector Machine (SVM) classification algorithm. Another study [14] focused on predicting CKD using the same dataset. This dataset encompassed 24 attributes along with a singular target variable. To construct their predictive model, they employed two ML algorithms, namely K-Nearest Neighbors (KNN) and Naïve Bayes. Impressively, KNN yielded the highest accuracy at 97%, while Naïve Bayes achieved a commendable accuracy of 91%. Until now, various feature selection algorithms have been used to identify the most effective algorithm for extracting the most important features for predicting chronic kidney disease. Since many datasets have imbalanced classes, class balancing is necessary to improve the performance of classifier models. Additionally, missing data values in datasets must be addressed to improve the classification task. In this study, synthetic minority oversampling (SMOTE) is used for data balancing and K-Nearest Neighbors (KNN) imputer is used to handle missing values for CKD detection. Recently, Nimmagadda et al. [15] also analyzed KNN, SVM and regression models to detect CKD. The ensemble model has been employed by researchers in disease diagnosis [16,17].

In this scenario, computer-aided diagnosis (CAD) can play a vital role in achieving the initial stage prognosis of the disease. Machine learning (ML), a subdomain of artificial intelligence (AI) can be of great help in its diagnosis [18]. Such systems aim to assist clinical decision-makers in accurately classifying diseases. In this study, an ML-based method is introduced for diagnosing CKD. The primary contributions of this study as

- This study offers an accurate prediction model for the precision model to help kidney disease patients. An ensemble model using extreme gradient boosting (XGB), random forest (RF), and extra tree classifier (ETC) called TrioNet offers highly accurate prediction of CKD.
- The proposed model provides point-of-care testing to help medical experts make faster diagnoses and timely treatment which greatly improves the chances of patients' survival.
- KNN imputer in the proposed approach resolves the missing values problem while SMOTE helps equalize the class-imbalance problem.
- Machine learning models such as decision tree (DT), RF, gradient boosting machine (GBM), Gaussian Naive Bayes (GNB), ETC, support vector classifier (SVC), logistic regression (LR), and stochastic gradient descent (SGD), have been employed to conduct a comparative analysis. The proposed model's effectiveness is evaluated by contrasting its performance with state-of-the-art techniques, considering metrics such as accuracy, precision, recall, and F1 score to comprehensively analyze its capabilities.

The remaining sections of the paper follow this structure: Section 2 discusses related works that utilize ML in the context of CKD health conditions. Section 3 elaborates on the methods and the dataset employed for analysis. Section 4 presents the proposed appraoch. Section 5 is dedicated to presenting and discussing the study findings. Section 6 provides the conclusion and outlines potential future directions.

2 Related Work

Currently, there is a growing interest in developing procedures and equipment for predicting and monitoring different diseases. In this section, recent works are explored that have utilized ML methods for forecasting the risk of CKD.

For example, in [19], ML models were employed for predicting CKD. The pre-processing phase involved handling missing values and selecting relevant features. The study utilized a total of 11 ML models, including tree-based, regression-based, and statistical models. Notably, among these models, five achieved an impressive 100% accuracy on the University of California, Irvine (UCI) dataset.

These high-performing models were DT, AdaBoost (ADA), ETC, XGB, and RF. For CKD diagnosis, Zahid et al. [20] proposed an ML model that incorporated selected predominant features. Similar to the previous study, this research also implemented pre-processing and feature selection steps. The results revealed that the KNN model outperformed other learning models, achieving an impressive accuracy score of 99.50%.

The study [21] introduced a hybrid model designed for predicting CKD. The authors effectively tackled challenges related to missing values and class imbalance by leveraging the capabilities of RF. The study incorporated a variety of ML models, including GNB, GBM, Naive Bayes (NB), RF, DT, and their hybrid model. These models were applied to the UCI chronic kidney disease dataset. Notably, the hybrid ML model achieved a remarkable accuracy score of 100%. Another hybrid ML model aimed at CKD prediction, built upon Apache Spark, was proposed by Abdel-Fattah et al. [22]. The authors conducted experiments using both full feature sets and features selected through Relief-F. The outcomes demonstrated that the machine learning models DT, SVM, and GBM achieved an impressive accuracy of 100%.

Venkatesan et al. [23] unveiled an ML-based system for the early detection of CKD. In the pre-processing step, the authors applied a class imbalance technique and handled missing values using the K-NN imputer. The results of the study revealed that the ML model XGBoost achieved an impressive accuracy score of 98% for detecting CKD at an early stage. For kidney disease prediction using image data, Kumar et al. [24] proposed a system that employed ML models for recognition and prediction. The approach combined a fuzzy deep neural network with a deep learning approach. The proposed system achieved a high accuracy score of 99.23% for accurately predicting kidney diseases using image data. Dritsas et al. [25] utilized a variety of ML models for the timely prediction of CKD. With the application of pre-processing techniques, the authors addressed the issue of non-uniform class distribution using the SMOTE oversampling technique. Among the 15 ML models used in the study, the rotational forest model stood out with an accuracy score of 99.2% for predicting CKD.

In [26], the experiments were conducted using a CKD dataset comprising 25 attributes sourced from the UCI ML repository. To diagnose CKD, three machine learning models were employed DT, RF, and SVM. These models achieved prediction accuracies of 99.16%, 94.16%, and 98.3%, respectively. In reference to the discoveries outlined in [27], the CKD dataset sourced from the UCI repository was also employed to assess a range of classifiers. The classifiers utilized in this investigation encompassed artificial neural network (ANN), chi-square automatic interaction detector, linear SVM with both L1 and L2 penalties, LR, C5.0, and random tree (RT). Notably, the linear SVM, when utilizing the L2 penalty and integrating SMOTE for input data, achieved an accuracy of 98.86%. By employing a combination of SMOTE and lasso techniques for feature selection, the linear SVM obtained a similar accuracy of 98.46%. Furthermore, employing a deep neural network on the same dataset resulted in an outstanding accuracy of 99.6%.

Swain et al. [28] proposed a robust approach for CKD classification, incorporating several data pre-processing techniques. For feature selection, the authors utilized the Chi-square test and addressed class imbalances using SMOTE. The classification phase involved two ML models SVM and RF. The results represented the highest accuracy of 99.33% for CKD classification by the SVM model. In the research conducted by Almasoud et al. [29], the primary emphasis was placed on the identification of CKD through the utilization of minimal predictor variables within ML algorithms. Four distinct ML models were utilized RF, GB, LR, and SVM. Among these, the GB algorithm exhibited superior performance, boasting an impressive accuracy of 99% for the precise detection of CKD.

