ABSTRACT

Osteoporosis is a bone disease for which the discovery of conventional drugs is time-consuming and costly. This study focuses on predicting the bioactivity of Cathepsin K enzyme inhibitor compounds, which are important targets for osteoporosis therapy. The lifetime risk of osteoporosis in women and men ranges from 40%–50% and 13%–22%, respectively. However, the conventional drug discovery process is often time-consuming and expensive. Although previous studies have applied machine learning, there is still room for improvement in performance, particularly in the feature selection process from high-dimensional chemical data. Feature selection optimization is key to building more efficient and accurate predictive models. The proposed solution is to build a predictive model by integrating the Glowworm Swarm Optimization (GSO) algorithm for feature selection and XGBoost as the classification model. The Cathepsin K inhibitor dataset was processed and divided into four bioactivity classes. The GSO algorithm was applied to select the most relevant and informative subset of molecular descriptor features from hundreds of initial features. Subsequently, the XGBoost classification model was optimized through hyperparameter tuning to maximize its predictive performance based on the features selected by GSO. Testing results showed that the best-optimized GSO-XGBoost model achieved peak performance with an accuracy of 0.84 and a weighted average F1-Score of 0.84. This study contributes an effective framework for combining GSO and XGBoost for bioactivity prediction, which has the potential to enhance efficiency in the early stages of drug discovery for bone diseases.

Keywords: Osteoporosis, Cathepsin K, GSO, XGBoost, Inhibitors Bioactivity, Prediction.