Journal of Theoretical and Applied Information Technology
31st March 2024. Vol.102. No 6
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A NEW APPROACH FOR DETECTING KINDS OF CHRONIC KIDNEY DISEASES BASED ON DATA MINING APPROACHES

JAMALALDEEN SALMAN, SARA ELHISHI, SAMIR ABDELRAZEK, HAZEM ELBAKRY
Department of Information Systems, Faculty of Computers and Information, Mansoura University, Egypt
E-mail: Jamalaldeen26@gmail.com, sara_shaker2008@mans.edu.eg, samir.abdelrazek@mans.edu.eg, elbakry@mans.edu.eg

ABSTRACT

Chronic Kidney Disease (CKD) is a potentially fatal condition that can last a person's whole life and is caused by either kidney cancer or impaired kidney function. It is possible to stop or limit the progression of this chronic condition to the point when dialysis or surgical intervention are the only possibilities for saving the life of a patient. Earlier detection and treatment can reduce the likelihood of this occurring. This study investigates how Machine Learning (ML) techniques can be used to detect different kidney disease kinds. ML algorithms have been a driving force in the detection of abnormalities in various physiological data and are being used successfully in various classification tasks. In this study, we employed six different supervised- machine learning algorithms such as Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), K-Nearest-Neighbor (KNN), XGBoost, and Random Forest (RF) in two different experiments for detecting kidney disease types. Each experiment used a different dataset taken from Kaggle (CKD dataset) and Alliance website (kidney disease genes dataset) for building binary and multi classification kidney disease models. SVC and KNN achieved 99.00%, 99.21% accuracy and recall, respectively for first experiment. While KNN and LR achieved 62.22% and 83.33% accuracy and precision, respectively for second experiment with adequate robustness, and our findings imply that KNN can also be used to detect similar diseases.

Keywords: *Machine Learning, chronic kidney disease, KNN, RF, LR, DT, SVM, XGBoost*

1 INTRODUCTION

CKD is a disorder in which the kidneys are so broken that they are unable to filter blood effectively as they should. The primary function of the kidneys is to Eliminate waste and extra water from the circulation, this is precisely how urine is produced. CKD indicates that the body has accumulated waste. This illness is referred to as chronic since the damage occurs gradually with time[1]. You may face a variety of health problems as a result of CKD. Diabetes, heart disease, and hypertension are only three of the more disorders that can contribute to CKD. Along with these major health issues, gender, and age influence who gets CKD [2]. According to current statics CKD affects people at different levels of age around the world. Among people 65 and older, more than 38% are susceptible to CKD. While those aged (45 -64) represent 12.4% of people affected by CKD. Young people represent a small percentage (6%) of getting affected with CKD [3]. For detecting kidney disease, healthcare practitioners use two fundamental ways. Initially, urine and blood tests are done to identify if a person has CKD, then a blood test can evaluate kidney performance (function), commonly called the Glomerular Filtration Rate (GFR). GFR values of 15 to 60 suggest insufficient renal function. Finally, a GFR of 15 or less suggests renal failure. Urinalysis, the second technique, checks for albumin, which can enter the urine if the kidneys are not functioning properly [4]. Early detection is critical for lowering the death rate among those with CKD. Late diagnosis of this illness frequently results in renal failure, necessitating kidney transplantation or dialysis [5]. As for the expanding number of CKD patients, a scarcity of specialists has resulted in expensive diagnostic and treatment expenses. Computer-assisted diagnostic systems, particularly in developing nations, are required to support radiologists and physicians in making diagnostic decisions [6]. In this and such cases, computer-aided diagnostics can play a vital role in determining the disease's prognosis early and efficiently. ML can be used to accurately identify a disease. With the aid of these technologies, clinical decision-makers may classify diseases more precisely. This study suggests two different kidney diseases classification experiments based on six

different machine learning algorithms. Each experiment used a different dataset; the UCI CKD dataset and kidney disease dataset genes and employing supervised-learning techniques such as LR, DT, SVM, KNN, XGBoost, and RF. Several techniques, including missing-value imputation, data cleaning, data balancing using Synthetic Minority Oversampling Technique (SMOTE) technique, and feature scaling, were used to improve the experiment model's scalability. Our proposed work's efficiency according to testing accuracy was compared to other different studies. Our paper contributions are represented as follows:

- ✓ Our study built 2 different experiments; binary and multi classification experiment based on six different ML algorithms and 2 different datasets (UCI CKD, kidney disease genes) for detecting kidney diseases.
- ✓ The study proposed the second experiment using kidney diseases genes dataset that indicates kidney disease type based on patient genes for detecting four different kinds of kidney diseases.
- ✓ The study solved class imbalance problem founded in the two datasets by applying oversampling techniques such as SMOTE.
- ✓ There are no other studies applied kidney disease genes dataset in our second experiment. The study provided better performance in detecting kidney disease types using this dataset.

