Analyzing RNA-Seq gene expression data using deep learning approaches for cancer classification

ABSTRACT

Ribonucleic acid Sequencing (RNA-Seq) analysis is particularly useful for obtaining insights into differentially expressed genes. However, it is challenging because of its high-dimensional data. Such analysis is a tool with which to find underlying patterns in data, e.g., for cancer specific biomarkers. In the past, analyses were performed on RNA-Seq data pertaining to the same cancer class as positive and negative samples, i.e., without samples of other cancer types. To perform multiple cancer type classification and to find differentially expressed genes, data for multiple cancer types need to be analyzed. Several repositories offer RNA-Seq data for various cancer types. In this paper, data from the Mendeley data repository for five cancer types are analyzed. As a first step, RNA-Seq values are converted to 2D images using normalization and zero padding. In the next step, relevant features are extracted and selected using Deep Learning (DL). In the last phase, classification is performed, and eight DL algorithms are used. Results and discussion are based on four different splitting strategies and k-fold cross validation for each DL classifier. Furthermore, a comparative analysis is performed with state of the art techniques discussed in literature. The results demonstrated that classifiers performed best at 70–30 split, and that Convolutional Neural Network (CNN) achieved the best overall results. Hence, CNN is the best DL model for classification among the eight studied DL models, and is easy to implement and simple to understand.