Automated Chronic Kidney Disease Detection Model with Knearest Neighbor

Tsehay Admassu Assegie
Department of Computer Science, Aksum University
Axum, Ethiopia
Email: tsehayadmassu2006 [AT] gmail.com

***Abstract*— chronic kidney disease is one of the most common disease in the world today. Kidney disease causes death if the patient is not threated at early stage. One of the challenge in kidney disease treatment is accurate identification of kidney disease at an early stage. Moreover, detecting kidney disease requires experienced nephrologist. However, in developing nations lack of medical specialist or nephrologist for identifying chronic kidney disease makes the problem more challenging. As alternative solution to kidney disease identification, researchers have developed many intelligent models using K-nearest Neighbors (KNN) algorithm. However, the accuracy of the existing KNN model has scope for improvement. Thus, this study proposed KNN based model for accurate identification of kidney disease at early stage. To develop optimized KNN model, we have employed error plot to find most favorable K value to obtain more accurate result than the existing models. To conduct experiments, study employed kidney disease dataset collected form publically available Kaggle data repository for training and testing the proposed model. Finally, we have evaluated the proposed model against predictive accuracy. The experimental result on the proposed model appears to prove that the predictive accuracy of the model is 99.86%.**

Keywords. Chronic kidney disease, Kidney disease; kidney disease detection; KNN; machine learning.

I. INTRODUCTION

In recent years, kidney disease has become one of the major health issues in the world [1]. Chronic kidney disease has become major health issue due to the wide spreading nature of the disease and the increased number of patients' suffering from chronic kidney disease. One of the problems with chronic kidney disease is that most of the patients' dose not notice until they get worthy affected. Hence, identification of chronic kidney disease in an early stage is crucial to avoid the complications at a severe stage and ultimately save human life.

Due to the challenge in identification of chronic kidney disease, the disease has become one of the world's deadliest diseases. The report shows that there are roughly 2.5 to 11.25 million cases chronic kidney disease worldwide [2-13]. Different machine learning algorithms such as support vector machine (SVM) [2] and boosting classifiers [4] are applied to kidney disease data repository to create predictive model with acceptable level of accuracy to identify chronic kidney disease as early as possible. However, chronic kidney disease identification is remained challenging task as the existing method need much of research effort for accurate identification of chronic kidney disease.

In recent years, machine-learning model is employed in the automation of disease diagnosis and disease identification process of chronic kidney disease in the healthcare centers. Most of the researchers applied supervised machine-learning algorithm to develop a model for identifying chronic kidney disease [5-6].

In [7-8], support vector machine (SVM) based chronic kidney transplant rejection prediction model is developed. A preliminary literature review shows that machine-learning model is significantly important to chronic kidney disease detection and promising performance is obtained. Machine-learning model is important for reducing mortality rate caused by chronic kidney disease by assisting in chronic kidney disease detection. Thus, this study is aimed at answering the following research questions:

1) How to determine an appropriate or optimal Kvalue to develop optimal KNN based model for chronic kindye disease detection?
2) What is the accuracy of KNN model on chronic kidney disease detection?

II. RELATED WORK

Many researches has been conducted on the problem of kidney disease detection using machine-learning model. In [9] the researchers conducted comparative study on the performance of machine learning model on chronic kidney disease detection. The researchers compared the performance of Artificial Neural Network (ANN) and support vector machine (SVM) on chronic kidney disease detection. The study showed that ANN has better predictive accuracy as compared to SVM model.

In another study [10], the researchers employed Naïve Bayes algorithm to implement automated intelligent decision support model for chronic kidney detection. The researchers used chronic kidney disease dataset collected from University of California Irvine (UCI) and conducted experiment on Naïve Bayes model. The evaluation on the performance of the developed model on chronic kidney disease detection shows that better predictive accuracy and promising result is obtained.

In another study [11], a model for chronic kidney disease detection is developed using an ensemble method. The model is implemented using decision tree algorithm and J48 supervised learning algorithm as base classifier and adaptive boosting as ensemble classifier. The researchers evaluated the implemented model and result shows that adaptive boosting performed well as compared to decision tree algorithm in terms of accuracy on UCI chronic kidney disease dataset.