The remainder of our paper is structured as follows. section 2 represents the related work, section 3 represents models' framework, section 4 represents models evaluation and discussion, and section represents conclusion and future work.

2 RELATED WORK

In the literature, various methodologies, and solutions for the chronic renal disease classification problem have been used and developed. The suggested study incorporates existing knowledge and contributes to improving the already feasible outcomes in the area of predicting chronic renal disease.

[7] developed a ML model to detect CKD using the same our CKD dataset. SVM and RF achieved 99.33%, 98.67% accuracy, respectively. proposed 12 ML models based on UCI CKD dataset. XGBoost classifier achieved the best performance accuracy. It achieved 98.30% accuracy and 98% sensitivity. While [8] proposed deep neural network model based on CKD dataset for detecting kidney disease. The proposed network model achieved 100% accuracy.

[9] proposed decision support system (DSS) to detect Chronic Renal Failure (CRF) based on Artificial Neural Network (ANN), DT, and Naïve Base (NB). DT achieved the best result (92%) accuracy compared to other algorithms. [10] created a decision support system similar to [9] for detecting CRF based on ANN and other various ML algorithms such as DT and NB. KNN achieved the best accuracy (94%). [11]proposed machine learning model for detecting kidney disease from CKD dataset. The proposed model was based on SVM, NB, DT, and KNN. DT achieved the best result with 99.75% accuracy. [12] This study provided a method for detecting CKD successfully by integrating an information-gain-based feature selection method using an AdaBoost classifier that is cost-sensitive. The proposed model achieved 99.8% accuracy and sensitivity. In [13], a machine learning-based healthcare support system for CKD prognosis was launched. The DT algorithm performed the best. The imputation of missing data, however, was inconsistent because it was done without taking the presence of outliers into consideration, and no data balancing was done either. The proposed model achieved 97.5% accuracy. Similarity in [14] health care support system was built based on machine learning algorithms for detecting kidney disease. RF achieved the highest accuracy (99%). Similarly, the imputation was carried out without taking into account outliers in the numerical features, and the training was carried out on the imbalanced dataset. The model was not trained with hyperparameter adjustment, and the resulting models were not cross validated.

Following a review of the literature on the various researchers' contributions to the CKD prediction, the following research needs were identified:

- ✓ During the data preprocessing phase, none of the articles in the literature evaluated the occurrence of outliers in numerical characteristics. As a result, such features' imputed values tend to diverge from the general central trend.
- ✓ Most of the related work built their models based on imbalanced datasets, resulting in biased models.
- ✓ None of researches in the literature used our second dataset (kidney disease genes dataset) for making multi classification model to detect different kidney diseases.
- ✓ The study proposed the second dataset for

detecting what type of kidney diseases the patient has. So, the benefits of the second dataset are not only detecting that the patient has kidney disease or not regardless of its type.

Table 1 shows a comparison between our proposed models and different related work on the same dataset.

Table 1: A comparison between our proposed models and different related work on the same dataset.

Paper	Model	Dataset	Evaluation measures
[7]	SVM RF	Chronic kidney disease	99.33% accuracy
[15]	XGBoost	Chronic kidney disease	98.30% accuracy 98.00% sensitivity
[8]	Deep neural network	Chronic kidney disease	100.00% accuracy
[9]	DT	Chronic renal failure	92.00% accuracy
[10]	KNN	Chronic renal failure (Jordan prince Hamza dataset)	94.oo% accuracy
[11]	DT	Chronic kidney disease	99.75% accuracy
[12]	AdaBoost	Chronic kidney disease	99.8% accuracy 99.8% sensitivity
[13]	DT	Chronic kidney disease	97.5% accuracy
[14]	RF	Chronic kidney disease	99.00% accuracy
The proposed model	SVC KNN	Chronic kidney disease	99.00% accuracy 99.21% recall
The proposed model	KNN Logistic Regression (LR)	Kidney disease genes dataset	61.22% KNN accuracy 83.33% LR precision

3 MODELS' FRAMEWORK

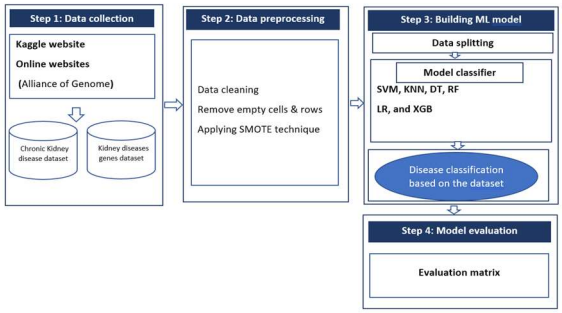


Figure 1: Kidney Diseases Classification Framework Based On Machine Learning Algorithms.

As shows in models' framework, two different experiments were built based on two different datasets: chronic kidney disease and kidney disease genes dataset for detecting whether patient has a kidney disease or not based on the first dataset, and detecting what type of kidney diseases he has based on the second dataset. The framework has four main steps, data collection, data preprocessing, building ML model, and model evaluation. These main steps are illustrated in detail in the following sections.

