Analysis of Chronic Kidney Disease Dataset by Applying Machine Learning Methods

Yedilkhan Amirgaliyev
Institute of Information and Computing Technologies (IICT), Almaty, Kazakhstan
amir_ed@mail.ru

Shahriar Shamiluulu
Faculty of Engineering and Natural Sciences, Suleyman Demirel University, Kazakhstan
shahriar.shamiluulu@sdu.edu.kz

Azamat Serek
Faculty of Engineering and Natural Sciences, Suleyman Demirel University, Kazakhstan
140107073@stu.sdu.edu.kz

***Abstract*— Currently, there are many people in the world suffering from chronic kidney diseases worldwide. Due to the several risk factors like food, environment and living standards many people get diseases suddenly without understanding of their condition. Diagnosing of chronic kidney diseases is generally invasive, costly, time-consuming and often risky. That is why many patients reach late stages of it without treatment, especially in those countries where the resources are limited. Therefore, the early detection strategy of the disease remains important, particularly in developing countries, where the diseases are generally diagnosed in late stages. Finding a solution for above-mentioned problems and riding out from disadvantages became a strong motive to conduct this study.**

In this research study, the effects of using clinical features to classify patients with chronic kidney disease by using support vector machines algorithm is investigated. The chronic kidney disease dataset is based on clinical history, physical examinations, and laboratory tests. Experimental results showed over 93% of success rate in classifying the patients with kidney diseases based on three performance metrics i.e., accuracy, sensitivity and specificity.

***Keywords*—Chronic kidney disease, Support vector machine, Biomedical engineering.**

I. INTRODUCTION

Chronic Kidney Disease (CKD) is one of the types of kidney disease, which results in a gradual loss of kidney function. This phenomenon can be observed over a period of months or years due to several living conditions of patients [5]. The CKD is also called a chronic kidney failure where according current medical statistics the 10% of the population worldwide is affected by CKD [1-2]. There were approximately 58 million deaths in the year of 2005 worldwide. Where according to the World Health Organization (WHO) 35 million attributed to chronic diseases. Currently it is estimated that one in five men, and one in four women aged 65 through 74 are going to be affected by CKD worldwide. According the 2010 Global Burden of Disease study, CKD was ranked 27th in the list of causes of total number of deaths worldwide in 1990, but unfortunately rose to 18th in 2010 due to the above factors. This degree of movement up the list was second only to that for HIV and AIDs [2].

Diagnosing CDK usually starts with clinical data, lab tests, imaging studies and finally biopsy. Although biopsy is the standard diagnosing test, it has many disadvantages, such as being invasive, costly, time-consuming and sometimes risky. For example; when a biopsy is performed, the patient may face infection, the scare of surgery and misdiagnosis. Imaging studies (mammogram, sonogram, and MRI of the kidney) has been used for many years to detect the disease. But using them has some limitation; more expressly is exposure effects of radiation. Besides being risky, the data provided by imaging is insufficient to diagnose CDK [2].

The automated diagnosis of different diseases has attracted many researchers. Comparison of the methods and accuracy of previous studies including this study has been summarized herein. There are several research studies has been conducted so far [3-6].

In this study [3] the authors by using three supervised machine learning algorithms i.e., Decision trees (DT), Logical Regression (LR) and Artificial Neural Networks (ANN) performed classification for Kidney dialysis data. The tool named Tanagra used to perform the classification. For the classifiers evaluation, the 10-fold cross validation is used. The experimental results showed that ANN outperformed by 93.8% remaining algorithms.

In another study [4], the researchers tried to predict the Long Term Kidney Transplantation Outcome. They have performed comparative analysis between an ANN and LR algorithms. The comparative analysis has been implemented based on performance metrics like accuracy, sensitivity and specificity. During the study for the kidney transplant recipients prediction of kidney rejection which was based on ten training and validating datasets. The experimental results showed that, ANN can be considered a useful supportive algorithm in the prediction process of the defined problem. In summary, the ability of predicting kidney rejection (sensitivity) was 38% for LR versus 62% for ANN. The ability of predicting no-rejection (specificity) was 68% for LR compared to 85% of ANN.

In one more research study [5], the researchers developed a software tool that demonstrates the capabilities of ANN for classification of patients' health status which potentially leading to End Stage of Kidney Disease (ESKD). The classifier is based on an ensemble of ten ANN networks. It has been trained by using data collected in a period of 38 years at University of Bari. The tool has been improved and made

derivable both as a mobile application and as a web application. The tool is important for clinical needs based on the largest cohort worldwide.

In another study [6], the authors have used various data preprocessing, data transformations, and data mining approaches to understand the insides of the interaction between various clinical parameters and patient survival, which are on kidney dialysis. Two different data mining algorithms were applied for the knowledge extraction in the form of decision rules. The extracted rules were used to predicts survival of new unseen patients. Data mining algorithms identified the important medical parameters for carrying out the prediction process. The introduced new research concept have been implemented and tested using data that is collected at four dialysis sites. Presented approach reduces the effort and cost of selecting patients for clinical studies. Patients can be chosen based on the prediction results and the discovered vital parameters.