A chronic kidney detection model using decision tree model is developed [12]. Experiment on the implemented model shows that decision tree model has 93% accuracy on chronic kidney disease detection. Moreover, the model is found to be helpful in reducing complications resulting from the kidney disease due better performance on chronic kidney disease detection. Researchers have also conducted another study on the problem of kidney disease identification by applying adaptive boosting algorithm [13]. The researchers developed a model using adaptive boosting and decision tree algorithm. The implemented model is evaluated and result shows that ensemble model with adaptive boosting algorithm performs better as compared to the individual decision tree model. Overall, a predictive accuracy of 92.76% is achieved on chronic kidney disease detection using the developed model.

III. RESEARCH METHODOLOGY

The dataset used to implement a model for chronic kidney disease detection is collected from Kaggle data repository. The number of observations in the dataset is 400. The dataset consists of 250 observations of non-patient and 150 patient observations. The distribution among the classes in the dataset is demonstrated in figure 1. The dataset has 17 input features demonstrated in table 1 and a target or an output variable. To implement KNN model we have employed Python programming language with Jupyter Notebook environment.

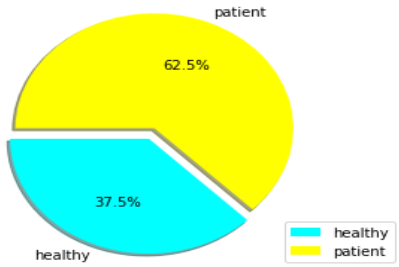

Figure 1. Kidney disease dataset class distribution

As demonstrated in figure 1, the class distribution of the chronic kidney disease positive (37.5%) and chronic kidney disease negative (67.5%) in the dataset.

TABLE I. : KIDNEY DISEASE DATASET FEATURE DESCRIPTION

Feature	**Decription**
Age	Age in years (continuous value)
Blood pressure (BP)	The blood pressure (Numeric)
Specific gravity (sg)	Numeric (continuous value)
Albumin (alb)	Numeric (continuous value 0 to 5)
Sugar (su)	Numeric (continuous value 0 to 5)
Blood glucose random (bgr)	Numeric (continuous value)
Blood urea (bu)	Numeric (continuous value)
Potassium (pot)	Numeric (continuous value)
Sodium (sod)	Numeric (continuous value)
Bacteria present (bac)	Nominal (1=Present, 0=Absent)
Pus cell (pc)	Nominal (1=Ab Normal, 0=Normal)
Pus cell clumps (pcc)	Nominal (1=Present, 0=Absent)
Haemoglobin (hemo)	Numeric (continuous value)
Serum creatinine (sc)	Numeric (continuous value)
Red blood cell count (rbc)	Nominal (1=Ab Normal, 0=Normal)
Hypertension (htn)	Nominal (1=Present, 0=Absent)
Appetite (apt)	Nominal (1=Poor, 0=Good)
Anaemia (an)	Nominal (1=Present, 0=Absent)

A. Corrrelation model

To investigate the relationship among chronic kidney disease dataset input feature and the class label or target variable, we have employed Pearson's correlation analysis. With Pearson's correlation, we have identified strongly correlated feature to chronic kidney disease input feature and target variable as demonstrated in figure 2. Pearson's correlation shown in figure 2 demonstrates the relationship between input feature and the target variable. As shown in figure 2, features such as haemoglobin (hemo), specific gravity (sg), pus cell (pc_normal), pus clumps (pcc_present), red blood cell count (rbc_normal) and sodium (sod) strongly correlated to the target feature. However, input features such as blood urea (bu), blood glucose random (bgc) and appetite (appet_poor) is negatively correlated to the target variable.

