Detection of chronic kidney disease in the north central province of Sri Lanka using effective feature selection and machine learning

D.V. Dissanayake [1,*], S. Sobana [1] and B. Yogarajah [1]

[1] Department of Physical Science, University of Vavuniya, Sri Lanka.
[*] Corresponding author email: deshandissanayake1996@gmail.com

Abstract: About 10% of the adult population globally is afflicted by chronic kidney disease (CKD), one of the top 20 causes of death worldwide. Sri Lanka, one of the countries most severely afflicted by CKD, has been identified as having a high prevalence of CKD in 10 out of 25 districts, including Anuradhapura. Patients frequently overlook the disease in the early stages of CKD since there are no obvious symptoms. Therefore, it is critical to identify CKD early to provide patients with prompt care and slow the disease's progression. The present manual methods for CKD detection have a number of drawbacks, including a lack of specialized doctors, high expenses for diagnosis and treatment, particularly in developing nations; a long detection period and low accuracy. In recent years, early CKD detection has been greatly aided by machine learning techniques. In this study, we developed a diagnosis system to detect chronic kidney diseases at any stage. Importantly, the data was collected from Anuradhapura district, Sri Lanka. Data preprocessing was performed, especially filling missing values using the *k*–nearest neighbor (KNN) imputation method with different *k* values (*k*= 3, 5, 7, 9, and 11). Stepwise forward and backward selection methods were used then for selecting the optimal features in the CKD dataset for all *k* values. To create the most accurate model using machine learning algorithms, four machine learning algorithms (Multiple Logistic Regression, Random Forest, Support Vector Machine, and KNN) were utilized and the accuracy compared with each k dataset. The Random Forest classification algorithm is given the highest accuracy of 96% and other evaluation metrics for the *k* values 3, 5, 7, and 9. The system is based on Sri Lankan data, which will aid those who can detect CKD early on.

Keywords: CKD diagnosis system, *K* value, Machine learning algorithm