Abstract No: PO-31

Exploring data mining avenues in β-Thalassemia carrier identification

G. K. Subasinghe[1*], N. V. Chandrasekara[1] and A. P. Premawardhena[2]

[1]Department of Statistics & Computer Science, Faculty of Science, University of Kelaniya, Sri Lanka
[2]Department of Medicine, Faculty of Medicine, University of Kelaniya, Sri Lanka
gayathri.subasinghe97@gmail.com*

Thalassemia is a genetic blood disorder that affects the production of haemoglobin and is a global health problem. In comparison to many other nations in the region, Sri Lanka also has a high prevalence of thalassemia. The traditional methods for identifying thalassemia carriers, such as genetics and blood tests, are expensive and time-consuming and may not be available to all demographic groups. Nevertheless, the use of data mining models for thalassemia carrier detection is still in its infancy, and there are few studies on its efficacy. Therefore, it is vital to investigate the efficacy and accuracy of data mining approaches for detecting thalassemia carriers, as well as the viability of employing these methods in clinical practice. Thus, the objective of this study is to develop a time-efficient model to detect the β-thalassemia carriers, which can reduce the time to take a decision and develop the built model as a decision support tool. Also, the earlier detection will help individuals to refer to necessary treatments further. This study is carried out with the data obtained from Hemal's Adolescent and Adult Thalassemia Care Centre, Mahara, one of the treatment centres for thalassemia. As the study population, 343 individuals' data values were considered from August 2019 to December 2019. When processing the dataset, 112 (36%) individuals were declared as β-thalassemia carriers, whereas 200 (64%) were identified as β-thalassemia non-carriers. Eight blood parameters, such as RBC, HGB, HCT, MCV, MCH, MCHC, RDW and HbA2 were identified by revealing the literature and the Chi-square and Mann-Whitney U tests were used to identify the association between the variables at 5% level of significance. A random over-sampling technique was used to overcome the class-imbalanced problem in the dataset, and based on that, model fitting was performed under the two data selection methods, i.e., Method 1: Model fitting before handling the class imbalance problem and Method 02: Model fitting with random over-sampling technique. Then 80% of the data was used for training the models, and 20% of the data was used for the evaluation. Support Vector Machine (SVM) and Probabilistic Neural Network (PNN) models were used to detect the β-thalassemia carriers. In comparison among methods, the better-performing models were given under Method 2, and the PNN model fitted under Method 2 (PNN Model 2) exhibits 98.75% overall classification accuracy. Here, the PNN model's network architecture consisted of eight nodes in the input layer, 320 nodes in the pattern layer, two nodes in the summation layer, and two nodes in the output layer. Further, the fitted PNN Model 2 can be utilised as a cost-effective and time-saving option to detect β-thalassemia carriers in a few seconds with acceptable accuracy and can be implemented as a decision support tool. However, it is recommended to get advice from a medical doctor for further investigation.

Keywords: Class-imbalance, Support Vector Machine, Probabilistic Neural Network, β- thalassemia carriers