4 DATA COLLECTION

The study developed two experiments with the same ML algorithms. Each experiment used a specific dataset, the first experiment applied CKD dataset, while the second applied kidney disease genes. This dataset description is highlighted in the next subsections.

A. Chronic Kidney Disease (CKD) Dataset

The CKD dataset was obtained from Kaggle website [16]and includes 400 samples. These 400 samples contain 250 chronic kidney disease (ckd) samples and 150 not chronic kidney disease (notckd) samples. Table 2 illustrates independent variables of CKD dataset.

B. Kidney Disease Genes Dataset

Kidney disease genes dataset collected online from Alliance of Genome website [17]. It contains 242 samples and 6 features (species id, species name, gene id, gene symbol, gene association, disease name). Table 2 provides a description of kidney disease genes dataset.

Journal of Theoretical and Applied Information Technology
31st March 2024. Vol.102. No 6
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2: Features Description Of Chronic Kidney Disease Dataset

Feature name	Feature abbreviation and measures
Age (numerical)	Age in year
Blood pressure (numerical)	Bp in mm/Hg
Specific gravity (nominal)	Sg-(1.055, 1.1010, 1.015, 1.020, 1.025)
Albumin(nominal)	Al-(0,1,2,3,4,5)
Sugar (nominal)	Su -(0,1,2,3,4,5)
Red blood cells (nominal)	Rbc-(normal - abnormal)
Pus cells (nominal)	Pc- (normal- abnormal)
Pus cells clumps (nominal)	Pcc- (present- not present)
Bacterial (nominal)	Ba – (present- not present)
Blood glucose random (numerical)	Bgr in mgs/dl
Blood urea (numerical)	Bu in mgs/dl
Serum creatinine(numerical)	Sc in mgs/dl
Sodium (numerical)	Sod in mEq/L
Potassium (numerical)	Pot in mEq/L
Hemoglobin (numerical)	Hemo in gms
Packed cell volume (numerical)	
White blood cell count (numerical)	Wc in cells /cumm
Red blood cell count (numerical)	Rc in millions/cmm
Hypertension (nominal) htn	Yes, no
Diabetes mellitus (nominal)	Dm – (yes, no)
Appetite (nominal)	Appet -(good, poor)
Coronary Artery disease (nominal)	Cad- (yes, no)
Pedal Edema (nominal)	Pe- (yes, no)
Anemia (nominal)	Ane- (yes, no)

As observed from table 2, chronic kidney diseases dataset contains 24 independent variables nominal or numerical and one dependent class variable [target class] (ckd and notckd).

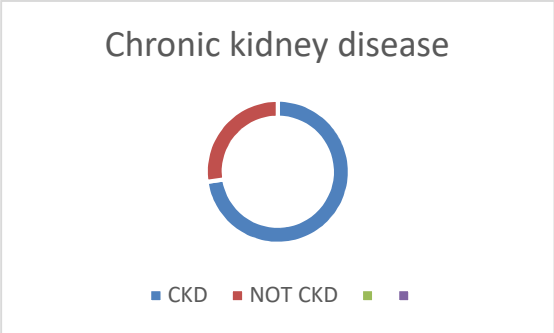
Table 3: Kidney Disease Genes Dataset Samples Description.

Disease name	Clear cell renal cell carcinoma	Acute kidney failure	Chronic kidney disease	Cystic Kidney disease	Total
Number of samples	41	119	70	12	242

As observed from table 3, This dataset classifies kidney disease into 4 classes (clear cell renal cell carcinoma, acute kidney failure, Cystic Kidney disease, and chronic kidney disease) based on the types of genes in patient and whether this gene association (is implicated in or marker for) the kidney disease.

Data Preprocessing

The next step in building our model was the data preprocessing step. One of the most important processes in building a ML model is data



preprocessing. The better the improvement of data preprocessing the better model's performance evaluation. The study model has three main steps data cleaning, encoding features, and solving imbalanced data issues. These three main steps are represented in the next sections.

A. Chronic Kidney Disease (CKD) Dataset

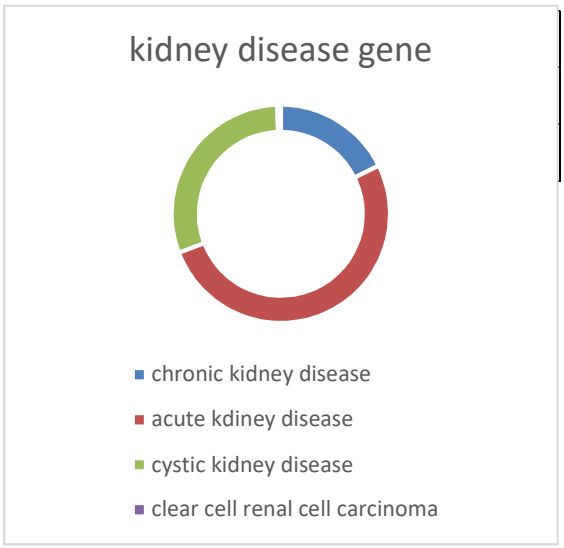
The data cleaning step was a great step for both chronic kidney disease dataset and kidney disease genes dataset. In chronic kidney disease dataset, there are some important phases in data cleaning that are represented as follows.