In the work presented in [30], Yashfi et al. introduced an approach for predicting risks associated with CKD using ML algorithms. The analysis involved data from CKD patients, and they employed ANN and RF for predictions. By selecting 20 features from a total of 25, ANN and RF were implemented, with RF achieving the highest accuracy of 97.12% for CKD risk detection. In a comparative study by Rady et al. [31], diverse algorithms, including SVM, radial basis function (RBF), multilayer perceptron (MLP),and probabilistic neural networks (PNN), were examined for predicting stages of kidney disease. Despite working with a limited dataset and a restricted feature set, the results highlighted that PNN outperformed other algorithms, achieving the highest accuracy of 96.7%. Furthermore, in [32], the authors utilized an ML technique for the prognosis of CKD. The study incorporated three ML models LR, DT, and SVM, with the application of the bagging technique to enhance the prediction accuracy. The findings demonstrate that the DT model achieved a noteworthy accuracy of 97.23% in predicting CKD.

The prior works in the field of CKD prediction using ML techniques have made significant contributions, but they also face several challenges and limitations. Many CKD datasets suffer from class imbalance, where the number of positive CKD cases is significantly lower than negative cases. This imbalance can lead to biased models and reduced predictive performance. Inherent issues with data quality, such as missing values, outliers, and inconsistencies, can affect the accuracy of predictive models. Handling missing data is a common challenge that prior works have had to address. Addressing these challenges and limitations is essential for advancing the field of CKD prediction using ML, as it would lead to more reliable and clinically relevant models. This paper addresses both challenges to get improved results in detecting CKD. For a comprehensive overview of the related studies, a comparative summary is shown in Table 1.

Table 1: Related work brief summary

Ref.	Classifiers	Dataset	Year	Accuracy achieved	K-NN imputer	SMOTE
[19]	DT, XGB, RF, ETC, ADA, K-NN, NN, SVC linear, LR, SVC RBF, GNB	UCI	2020	100% DT, RF, XGB, ETC, ADA	No	No
[20]	K-NN, SVM, RF, Bagging	UCI	2023	99.50% K-NN	No	No
[21]	GNB, GBM, NB, RF, DT and Hybrid model	UCI	2023	100% hybrid model	No	No
[22]	SVM, RF, DT, GBM, LR,NB	UCI	2022	DT 100% GBM 100%	No	No
[23]	LR, RF, K-NN, DT, SVM, XGBoost	UCI	2023	98% XGBoost	Yes	Yes
[24]	Hybrid on Fuzzy Deep Neural Network, Traditional Radioimmunoassay Method	Changhua Christian Hospital, Taiwan	2023	99.23% Hybrid on Fuzzy Deep Neural Network	No	No

(Continued)

Table 1 (continued)

Ref.	Classifiers	Dataset	Year	Accuracy achieved	K-NN imputer	SMOTE
[25]	SDG, NB, MLP, K-NN, J48, LR model tree, RF, RT, SVM, DT, RotF, ADA, LR, Stacking, Soft Voting	UCI	2022	99.2% Rotation Forest	No	Yes
[26]	RF, DT and SVM	UCI	2019	99.16% RF	No	No
[27]	LR, C5.0, ANN, linear SVM with penalty L1 and L2, chi-square automatic interaction detector, and random tree (RT).	UCI	2021	98.86% linear SVM with penalty L1 and L2 99.6% DNN	No	Yes
[28]	SVM and RF	UCI		99.33% SVM	No	Yes
[29]	LR, SVM, RF, GB	UCI	2019	99% GB	No	No
[30]	RF, ANN	UCI	2020	97.12% RF	No	No
[31]	PNN, MLP, SVM, and RBF	UCI	2019	96.7% PNN	No	No
[32]	LR, DT, SVM	UCI	2022	97.23% DT	No	No

3 Materials and Methods

This section of the research delineates the materials and methods employed in predicting kidney disease. It encompasses a succinct overview of the chronic kidney disease dataset, the data preprocessing steps, a description of ML models utilized, the proposed methodology, and the evaluation parameters employed for assessing the performance of the ML models.

3.1 Dataset

This research utilizes a dataset containing information on patients diagnosed with CKD, which was collected from India's Apollo Hospital over a two-month period in 2015. This dataset originates from the UCI data repository [33]. The dataset consists of 400 observations, which may include instances with incomplete or missing values. Among these observations, there are records for 250 CKD patients and 150 individuals without CKD. This breakdown represents 62.5% associated with CKD cases and 37.5% with non-CKD cases. The age of the individuals in the dataset ranges from 2 to 90 years. The CKD dataset consists of 24 features, which include 11 numeric features and 13 nominal features. Moreover, the 25th feature serves as an indicator for CKD classification. As demonstrated in Table 2.

3.2 Data Preprocessing

The CKD dataset presented challenges related to outliers and missing data, which necessitated a thorough cleaning during the preprocessing phase. Additionally, the model's performance is biased due to an inherited imbalance in the dataset class. The pre-handling phase involved addressing these issues through missed imputation and estimation values, removing outliers that introduced noise, and balancing the dataset [33,34]. The problem of class imbalance is addressed through the implementation of SMOTE. Initially, the dataset contained numerous missing values, with only a complete record for 158 patients available. During the preprocessing stage, significant attention is given to imputing these missing values, primarily through the utilization of the KNN imputer. After the completion of this imputation process, the dataset is carefully balanced. It encompassed 250 instances of patients with CKD and 150 instances of patients without CKD, resulting in an uneven distribution that could potentially affect the performance of the model. To mitigate this, dataset balancing is carried out as a crucial preprocessing step.

Table 2: Attributes details of the dataset

Attribute	Symbol	Unit	Descritption	Type	Missing		Valid number
					Number	Percentage	
Age	age	years	Patient's age	Numerical	9	2.3%	391
Anemia	ane	Yes, No	Does patient has anemia or not	Nominal	1	0.3%	399
Albumin	al	0,1,2,3,4,5	level of Albumin in the blood	Nominal	46	11.5%	354
Appetite	appet	Good, poor	Patient's appetite	Nominal	1	0.3%	399
Blood pressure	bp	mm/Hg	Patients'BP	Numerical	12	3.0%	388
Blood glucose	bgr	mgs/dl	blood glucose random count	Numerical	44	11.0%	356
Blood urea	bu	mgs/dl	Urea level in patients' blood	Numerical	19	4.8%	381
Bacteria	ba	Present, not present	Bacterial presence in the blood	Nominal	4	1.0%	396
Coronary artery disease	cad	Yes, No	Does the patient has coronary artery disease or not	Nominal	2	0.5%	398
Diabetes mellitus	dm	Yes, No	Is a patient diabetic or not?	Nominal	2	0.5%	398
Hemoglobin	hemo	gms	hemoglobin level in the blood	Numerical	52	13.0%	348
Hypertension	htn	Yes, No	Does the patient has hypertension on not	Numerical	2	0.5%	398
Pus cell	pc	normal, abnormal	pus cell count of patient	Nominal	65	16.3%	335

(Continued)

Table 2 (continued)

