Chronic Kidney Disease Classification Using Machine Learning Classifier

K. Navaz[1],Shaik Suhail[2], P. Shobha[3], L. Sudharshan Reddy[4], A. Yuva Teja[5]
1,2,3,4,5Department of CSE, Annamacharya Institute of Technology & Sciences, Tirupati-517520

Abstract
Chronic kidney disease is one of the major health issues (CKD). Poor dietary habits, low water consumption, and a lack of health awareness are all contributing to its daily rise. A disorder known as chronic kidney disease (CKD) occurs when the kidneys sustain damage over time from a number of factors, leading to a gradual and irreversible loss of kidney function. The health sector has been significantly impacted by technological developments like machine learning (ML), which allows for more precise diagnosis and effective treatment of many chronic conditions. This study investigates several ML methods to forecast kidney disease. The goal is to determine which machine learning classifier would be the most effective one for CKD detection and prediction.

Keywords: Chronic Kidney Disease, Technological developments, Machine Learning, Kidney

1. Introduction

Chronic kidney disease is among the leading causes of death globally. A recent medical report states that approximately 324 million people suffer from CKD globally [1]. The glomerular filtration rate (GFR) is a widely used CKD screening test [2]. Though CKD affects people worldwide, it is more prevalent in developing countries [3]. Meanwhile, early detection is vital in reducing the progression of CKD. However, people from developing countries have not benefitted from early-stage CKD screening due to the cost of diagnosing the disease and limited healthcare infrastructure. While the global prevalence of CKD is reported to be 13.4% [4], it is said to have a 13.9% prevalence in Sub-Saharan Africa [5,6]. Another study reported a 16% pooled prevalence of CKD in West Africa [7], the highest in Africa. Numerous research works have specified that CKD is more prevalent in developing countries [8]. Notably, it is reported that 1 out of every 10 persons suffers from CKD in South Asia, including Pakistan, India, Bhutan, Bangladesh, and Nepal [3].

Therefore, several researchers have proposed machine learning (ML)-based methods for the early detection of CKD. These ML methods could provide effective, convenient, and low-cost computer-aided CKD diagnosis systems to enable early detection and intervention, especially in developing countries. Researchers have proposed different methods to detect CKD effectively using the CKD dataset [9] available at the University of California, Irvine (UCI) machine learning repository. For example, Qin et al. [9] proposed an ML approach for the early detection of CKD. The approach involved using the k-Nearest Neighbours (KNN) imputation to handle the missing values in the dataset. After filling the missing data, six ML classifiers were trained and tested with the pre-processed data. The classifiers include logistic regression, SVM, random forest, KNN, naïve Bayes, and a feed-forward neural network. Due to the misclassification of these classifiers, the authors developed an integrated classifier that uses a perceptron to combine the random forest and logistic regression classifiers, which produced an enhanced accuracy of 99.83%.

2. Related Work

2.1 Material and Methods:

2.1.1 Dataset:

This study uses a dataset of medical test results and records from 400 patients, which includes 250 patients with chronic kidney disease (CKD) and 150 patients without CKD. The dataset was prepared in 2015 by Apollo Hospitals, Tamil Nadu, India, and is publicly available at the University of California, Irvine (UCI) machine learning repository. The dataset has 24 factors, such as blood pressure and white blood cell count, and a classification variable indicating whether the patient has CKD or not. It's important to note that the dataset is imbalanced, with more patients having CKD than not.

2.2 Information Gain:

Feature selection is an important step in building accurate predictive models. By removing attributes that are less useful or unrelated to the target variable, the computational cost is reduced, and the model's performance is improved. In this study, the information gain (IG) technique is used for feature selection. IG is a filter-based method that evaluates the ability of predictor variables to classify the dependent variable

$$IG(X|Y) = H(X) - H(X|Y),$$
$$H(X) = -\sum_{x\in X} P(x) log_2(x),$$
$$H(X|Y) = -\sum_{x\in X} P(x) \sum_{y\in Y} P(x|y) log_2\big(P(x|y)\big),$$

3. Architecture

3.1 Dataset

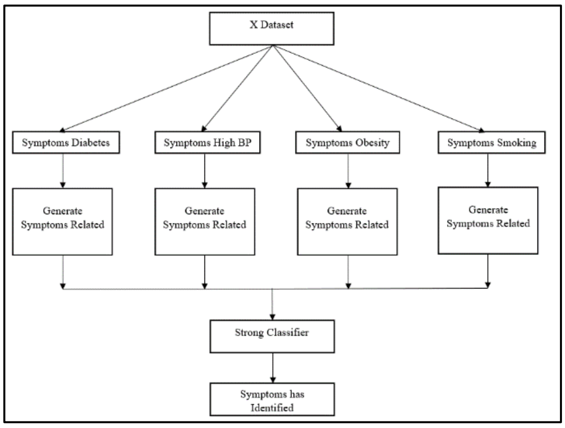

Figure 1: The architecture of the Simple Proposed System

This study uses a dataset of medical test results and records from 400 patients, which includes 250 patients with chronic kidney disease (CKD) and 150 patients without CKD. The dataset was prepared in 2015 by Apollo Hospitals, Tamil Nadu, India, and is publicly available at the University of California, Irvine (UCI) machine learning repository. The dataset has 24 factors, such as blood pressure and white blood cell count, and a classification variable indicating whether the patient has CKD or not. It's important to note that the dataset is imbalanced, with more patients having CKD than not.

