ABSTRACT

Heart diseases are a major health concern globally and early detection of the disease plays a crucial role in reducing mortality rates and improving patient outcomes. With the advancements in machine learning, there is an increasing focus on utilizing these techniques to enhance the prediction and diagnosis of heart diseases, even before symptoms manifest. Machine learning (ML) research in this field aims to develop accurateand efficient models that can assist in early detection, risk assessment, and clinical decision-making. There are a variety of ML methods making use of stand-alone classifiers and hybrids. However, the results from these models vary considerably between various cardiac datasets and/or they are not modeled using both low and high dimensional data. With increased dimensionality of data, it becomes imperative to address shortcomings of existing feature selection and prediction approaches in handling such datasets with complicated feature relationships and significant degrees of redundancy. This research identifies and addresses issues with existing feature selection and classification methods and proposes novel and improved feature selection and classification techniques towards enhancing and improving heart disease prediction performance.

In the first stage of work, feature selection using Feature Importance (FI) ranking of Gradient Boosting algorithms is done and a significant reduction in the search space of feature subsets is identified. Next, this research work proposes a novel feature selection algorithm called ModifiedBoostARoota (MBAR), which identifies the risk parameters that strongly contributes to the prediction of heart disease. This algorithm incorporates CatBoost as the base model and utilizes a novel feature elimination process.

In the second phase of the work, a novel Super Learner Ensemble Model (SLEM) is proposed to perform on features selected by MBAR. The SLEM model is an integration of diverse ML base models selected by repeated stratified k-fold cross validation. A meta learner logistic regression is employed to learn from the predictions of the base classifiers. By backward elimination method, an optimal combination of classifiers in SLEM was identified as Catboost and Decision Tree, in order to improve the classification time complexity and performance. The performance of the SLEM model improved when used on features selected by MBAR compared to its performance on datasets with no feature selection.

In the third phase of the work, to further improve the classification by selecting an optimal combination of base classifiers in the Super Learner Ensemble Model (SLEM) model, a new Optimized Super Learner Ensemble Model (OSLEM) is proposed. OSLEM utilizes the Whale Optimization Algorithm and pairwise divergence measure to select an optimal base classifier combination. The performance of the final proposed model where ModifiedBoostARoota algorithm is used for feature selection and Optimized Super Learner Ensemble Model (OSLEM) is used for classification was analyzed on both lowand high dimensional heart datasets. The proposed model presented high performance in terms of precision, recall, f1-score, specificity and accuracy, when compared with existing models across the heart datasets. The proposed model improved prediction accuracy while being robust against overfitting.

Overall, the contributions from this research work would improve the accuracy, efficiency, and decision-making support of heart disease prediction systems, ultimately benefiting clinical practice.