International Journal of Cyber-Security and Digital Forensics (IJCSDF) 8(3): 188-193
The Society of Digital Information and Wireless Communications (SDIWC), 2019 ISSN: 2305-0011

Computer Aided Diagnosis for Chronic Kidney Disease Using Data Mining

Mohammad Ashraf Ottom
Computer Information Systems Department
Yarmouk University
Irbid, JORDAN
ottom.ma@yu.edu.jo

Khalid M. Nahar
Computer Science Department
Yarmouk University
Irbid, JORDAN
khalids@yu.edu.jo

***Abstract*—Chronic Kidney Disease (CKD) is a gradual disorder of kidney tasks over time. CKD early detection could reduce the impacts and harms by offering the necessary treatment. Nowadays, data mining can assist doctors to diagnose and predict diseases. In this paper, we shown that data mining techniques is a successful tool for diagnosing CKD using well know classification techniques such as naïve bays, and positively enhanced the diagnostic process. Since digital medical records could have redundant and unnecessary features, the paper also utilized features selection techniques to identify the most useful features that improves the diagnosis. The experiments showed that data mining techniques are capable of predicting and diagnosing CKD. In addition, features selection techniques such as CorrelationAttributeEval and CfsSubsetEval can assist to achieve better prediction accuracy. The experiments also showed that naïve bays classification technique performed better in collaboration with features selection techniques.**

Keywords—Computer Aided Diagnosis; Chronic Kidney Disease; Classification Techniques; Features Selection Techniques.

I. INTRODUCTION

Human kidneys are considered to be the blood filters. Kidneys, filter the blood from poison and detoxicate its contents. When the kidney starts functioning improperly, unuseful substances will drain to the urine poisoning the blood, which results in malfunctioning of human health. Consequently, kidney may stop functioning leading for Chronic Kidney Disease (CDK), causing kidney failure and accumulationunwanted substances in blood[1].

Rapid improvements in information technology, disk and cloud storage capability and healthcare information systems evolution, produce the Electronic Health Record (EHR) or Electronic Medical Record (EMR). EHR is an electronic form of a patient's medical history that is equivalent to the traditional hard copy of patient's medical record. The existence of EHR provides more options for data processing using the revolution of data mining techniques.

Data mining is rapidly growing field in information technology. It is used in many domains such as; financial forecasting, weather forecasting, insurance and health care.etc. The main objective of data mining in healthcare sector is to extract knowledge from high volume of data. Patient medical data and computational techniques can be used to answer clinical questions, modelling disease spread and real time identification of emergencies, discover disease patterns and record disease outbreaks over the world, enhance the quality of care and services for patients and provide swift prediction and diagnosis of some diseases.

The rest of our paper includes a section for Data mining techniques or classifications (Section 2). Another section (section 3) which relates our work to others. In addition, section 4 details our methodology, and lastly the discussion and conclusion in section 5.

II. DATA MINING CLASSIFICATION TECHNIQUES

A. *Support Vector Machine SVM*

Support Vector Machine (SVM) is a classifier which mainly separate the hyperplane. In *supervised*

learning, SVM produces an optimal strait line that separate between categorizes as shown in figure 1.In fact, SVM algorithm finds a line with maximum margins between categorizes, hence, the optimal separating hyperplane *maximizes* the margin of the training data [2]. SMO (Sequential Minimal Optimization) algorithm is the SVM implementation in WEKA, it isimplemented by Platt (Platt 1998) and it is one of the most efficient solutions for the SVM algorithm. It is based on solving a series of small quadratic problems, where in each iteration only two variables are selected in the working set to avoid a time consuming [3].

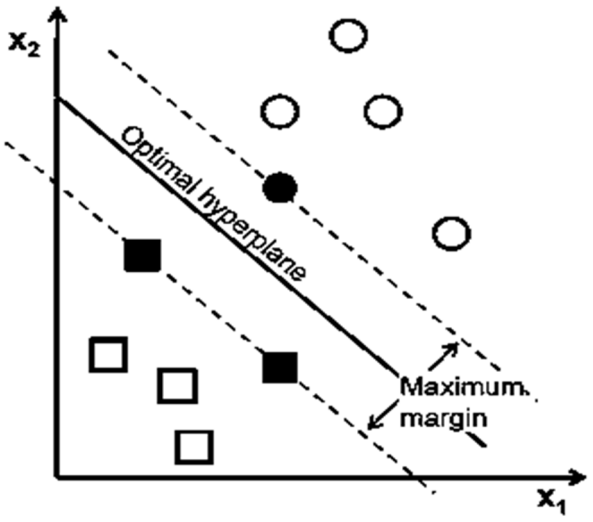

Figure 1: Line separation between labels by Support Vector Machine.

