Early Detection of Chronic Kidney Disease by Using Machine Learning Algorithms

**Amirgaliyev Ye. N. 1,2\*, Shamiluulu Sh. 1,2**

*1Institute of Information and Computing Technologies (IICT), Almaty, Kazakhstan*

2Department of Computer Science,

*Suleyman Demirel University, Kaskelen, Kazakhstan,* ***040900***

*E-mails:* yedilkhan.amirgaliyev@sdu.edu.kz, shahriar.shamiluulu@sdu.edu.kz

**Abstract**

Currently, due to many environmental factors like food, living standards many people might develop severe diseases without understanding of their condition. One of such diseases is a chronic kidney disease, which effects more than 35 million people worldwide according to World Health Organization. Diagnosing of chronic kidney diseases is generally invasive, costly, time-consuming and often risky for that reason preventive methods are advised in order to avoid developing such disease. In order to tackle this problem we propose a hybrid model based on seven machine-learning algorithms, which might help to identify the disease at an early stage. The experimental results showed that over 80% of success rate in classifying the patients with kidney diseases based on three performance metrics i.e., accuracy, sensitivity and specificity.

**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].

**II - MATERIALS AND METHODS**

**A. Dataset information**

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 refer Table 1.

|  |  |  |  |
| --- | --- | --- | --- |
| No | Feature Id | Description | Type |
| *1* | AGE | Patient age | Numerical |
| *2* | BP | Blood pressure | Numerical |
| *3* | SG | Specific Gravity | Nominal |
| *4* | AL | Albumin concertation in blood | Nominal |
| *5* | SG | Sugar level in blood | Nominal |
| *6* | RBC | Red blood cells | Nominal |
| *7* | PC | Pus cell | Nominal |
| *8* | PCC | Pus cell clumps | Nominal |
| *9* | BA | Presence of bacteria | Nominal |
| *10* | BGR | Blood glucose random | Numeric |
| *11* | BU | Blood urea | Numeric |
| *12* | SC | Serum creatinine | Numeric |
| *13* | SOD | Sodium concentration in blood | Numeric |
| *14* | POT | Potassium level in blood | Numeric |
| *15* | HEMO | Hemoglobin level | Numeric |
| *16* | PCV | Packed cell volume | Numeric |
| *17* | WC | White blood cell count | Numeric |
| *18* | RC | Red blood cell count | Numeric |
| *20* | HTN | Hypertension | Nominal |
| *21* | DM | Diabetes mellitus | Nominal |
| *22* | CAD | Coronary artery disease | Nominal |
| *23* | APPET | Appetite | Nominal |
| *24* | ANE | Anemia | Nominal |
| *25* | TRG | Class target feature | Nominal |

*Table 1.* Patient data features

There are 11 numerical and 14 nominal attributes. The nominal attributes were mapped to numerical values. 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. Exploratory data analysis**

During the feature engineering process, we have performed the spearman correlational analysis. The process revealed that the non-linear relationship between features see Figure 1. The aim is to select features that are less correlated (r ≤ 0.3) between each other as well as has less correlation to a target attribute.