1. Removing duplicated columns such as id and age columns from the dataset.
2. Dealing with rendering problems. Some values in chronic kidney disease were incorrect such as value (ckd\t) in target class, we replace it with (ckd) value. In dm and cad features, there are some samples that have (t\no) value, we replace it to (no) value.
3. Dealing with missing values. The chronic kidney disease dataset contains samples with missing values, such as age. The study created a mean function to deal with these missing values, this function gives a mean age for missing values.

For kidney disease genes dataset, the dataset has more than 10 features. We selected only 6 features and removed other features such as gene name feature that didn't have any effect on model performance.

B. Encoding Features

As observed from the dataset section, both chronic kidney disease and kidney disease genes datasets have some nominal features and other numerical.



Models of machine learning cannot work well using these categorical data. Feature encoding was a crucial step for converting this data from categorical to numerical for dealing with machine learning algorithms[18]. The technique was used here, was label encoder technique that assign 0, 1…4 to each nominal values[19]. Also, assign numerical variables for target class, o for ckd and 1 for not ckd. For kidney disease genes dataset, assigning 0 For clear cell renal cell carcinoma disease, 1 for acute kidney failure, 2 for cystic kidney disease, and 3 for chronic kidney disease.

C. Solving Imbalanced Dataset

As shown from data collection section and table 3, there was not balanced distribution for target classes in both the 2 datasets, figure 2 illustrates labels distribution for both datasets.

Figure 2: Class label distribution of chronic kidney disease and kidney disease genes datasets.

As shown from figure 2, there is an imbalanced data issue in both datasets. Due to this issue, the majority class dominates the minority class. Thus, the classifiers' performance is unreliable since they frequently fall into the majority class [20]. The study applied synthetic minority oversampling technique (SMOTE). SMOTE is a form of over sampling technique that increases dataset size by generating a new examples of minority class [21]. SMOTE in our study overcame the imbalanced data issue. Table 4 presents number of training data for chronic kidney disease dataset before and after SMOTE.

Table 4: The number of training data for chronic kidney disease dataset before and after SMOTE.

As shown from table 4, Training chronic kidney disease Dataset size before SMOTE contained 187 ckd class and 113 notckd class. While after SMOTE, dataset contained 187 training samples in both classes. For kidney disease genes dataset contained 193 training samples before applying SMOTE. While after applying SMOTE it contained 372 training samples. Table 5 presents number of training data for kidney disease genes dataset before and after SMOTE.

Table 5: The Number Of Training Data For Kidney Disease Genes Dataset Before And After SMOTE.

	Label 0	Label 1	Label 2	Label 3	Total training data
Before SMOTE	93	62	29	9	193
After SMOTE	93	93	93	93	372

5 BUILDING ML MODELS (METHODOLOGY)

The study applied two different experiments with the same ML algorithms and their parameter setup. each experiment used a different dataset. first experiment detects whether patient has kidney disease or not based on CKD dataset regardless of disease type. While the second experiment detects what type of kidney disease patient has based on kidney disease genes dataset. The study applied six traditional machine learning algorithms for each experiment. These algorithms were KNN, DT, RF, SVC, LR, and Xtreme Gradient Boosting (XGB). Each of these algorithms in each experiment was built two times, the first time without applying SMOTE technique and the second time employing the SMOTE technique. Dataset in first experiment was split into 75% training and 25% testing. While kidney disease genes dataset in the second experiment was split into 80% training and 20% testing. The details of each algorithm applied in the study are being discussed in the next sections.

KNN Algorithm

KNN is a supervised ML algorithm, which is used frequently in regression and classification issues [22]. KNN uses no assumptions about the underlying data because it is a non-parametric method. As a result of saving the training dataset rather than instantly learning from it, the method is also referred to as a lazy learner. Instead, it stores the data

throughout the training phase and classifies new data into a category that is quite close to the new data [23].

Our study applied 5 N- Neighbors and 30 leaf size parameters for KNN algorithms. KNN achieved better classification performance for the two experiments whether applying the SMOTE technique or not. The results of the KNN algorithm are represented in Table 6.

DT Algorithm

The DT is recognized as the most effective and efficient ML method, and it is successfully applied to address issues in the real-time AI field. To build DT from the supplied trained set and guarantee the classification of query samples, a variety of approaches were presented. This algorithm simulates a tree in its work. It divides the data into subsets recursively, according to the most important attributes at each node of the tree [24].

