Ministry of Higher Education and
Scientific Research

University of Diyala

College of science
Department of Computer Science

A comparision for Chronic Kidney Diseases Classification Based on Machine Learning Approaches

Thesis

Submitted to the Department of Computer Science\ College of Sciences\ University of Diyala in a Partial Fulfillment of the Requirements for the Degree of Master in Computer Science

By

Noor Saud Abd

Supervised By

Dr. Dhahir Abdulhade Abdulah
Dr. Prof.

Amer Dawood Majeed
Dr. Prof.

2021 AC 1442 AH

بِسمِ ٱللَّهِ ٱلرَّحمنِ ٱلرَّحِيمِ

﴿ نَرْفَعُ دَرَجَاتٍ مَّن نَّشَآءُ وَفَوْقَ كُلِّ ذِي عِلْمٍ عَلِيمٌ ﴾

صَدقَ ٱللَّهُ العَظيم

سورة يوسف

الآية (٧٦)

Acknowledgment

First of all, praise is to GOD, the lord of the whole creation, on all the blessing was the help in achieving this research to its end. I wish to express my thanks to my supervisors, **prof. Dr. (Dhahir Abdulhade Abdulah) and prof. Dr. (Amer Dawood Majeed).**

for supervising this research and for the generosity, patience and continuous guidance throughout the work. It has been my good fortune to have the advice and guidance from them. My thanks to the academic and administrative staff at the Department of the computer sciences.

I would like to express my gratitude to my family who were unlimited support and patience.

Noor Saud Abd

Dedication

To …

Our Prophet Mohammed
Peace be Upon Him (PBH)

My dear parents

My friends

I produce this work with all
my love …

Noor Saud

Abstract

Although chronic kidney disease has been known for many centuries, some people still die from it to this day, but with the development of medicine and diagnostic and therapeutic techniques, it has become possible to avoid and treat some conditions if the disease is diagnosed early.

Data mining is a broad scientific field concerned with discovering patterns and relationships between data elements. The algorithms used in this field help to increase the ease and effectiveness of the decision.

Classification is one of the major issues in knowledge discovery and decision-making within data mining techniques. There are many algorithms used to build classes.

In this thesis, the proposed system was to diagnose chronic kidney disease using classification and comparison algorithms, where the dataset was download from the UCI repository and a set of machine learning algorithms was applied, which was Naïve Bayes algorithm, the accuracy result was 99 %, it was better than the rest of the algorithms because it suits the nature of The distributed dataset is a normal distribution. In addition to the artificial neural network 97%, the supporting vector machine 83% as well as the hybrid algorithm experiment (GA and SVM), the optimization algorithm for selecting the best features were combined with the classification algorithm and the accuracy was 95%.

List of Contents

List of Figures

List of Tables

List Of Abbreviation

Abbreviations	Meaning
ANN	Artificial neural network
AI	Artificial intelligence
Avg	Average
CKD	Chronic Kidney diseases
FP	The number of false positives
FN	The number of false negatives
GA	Genetic Algorithm
KDD	Knowledge discovery in databases
ML	Machine learning
MLP	Multiple Decision Boundaries
NB	Naïve Bayes
RBF	Radial Basis Function
SVM	Support Vector Machine
TP	The number of true positives

TN	The number of true negatives
μ	Mean Value
σ	Standard Deviation

Table of Symbols

Symbol	Meaning
*	Multiplication operation
+	Addition operation
/	Division operation
-	Subtraction operation
=	Equality sign
θ	Theta
∑	Summation - sum of all values in range of series
δ	Delta
σ	Sigma
%	Percent sign
‖ ‖	Norm

List of Algorithms

Chapter one

Introduction

Chapter one

Introduction

1.1 Overview

A pair of kidneys is a vital organ for its proper functioning in the human body. Its function is to filter the blood, remove waste products, and control fluid balance in the body and urine formation. Chronic Kidney disease (CKD) is a condition in which kidney function has altered the ability to function properly decreases, which leads to elevation the number of waste products in the blood that make the human body sick in the long term [1].

Chronic kidney disease (CKD) is a global health problem with a high death rate, Researches have reported that CKD results in the tens of thousands human death around the world. It is therefore of great interest to find out ways to diagnose the CKD early-stage [2].