Attribute	Symbol	Unit	Descritption	Type	Missing		Valid number
					Number	Percentage	
Pus cell clumps	pcc	present, not present	Pus cell clumps presence in the blood	Nominal	4	1.0%	396
Potassium	pot	mEq/L	potassium level in the blood	Numerical	90	22.5%	310
Packed cell volume	pcv		packed cell volume in the blood	Numerical	71	17.8%	329
Pedal Edema	pe	yes, no	Does patient has pedal edema or not	Nominal	1	0.3%	399
Red blood cell count	rc	millions/cmm	red blood cell count of the patient	Nominal	106	26.5%	294
Red blood cells	rbc	normal, abnormal	Patients' red blood cells count	Nominal	152	38%	248
Specific gravity	sg	1.005, 1.010, 1.015, 1.020,1.025	The ratio of the density of urine	Nominal	47	11.8%	353
Sugar	su	0,1,2,3,4,5	Sugar level of the patient	Nominal	49	12.3%	351
Serum creatinine	sc	mgs/dl	serum creatinine level in the blood	Numerical	18	4.5%	382
Sodium	sod	mEq/L	sodium level in the blood	Numerical	88	22.0%	312
White blood cell count	wc	cells/cumm	white blood cell count of the patient	Numerical	106	26.5%	294
Class		CKD, not CKD	Does the patient has kidney disease or not	Nominal			

3.3 SMOTE (Synthetic-Minority-Oversampling-Technique)

To handle the class imbalance problems present in the CKD dataset this study makes use of SMOTE. Oversampling is a technique that entails increasing the number of instances in the minority class in comparison to the majority class, effectively enhancing the presence and representation of the underrepresented minority class.

This approach can contribute to enhancing model accuracy by generating a richer set of features for training purposes [34]. SMOTE is a contemporary methodology specifically designed to address overfitting concerns within imbalanced datasets. It functions by randomly selecting instances from the minority class and identifying their K-nearest neighbors within that same class. This process aids in the creation of synthetic samples that contribute to a more balanced representation of classes and, subsequently more robust model training [35]. Based on these neighbors, new synthetic instances are constructed, these synthetic instances are the same as the original instances as these synthetic instances are developed using existing features. This process aims to create a more balanced representation of the minority class without duplicating existing data.

Although SMOTE proves effective in addressing class imbalance, it may not be the optimal choice for high-dimensional datasets due to the potential introduction of additional noise. However, in the context of this study, SMOTE is employed to create a new training dataset, effectively tackling the class imbalance issue and enabling more robust model training.

3.4 KNN Imputer

In contemporary times, gathering data from various origins forms the bedrock for analysis, insights generation, theory validation, and other endeavors [36]. Nevertheless, it is crucial to acknowledge the potential for human errors during data extraction, leading to an issue of incomplete information. Consequently, efficiently handling missing values becomes a critical step in data preprocessing. The choice of imputation method to address these gaps significantly influences model performance. One noteworthy approach for missing values imputation is the KNN imputer provided by sci-kit-learn. It is the best alternative to orthodox imputation approaches. By powering the Euclidean distance matrix, the KNN imputer identifies the nearest neighbors to the observations containing missing values. The imputation process involves weighing available data by discounting the missing values while computing the Euclidean distance.

3.5 Eliminating Missing Values from Dataset

The subsequent data management technique involves the removal of instances containing missed values. In the second experimental set, this method is implemented, where all records with any missing values are excluded from the analysis. This means that any observation in the dataset that has at least one field with missing information is entirely removed from the dataset.

3.6 Machine Learning Models

ML is of great importance in improving the precision and effectiveness of predicting chronic kidney disease. Various ML algorithms are available for classifying CKD. Python's Scikit-learn library offers a diverse range of freely available ML classifiers, extensively used by a substantial user base, making significant contributions to the study. In this study, the Scikit-learn library is utilized to implement a range of machine learning models, including SVM, KNN, XGBoost, Stochastic Gradient Decent Classifier (SGDC), NB, LR, RF, and DT.

3.6.1 Random Forest

RF, an ensemble classifier, achieves remarkable prediction accuracy by aggregating multiple weak learners. It utilizes bootstrap bagging to train numerous DTs (Decision Trees) with different bootstrap samples [37]. A training set is sub-sampled to a bootstrap sample, during this process, maintaining the sample size of the original dataset. Predictions are made based on DT by RF and other ensemble classifiers [38]. Identifying the root node attribute is a significant challenge in constructing these decision trees at each stage of the tree-building process. Let p be the ultimate decision obtained through a majority vote from the ensemble of DT, where $T1(y)$, $T2(y)$, $T3(y)$, and $Tm(y)$ represent the number of decision trees participating in the prediction process, it is denoted as

$$p = mode\,\{T_1(y), T_2(y), \cdots, T_m(y)\} \tag{1}$$

$$p = mode\left\{\sum_{m=1}^{m}[T_m(y)]\right\} \tag{2}$$

3.6.2 Gradient Boosting Machine

GBM classifier is an ML algorithm that combines multiple weak learners to create a powerful predictive model [39,40]. GBM is loss function-dependent, which can be customized to suit specific needs. The GBM backs various common loss functions, as long as they are differentiable. In classification, logarithmic loss is commonly used, while regression often utilizes squared errors. One advantage of gradient boosting is that new loss function derivation is not needed each time the algorithm is employed. Instead, the system can be applied with any differentiable loss function. Higher accuracy can be achieved with GMB by fine-tuning multiple hyperparameters. For instance, the parameter n = 100 indicates the prediction contributing tree numbers to be 100. The final forecast is the aggregation of individual forecasts of 100 DTs through voting. Another crucial hyper-parameter is "max depth," which is often set to 60, allowing each DT 60 levels of depth at maximum. This control over the tree's depth helps prevent overfitting and enhances generalization capabilities.

3.6.3 Logistic Regression

LR stands out as the most prevalent technique employed for binomial classification tasks [41]. This method is distinguished by its utilization of the logistic equation or sigmoid function. An S-shaped curve denoting a Sigmoid function effectively transforms any input value into an output value that ranges between 0 and 1.

$$f(x) = \frac{1}{(1 + e^{-value})} \tag{3}$$

The base of the natural logarithms is denoted as e, and the variable 'value' represents the actual numerical value that needs to undergo conversion. The following graph displays the transformation of numbers ranging from –5 to 5 using the logistic function, resulting in values within the range of 0 to 1.

$$y = \frac{e^{b0+b1\times x}}{(1 + e^{(b0+b1\times x)})} \tag{4}$$

In the given context, $b0$ represents the intercept or bias, y stands for the projected performance, and $b1$ represents the single input value x associated coefficient. Each raw data column possesses a corresponding coefficient b, interrelated with it (a constant actual value) and prerequisites to be learned from the training data.

3.6.4 Extra Tree Classifier

Extremely Randomized Trees, abbreviated as ETC, is an ensemble learning technique primarily employed for classification tasks [42]. Much like the RF classifier, ETC also employs the strategy of creating multiple decision trees and then aggregating their predictions to enhance the model's precision and resilience. Yet, what sets ETC apart from RF is its approach to constructing decision trees [43]. Unlike RF, which picks a subset of features and evaluates the optimal split point at each node, ETC takes a distinct path. ETC chooses split points randomly across the feature space, resulting in quicker training times and reduced susceptibility to overfitting [44].