DT achieved better classification performance in the two experiments whether applying the SMOTE technique or not. The results of the DT algorithm are represented in Table 7.

LR Algorithm

The LR is used when it is necessary to forecast the absence or presence probabilities of a specific disease, or outcome based on a collection of autonomous explanatory parameters of different types, such as categorical continuous, or discrete parameters[25].

The study employed c=1, penalty = L2, and solver= sag parameters for LR algorithm in building the model. LR achieved better classification performance for the two experiments whether applying the SMOTE technique or not. The results of the LR algorithm are represented in Table 8.

RF Algorithm

The RF is an ensemble classification that refers to a new technique that makes use of several classifiers. The objective is to eliminate feature subset conflicts by incorporating ensemble classification, which is frequently more accurate than other ensembles. All trees in the forest are dependent on the outcomes of individual experiments with arbitrary vector values that have a similar distribution since RF is an integration of tree predictors [26].

The study used 30 N-estimators and 5 min-sample-leaf parameters for RF classifiers. RF achieved better classification performance for the two experiments whether applying the SMOTE technique or not. The results of the RF algorithm are represented in Table 9.

SVC Algorithm

The SVM has become the most widely used data learning technique in recent years. It is typically applied to solve binary pattern categorization issues. In an infinitely dimensional space, the binary SVM generates a collection of hyperplanes that can be divided into linear and non-linear SVM types of representations[27].

The study used C=1.8, kernel= rbf, Gamma= scale, and degree= 1 parameters for SVC algorithm. SVC achieved better classification performance along the two experiments whether applying the SMOTE technique or not. The results of the SVC algorithm are represented in Table 10.

XGB Algorithm

For speed and performance, XGBoost, a gradient-boosted decision tree implementation, was developed. The XGBoost technique has recently dominated Kaggle competitions and applicable machine learning challenges for structured or tabular data. A distributed gradient boosting toolkit called XGBoost has been designed to be rapid and scalable while training machine learning models. This ensemble learning method combines the predictions from several weak models to produce a stronger prediction. Because of the way it handles missing values, real-world data with missing values can be handled with little to no pre-processing.

The study used 100 N-Estimator and 0.1 learning rate parameters for XGB algorithm. XGB achieved better classification performance whether applying the SMOTE technique or not. The results of the XGB algorithm are represented in Table 11.

6 MODELS' EVALUATION AND DISCUSSION

As was mentioned in methodology section, the study employed two different experiments; six different machine learning models were adopted in each experiment for detecting kidney diseases. Two different datasets with different samples and different features were used separately in each experiment. Both datasets had imbalanced data issues, the study solved it by adopting SMOTE technique that increase the numbers of instance in minority class. The study also applied feature

encoding technique for converting categorical data into numerical data since some ML algorithms cannot work well for categorical data as we explained in feature encoding section. The proposed study models achieved better performance for detecting kidney disease than other studies as described in literature reviews section. Which means that our model can give better performance in detecting different kinds of kidney disease with different features. The results of each algorithm for the two experiments whether applying SMOTE technique or not in the following sections.

KNN Algorithm

Table 6: Performance Evaluation Of KNN Algorithm On Chronic Kidney Disease And Kidney Diseases Genes Datasets.

SMOTE				
Dataset	Accuracy	Precision	Recall	F1-Score
Exp1) chronic kidney disease	99.00%	99.22%	98.65%	98.92%
Exp2) kidney disease genes	61.22%	58.33%	58.33%	58.33%
Without SMOTE				
Dataset	Accuracy	Precision	Recall	F1-Score
Exp1) chronic kidney disease	99.00%	99.22%	98.65%	98.92%
Exp2) kidney disease genes	59.18%	66.67%	50.00%	57.14%

DT Algorithm

Table 7: Performance Evaluation Of DT Algorithm On Chronic Kidney Disease And Kidney Diseases Genes Datasets.

SMOTE				
Dataset	Accuracy	Precision	Recall	F1-Score
Exp1) chronic kidney disease	98.00%	98.46%	97.30 %	97.83 %
Exp2) kidney disease genes	51.02%	66.67%	66.67 %	66.67 %
Without SMOTE				
Dataset	Accuracy	Precision	Recall	F1-Score
Exp1) chronic kidney disease	98.00%	98.46%	97.30 %	97.83 %
Exp2) kidney disease genes	61.22%	64.29%	75.00 %	69.23 %

LR Algorithm

Table 8: Performance Evaluation Of LR Algorithm On Chronic Kidney Disease And Kidney Diseases Genes Datasets.

SMOTE				
Dataset	Accuracy	Precision	Recall	F1-Score
Exp1) chronic kidney disease	96.00%	95.12%	96.83 %	95.80 %
Exp2) kidney disease genes	42.86%	55.56%	83.33 %	66.67 %
Without SMOTE				
Dataset	Accuracy	Precision	Recall	F1-Score
Exp1) chronic kidney disease	97.00%	96.25%	97.62 %	96.83 %
Exp2) kidney disease genes	57.14%	66.56%	83.33 %	66.67 %

RF Algorithm

Table 9: Performance Evaluation Of RF Algorithm On Chronic Kidney Disease And Kidney Diseases Genes Datasets.