The mounting evidence indicates that this negative outcome and be prevented through early detection and proper management**.** Although human decision-making may be perfect, it gets poor when needs to classify a large amount of data. It will also reduce the accuracy and efficiency of decisions when humans are put under enormous work and pressure. So the use of machine learning predictive modeling applications in diagnosing chronic diseases to help doctors in disease data accurately, and quickly to save time, effort, and better performance without errors as possible [3].

Researchers are trying to use different types of machine learning (ML), which is a subset of artificial intelligence techniques, to diagnose different types of diseases early [2].

Deducing patterns from data both artificial intelligence and machine learning differ from traditional statistical methods, for example, they focus on prediction and categorization from high-dimensional data, rather than inference. Successful machine learning requires strong data that it can learn from. This data must be abundant enough to enable the model to be robust and generalizable to non-visual data [4].

Data mining techniques support the machine learning process, and they are widely used in different applications. The main function of data mining is to apply different techniques, algorithms, and methods for extracting any specific patterns or information from big stored data and convert it to a comprehensible form for future uses [5].

Data mining is one of the most encouraging areas of research with the goal to find useful information from large data sets and categorizing valid and unique patterns in the data. There are various data extraction techniques such as clustering, classification, regression, correlation analysis, etc. [6].

In this thesis, we used some algorithms to extract data and to discover kidney failure in a group of patients. By taking several analyzes, which are entered as input to the algorithms program, it gives us a result as to whether the patient needs a new kidney transplant or can be treated in other ways.

1.2 Related Work

This section reviews some of the earlier studies and explains the different techniques that are utilized to diagnose chronic kidney diseases :

- **V Kunwar, et al. 2016 [7].**

They have performed the prediction and diagnosis of chronic kidney disease with the use of the data extraction classifiers, which are: the Naïve Bayesian and the ANNs. The performances of those algorithms has been compared with the use of the Rapid miner tool. Obtained results have shown that the naïve Bayesian has been the best classifier with an accuracy of up to 99%. In comparison with the ANNs with 72.73% accuracy. They took into account some of factors considering the age, diabetes, RBC count, blood pressure, and so on. May be a job that is extended through looking at other criteria like the food type living conditions and work environment, availability of hygiene water, environmental and other factors to detect the kidney diseases.

- **N Tazin, et al, Conference on Medical, 2016 [8].**

They used a SVMs classification, decision tree, K-Nearest Neighbor (KNN) Algorithm, Chronic kidney disease has been analyzed and predicted for different classifiers: Naïve Bayes, SVM, KNN and Decision tree. To compare the performance of these classifier algorithms, WEKA tool has been used. From the performance result it is observed that decision tree algorithm provides the highest accuracy of 99%. The second most accurate classifier is SVM with accuracy of 97.75%. It is also observed that implementation of ranking algorithm increases the performance for predicting CKD but correct number of attributes must be selected. Some major factors like age, RBC, blood pressure etc. have been considered for classification. Other parameters like nutrition, accommodation status, clean water availability, surroundings can be considered for detection of CKD. In future, performance of other classifiers like ANN, Fuzzy logic can be compared using the WEKA tool for similar situation and dataset.

- **A Subasi, et al , 2017 [9].**

In this study, different machine learning methods are used for diagnosis of CKD. This study showed that Random Forest classifier results in very high performances during classification tasks. This method is able to classify two different classes with perfect classification rate, 100%. Precision for RF are also perfect, 1.0 We also applied ANN, k-NN, SVM and C4.5 decision tree methods and obtained high performance results, especially for C4.5 decision tree classifier. For diagnosis of CKD, we propose the usage of RF machine learning tool since it results in high classification accuracy rate, time needed for training and testing is low.

- **T Balakrishna, et al , 2017 [10].**

They have conducted a performance evaluation of a suggested approach of the Random Forest Classification that has been implemented on the Chronic Kidney Disease Data-set that exhibited higher efficiency compared to some Existing approaches, such as the J-48, REP Tree and Naïve Bayesian. With those suggested techniques, a new system of diagnosis has been advanced for chronic kidney diseases, providing assistance to end-users there through the reduction of patient queue near a doctor. The whole dataset has been converted into tables and inserted into My Sql Data-base.

