Heart disease prediction using effective feature selection and machine learning – A Sri Lankan study

S. Weerakkodi [1,*] and S. Sobana [1]

[1]   Department of Physical Science, University of Vavuniya, Sri Lanka.

*   Corresponding author email: sathsaraweerakkodi27@gmail.com

Abstract: One of the main causes of death worldwide is heart disease. The latest WHO figures show that 26,304 deaths from coronary heart disease, or 23% of all fatalities, occurred in Sri Lanka in 2020. The diagnosis and treatment of heart disease in its early stages is very important since some types of heart diseases are preventable and considering heart disease's early stages do not have any obvious signs. Importantly, early identification of heart disease enables patients to get treatment right away to halt the disease's progression. The cost of diagnosing and treating heart disease is relatively significant, especially in developing countries. In this paper, we developed a system to classify the chances of heart diseases at an earlier stage using efficient feature selection and machine learning techniques based on the newly collected Sri Lankan data. The medical data of heart disease patients and normal subjects with 13 features was collected from the Kandy general hospital, Sri Lanka, with ethical approval. Four feature selection techniques, backward feature elimination, forward feature selection, recursive feature elimination, and exhaustive feature selection and six classification approaches, *K*-nearest neighbor, support vector machine, Gaussian naive Bayes, random forest, logistic regression, and decision tree, applied to the dataset. Different evaluation criteria (precision, accuracy, F1-score, and recall) are used to analyze the effectiveness of the feature selection methods and machine learning techniques. The feature subset selected by the recursive feature elimination technique has achieved the highest classification accuracy of 95%, precision of 100%, sensitivity of 91%, and f-measure of 95% with the logistic regression classifier. This system will help people identify heart disease at earlier stages. Heart health and quality of life will be enhanced through the elimination of heart disease risk factors and the identification and treatment of these factors.

Keywords: Classification approach, Feature selection technique, Heart disease prediction, Logistic regression