SMOTE				
Dataset	Accuracy	Precision	Recall	F1-Score
Exp1) chronic kidney disease	99.00%	99.22%	98.65 %	98.92 %

Exp2) kidney disease genes	53.06%	61.54%	66.67 %	64.00 %
Without SMOTE				
Dataset	Accuracy	Precision	Recall	F1-Score
Exp1) chronic kidney disease	99.00%	99.22%	98.65 %	98.92 %
Exp2) kidney disease genes	57.14%	60.00%	75.00 %	66.67 %

SVC Algorithm

Table 10: Performance Evaluation Of SVC Algorithm On Chronic Kidney Disease And Kidney Diseases Genes Datasets.

SMOTE				
Dataset	Accuracy	Precision	Recall	F1-Score
Exp1) chronic kidney disease	99.00%	98.68%	99.21 %	98.93 %
Exp2) kidney disease genes	53.06%	50.00%	50.00 %	50.00 %
Without SMOTE				
Dataset	Accuracy	Precision	Recall	F1-Score
Exp1) chronic kidney disease	99.00%	98.68%	99.21 %	98.93 %
Exp2) kidney disease genes	53.06%	50.00%	50.00 %	50.00 %

XGB Algorithm

Table 11: Performance Evaluation Of XGB Algorithm On Chronic Kidney Disease And Kidney Diseases Genes Datasets.

SMOTE				
Dataset	Accuracy	Precision	Recall	F1-Score
Exp1) chronic kidney disease	98.00%	97.85%	97.85 %	97.85 %
Exp2) kidney disease genes	55.10%	61.64%	66.67 %	64.00 %
Without SMOTE				
Dataset	Accuracy	Precision	Recall	F1-Score
Exp1) chronic kidney disease	97.00%	97.05%	96.50 %	96.76 %
Exp2) kidney disease genes	61.22%	69.23%	75.00 %	72.00 %

As shown from the previous results, machine learning algorithms achieved better performance for classifying kidney diseases into two classes (CKD, NOT-CKD). For chronic kidney disease dataset, KNN, and RF, achieved the best performance for each of accuracy and precision whether applying SMOTE technique or not. They achieved 99.00% accuracy and 99.22% precision. But SVC has the superior performance according to Recall and F1-Score. It achieved 99.21% a Recall and 98.93% accuracy whether SMOTE or not. For second experiment using kidney disease genes dataset, machine learning algorithms had the ability of classifying kidney diseases into four groups (clear cell renal cell carcinoma, acute kidney failure, Cystic Kidney disease, and chronic kidney disease). According to applying SMOTE technique also, KNN achieved the best performance. It achieved 61.22% accuracy. While LR had the better performance for both Recall and F1-Score, it achieved 83.33%, and 66.67%, respectively. DT achieved the higher precision; it achieved 66.67% precision. For second experiment using kidney disease genes dataset without applying SMOTE technique, XGB model achieved the best performance for each of accuracy, precision, and F1-Score. It achieved 61.22%, 69.23, 72.00%, accuracy, precision, and F1-Score,

respectively. While LR achieved the higher precision; it achieved 83.33% Recall. Since SMOTE resolve data imbalanced issues and gives more accurate performance, the study concentrated more on models results after applying SMOTE technique for solving data imbalance issue. Figure 3 and Figure 4 represent evaluation measures of ML models after applying SMOTE for both datasets.

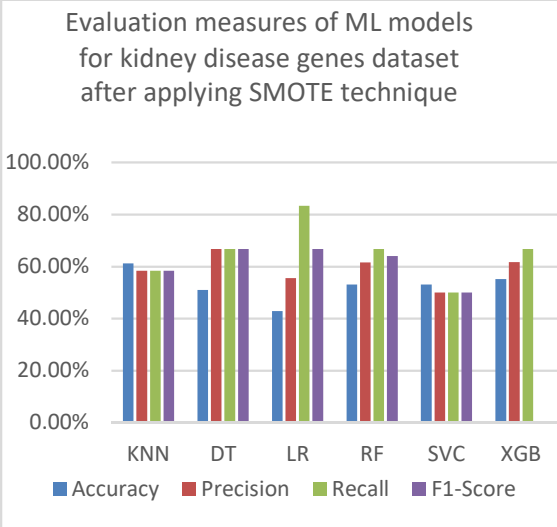


Figure 3: Comparison Of Evaluation Measures Of ML Models For Kidney Disease Genes Dataset After Applying SMOTE Technique.