B. *Naïve Bayes NB*

In data mining, Naïve Bayes classifier is a probabilistic classifier based on Bayes theorem.Naïve Bayes classifier assume the independency between problem features, it implements the idea that the existence of a certain feature of an object does not depend to the existence of any other features.In addition, Naïve Bayes classifier process all features independently where no feature depends on others features values [4].

$$p(c_j|d) = \frac{p(d|c_j)\,p(c_j)}{p(d)} \qquad (1)$$

where $p(c_j|d)$ is the probability of instance d being in class c_j, $p(d|c_j)$ is the probability of having instance d given class c_j, $p(c_j)$ is the probability of occurrence of class c_j and $p(d)$ is the probability of instance d occurring.

C. *Artificial Neural Networks ANN*

Artificial Neural Networks (ANN) are inspired by human biological nerve system. Artificial Neural Network is a collection of many artificial neurons that are connected. The main purpose of ANN is to map the inputs into meaningful outputs. Figure 2 shows the architecture of feedforward ANN.

Each input from the input layer is to feed each node in the hidden layer, and the hidden layer feed the next layer until reaching the output layer. ANN, in most cases, consist of multiple hidden layers to pass through before ultimately reaching the output layer. The ANN in Figure 2 is called feed forward ANN because the signals are passed through the layers of the neural network in a single forward direction. However, ANN can be feedback networks where the architecture allows signals to travel in both directions [5].

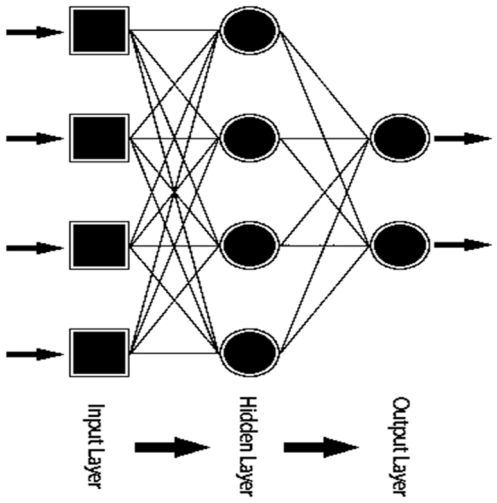

Figure 2 Feedforward Artificial Neural Network.

D. *k Nearest Neighbors kNN*

k Nearest Neighbors (kNN) classification is regarded as one of the top data mining classification tools. Simplicity and efficiency are the main reason of kNN popularity. The main idea of kNN methodology is to predict the class label of testedinstances using the distance between the tested instance and the training neighbors instances. kNN decides the class label for the tested instance by identifying the most frequent class label among the training data (the lesser distance between tested instance and training instances). The distance is determined by the distance metric such as Euclidean distance (equation 2), Manhattan distance (equation 3) or Minkowski distance (equation 4)[6].

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (2)$$

$$d(x,y) = \sqrt{\sum_{i=1}^{n}|x_i - y_i|} \quad (3)$$

$$d(x,y) = \left(\sqrt{\sum_{i=1}^{n}|x_i - y_i|^r}\right)^{\frac{1}{r}} \quad (4)$$

III. Related Work

Data mining techniques proved the capibility of predicting diseases such as breast cancer and heart disease. Many researchers in data mining fields are working continousaly to improve classification process and produce better results in regards to accuracy, noise tolierance, missing values handling and execution time. Features selection techniques engaged successfully in disease prediction by reducing the number of features that map inputs to output. Jena and Kam [7] utilized data mining techniques to predict chronic kidney disease using Naive Bayes, Multilayer Perceptron, SVM, J48, Conjunctive Rule and Decision Table. The study concluded that Multilayer Perceptron was the superior classifier in regards to classification accuracy. Another study [8] aimed to predict early detection of chronic kidney disease using data mining techniques.They found that classification algorithms Naïve Bayes and Decision tree produced the highest accuracy (91%) using Decision tree. A research group [9] used data mining techniques to specify the most important factors to diagnose chronic kidney diseaseusing Iranian patients suffered chronic kidney disease CKD. They used data mining tools to discover the hidden rules and relationships between features the dataset, also, they utilized neural network to predict the status of patients with CKD. A recent experiments [10] were conducted to predict CKD. The dataset obtained from UCI Machine Learning repository, then, the dataset applied to six data mining algorithms, namely: Random Forest (RF) classifiers, Sequential Minimal Optimization (SMO), NaiveBayes, Radial Basis Function (RBF), Multilayer Perceptron Classifier (MLPC), and SimpleLogistic (SLG). Classifers training and testing obtained using ten-fold cross validation. The results showed that the RF classifier outperforms other classifiers in regards to accuracy, sensitivity and specificity. Additional study [11]employed data mining algorithms on medical dataset to extract decision making rules that can be used to diagnose CKD. The study used C4.5 decision tree algorithm to obtain a set of diagnosis rules for CKD. The C4.5 algorithm using 3-fold cross validation obtained high prediction accuracy. Celik, Atalay and Kondiloglu [12] examined CKD dataset using some data mining techniques such as Support Vector Machine and Decision Tree algorithm. In the classification stage, the experiment showed that Decision Tree has been more successful than the Support Vector Machine. In addition, the study showed that Decision Tree obtained better results than the Support Vector Machine for the early diagnosis of CKD.