During the literature review, we have identified that there are less studies performed in CKD by using SVMs. But there are several studies where SVM has been used abundantly. For instance in these studies researchers used SVMs [7-9] as a classification model for detecting and diagnosing malignant and benign tumors based on MRI features, ultrasound feature and mammographic features. This result let us try to create and test SVM classifier over CDK dataset.

In this study, by using the machine learning techniques, we have used cheap, simple and noninvasive tests that can be performed easily. The data has been obtained from dataset was obtained from UCI machine learning repository for CDK patients [13] the details are provided further. By this strategy, we hope to produce "down-staging" (increasing in the proportion of CDK detected at an early stage) of the disease to stages that are more amenable to curative treatment.

In order to improve the accuracy of CDK classification as positive and negative, the performance of Support Vector Machine (SVM) was evaluated. The SVM is a flexible classifier algorithm that has been proposed as an effective statistical learning method for pattern recognition [7, 8], which is based on finding optimal hyperplane to separate different classes mapping input data into higher-dimensional feature space. The SVM has been used for many applications such as object recognition and face detection [11].

II. Materials and Methods

A. Data Collection

The CKD has been obtained from UCI machine learning repository [13]. In total it contains 400 cases, out of which 250 of the cases are patients with CDK and the rest 150 are not. The target variable indicates whether a patient has a CDK or not. There are 25 attributes where 24 are clinical features and remaining is a target attribute. The features are divided into three parts clinical history, physical examination and lab tests. According to the properties of the attributes, the target attribute was classified into negative (expressed by "no disease") and positive (expressed by "presence of disease").

B. Support Vector Machine (SVM)

The SVM is a supervised learning algorithm that is used for data classification and regression [11]. It searches for a best hyperplane which separate between classes. The best hyperplane is considered the one which leaves the maximum margin between the two distinct classes. The margin is defined as the width of the hyperplane from the closest point of the two distinct classes. Bounds between data sets and hyperplane are called support vectors [8,9]. The hyperplane can be found by:

$$g(x) = w^T x + b \quad (1)$$

- " x " : refers data points
- " w ": is a coefficient vector
- " b " : is offset from the origin

In the case of linear SVM $g(x) \geq 0$ for the closest point on the one of the class, $g(x) < 0$ for the closest point belongs to another class. The margin between support vectors is defined by:

$$d = \frac{2}{||w||} \quad (2)$$

For better separation, the margin d should be maximized. The data points which are also called support vectors are the only ones which are needed to solve classification or prediction problem. The remaining data points in the dataset could be taken-out and the same solution should be obtained.

C. Classifier's Performance Measures

The confusion matrix commonly is used to evaluate the classifier, which measures the quality of the classification process. In addition, there are also various standard evaluation measures for correct and incorrect classification results of the classifier. The most common measure to evaluate the performance is accuracy. It is defined as the proportion of the total number of instances that were correctly classified.

Another evaluation metric is *sensitivity*, is the mean proportion of actual true positives that are correctly identified. On the other hand, *specificity* is the mean proportion of true negatives which are identified correctly.

These performance metrics are calculated according to the data in the confusion matrix which are mentioned in [12].

D. Simulated Program

The experimental studies mainly has been done by using the Python and Scikit-learn machine learning framework has been used in this study that contains a large number of algorithms for data preprocessing, feature selection, classification, clustering, and finding the associative rules [12]. Meanwhile we have also used machine learning tools like WEKA for the validation purposes.

III. Results and Discussions

In this study, the 25 original features of the CDK data are used for classification. The accuracy, sensitivity, specificity of 25 features has been performed using 10-fold cross-validation. To construct the SVM classifier proper kernel function and its

parameters has been chosen. Generally Sequential minimal optimization (SMO) is used for training SVMs, which is implemented by the popular application to find proper hyperplane for SVMs [11, 12]. Such implementation replaces all missing values and transforms nominal features into binary ones.

In the present study, the accuracy evaluation of SVM has been computed for linear, where the complexity parameter is set to 1.0.

SVM have been tested and trained to find out maximum accuracy adjusting their parameter. The performance measures such as accuracy, sensitivity, specificity of the classifiers are compared to each other. The parameters of the classifiers which provide maximum accuracy are selected to be compared to the other classifiers.

10-fold cross validation with 25 features used to compare the performances of the classifier. In input data of the classifier, and the test data are compared to the original class label to find out TP, TN, FP, and FN values. These values for classifiers are provided in the form of confusion matrix in Table 1.

Sensitivity refers successfully identified positive samples in cancer classification. Thus, higher sensitivity means the higher diagnostic capability of CDK patients and it can be used to help physicians to diagnose disease more correctly. The accuracy, sensitivity and specificity measures in this study are given in Table 2 to compare the effect of the feature using pattern recognition tools.