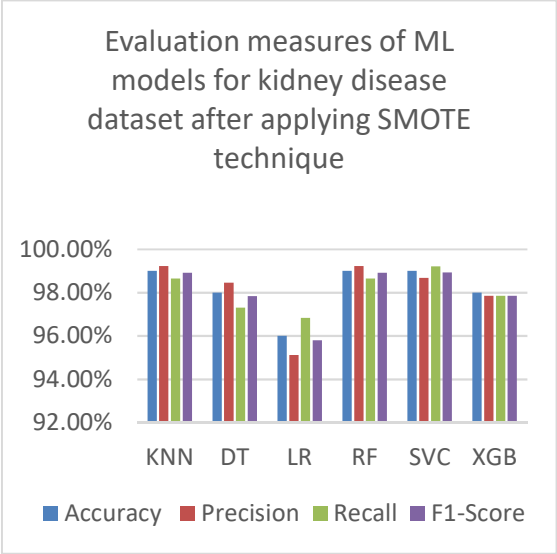


Figure 4: Comparison Of Evaluation Measures Of ML Models For Chronic Kidney Disease Dataset After Applying SMOTE Technique.

Our paper contributions are represented as follows:

Our study built 2 different experiments; binary and multi classification experiment based on six different ML algorithms and 2 different datasets (UCI CKD, kidney disease genes) for detecting kidney diseases. The study proposed the second experiment using kidney diseases genes dataset that indicates kidney disease type based on patient genes for detecting four different kinds of kidney diseases. The study solved class imbalance problem founded in the two datasets by applying oversampling techniques such as SMOTE. There are no other studies applied kidney disease genes dataset in our second experiment. The study provided better performance in detecting kidney disease types using this dataset.

7 CONCLUSION AND FUTURE WORK

This work began with a thorough investigation of the performance of several approaches for detecting kidney diseases. Following this analysis, proper data preprocessing processes were taken to deal with issues in the CKD and kidney disease genes datasets, such as imbalanced data, missing values, feature encoding, and the existence of outliers. The main novelty of the study is applying over sampling technique such as SMOTE for solving data imbalance issue and building six different machine learning models based on two different experiments with two different datasets. The first dataset is UCI CKD dataset that indicates whether patient has a chronic kidney disease or not. While the other dataset is kidney disease genes dataset that indicated different kinds of kidney disease based on patient genes. So, the study model can indicate what type of kidney diseases a patient has. The study models achieved better performance in detecting kidney disease types with different features. But the study has limitations such as the small sample size of both datasets. And the limited kinds of kidney datasets. So, our future work is increasing dataset size and applying deep learning models for classifying different kinds of kidney diseases, and achieving better performance.

REFERENCES

[1] Z. Chen, X. Zhang, and Z. Zhang, "Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models," *Int. Urol. Nephrol.*, vol. 48, no. 12, pp. 2069–2075, Dec. 2016, doi: 10.1007/S11255-016-1346-4/FIGURES/1.

[2] A. Subas, E. Alickovic, and J. Kevric, "Diagnosis of chronic kidney disease by using random forest," *IFMBE Proc.*, vol. 62, pp. 589–594, 2017, doi: 10.1007/978-981-10-4166-2_89/COVER.

[3] "Chronic Kidney Disease in the United States, 2023." https://www.cdc.gov/kidneydisease/publications-resources/CKD-national-facts.html (accessed Jun. 22, 2023).

[4] "Chronic Kidney Disease (CKD) - NIDDK." https://www.niddk.nih.gov/health-information/kidney-disease/chronic-kidney-disease-ckd (accessed Jun. 22, 2023).

[5] G. G. Garcia, P. Harden, and J. Chapman, "The global role of kidney transplantation," *Kidney Blood Press. Res.*, vol. 35, no. 5, pp. 299–304, 2012, doi: 10.1159/000337044.

[6] E. M. Senan *et al.*, "Diagnosis of Chronic Kidney Disease Using Effective Classification Algorithms and Recursive Feature Elimination Techniques," *J. Healthc. Eng.*, vol. 2021, 2021, doi: 10.1155/2021/1004767.

[7] D. Swain *et al.*, "A Robust Chronic Kidney Disease Classifier Using Machine Learning," *Electron. 2023, Vol. 12, Page 212*, vol. 12, no. 1, p. 212, Jan. 2023, doi: 10.3390/ELECTRONICS12010212.

[8] V. Singh, V. K. Asari, and R. Rajasekaran, "A Deep Neural Network for Early Detection and Prediction of Chronic Kidney Disease," *Diagnostics*, vol. 12, no. 1, pp. 1–22, 2022, doi: 10.3390/diagnostics12010116.

[9] A. Y. Al-Hyari, A. M. Al-Taee, and M. A. Al-Taee, "Clinical decision support system for diagnosis and management of Chronic Renal Failure," *2013 IEEE Jordan Conf. Appl. Electr. Eng. Comput. Technol. AEECT 2013*, 2013, doi: 10.1109/AEECT.2013.6716440.