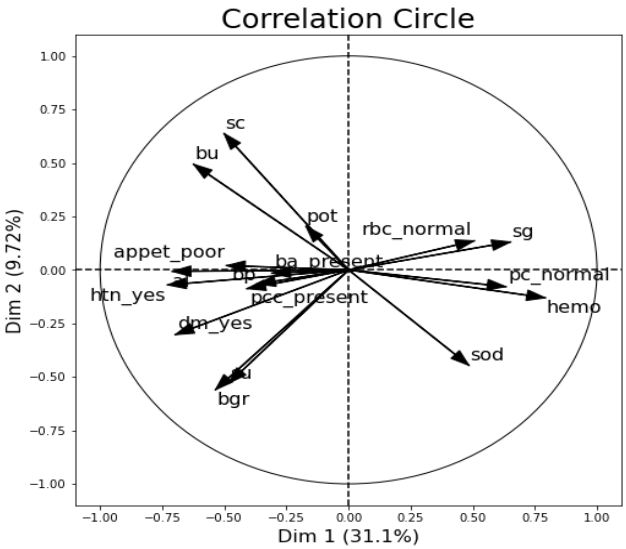

Figure 2. Kidney disease feature correlation

IV. EXPERIMENTAL RESULT AND DISCUSSIONS

In this section predictive accuracy, confusion matrix and learning curve is used to analyze the performance of the KNN model on the test set.

A. Predictive accuracy of the proposed model

The accuracy of KNN model on chronic kidney disease prediction is evaluated on test set. The K values that gives the optimum prediction accuracy is determined using error plot against various K values as demonstrated in figure 3. The optimum K value that yields the maximum possible accuracy on chronic kidney disease detection the KNN model is shown in figure 3. As shown in figure 3, the error rate varies from 0.13 to 0.26 for K values in the range 0 to 40. Thus, error rate plot against K value is helpful to visualize the optimal K value for optimization of KNN model. As shown in figure 3, the proportion of misclassification is low when the K value is 6. An accuracy of 99.86 is achieved with K value of 6 (K=6).

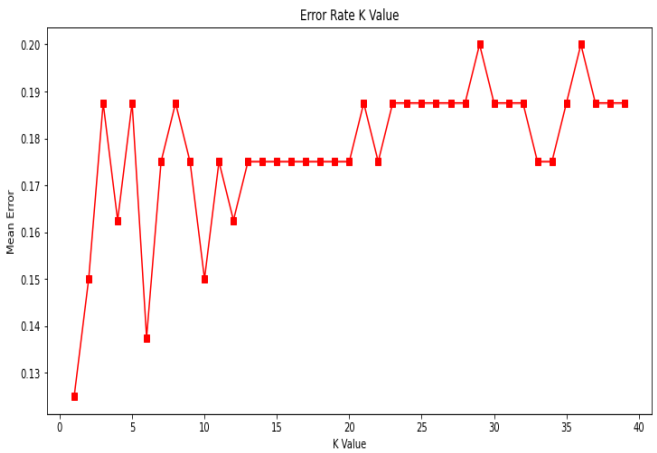

Figure 3. K value vs error rate

We employed learning curve for analyzing the KNN model on chronic kidney disease detection. The learning curve of KNN model shows that the training and test converges together as we observe in figure 4. Thus, we conclude that the performance of the KNN model tends to improves much with larger number of training samples. Moreover, the learning curve for KNN, demonstrated in figure 4, shows that the model is not suffering from overfitting or under fitting as both the training and test accuracy tends to increase with an increase in the number of observations.

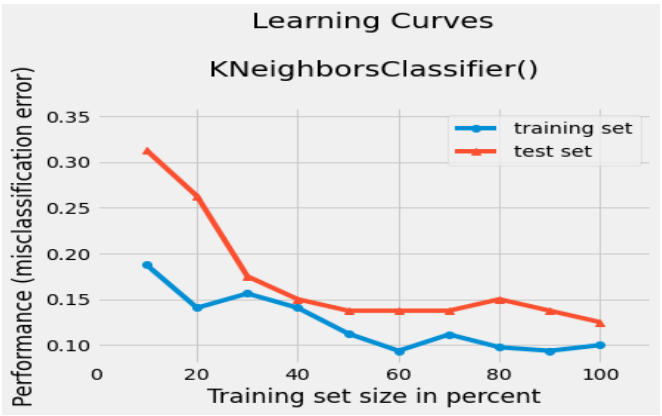

Figure 4. Learning curve for KNN model

B. Cummulative gain curve

In addition to learning curve and accuracy, we have employed cumulative gain curve for analyzing the KNN model on chronic kidney disease detection. The cumulative gain curve of the KNN model is demonstrated in figure 5. The cumulative gain curve shows the actual class and predicted score. The cumulative gain curve shows the predictive positive rate against the true positive rate or sensitivity of KNN Model. As we observe from figure 5, the KNN model detects positive class (class 1) with better accuracy as compared to the negative class (class 0).