IV. The Proposed Method

Chronic Kidney Disease dataset is a publicly available dataset, published on UCI Machine Learning Repository. Table I describes attributes, description and type of each attributes. The dataset is a real data, consist of 400 instances, 25 features, and donated for research purposes in 2015.

TABLE I CHRONIC KIDNEY DISEASE DATASET DESCRIPTION

No	Attribute	Description	Type
1	AGE	Age	numerical
2	BP	blood pressure	numerical
3	SG	specific gravity	nominal
4	AL	albumin	nominal
5	SU	sugar	nominal
6	RBC	red blood cells	nominal
7	PC	pus cell	nominal
8	PCC	pus cell clumps	nominal
9	BA	bacteria	nominal
10	BGR	blood glucose	numerical
11	BU	blood urea	numerical
12	SC	serum creatinine	numerical
13	SOD	sodium	numerical
14	POT	potassium	numerical
15	HEMO	haemoglobin	numerical
16	PCV	packed cell volume	numerical
17	WC	white blood cell count	numerical
18	RC	red blood cell count	numerical
19	HTN	hypertension	nominal
20	DM	diabetes mellitus	nominal
21	CAD	coronary artery disease	nominal
22	APPET	appetite	nominal
23	PE	pedal edema	nominal
24	ANE	anaemia	nominal
25	CLASS	class	nominal

The proposed approach is a hybrid method that combines features selection techniques with data mining classifiers to diagnose diseases such as Chronic Kidney Disease (CDK). Figure 3 describes the proposed research model. We applied the original CKD dataset on a well know classifications techniques, Naïve Bayes, Neural Networks, Support Vector Machine and k Nearest Neighbors. The results of classification have been recorded as a benchmark for later comparison. Then, we applied features selection techniques on the original CKD dataset using a popular technique (CorrelationAttributeEval and CfsSubsetEval). The purpose of features selection is to obtain a set of features that best predict the class label. Then, we compared the classification results on the original dataset with classification results on the newly constructed dataset in regards to classifier accuracy, precision, recall and f-measure. The experiments performed using data mining software WEKA, which provides the environment for experimenting data mining techniques.

TABLE II CLASSIFIERS RESULTS ON ORIGINAL CKD DATASET

	NB	SVM	kNN	ANN
Accuracy	0.950	0.978	0.918	0.998
Precision	0.956	0.979	0.919	0.998
Recall	0.950	0.978	0.918	0.998
F-Measure	0.951	0.978	0.918	0.998

Table II shows classification results of applying CKD dataset on data mining techniques. Artificial Neural Networks produced the best results in regards to accuracy, precision, recall and f-measure with classification accuracy of 0.998. On the other hand, NB, SVM and kNN performed well with classification accuracy of 0.950, 0.978 and 0.918 respectively.

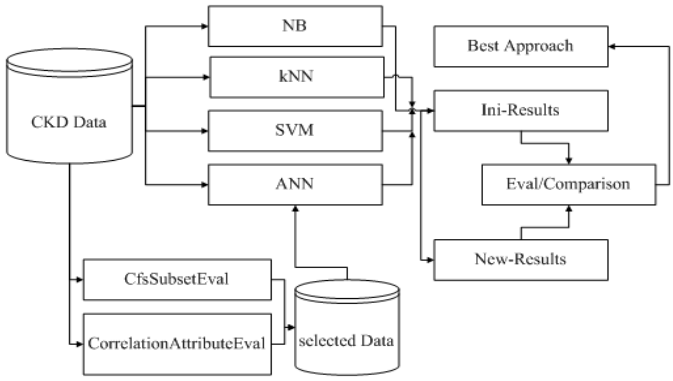

Figure 3 Proposed Research Model

TABLE III CLASSIFIERS RESULTS USING CAEVA AND CLASSIFIERS ON CKD DATASET

	CAEvaNB	CAEvaSVM	CAEvakNN	CAEvaANN
Accuracy	0.953	0.978	0.918	0.998
Precision	0.958	0.979	0.922	0.998
Recall	0.953	0.978	0.918	0.998
F-Measure	0.953	0.978	0.918	0.998

Table III shows the classification results of hybrid system using CorrelationAttributeEval

features selection technique and data mining classifiers. Features selection technique contributed successfully in classification accuracy, precision, recall and f-measure using naïve bays classifier. However, CorrelationAttributeEval did not affect classification results for the rest of classifiers.