[10] R. Ani, G. Sasi, U. R. Sankar, and O. S. Deepa, "Decision support system for diagnosis and prediction of chronic renal failure using random subspace classification," *2016 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2016*, pp. 1287–1292, Nov. 2016, doi: 10.1109/ICACCI.2016.7732224.

[11] N. Tazin, S. A. Sabab, and M. T. Chowdhury, "Diagnosis of chronic kidney Disease using effective classification and feature selection technique," *1st Int. Conf. Med. Eng. Heal. Informatics Technol. MediTec 2016*, Jan. 2017, doi: 10.1109/MEDITEC.2016.7835365.

[12] S. A. Ebiaredoh-Mienye, T. G. Swart, E. Esenogho, and I. D. Mienye, "A Machine Learning Method with Filter-Based Feature Selection for Improved Prediction of Chronic Kidney Disease," *Bioengineering*, vol. 9, no. 8, 2022, doi: 10.3390/bioengineering9080350.

[13] "(5) (PDF) Increasing Accuracy of C4.5 Algorithm by Applying Discretization and Correlation-based Feature Selection for Chronic Kidney Disease Diagnosis." https://www.researchgate.net/publication/339972964_Increasing_Accuracy_of_C45_Algorithm_by_Applying_Discretization_and_Correlation-based_Feature_Selection_for_Chronic_Kidney_Disease_Diagnosis (accessed Jun. 22, 2023).

[14] S. Shankar, S. Verma, S. Elavarthy, T. Kiran, and P. Ghuli, "Analysis and Prediction of Chronic Kidney Disease," *Int. Res. J. Eng. Technol.*, no. May, pp. 4536–4541, 2020, [Online]. Available: www.irjet.net

[15] M. A. Islam, M. Z. H. Majumder, and M. A. Hussein, "Chronic kidney disease prediction based on machine learning algorithms," *J. Pathol. Inform.*, vol. 14, p. 100189, Jan. 2023, doi: 10.1016/J.JPI.2023.100189.

[16] "Chronic KIdney Disease dataset | Kaggle." https://www.kaggle.com/datasets/mansoordaku/ckdisease (accessed Apr. 17, 2023).

[17] "kidney disease | Alliance of Genome Resources." https://www.alliancegenome.org/disease/DOID:557#associated-genes (accessed Apr. 17, 2023).

[18] H. Nugroho, N. P. Utama, and K. Surendro, "Smoothing target encoding and class center-based firefly algorithm for handling missing values in categorical variable," *J. Big Data*, vol. 10, no. 1, pp. 1–18, Dec. 2023, doi: 10.1186/S40537-022-00679-Z/FIGURES/15.

[19] D. Jiang, W. Lin, and N. Raghavan, "A novel framework for semiconductor manufacturing final test yield classification using machine learning techniques," *IEEE Access*, vol. 8, pp. 197885–197895, 2020, doi: 10.1109/ACCESS.2020.3034680.

[20] R. Ghorbani and R. Ghousi, "Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques," *IEEE Access*, vol. 8, pp. 67899–67911, 2020, doi: 10.1109/ACCESS.2020.2986809.

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/JAIR.953.

[22] R. Devika, S. V. Avilala, and V. Subramaniyaswamy, "Comparative study of classifier for chronic kidney disease prediction using naive bayes, KNN and random forest," *Proc. 3rd Int. Conf. Comput. Methodol. Commun. ICCMC 2019*, pp. 679–684, Mar. 2019, doi: 10.1109/ICCMC.2019.8819654.

[23] "View of KNN based Detection and Diagnosis of Chronic Kidney Disease." http://annalsofrscb.ro/index.php/journal/article/view/2828/2352 (accessed Apr. 27, 2023).

[24] A. Trabelsi, Z. Elouedi, and E. Lefevre, "Decision tree classifiers for evidential attribute values and class labels," *Fuzzy Sets Syst.*, vol. 366, pp. 46–62, 2019, doi: 10.1016/j.fss.2018.11.006.

[25] G. Antonogeorgos, … D. P.-I. journal of, and undefined 2009, "Logistic regression and linear discriminant analyses in evaluating factors associated with asthma prevalence among 10-to 12-years-old children: divergence," *hindawi.com*, Accessed: Apr. 27, 2023. [Online]. Available: https://www.hindawi.com/journals/ijpedi/2009/952042/

[26] S. Tian, X. Zhang, J. Tian, and Q. Sun, "Random forest classification of wetland landcovers from multi-sensor data in the arid region of Xinjiang, China," *Remote Sens.*, vol. 8, no. 11, pp. 1–14, 2016, doi: 10.3390/rs8110954.

[27] M. Gokiladevi, S. Santhoshkumar, and V. Varadarajan, "MACHINE LEARNING ALGORITHM SELECTION FOR CHRONIC KIDNEY DISEASE DIAGNOSIS AND CLASSIFICATION," *Malaysian J. Comput. Sci.*, vol. 2022, no. Special Issue 1, pp. 102–115, Mar. 2022, doi: 10.22452/MJCS.SP2022NO1.8.