3.6.5 Support Vector Classifier

SVC is a popular ML technique that aims to identify a hyperplane in N-dimensional space for classifying data points [45]. The primary goal of this algorithm is to find the hyperplane that maximizes the margin between classes. The dimensionality, represented by N, varies depending on the number of features. While comparing two features is relatively straightforward, dealing with multiple features for classification can be more complex. By maximizing the margin, the accuracy of prediction is enhanced by SVC. Fuzzy SVC has been utilized by researchers in pathological brain detection [46].

3.6.6 Decision Tree

DT is a widely acclaimed ML algorithm used comprehensively for solving classification and regression tasks [47,48]. The key challenge in constructing a decision tree lies in selecting leveler root nodes, which is known as 'attribute selection'. Two prominent techniques commonly used for attribute selection include 'information gain' and 'Gini index (GI)'. The GI can be calculated using the following equation:

$$Gini = 1 - \sum_{i=1}^{classes} p(i|t)^2 \tag{5}$$

The DI is a metric utilized to measure the impurity of data within a dataset.

3.6.7 K-Nearest Neighbour

KNN is an ML model that does not rely on a dependent variable to make predictions for a given dataset [36]. Instead, the model is provided with ample training data, and it determines a data point's neighborhood. The KNN model calculates the distance between the nearest neighbors of a new data point and the data point itself. The 'K' value specifies the number of nearest neighbors to consider, and it determines the majority of the neighbors' votes. For instance, if 'K' = 1, the new data point is assigned to the class of the closest neighbor.

3.6.8 Gaussian Naive Bayes

The GNB model utilizes Bayes' theorem. The output of an event is predicted using unconditional probabilities in GNB [49,50]. If a sample is classified into k categories represented by $k = \{c1, c2, ..., ck\}$, the resulting output is denoted as c. The GNB function is expressed as under, where the class is represented by c and the sample by d

$$P(c|d) = (P(c) \times P(d|c))/P(d) \tag{6}$$

In the above equation, the probability of class c is represented by $P(c|d)$, given the sample d. The prior probability of class c is denoted by $P(c)$ while the likelihood of observing sample d of a given class c is by $P(d|c)$ and $P(d)$ is the probability of observing sample d.

3.6.9 Stochastic Gradient Decent Classifier

SGDC functions by amalgamating concepts from both LR and SVM [34]. It adopts the convex loss function characteristic of LR, endowing it with robust classification capabilities. Particularly well-suited for multiclass classification tasks, SGDC employs the one *vs*. all (OvA) strategy to combine multiple classifiers. One of SGDC's noteworthy advantages lies in its proficiency in handling large datasets, processing one example per iteration. Due to the regression-based approach, SGDC is both straightforward to implement and comprehend. However, achieving optimal performance necessitates careful parameter tuning for SGDC. Additionally, it is important to note that SGDC is highly sensitive to feature scaling, which should be taken into account for improved results.

3.6.10 XGBoost

XGBoost stands as a rapid supervised learning algorithm, chosen for its precise and reliable classification capability [42]. Its efficacy in this task is attributed to the presence of regularized learning capabilities, which aid in achieving smoother final weights and prevent overfitting issues.

3.7 Evaluation Metrics

ML model's effectiveness is measured based on metrics such as precision, accuracy, recall, and the F1 score.

Accuracy is a metric that represents the proportion of correct predictions made by the classifiers compared to the total predictions on the test data. Its highest possible value is 1, which indicates that classifier predictions are 100% correct, while the lowest possible value is 0, meaning all the predictions are false. It is calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

where TN is a true negative, TP is a true positive, FP is a false positive and FN is a false negative.

A positive predictive value, commonly known as precision, measures the value of True Positive (TP) instances among all the instances that are classified as true positives and False positives (FP). A precision score of 1 indicates that all instances considered as positive are actually positive. However, it is noteworthy that this metric does not affect positive instances mistakenly predicted as positive, even if they are labeled as negative.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

Recall, often referred to as sensitivity, signifies the proportion of positive classified instances out of all the actual positive instances. The recall metric is defined as follows:

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

Classifiers' performance is not evaluated as a definitive representation by "recall" and "precision" results, so the F1 score is of utmost importance because it combines both recall and precision, providing a balanced evaluation with a value between 0 and 1. It is calculated using the harmonic mean of recall and precision and is calculated using

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{10}$$

4 Proposed Approach

Missing values play a vital role in assessing learning models' performance. For any research aiming for good performance, it is necessary to handle missing values. In this study, we utilized KNN imputer to handle this problem. For the experimentation, the dataset is divided into the ratio of 70:30, this means that 70% of the data is used for the training of the models and the remaining 30% used for the testing of the models. For the efficient prognosis of kidney disease, this study proposed an ensemble learning technique that consists of three machine learning models (RF, ETC, and XGB). We called this model TrioNet. In the ensemble learning approach predictions from the different models are combined together to attain a higher value of accuracy. In this ensemble learning framework, each model has its own strength and weakness that plays a vital role in increasing the efficiency of the prediction. The proposed system architectural diagram is shown in the Fig. 1.

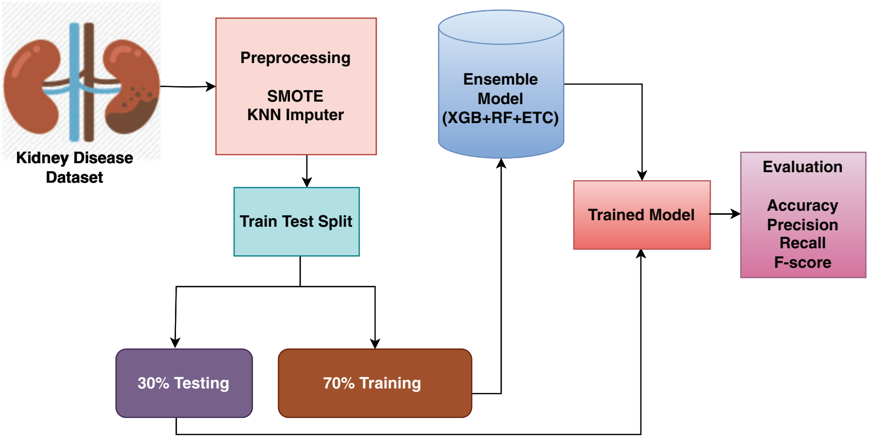

Figure 1: Workflow of TrioNet for chronic kidney disease detection

The TrioNet works by aggregating predictions from three different machine learning algorithms. In general, building such an ensemble entails training numerous models on the same dataset and then combining their predictions. This process requires individually training the XGB, RF, and ETC models on the same dataset. Each model generates an estimated probability for the target variable's various classifications. These probabilities may be added together to provide the final prediction for each instance in the dataset. A popular method for combining predictions is to take an average of all predicted probabilities. To create accurate and robust predictions, the proposed ensemble model capitalizes on the particular characteristics of three separate machine-learning algorithms. We can increase the model's generalizability and decrease overfitting by training all models on the chronic kidney disease dataset and integrating their predictions. The following is an outline of Algorithm 1, which offers a full explanation of how the proposed ensemble model operates:

$$\widehat{p} = argmax\left\{\sum_{i}^{n} XGB_i, \sum_{i}^{n} RF_i, \sum_{i}^{n} ETC_i\right\}. \tag{11}$$