TABLE IV CLASSIFIERS RESULTS USING CSEVAL AND CLASSIFIERS ON CKD DATASET

	CSEval NB	CSEval SVM	CSEvalkNN	CSEval ANN
Accuracy	0.955	0.983	0.948	0.998
Precision	0.960	0.983	0.952	0.998
Recall	0.955	0.983	0.948	0.998
F-Measure	0.955	0.983	0.948	0.998

Table IV shows the classification results of hybrid system using CfsSubsetEval featuers selection technique and data mining classifiers. It shows that CfsSubsetEval features selection technique combined with classifers have improved classification accuracy, precision, recall and f-measure using naïve bays classifier, SVM and kNN. However, CorrelationAttributeEval combined with ANN did not affect classification results.

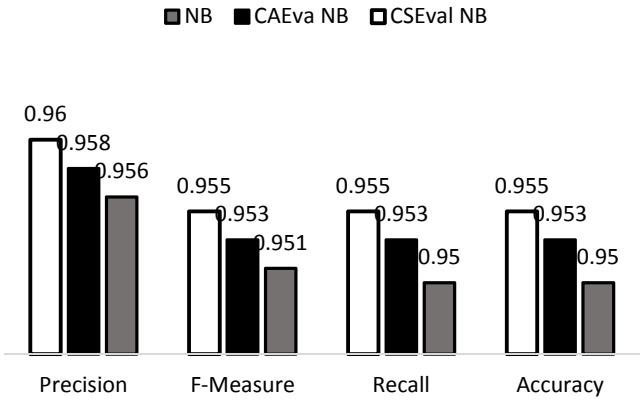

Figure 4 classification results using NB, CAEva NB and CSEval NB

Figure 4 illustrate classification results using naïve bays classifier, a hybrid approach using naïve bays and CorrelationAttributeEval (CAEvaNB), and a hybrid method using naïve bays and CfsSubsetEval (CSEvalNB). The graph shows improvement in classification accuracy using CAEvaNBand CSEvalNB with a maximum classification accuracy 0.955.

V. DISCUSSION AND CONCLUSION

The performed experiments of data mining techniques on CKD dataset proved that data mining techniques are capable of predicting and diagnosing disease. Since diseases datasets could contain huge number of features, and may contain less important features which affect the prediction process; features selection techniques such as CorrelationAttributeEval and CfsSubsetEval can enhance the prediction accuracy. The experiments also showed that naïve bays performed better when applying features selection techniques. However, some classifiers obtained the same performance on full dataset as the newly constructed dataset, which contain less number of features.

REFERENCES

[1] A. A. Razmaria, "Chronic Kidney Disease," *JAMA*, vol. 315, no. 20, p. 2248, May 2016.

[2] A. R. Statnikov, *A gentle introduction to support vector machines in biomedicine*. World Scientific, 2011.

[3] J. Wang, A. Lu, and X. Jiang, "An Improved SMO Algorithm for Credit Risk Evaluation*," in *Australasian Data Mining Conference*, 2015, pp. 169–176.

[4] E. Keogh, "Naïve Bayes Classifier." 2015.

[5] J. LEE, "Introduction to Artificial Neural Networks," 2013.

[6] S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng, "Learning k for kNN Classification," *ACM Trans. Intell. Syst. Technol. Artic.*, vol. 8, no. 19, 2017.

[7] L. Jena and N. Ku Kamila, "Distributed Data Mining Classification Algorithms for Prediction of Chronic-Kidney-Disease," *Int. J. Emerg. Res. Manag. &Technology*, no. 4, pp. 2278–9359, 2015.

[8] K. R. A. Padmanaban and G. Parthiban, *Indian journal of science and technology IndJST*, vol. 9, no. 29. 2016.

[9] M. Ghazisaeedi, S. Tahmasebian, M. Langarizadeh, M. Mokhtaran, M. Mahdavi-Mazdeh, and P. Javadian, "Journal of Renal Injury Prevention Applying data mining techniques to determine important parameters in chronic kidney disease and the relations of these

parameters to each other," *J Ren. Inj Prev*, vol. 6, no. 2, pp. 83–87, 2017.

[10] M. Kumar and M. Kumar, "Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm Running Title: Prediction of Chronic Kidney Disease," *Int. J. Comput. Sci. Mob. Comput.*, vol. 525, no. 22, pp. 24–33, 2016.

[11] A. A. Serpen, "Diagnosis Rule Extraction from Patient Data for Chronic Kidney Disease Using Machine Learning," *Int. J. Biomed. Clin. Eng.*, vol. 5, no. 2, pp. 64–72, Jun. 2016.

[12] E. Celik, M. Atalay, and A. Kondiloglu, "The Diagnosis and Estimate of Chronic Kidney Disease Using the Machine Learning Methods," no. 4, pp. 27–31, 2016.

