Abstract

Cancer has become one of the deadliest diseases in the world, mainly caused by the accumulation of somatic and inherited mutations. However, this phenomenon can be traced back to the molecular level, specifically, to proteins. Proteins are molecules responsible for various bioprocesses in the human body through their interactions with other molecules. Abnormalities in these interactions can lead to various undesirable outcomes, including disease and cancer. Peptides have the potential to serve as molecules that can be used in protein interactions to treat cancer. However, identification of peptides corresponding to target proteins in the laboratory is time-consuming and expensive. Therefore, there is a need for computational methods to aid identification. TabNet, a deep learning-based computational method was used in this study. For comparison purposes, we selected techniques from ensemble learning, including Random Forest and Extreme Gradient Boosting, along with methods from deep learning such as Convolutional Neural Network and Stacked Autoencoder-Deep Neural Network. Predictions are performed on a multi-feature peptide-protein interaction dataset, and the features include position-specific scoring matrices, intrinsic disorder, amino acid sequence, and physicochemical properties. Among our selected metrics, we found that TabNet achieved a better score in AUC of 0.7 and lower false negatives compared to other models.

***Keywords*: Cancer, Peptide-Protein Interaction, TabNet**