This paper presents a detailed predictive analysis of heart disease conditions of some patients to determine the possible risk factors that are suggestive of whether a patient in the sample has heart disease or not. The predictive strength of the identified risk factors was further determined using suitable classification methods. Two independent (but similar) published heart disease data, the Cleveland (training) and the Statlog (validation), were considered in this paper. Detail exploratory analysis using the Chi-square test of independence was performed to get more insight about the features that are useful in determining whether a patient has heart disease or not before employing ten standard machine learning techniques for the class prediction and the results showed that some bio-clinical categorical variables are strongly associated with the heart disease conditions of the patients in the two data sets (p < 0.001). We then randomly partitioned the Cleveland data into 70% (208) training samples for building the models and 30% (89) test samples for evaluating their predictive performances over 200 replications. These results were validated on the Statlog data in 10-fold cross-validation. Results from this study indicated that the Support Vector Machine (SVM) yielded the best predictive performances with 85% Accuracy, 82% Sensitivity, 87% Specificity, 87% Precision, 91% Area under the ROC curve (AUC), and 38.5% LogLoss value.

The result presented in this work is original. More importantly, the prediction results provided by the best-chosen classifier, SVM, on the Cleveland (training) data were all consistent with the validation results obtained from the Statlog data. Therefore, the strength of this work lies in the ability of the SVM classifier to reproduce the prediction results obtained when applied on similar independent heart disease data sets.