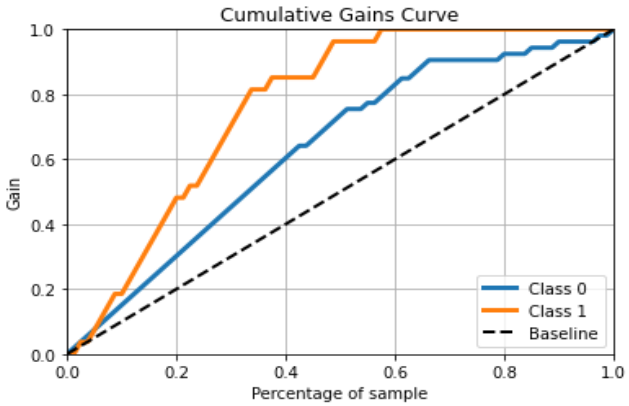

Figure 5. Cummulative gain curve

As cumulative gain curve does not show the number of incorrect and correct predictions made by the KNN model, we

have employed confusion matrix to quantify the performance of KNN model. As shown in figure 6. As shown in figure 6, the model misclassified 10 observations and correctly detected 70 observations. Thus, KNN model performs well on chronic kidney disease detection. In the experiment, we have employed 80% of the dataset or 320 observations for training and 20% of the dataset or 80 observations for testing.

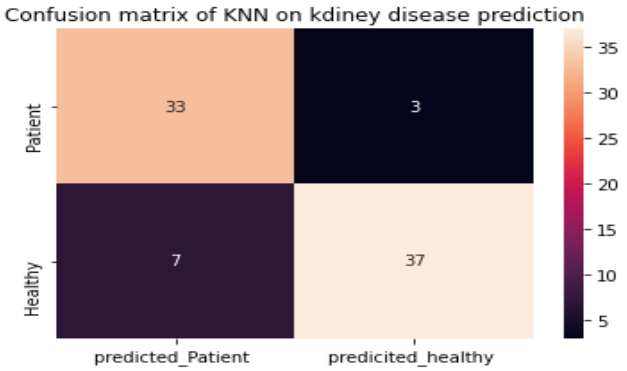

Figure 6. Confusion matrix

In addition to confusion matrix, we have employed receiver operating characteristic curve (ROC) to evaluate the performance of KNN model. The receiver operating characteristic curve for the proposed model is demonstrated in figure 7. As demonstrated in figure 7, the area under curve is 0.92 for both classes (negative and positive class or class 0 and class 1 respectively). Hence, the model performs well on chronic kidney disease detection.

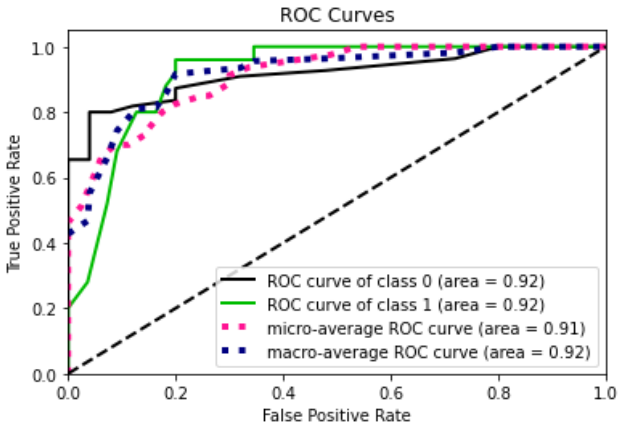

Figure 7. Reciver oprating chracteritic curve (ROC)

V. COMPARATIVE STUDY

The exisitn methods are compared with the KNN model implemented in this study. We have compared the KNN model with the recently published chonic kidney disease detection model developed with different machine learning algorithms. Table 2, shows comparison of the existing methods and the developed KNN model.