3.2 Symptoms

Kidney disease is a condition where the kidneys are unable to function properly, and it is often linked to risk factors such as high blood pressure, obesity, smoking, and diabetes. High blood pressure can damage the small blood vessels in the kidneys, reducing blood flow and impairing kidney function. Additionally, obesity can increase the risk of kidney disease by causing inflammation and hormonal changes that may lead to kidney damage. Smoking can also contribute to impaired kidney function by damaging blood vessels and reducing blood flow. Finally, diabetes is a significant risk factor for kidney disease, as high blood sugar levels can cause progressive damage to the kidneys over time.

3.3 Strong Classifier

A strong classifier is a type of machine learning algorithm that is used to make predictions or classifications based on a set of input features. In the context of a flowchart, a strong classifier may be used as part of a simple architecture to help make decisions or perform tasks based on user input. The use of a strong classifier in a flowchart can help streamline decision-making processes and improve the user experience by providing more personalized and effective guidance.

4. Algorithm

4.1 AdaBoost Algorithm

The AdaBoost algorithm is a machine learning technique based on the boosting concept, which aims to convert weak learners into strong learners. It was introduced by Freund and Schapire [10] and works by iteratively training multiple learning classifiers using the same dataset. In this process, weak learners are trained and then combined to form a strong classifier.

Algorithm 1: Conventional AdaBoost technique

Input: training dataset $S = \{(x_1,y_1), \ldots, (x_2,y_2), \ldots, (x_n,y_n)\}$, base learner h, the number of training rounds T.

Output: the final strong classifier H.

Procedure:

```
1.  for $i = 1 : 1 : n$
1.  for $i = 1 : 1 : n$
2.      compute the weight of the sample $x_i$: $D_1(i) = \frac{1}{n}$
3.  end for
4.  for $t = 1 : 1 : T$
5.  select a training date subset $X$ from $S$, for $h$ using X to get a weak classifier $h_t$, compute the classification error $\varepsilon_t$: $\varepsilon_t = P[h_t(x_i) \neq y_i] = \sum_{i=0}^{n} D_t(i) I[h_t(x_i) \neq y_i]$ where $h_t(x_i)$ denotes the predicted label of $x_i$ using the weak classifier $h_t$, and $y_i$ denotes the
6.  actual label of $x_i$.
7.      compute the weight of $h_i$ : $\alpha_t = \frac{1}{2} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$
8.      update the weights of all the instances in $S$: for $i = 1 : 1 : n$ $D_{t+1}(i) = \frac{D_t(i)}{Z_i} exp\left(-\alpha_t y_i h_t(x_i)\right)$ where $Z_t$ is a normalization factor and is calculated as: $Z_t = \sum_{i=1}^{n} D_t(i) exp\left(-\alpha_t y_t h_t(x_i)\right)$
9.      end for
10.         end for
11. Assuming $H(x)$ is the class label for an instance $x$; after the iterations, the final classifier $H$ is obtained as: $H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$
```

l of enhancing its performance.

5. Result:

In order to establish a baseline for comparison with the proposed cost-sensitive AdaBoost (CS AdaBoost), this research also implemented several well-known classifiers, including logistic regression, decision tree, XGBoost, random forest, and SVM, along with the traditional AdaBoost presented in Algorithm 1. Both the complete and reduced feature sets were used to train the classifiers, in order to demonstrate the impact of feature selection.

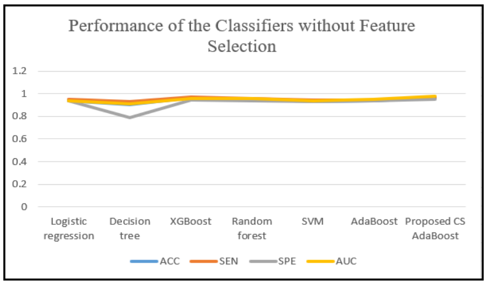

Figure 2: Performance of the Classifiers with Feature Selection

Table 1: Performance of the Classifiers without Feature Selection

Classifier	ACC	SEN	SPE	AUC
Logistic regression	0.940	0.948	0.935	0.940
Decision tree	0.902	0.932	0.790	0.910
XGBoost	0.963	0.974	0.942	0.960
Random forest	0.952	0.955	0.940	0.960
SVM	0.927	0.943	0.930	0.940
AdaBoost	0.940	0.941	0.935	0.950
Proposed CS AdaBoost	0.965	0.965	0.950	0.980