Algorithm 1: Ensembling of XGB, RF, and ETC.

```
Input: input data $(x, y)_{i=1}^{N}$
$M_{XGB}$ = Trained_XGB
$M_{RF}$ = Trained_RF
$M_{ETC}$ = Trained_ETC
for $i = 1$ to $M$ do
    if $M_{XGB} \neq 0$ & $M_{RF} \neq 0$ & $M_{ETC} \neq 0$ & $training\_set \neq 0$ then
        $ProbXGB - 1 = M_{XGB}.probability(CKD - class)$
        $ProbXGB - 2 = M_{XGB}.probability(NonCKD - class)$
        $ProbRF - 1 = M_{RF}.probability(CKD - class)$
        $ProbRF - 2 = M_{RF}.probability(NonCKD - class)$
        $ProbETC - 1 = M_{ETC}.probability(CKD - class)$
        $ProbETC - 2 = M_{ETC}.probability(NonCKD - class)$
        Decision function $= max(\frac{1}{N_{classifier}} \sum_{classifier}$
        $(Avg_{(ProbXGB-1, ProbRF-1, ProbETC-1)}$
        $, (Avg_{(ProbXGB-2, ProbRF-2, ProbETC-2)}$
    end if
    Return final label $\widehat{p}$
end for
```

Within the proposed ensemble model, the prediction probabilities for individual test samples, denoted as $\sum_i^n XGB_i$, $\sum_i^n RF_i$, and $\sum_i^n ETC_i$, originate from the XGB, RF, and ETC models, respectively, as shown in Fig. 2. Subsequent to acquiring these probabilities from the XGB, RF, and ETC models for each test case, they undergo assessment using the soft voting criterion. This procedure is elucidated in the illustration of the voting architecture presented in the figure outlining the proposed approach. The ensemble model computes the greatest average probability among the classes to get the final class prediction. This entails combining the estimated probabilities from the XGB, ETC and RF classifiers. Eq. (12) defines that the class that has the high probability score will be considered as the final label:

$$VC(XGB + RF + ETC) = argmax(g(x)) \tag{12}$$

5 Results and Discussions

This section discusses the results and discussions on CKD detection. This section of the study also contains a comparison of the recent state-of-the-art models with the proposed approach.

5.1 Experimental Results of Learning Models by Removing Missing Values

During the initial phase of the experiments, the main focus is on dealing with any missing values present in the dataset. These occurrences of missing data are removed from the dataset. Following this, the adjusted dataset is employed to train machine learning models. Table 3 displays the outcomes of ML models using the dataset with missing values removed.

Results indicate that the XGB, ETC, and RF classifiers achieve a high accuracy score by attaining 83.14%,82.89%, and 81.34%, respectively. RF displayed precision, recall, and F1 scores of 90.56%, 89.52%, and 90.54%, respectively. ETC demonstrated 90.52% precision, recall, and F1 score. In a similar vein, XGBoost attained 89.58%, 89.88%, and 89.67% for precision, recall, and F1 scores, respectively. In contrast, LR demonstrated lower effectiveness, yielding an accuracy rate of 73.74%, a precision of 86.33%, a recall of 88.45%, and an F1 score of 87.14%. On the other hand, the proposed TrioNet outperformed other models, attaining an accuracy of 89.39%, a recall of 95.12%, a precision of 93.63%, and an F1 score of 94.46%. The superior performance of TrioNet, an ensemble model using XGB, RF, and ETC, can be attributed to its ability to combine the diverse strengths of these three models, reduce overfitting, improve robustness, balance precision, and recall, and create a more comprehensive understanding of the data. This ensemble approach effectively predicts kidney disease, making it a reliable and accurate choice for such medical applications.

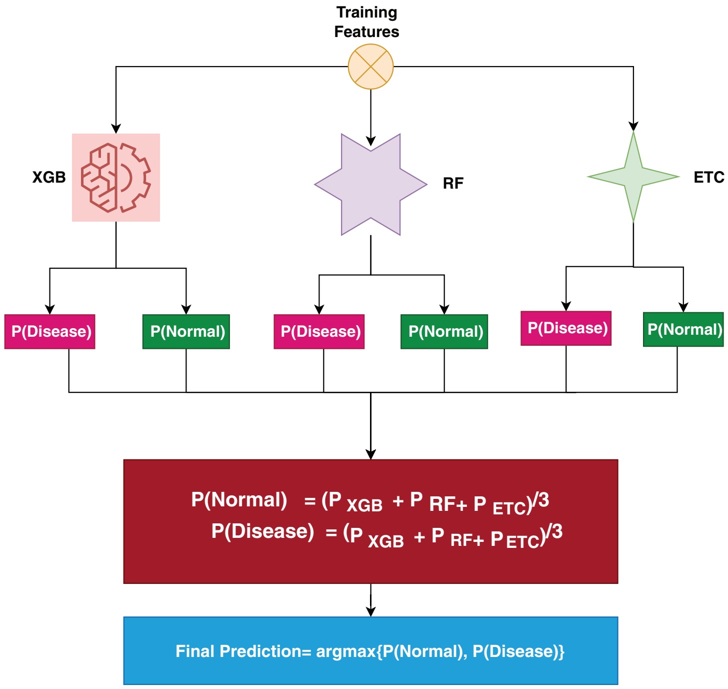
Training Features
XGB
RF
ETC
P(Disease)
P(Normal)
P(Disease)
P(Normal)
P(Disease)
P(Normal)
P(Normal) = (P XGB + P RF+ P ETC)/3
P(Disease) = (P XGB + P RF+ P ETC)/3
Final Prediction= argmax{P(Normal), P(Disease)}

Figure 2: Ensemble architecture

Table 3: Classification report of all learning models by removing missing values (in %)

Model	Accuracy	Precision	Recall	F1 score
LR	73.74	86.33	88.45	87.14
DT	77.41	87.14	89.53	88.76
RF	81.34	89.52	90.56	90.54
SGD	78.94	86.72	88.87	87.65
ETC	82.89	90.52	90.52	90.52
XGB	83.14	89.58	89.88	89.67
SVC	79.52	86.42	91.43	88.42
GNB	75.82	84.43	85.20	84.33
TrioNet	89.39	93.63	95.12	94.46

5.2 Experimental Results of Learning Models Using KNN-Imputed Values

The KNN imputer is used to tackle the issue of missing values that are present in the CKD dataset. The presence of missing values in the dataset is detected during the data preprocessing phase and this thing greatly affects the performance of all learning models. KNN imputer imputes the missing values by calculating the Euclidean distance of neighboring values. After applying the KNN imputer, the resultant dataset is used to train and evaluate the machine learning models. ML model's performance is shown in the Table 4.