TABLE II. : COMPARTIVE STUDY

Autho	**Algorihtm**	**Accuracy**
[12]	Decision tree	93%
[13]	Adaptive boosting	92.76%
[14]	Naive Bayes	76.8%
[15]	Random forest	99.3%
[16]	ANN	95.3%
Proposed	KNN	99.84%

VI. CONCLSUION

In this study, we have proposed a machine-learning model for predicting chronic kidney disease by employing K-Nearest Neighbor (KNN) algorithm. Moreover, we have analyzed the predictive performance of KNN model on chronic kidney disease detection. The KNN model is optimized using error plot to visualize and determine the optimal K value. With error plot, we have obtained good result on chronic kidney disease detection using KNN model. The performance of the proposed model is evaluated using accuracy, cumulative gain curve and confusion matrix as performance measure. Experiment shows that the model has better accuracy on detection of chronic kidney disease. Overall, an accuracy of 99.86% is achieved using the KNN model. As a future work, we recommend researchers to extend this work with other optimization method to improve the accuracy of KNN model using other datasets and optimization approaches.

REFERENCES

[1] Xun Li, Yao Wang, Chengxuan Wang, Sanqing Hu, Ying Xu, Fei Han, Jianghua Chen, Prediction of Renal Transplant Rejection and Acute Tubular Necrosis in Renal Transplant Based on SVM, 2012 5th International Conference on BioMedical Engineering and Informatics.

[2] Njoud Abdullah Almansour, Hajra Fahim Syed, Nuha Radwan Khayat, Rawan Kanaan Altheeb, Renad Emad Juri, Jamal Alhiyafi, Saleh Alrashed, Sunday O. Olatunji, Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study, Computers in Biology and Medicine 109 (2019).

[3] Merve Doğruyol Başar, Aydın Akan, Chronic Kidney Disease Prediction with Reduced Individual Classifiers, Electrical (2018).

[4] Arif-Ul-Islam, Shamim H Ripon, Rule Induction and Prediction of Chronic Kidney Disease Using Boosting Classifiers, Ant-Miner and J48 Decision Tree, International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, IEEE, (2019)..

[5] Vikas Chaurasia, Saurabh Pal, B.B. Tiwari, Chronic Kidney Disease: A Predictive model using Decision Tree, International Journal of Engineering Research and Technology. ISSN 0974-3154 Volume 11, Number 11 (2018).

[6] Komal Kumar N, R. Lakshmi Tulasi, Vigneswari D, An ensemble multi-model technique for predicting chronic kidney disease, International Journal of Electrical and Computer Engineering (IJECE) Vol. 9, No. 2, April, (2019).

[7] Assegie, T.A, Sushma S.J, A Support Vector Machine and Decision Tree Based Breast Cancer Prediction, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume9 Issue-3, February, (2020).

[8] Assegie, T.A, Nair, P.S, The Performance Of Different Machine Learning Models On Diabetes Prediction, International Journal Of Scientific & Technology Research Volume 9, Issue 01, January (2020).

[9] Assegie, T.A, Sushma S.J, Prasanna Kumar S.C, Weighted Decision Tree Model for Breast Cancer Detection, Technology Reports of Knsai University, Volume 62, Issue 03, January (2020).

[10] Dilip Singh Sisodia, Akanksha Verma, Prediction Performance of Individual and Ensemble learners for Chronic Kidney Disease, Proceedings of the International Conference on Inventive Computing and Informatics, IEEE, (2017).

[11] Assegie, T.A, R. Lakshmi Tulasi, N. Komal Kumar, Breast cancer prediction model with decision tree and adaptive boosting, IAES International Journal of Artificial Intelligence (IJ-AI) Vol. 10, No. 1, March 2021.

[12] Assegie, T.A, Support Vector Machine And K-Nearest Neighbor Based Liver Disease Classification Model, Indonesian Journal of Electronics, Electromedical, and Medical Informatics (IJEEEMI) Vol. 3, No. 1, February 2021.

[13] Assegie, T.A, An optimized K-Nearest Neighbor based breast cancer detection, Journal of Robotics and Control (JRC) Volume 2, Issue 3, May 2020.

[14] Kashi Sai Prasad, N. Chandra Sekhar Reddy, B. N. Puneeth, A Framework for Diagnosing Kidney Disease in Diabetes Patients Using Classifcation Algorithms, SN Computer Science, 2020.

[15] Asif Salekin, John Stankovic, Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes, IEEE, 2016.

[16] Ravinra BV, N. Sriraam, M. Geetha, Chronic kdiney detection using back propagation neural network classsiferir, IEEE, 2019.