- **EHA Rady, AS Anwar- Informatics in Medicine Unlocked, 2019 [11].**

They proposed a system of application of probabilistic neural networks (PNN), Multilayer Perceptron (MLP), Radial Basis Function (RBF), and Support Vector Machine (SVM) algorithms were compared. The results showed that the PNN algorithm provides the best classification and prediction performance for determining the stage of severity in CKD. Probabilistic neural networks gave the algorithm the highest overall classification accuracy of 96%, compared to other algorithms in staging CKD patients. On the other side, a multilayer preceptor requires a minimum execution time (3 s) while a probabilistic neural network requires 12 s to finish the analysis. These algorithms were compared with classification accuracy based on the categorized stages of CKD patients, time to build the model, and time to test the model. The probabilistic neural network (PNN) algorithm provides better classification accuracy and predictive performance to predict the stages of CKD.

- **J Qin, et al , IEEE Access, 2019 [3].**

They studied six machine learning algorithms (logistic regression, random forest, k-nearest neighbor, support vector machine, naive Bayes classifier, and feed forward neural network) that were used to establish the models. Among these machine learning models, random forest achieved the best performance with 99.75% diagnosis accuracy. By analyzing the misjudgments generated by the models.

- **R Devika, et al , 2019 [12]**

They examined the performance of Naive Bayes, K-Nearest Neighbor (KNN), and Random Forest classifier based on its accuracy, preciseness, and execution time for CKD prediction. The result after conducted research is that the performance of the Random Forest classifier (99.84%) is better than Naive Bayes (99.63%) and KNN (87.78%) .

- **B Khan, et al , 2020 [2].**

They employed experiential analysis of Machine learning techniques for classifying the kidney patient dataset as CKD or NOTCKD. Seven ML techniques together with NBTree, J48, Support Vector Machine, Logistic Regression, Multi-layer Perceptron, Naïve Bayes, and Composite Hypercube on Iterated Random Projection (CHIRP), and the outcomes accomplished of MAE are 0.0419 for NB, 0.035 for LR, 0.265 for MLP, 0.0229 for J48, 0.015 for SVM, 0.0158 for NBTree and 0.0025 for CHIRP. In addition to, experimental results using accuracy revealed 95.75% for NB, 94.50% for LR, 95.25% for MLP, 94% for J48, 95.25% for SVM, 98.75% for NBTree, and 99.75% for CHIRP. The final outcomes show that CHIRP performs well in terms of diminishing error rates and improving accuracy.

1.3 Problem Statement

Chronic kidney disease (CKD) is prevalent in all middle and low-income areas, and many high-income areas as well. Chronic kidney disease may be a burden for people, families, and society to seek ongoing treatment and ongoing dialysis, or a new kidney transplant. Early detection of chronic kidney disease is critical.

Considering that the medical field is one of the most important fields of knowledge throughout the ages because it is directly related to human health and life, so it is necessary to pay attention to extracting knowledge and accuracy in diagnosis to the extent that reduces error as much as possible to avoid making a wrong decision that harms the individual, so the need to use Effective techniques in diagnosis, including machine learning algorithms (a branch of artificial intelligence) that mine data and help the doctor make the appropriate decision accurately and quickly.

1.4 Aim of Thesis

This thesis aims to apply some data mining algorithms and how to use them in diagnosing kidney diseases, determine the best ones in terms of diagnostic accuracy of the results, and compare them with each other. To know the best in terms of accuracy of diagnosis and speed of results, to help doctors recognize kidney failure and prevent errors.

1.5 Thesis Outline

The rest of the chapters in the present thesis will be organized in the following manner:

Chapter 2: Theoretical Background

This chapter clarifies the definition and the types of data mining in addition to Knowledge discovery in databases (KDD) steps, it also explains the Data mining techniques used for classification.

Chapter 3: The Proposed System

Which includes a description of the suggested system of classification with its implementation and design.

Chapter 4: Experimental Results and Evaluation

Which shows the implementation results of the proposed system steps and evaluates these results.

Chapter 5: Conclusions and Suggestions for Future Work

This includes the main points that have been concluded from the present study. In addition to that, it presents some of the suggestions for future works.