Table 4: Classification report of all learning models using KNN imputer (in %)

Model	Accuracy	Precision	Recall	F1 score
LR	70.24	77.45	79.17	78.26
DT	70.37	75.48	76.37	75.18
RF	75.59	78.61	79.99	78.36
SGD	72.28	75.32	76.52	75.43
ETC	78.29	81.49	80.19	80.87
XGB	79.64	83.63	82.63	82.07
SVC	73.48	79.08	81.28	80.34
GNB	70.25	76.48	78.42	77.45
TrioNet	85.84	89.78	87.64	88.34

The experimental findings indicate that XGB, ETC, and RF achieved accuracies of 79.64%, 78.29%, and 75.59%, respectively. However, the suggested voting classifier surpassed all of them, achieving an impressive accuracy rate of 85.84%. Furthermore, TrioNet achieves a precision of 89.78%, a recall value of 87.64%, and an F1 score of 88.34%. The regression model LR shows an accuracy of 70.24%, which is lower than all models.

5.3 Experimental Results of Learning Models Using SMOTE Dataset

In the third set of experiments, the class imbalanced problem that is present in the dataset is treated using the SMOTE. In the preprocessing of the dataset, it has been observed that there is a class imbalanced problem. There are a total of 400 instances in the dataset and out of these 250 belong to the CKD patients class and 150 belong to the non-CKD class. it is clear from the fact that came from the preprocessing of the dataset that 62.5% instances belong to the CKD class and 37.5% belong to the non-CKD class. This class imbalanced problem is handled with the SMOTE. SMOTE is a widely used method used for the oversampling of data. training of the machine learning models done on the dataset obtained after applying SMOTE. Performance of the machine learning models using SMOTE balanced data is shown in the Table 5.

Table 5: Classification report of all learning models using SMOTE (in %)

Model	Accuracy	Precision	Recall	F1 score
LR	86.58	89.21	93.27	92.24
DT	87.28	90.34	94.57	92.44

(Continued)

Table 5 (continued)

Model	Accuracy	Precision	Recall	F1 score
RF	92.85	94.34	96.16	95.58
SGD	88.99	93.48	95.19	94.37
ETC	93.54	95.37	97.12	96.24
XGB	93.34	94.27	94.18	94.24
SVM	89.72	90.28	91.29	90.88
GNN	86.48	89.55	90.24	90.13
TrioNet	94.88	97.27	98.59	97.96

The results of the learning models show that the proposed stacked ensemble model outperformed other models in terms of accuracy value and achieved an accuracy score of 94.88%. Similarly, the performance of the individual learning model also improved as, ETC achieved 93.54%, RF 92.85%, and XGB achieved a 93.34% accuracy score on the balanced dataset. Tree-based model DT achieved an accuracy value of 87.28%. The two low performers on the balanced dataset are probability-based and regression-based methods which are GNB and LR achieved accuracy scores of 86.58% and 86.48%, respectively. Nonetheless, TrioNet outperforms using the upsampled dataset, showing its efficiency in detecting CKD.

5.4 Experimental Results of Learning Models Using KNN-Imputed Features and SMOTE

In the final set of experiments, SMOTE is used to tackle the class imbalance problem, and KNN imputer is used to handle the missing values. Results of this set of experiments are shown in the Table 6. using both SMOT and KNN imputer handles the class imbalanced problem and missing values concurrently. this will potentially increase the efficacy of the learning models. Subsequently, machine learning models are trained and evaluated after the implementation of KNN imputer and SMOTE.

Table 6: Classification report of all learning models using KNN-imputed features and SMOTE (in %)

Model	Accuracy	Precision	Recall	F1 score
LR	94.29	93.45	95.79	94.67
DT	92.38	95.29	96.24	95.69
RF	96.49	97.29	98.49	97.98
SGD	94.68	96.27	97.68	97.15
ETC	96.58	96.94	97.58	97.46
XGB	97.32	97.65	98.48	98.10
SVM	94.29	93.29	93.86	93.58
GNN	91.15	90.32	91.28	91.08
TrioNet	98.97	98.92	99.84	99.76

Experiment findings indicate that using KNN imputer and SMOTE enhances model performance. For example, the proposed voting classifier outperforms all other learning models with an accuracy of 98.97%. Furthermore, with 98.92% precision, 99.84% recall, and F1 score of 99.76%. The trioNet model achieves the greatest values in terms of all evaluation metrics. GNB has the worst performance with an accuracy score of 91.15%.

It is clear that TrioNet, which combines XGB, RF, and ETC, performs admirably and is especially successful for CKD diagnosis, resulting in higher accuracy and evaluation metrics when compared to other models, including the GNB.

5.5 Comparison of ML Models

To evaluate the effectiveness of the KNN imputer and SMOTE, we conducted a comparison across four scenarios: (i) deletion of missing values, (ii) application of the KNN imputer, (iii) utilization of SMOTE for up-sampling, and (iv) a combination of both SMOTE and the KNN imputer. Our assessment revealed that the fourth scenario, involving the initial use of the KNN imputer followed by SMOTE for up-sampling, exhibited significant improvement over the other three experiments. For a comprehensive performance analysis and clarity, please refer to Table 7 which displays the results of the machine learning models in all scenarios.

Table 7: Performance comparison of all four scenarios (in %)

Model	KNN imputer	Deleting missing values	SMOTE	KNN imputer + SMOTE
LR	70.24	73.74	86.58	94.29
DT	70.37	77.41	87.28	92.38
RF	75.59	81.34	92.85	96.49
SGD	72.28	78.94	88.99	94.68
ETC	78.29	82.89	93.54	96.58
XGB	79.64	83.14	93.34	97.32
SVC	73.48	79.52	89.72	94.29
GNB	70.25	75.82	86.48	91.15
TrioNet	85.84	89.39	94.88	98.97

5.6 Comparison with State-of-the-Art Techniques

In this section, a comparison with state-of-the-art CKD detection models is performed. This review took into account a number of recent research from the literature that served as comparative points. In the study by [30], an accuracy of 97.12% was achieved. Reference [32] demonstrated that a tree-based model, DT, achieved an accuracy of 97.23%, whereas in [31], the PNN model was employed, resulting in a 96.7% accuracy. Furthermore, references [23] and [27] reported impressive accuracy scores of 98% and 98.86%, respectively. Despite the high reported accuracy in these studies, the proposed models displayed commendable results, as outlined in Table 8. The results indicate the competitive performance of our proposed approach compared to the state-of-the-art models, underscoring its effectiveness in CKD detection.

