ABSTRACT

Tuberculosis (TB) remains one of the world's leading causes of death, and its diagnosis becomes even more complex when co-infected with HIV. In recent years, gene expression profiling from microarray data has emerged as a promising approach for TB detection. However, the inherent high dimensionality and limited sample size of such data pose serious challenges for machine learning models, often leading to unstable performance. Conventional ensemble methods, including Bagging, AdaBoost, and Uniform Weighted Ensemble, have been applied in this domain but tend to produce unbalanced results—achieving high precision while sacrificing recall—which can be critical in clinical decision-making where missing a positive case is unacceptable. To address these challenges, this study proposes a weighted ensemble classification framework optimized using the Firefly Algorithm (FA). The method combines multiple base classifiers under a bagging framework, assigning adaptive weights determined by FA to enhance predictive balance and robustness. The dataset, consisting of gene expression profiles from TB-HIV co-infected patients, was normalized and reduced in dimensionality through feature selection before training. Experimental evaluation revealed that the best baseline model, Bagging with Decision Tree, achieved an AUC of 0.7755 and perfect precision (1.0000) but a low recall (0.4286). In comparison, the proposed FA-optimized Decision Tree ensemble achieved superior and more balanced performance, with an AUC of 0.8367, accuracy of 0.7857, precision of 0.8333, and recall of 0.7143. These results demonstrate that integrating ensemble learning with metaheuristic optimization can effectively handle high-dimensional, small-sample biomedical data while maintaining clinically meaningful predictive performance.

Keywords: Tuberculosis detection, HIV-positive, gene expression data, weighted ensemble, Firefly Algorithm, bagging, classification