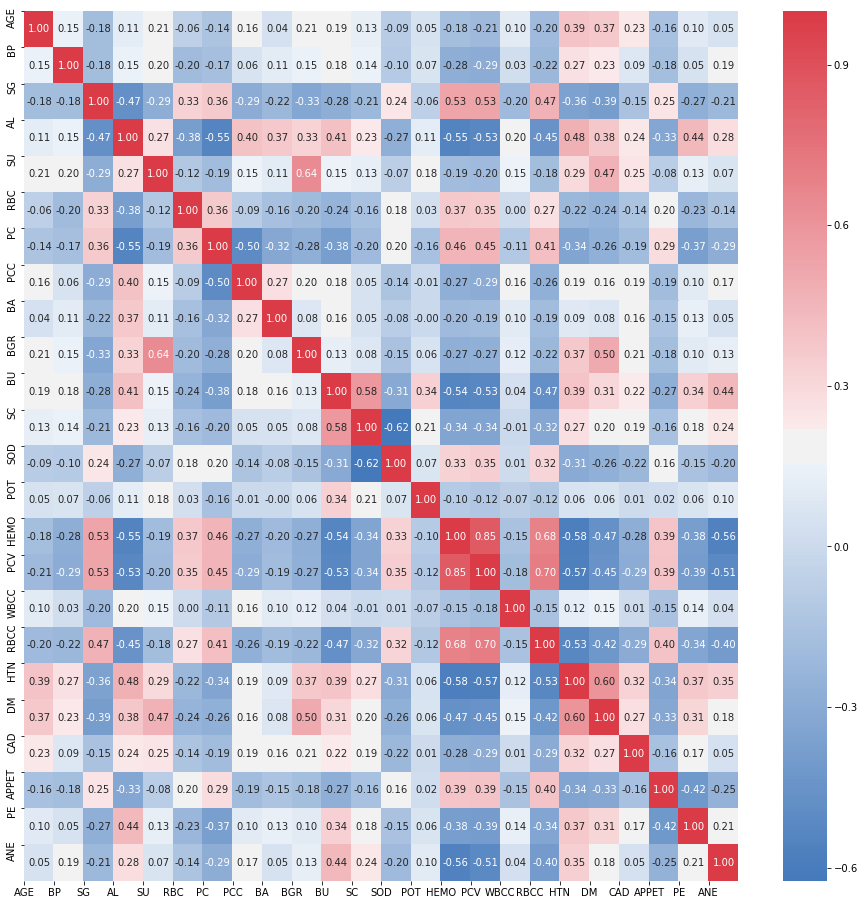


Figure 1. Correlation diagram of features

Based on the correlational analysis five attributes selected to train the hybrid model for the diagnosis of early chronic kidney disease which are provided in Table 2. Generally, these attributes are also related to other diseases but here we only consider CKD disorder.

|  |  |  |
| --- | --- | --- |
| No | Feature | Corr. value |
| *1* | BP | 0.27 |
| *2* | PCC | 0.22 |
| *3* | BA | 0.12 |
| *4* | POT | 0.03 |
| *5* | CAD | 0.21 |

Table 2. Features with low correlations

**C. Machine Learning Algorithms**

*i. K Nearest Neighbors*

The K Nearest Neighbors (KNN) algorithm is one of the simplest supervised learning algorithms. The logic behind this method is to find a predefined number of training samples closest in distance to the new point, and predict the label from these given data-points. Despite its simplicity, nearest neighbors has been successful in a large number of classification and regression problems. There are several distance metrics can be used like Euclidian, Manhattan or Hamming generally the Euclidean distance measure is used [1].

*ii. Decision Trees*

The Decision Tree (DT) is a graph that uses a branching method to illustrate every possible outcome of a decision for particular. The goal in building a tree is to identify a best splitting attribute which is being found by Entropy and Information Gain methods. The detailed theoretical background regarding decision trees for details [2].

*iii. Gaussian Naïve Bayes*

The Naïve-base (NB) is a classification technique based on Bayes’ Theorem. In general, the NB classifier assumes the presence of a particular feature in a class that is unrelated to the presence of any other feature. For example, a fruit may be considered to be an orange if it is orange, round, and about 10 cm in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as ‘Naive’. This method’s advantage is that Naive Bayes model is easy to build and particularly useful for very large data sets [1].

*iv. Logistic Regression*

The Logistic Regression (LR) is one of the strong algorithms used for classification. The LR model The distinctive feature of the model is that the outcome variable is dichotomous. The result is not bounded to a linear form. As a result, the created model can be used to classify a newly provided data via placing them in a model for the probability P, for more details [1].

*v. Multi-Layer Perceptron Neural Networks*

The Multi-layer perceptron (MLP) are processing devices, which are closely resemble a model of the neuronal structure of the mammalian cerebral cortex. Large MLPs might have hundreds or thousands of processor units, whereas a mammalian brain has billions of neurons with a corresponding increase in magnitude of their overall interaction and emergent behavior. Generally, the neural network has three components or layers. The first layer is called input layer, through which it gets data inside network on our case disease related attributes. The second layer is called hidden where all operations performed. The last layer called out where network make final decision regarding patient’s condition. The detailed information regarding neural networks provided in [1, 2].

*vi. Support Vector Machines*

The Support vector machines (SVM) is a supervised learning algorithm that is used for data classification and regression [11]. It searches for a best hyperplane which best 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].

*vii. Random Forest*

The Random Forest (RF) is a supervised ensemble-learning algorithm used for classification and regression tasks. Ensemble learning models aggregate multiple machine learning models, allowing for overall better performance. The logic behind this is that each of the models used is weak when employed on its own, but strong when put together in an ensemble. In the case of Random Forests, a large number of Decision Trees, acting as the “weak” factors, are used and their outputs are aggregated, with the result representing the “strong” ensemble.

Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." *R news* 2.3 (2002): 18-22.

**D. Performance Measures**

The classifiers evaluations are commonly evaluated based on the data in the confusion matrix. Several standard measures have been defined for correct and incorrect classification results of the matrix. The most common practical measure to evaluate the performance is accuracy, which is defined as the proportion of the total number of instances that were classified correctly.

*Recall* is the mean proportion of actual positives which are correctly identified. *Precision* is the mean proportion of positives which are relevant. *F-measure* is a harmonic mean of recall and precision. *TP rate* is a measure which shows the matching states of particular instances. *FP rate* is a measure that shows the mismatching states of particular instances. There are also additional performance metrics like positive and negative predictive values, and mean square error.