TABLE I. CLASSIFIER'S CONFUSION MATRIX

	Expected Value	
Actual Value	***Positive***	***Negative***
Positive	TP = 234	FN = 16
Negative	FP = 8	TN = 142

TABLE II. PERFORMANCE RESULTS OF SVM WITH LINEAR KERNEL

Total Number of Folds	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**Complexity Parameter (C)**
2	94.602	93.600	94.700	1.0
3	94.631	93.600	94.700	1.0
4	94.602	93.600	94.700	1.0
5	94.602	93.600	94.700	1.0
6	94.117	93.100	94.200	1.0
7	94.117	93.100	94.200	1.0
8	94.631	93.600	94.700	1.0
9	94.117	93.100	94.200	1.0
10	94.602	93.600	94.700	1.0

IV. CONCLUSION

The diagnosis of the CDK is a cumbersome problem. In the war zone the diagnosis process by using lab tests, imaging studies and biopsy might be a time-consuming, invasive, and costly. In this study, we propose automatic classification algorithm for kidney disease diagnosis based on clinical history, physical examinations, and laboratory tests, which are noninvasive, cheap and save. The performance measures of SVM classifier with linear kernels have been evaluated in order to find the best scores for sensitivity, specificity, and accuracy metrics. Experimental results showed that properly implemented classifier can reach the overall performance of 94.602%. The sensitivity value of SVM classifier with a linear kernel is 93.100%. Such computational studies are going to be vital for benefitting well beign of people and trying to identify diseases at early stages. In the future more compact and autonomous tools can be developed that could be used as a screening and also diagnosing mechanism as compare to other mentioned medical methods.

ACKNOWLEDGMENT

The authors thanks, Institute of Information and Computing Technologies in Almaty and Faculty of Engineering and Natural Sciences for providing needed resources for conducting this research study.

REFERENCES

[1] A. S. Levey, R. Atkins, and J . Coresh, "Chronic kidney disease as a global public health problem: approaches and initiatives - a position statement from Kidney Disease Improving Global Outcomes", *Kidney International*, vol. 72 no.3, pp. 247-259, Aug 2007.

[2] V . Jha, G. Garcia, and K. Iseki, "Chronic kidney disease: global dimension and perspectives", *Lancet,* vol. 382 no. 9888, pp. 260-272, Jul 2013.

[3] K.R Lakshmi, Y. Nagesh, and M. VeeraKrishna, "Performance Comparison of Three Data Mining Techniques for Predicting Kidney Dialysis Survivability", *International Journal of Advances in Engineering and Technology*, vol.7, no.1, pp. 242-254, March 2014.

[4] G. Caocci, R. Baccoli, R. Littera, S. Orrù, C. Carcassi and G. La Nasa, "Comparison Between an Artificial Neural Network and Logistic Regression in Predicting Long Term Kidney Transplantation Outcome", *Artificial Neural Networks Kenji Suzuki*, IntechOpen, DOI: 10.5772/53104, 2013.

[5] T. Di Noia, V. C. Ostuni, F. Pesce, G. Binetti, D. Naso, F. P. Schena, and E. Di Sciascio. "An end stage kidney disease predictor based on an artificial neural networks ensemble", *Expert Systems with Applications*, vol. 40, pp. 4438–4445, 2013

[6] A. Kusiak, B. Dixonb, and Sh. Shaha, "Predicting survival time for kidney dialysis patients: a data mining approach", *Computers in Biology and Medicine*, vol. 35, pp. 311–327, 2005

[7] J. Levman, T. Leung, P. Causer, D. Plewes, and A. L. Martel, "Classification of dynamic contrast-enhanced magnetic resonance breast lesions by support vector machines," *Medical Imaging,* IEEE Transactions on, vol. 27, pp. 688-696, 2008.

[8] N. H. Sweilam, A. Tharwat, and N. A. Moniem, "Support vector machine for diagnosis cancer disease: A comparative study," *Egyptian Informatics Journal,* vol. 11, pp. 81-92, 2010.

[9] E. Gumus, N. Kilic, A. Sertbas, and O. N. Ucan, "Evaluation of face recognition techniques using PCA, wavelets and SVM," *Expert Systems with Applications,* vol. 37, pp. 6404-6408, 2010.

[10] V. Vapnik, "The nature of statistical learning theory" *Springer Science and Business Media,* 2013.

[11] C.C. Chang and C.J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST),* vol. 2, pp. 27, 2011.

[12] K. Tufan, "Noninvasive diagnosis of atherosclerosis by using empirical mode decom-position, singular spectral analysis, and support vector machines," *Biomedical Research,* vol. 24, pp. 303-313, 2013.

[13] Soundarapandian P. (2015). UCI Machine Learning Repository[https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease]. Irvine, CA: University of California, School of Information and Computer Science.

[14] Sh. Shamiluulu, M.M. Boukar, Z. Yussupova. “Medical Tool for Assisting Patients in Kazakhstan Polyclinics”. *Proceedings:* 11th IEEE International Conference on Application of Information and Communication Technologies (ICECCO 2017). Abuja, Nigeria pp: 80-84.