Table 8: Performance comparison TrioNet with existing techniques (in %)

Ref.	Classifiers	Accuracy achieved
[23]	XGBoost	98%
[27]	linear SVM	98.86%
[29]	GB	99%
[30]	RF	97.12%
[31]	PNN	96.7%
[32]	DT	97.23%
Proposed approach	**TrioNet**	**98.97%**

5.7 Challenges and Future Directions

The research discusses the development of TrioNet, an ensemble learning technique for the prognosis of kidney disease. The study does not extensively address external validation or transferability of the proposed model. One direction for future research involves conducting clinical validation studies in collaboration with healthcare institutions. Real-world testing of TrioNet within clinical settings can provide valuable insights into its effectiveness and practical utility. It enables the model to be evaluated in the context of real patient cases and healthcare workflows, allowing us to assess its impact on clinical decision-making and patient outcomes. Consider integrating various data types (e.g., imaging, genetic, clinical notes) for a more comprehensive patient profile, which may lead to more accurate predictions. In Future research, TrioNet's performance will be evaluated on diverse datasets to ensure its applicability in different clinical settings. Real-world testing can provide insights into the model's effectiveness and practical utility.

6 Conclusions

Kidney disease, a prevalent health concern, encompasses a range of conditions affecting the vital filtration and regulatory functions of the kidneys. The growing incidence of kidney disease underscores the need for comprehensive research aimed at early detection, accurate diagnosis, and effective management strategies. This research introduces a comprehensive framework designed to accurately diagnose chronic kidney disease in patients, comprising two distinct segments. The initial phase involves dataset normalization through the utilization of KNN-imputed and SMOTE techniques. The proposed TrioNet model encompasses XGB, RF, and ETC models. The exceptional accuracy of 98.97% observed in the results underscores the capability of ensemble models to furnish a potential solution for the early identification of chronic kidney disease. A comparison against state-of-the-art models further highlights the superiority of the proposed model. The forthcoming trajectory of this research aims to extend the approach by creating a stacked ensemble that combines both machine learning and deep learning models. This expansion seeks to elevate the model's performance in handling datasets of greater dimensionality, thus yielding more generalized and robust outcomes.

Acknowledgement: This research was funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project Number PNURSP2023R333, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Funding Statement: This research was funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project Number PNURSP2024R333, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Author Contributions: study conception and design: Muhammad Umer, Nazik Alturki, Oumaima Saidani, and Amal Alshardan; data collection: Shtwai Alsubai, Abdulaziz Altamimi and Muhammad Umer; analysis and interpretation of results: Muhammad Umer, Marwan Omar, and Shtwai Alsubai; draft manuscript preparation: Muhammad Umer, Imran Ashraf, Nazik Alturki, Oumaima Saidani, and Amal Alshardan; funding: Abdulaziz Altamimi, Nazik Alturki. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: All the data related to the development of the proposed model can be found upon suitable request from the corresponding authors.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. NIH (2018). Your kidneys, how they work-niddk. https://www.niddk.nih.gov/health-information/kidney-disease/kidneys-how-they-work#::~:text=Each%20of% (accessed on 09/09/2023).
2. Foundation, N. K. (2020). Kidney disease: The basics. https://www.kidney.org/news/newsroom/factsheets/KidneyDiseaseBasics (accessed on 09/09/2023).
3. Reiser, J. (2011). Filtering new facts about kidney disease. *Nature Medicine, 17(1),* 44–45.
4. Dan, Z., Zhao, Y., Bi, X., Wu, L., Ji, Q. (2022). Multi-task transformer with adaptive cross-entropy loss for multi-dialect speech recognition. *Entropy, 24(10),* 1429.
5. Yang, Y., Tian, Z., Song, M., Ma, C., Ge, Z. et al. (2022). Detecting the critical states of Type 2 diabetes mellitus based on degree matrix network entropy by cross-tissue analysis. *Entropy, 24(9),* 1249.
6. Furman, D., Campisi, J., Verdin, E., Carrera-Bastos, P., Targ, S. et al. (2019). Chronic inflammation in the etiology of disease across the life span. *Nature Medicine, 25(12),* 1822–1832.
7. Koye, D. N., Magliano, D. J., Nelson, R. G., Pavkov, M. E. (2018). The global epidemiology of diabetes and kidney disease. *Advances in Chronic Kidney Disease, 25(2),* 121–132.
8. Care, U. (2020). Kidney (renal) failure. https://www.urologyhealth.org/urology-a-z/k/kidney-(renal)-failure (accessed on 09/09/2023).
9. Abdel-Kader, K. (2022). Symptoms with or because of kidney failure? *Clinical Journal of the American Society of Nephrology, 17(4),* 475–477.
10. Ruiz-Arenas, R., Sierra-Amor, R., Seccombe, D., Raymondo, S., Graziani, M. S. et al. (2017). A summary of worldwide national activities in chronic kidney disease (CKD) testing. *EJIFCC, 28(4),* 302–314.
11. Star, D. (2019). Over 35,000 develop kidney failure in bangladesh every year. https://www.thedailystar.net/city/news/18m-kidney-patients-bangladesh-every-year-1703665 (accessed on 09/09/2023).
12. Desk, P. A. E. (2020). Women affected at work and in businesses. https://en.prothomalo.com/bangladesh/roundtable/women-affected-in-workplaces-and-businesses (accessed on 09/09/2023).
13. Reshma, S., Shaji, S., Ajina, S. R., Vishnu Priya, S. R., Janisha, A. (2020). Chronic kidney disease prediction using machine learning. *International Journal of Engineering Research and, 9(7).* https://doi.org/10.17577/ijertv9is070092
14. Deepika, B., Rao, V., Rampure, D. N., Prajwal, P., Gowda, D. et al. (2020). Early prediction of chronic kidney disease by using machine learning techniques. *American Journal of Computer Science and Engineering Survey, 8(2),* 7.