**III - RESULTS AND DISCUSSIONS**

The experimental results show a promising outcomes related to predicting the chronic kidney disease on early stages from less correlated features. In order to decrease the problem of overfitting a 10-fold cross validation performed. From the Table 3, we can observe that Random Forest algorithm outperforms others.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| k-folds | NB | SVM | RF | DT | KNN | ANN | LR |
| *1* | 0.582 | 0.651 | 0.823 | 0.715 | 0.742 | 0.691 | 0.592 |
| *2* | 0.582 | 0.494 | 0.842 | 0.855 | 0.721 | 0.691 | 0.498 |
| *3* | 0.531 | 0.689 | 0.943 | 0.795 | 0.831 | 0.691 | 0.689 |
| *4* | 0.651 | 0.693 | 0.788 | 0.705 | 0.788 | 0.753 | 0.787 |
| *5* | 0.712 | 0.702 | 0.883 | 0.931 | 0.845 | 0.811 | 0.712 |
| *6* | 0.564 | 0.736 | 0.842 | 0.895 | 0.748 | 0.657 | 0.744 |
| *7* | 0.564 | 0.744 | 0.811 | 0.804 | 0.744 | 0.785 | 0.731 |
| *8* | 0.564 | 0.731 | 0.998 | 0.942 | 0.867 | 0.785 | 0.702 |
| *9* | 0.595 | 0.702 | 0.834 | 0.821 | 0.845 | 0.691 | 0.701 |
| *10* | 0.492 | 0.802 | 0.973 | 0.832 | 0.787 | 0.694 | 0.695 |
| *Avg. score* | 0.584 | 0.694 | 0.874 | 0.829 | 0.792 | 0.725 | 0.685 |

Table 3: 10-kfolds cross-validation results

In addition, there are similar patterns of how these algorithms performing classification during the cross-validation process. For instance RF, DT and KNN has somewhat similar learning path see Figure 2.

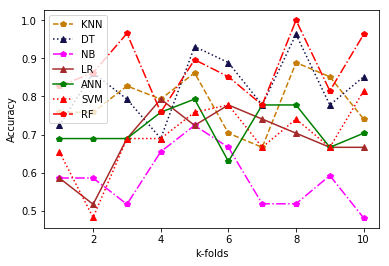


Figure 2: 10-folds cross-validation overview

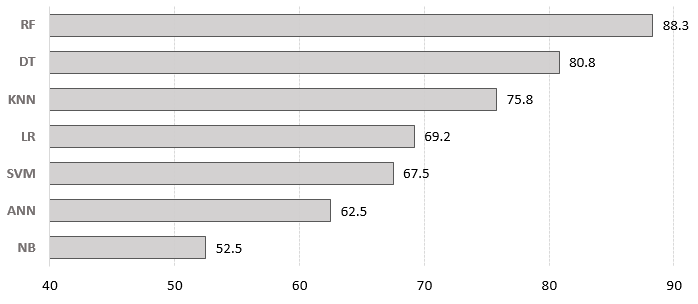


Figure 3. Accuracy parameter of the test process

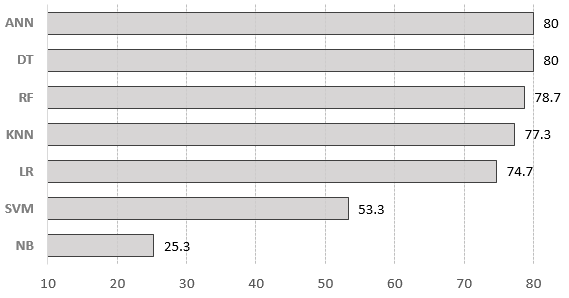


Figure 4. Sensitivity parameter of the algorithms

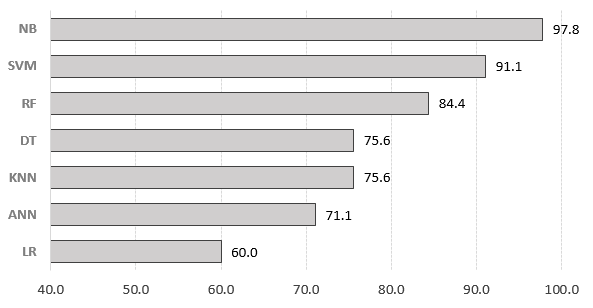


Figure 5. Specificity parameter of the algorithms

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithms | PPV | NPV | MSE |
| NB | 0.95 | 0.44 | 0.475 |
| SVM | 0.909 | 0.539 | 0.325 |
| RF | 0.894 | 0.704 | 0.192 |
| DT | 0.845 | 0.694 | 0.217 |
| KNN | 0.841 | 0.667 | 0.233 |
| ANN | 0.822 | 0.681 | 0.233 |
| LR | 0.757 | 0.587 | 0.308 |

Table 4. Positive predictive value, Negative Predictive Value

and Mean Squared Error

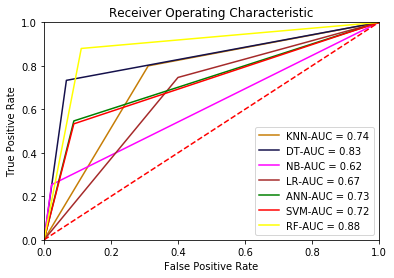


Figure 6. ROC curve for algorithms

**IV CONCLUSIONS**

Acknowledgements

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

**REFERENCES**

[1] Norvig P, Russell S. Artificial Intelligence: A Modern Approach, Prentice Hall. 2002.

[2] Dunham MH. Data mining: Introductory and advanced topics, Pearson Education, 2006.

[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.

[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.

[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.

[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.