Figure 1: Performance of the Classifiers without Feature Selection

Table 2: Performance of the Classifiers with Feature Selection

Classifier	ACC	SEN	SPE	AUC
Logistic regression	0.971	0.949	0.951	0.970
Decision tree	0.940	0.925	0.948	0.940
XGBoost	0.989	0.980	0.976	0.980
Random forest	0.977	0.981	0.973	0.980
SVM	0.954	0.957	0.961	0.960
AdaBoost	0.954	0.970	0.958	0.960
Proposed CS AdaBoost	0.998	1.000	0.998	0.990

6. Conclusion:

In this paper, a new method was suggested to improve the detection of chronic kidney disease using machine learning. The proposed method uses a combination of information gain-based feature selection and a cost sensitive AdaBoost classifier. The algorithms compared the performance of this new method with six other machine learning classifiers, including logistic regression, decision tree, random forest, SVM, XGBoost, and the traditional AdaBoost. The new method first used a technique called information gain (IG) to find out which attributes were most important in detecting kidney disease. Then, the machine learning classifiers were trained using both the important and all the available attributes. The results showed that using the important attributes improved the performance of the classifiers. This new method proposed in this research paper can help improve the accuracy of detecting chronic kidney disease, which is essential for early intervention and better patient outcomes

7. References:

1. Bhaskar, N.; Suchetha, M.; Philip, N.Y. Time Series Classification-Based Correlational Neural Network with Bidirectional LSTM for Automated Detection of Kidney Disease. IEEE Sens. J. 2021, 21, 4811–4818.
2. Sobrinho, A.; Queiroz, A.C.M.D.S.; Dias Da Silva, L.; De Barros Costa, E.; Eliete Pinheiro, M.; Perkusich, A. Computer-Aided Diagnosis of Chronic Kidney Disease in Developing Countries: A Comparative Analysis of Machine Learning Techniques. IEEE Access 2020, 8, 25407–25419.
3. Chothia, M.Y.; Davids, M.R. Chronic kidney disease for the primary care clinician. South Afr. Fam. Pract. 2019, 61, 19–23.
4. Qin, J.; Chen, L.; Liu, Y.; Liu, C.; Feng, C.; Chen, B. A Machine Learning Methodology for Diagnosing Chronic Kidney Disease. IEEE Access 2020, 8, 20991–21002.
5. Ebiaredoh-Mienye, S.A.; Esenogho, E.; Swart, T.G. Integrating Enhanced Sparse Autoencoder-Based Artificial Neural Network Technique and Softmax Regression for Medical Diagnosis. Electronics 2020, 9, 1963.

6. Chittora, P.; Chaurasia, S.; Chakrabarti, P.; Kumawat, G.; Chakrabarti, T.; Leonowicz, Z.; Jasin´ski, M.; Jasin´ski, Ł.; Gono, R.; Jasin´ska, E.; et al. Prediction of Chronic Kidney Disease—A Machine Learning Perspective. IEEE Access 2021, 9, 17312–17334.
7. Tadist, K.; Najah, S.; Nikolov, N.S.; Mrabti, F.; Zahi, A. Feature selection methods and genomic big data: A systematic review. J. Big Data 2019, 6, 79.
8. Pirgazi, J.; Alimoradi, M.; Esmaeili Abharian, T.; Olyaee, M.H. An Efficient hybrid filter-wrapper metaheuristic-based gene selection method for high dimensional datasets. Sci. Rep. 2019, 9, 18580.
9. Nikravesh, F.Y.; Shirkhani, S.; Bayat, E.; Talebkhan, Y.; Mirabzadeh, E.; Sabzalinejad, M.; Aliabadi, H.A.M.; Nematollahi, L.; Ardakani, Y.H.; Sardari, S. Extension of human GCSF serum half-life by the fusion of albumin binding domain. Sci. Rep. 2022, 12, 667.
10. Freund, Y.; Schapire, R.E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. J. Comput. Syst. Sci. 1997, 55, 119–139.
11. Akter, S.; Habib, A.; Islam, M.A.; Hossen, M.S.; Fahim, W.A.; Sarkar, P.R.; Ahmed, M. Comprehensive Performance Assessment of Deep Learning Models in Early Prediction and Risk Identification of Chronic Kidney Disease. IEEE Access 2021, 9, 165184–165206.
12. Elkholy, S.M.M.; Rezk, A.; Saleh, A.A.E.F. Early Prediction of Chronic Kidney Disease Using Deep Belief Network. IEEE Access 2021, 9, 135542–135549.