15. Nimmagadda, S. M., Agasthi, S. S., Shai, A., Khandavalli, D. K. R., Vatti, J. R. (2023). Kidney failure detection and predictive analytics for ckd using machine learning procedures. *Archives of Computational Methods in Engineering, 30(4),* 2341–2354.
16. Su, Q., Wang, F., Chen, D., Chen, G., Li, C. et al. (2022). Deep convolutional neural networks with ensemble learning and transfer learning for automated detection of gastrointestinal diseases. *Computers in Biology and Medicine, 150,* 106054.
17. Yu, X., Qin, W., Lin, X., Shan, Z., Huang, L. et al. (2023). Synergizing the enhanced rime with fuzzy K-nearest neighbor for diagnose of pulmonary hypertension. *Computers in Biology and Medicine, 165,* 107408.
18. Devarajan, G. G., Nagarajan, S. M., Amanullah, S. I., Mary, S. S. A., Bashir, A. K. (2023). AI-assisted deep NLP-based approach for prediction of fake news from social media users. *IEEE Transactions on Computational Social Systems,* 1–11. https://doi.org/10.1109/TCSS.2023.3259480
19. Ekanayake, I. U., Herath, D. (2020). Chronic kidney disease prediction using machine learning methods. *2020 Moratuwa Engineering Research Conference (MERCon)*, Moratuwa, Sri Lanka, IEEE.
20. Ullah, Z., Jamjoom, M. (2022). Early detection and diagnosis of chronic kidney disease based on selected predominant features. *Journal of Healthcare Engineering, 2023,* 3553216. https://doi.org/10.1155/2023/3553216
21. Khalid, H., Khan, A., Zahid Khan, M., Mehmood, G., Shuaib Qureshi, M. et al. (2023). Machine learning hybrid model for the prediction of chronic kidney disease. *Computational Intelligence and Neuroscience*, *2023*, 9266889. https://doi.org/10.1155/2023/9266889
22. Abdel-Fattah, M. A., Othman, N. A., Goher, N. (2022). Predicting chronic kidney disease using hybrid machine learning based on apache spark. *Computational Intelligence and Neuroscience*, *2022,* 9898831. https://doi.org/10.1155/2022/9898831
23. Venkatesan, V. K., Ramakrishna, M. T., Izonin, I., Tkachenko, R., Havryliuk, M. (2023). Efficient data preprocessing with ensemble machine learning technique for the early detection of chronic kidney disease. *Applied Sciences, 13(5),* 2885.
24. Kumar, K., Pradeepa, M., Mahdal, M., Verma, S., RajaRao, M. et al. (2023). A deep learning approach for kidney disease recognition and prediction through image processing. *Applied Sciences, 13(6),* 3621.
25. Dritsas, E., Trigka, M. (2022). Machine learning techniques for chronic kidney disease risk prediction. *Big Data and Cognitive Computing, 6(3),* 98.
26. Revathy, S., Bharathi, B., Jeyanthi, P., Ramesh, M. (2019). Chronic kidney disease prediction using machine learning models. *International Journal of Engineering and Advanced Technology, 9(1),* 6364–6367.
27. Chittora, P., Chaurasia, S., Chakrabarti, P., Kumawat, G., Chakrabarti, T. et al. (2021). Prediction of chronic kidney disease–A machine learning perspective. *IEEE Access, 9,* 17312–17334.
28. Swain, D., Mehta, U., Bhatt, A., Patel, H., Patel, K. et al. (2023). A robust chronic kidney disease classifier using machine learning. *Electronics, 12(1),* 212.
29. Almasoud, M., Ward, T. E. (2019). Detection of chronic kidney disease using machine learning algorithms with least number of predictors. *International Journal of Advanced Computer Science and Applications, 10(8)*. https://doi.org/10.14569/IJACSA.2019.0100813
30. Yashfi, S. Y., Islam, M. A., Sakib, N., Islam, T., Shahbaaz, M. et al. (2020). Risk prediction of chronic kidney disease using machine learning algorithms. *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, IEEE.
31. Rady, E. H. A., Anwar, A. S. (2019). Prediction of kidney disease stages using data mining algorithms. *Informatics in Medicine Unlocked, 15,* 100178.
32. Tekale, S., Shingavi, P., Wandhekar, S., Chatorikar, A. (2018). Prediction of chronic kidney disease using machine learning algorithm. *International Journal of Advanced Research in Computer and Communication Engineering, 7(10),* 92–96.

33. Machine Learning University (2015). Chronic_kidney_disease. https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease (accessed on 09/09/2023).
34. Umer, M., Sadiq, S., Missen, M. M. S., Hameed, Z., Aslam, Z. et al. (2021). Scientific papers citation analysis using textual features and smote resampling techniques. *Pattern Recognition Letters, 150,* 250–257.
35. Shafiq, M., Tian, Z., Bashir, A. K., Jolfaei, A., Yu, X. (2020). Data mining and machine learning methods for sustainable smart cities traffic classification: A survey. *Sustainable Cities and Society, 60,* 102177.
36. Juna, A., Umer, M., Sadiq, S., Karamti, H., Eshmawi, A. et al. (2022). Water quality prediction using KNN imputer and multilayer perceptron. *Water, 14(17),* 2592.
37. Breiman, L. (2001). Random forests. *Machine Learning, 45,* 5–32.
38. Iwendi, C., Bashir, A. K., Peshkar, A., Sujatha, R., Chatterjee, J. M. et al. (2020). Covid-19 patient health prediction using boosted random forest algorithm. *Frontiers in Public Health, 8,* 357.
39. Ashraf, I., Narra, M., Umer, M., Majeed, R., Sadiq, S. et al. (2022). A deep learning-based smart framework for cyber-physical and satellite system security threats detection. *Electronics, 11(4),* 667.
40. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29(5),* 1189–1232. http://www.jstor.org/stable/2699986 (accessed on 30/12/2023).
41. Besharati, E., Naderan, M., Namjoo, E. (2019). LR-HIDS: Logistic regression host-based intrusion detection system for cloud environments. *Journal of Ambient Intelligence and Humanized Computing, 10(9),* 3669–3692.
42. Umer, M., Sadiq, S., Nappi, M., Sana, M. U., Ashraf, I. et al. (2022). ETCNN: Extra tree and convolutional neural network-based ensemble model for COVID-19 tweets sentiment classification. *Pattern Recognition Letters, 164,* 224–231.
43. Kumar, V. A., Sharmila, S., Kumar, A., Bashir, A., Rashid, M. et al. (2021). A novel solution for finding postpartum haemorrhage using fuzzy neural techniques. *Neural Computing and Applications, 35,* 1–14. https://doi.org/10.1007/s00521-020-05683-z
44. Geurts, P., Ernst, D., Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning, 63,* 3–42.
45. Sarwat, S., Ullah, N., Sadiq, S., Saleem, R., Umer, M. et al. (2022). Predicting students' academic performance with conditional generative adversarial network and deep SVM. *Sensors, 22(13),* 4834.
46. Zhang, Y. D., Wang, S. H., Yang, X. J., Dong, Z. C., Liu, G. et al. (2015). Pathological brain detection in mri scanning by wavelet packet tsallis entropy and fuzzy support vector machine. *SpringerPlus, 4(1),* 1–16.
47. Manzoor, M., Umer, M., Sadiq, S., Ishaq, A., Ullah, S. et al. (2021). RFCNN: Traffic accident severity prediction based on decision level fusion of machine and deep learning model. *IEEE Access, 9,* 128359–128371.
48. Shahzad Farooq, S., Fiaz, M., Mehmood, I., Kashif Bashir, A., Nawaz, R. et al. (2021). Multi-modal data analysis based game player experience modeling using LSTM-DNN. *Computers, Materials & Continua, 68(3),* 4087–4108. https://doi.org/10.32604/cmc.2021.015612
49. Majeed, R., Abdullah, N. A., Faheem Mushtaq, M., Umer, M., Nappi, M. (2021). Intelligent cyber-security system for IoT-aided drones using voting classifier. *Electronics, 10(23),* 2926.
50. Rish, I. (2001). An empirical study of the naive bayes classifier. *JCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, vol. 3. Seattle, Washington, USA.