[BIOINFO 2023]

The Effectiveness of Transfer Learning in Predictive Modeling for Pediatric Cancers: Insights from VAECox Architecture

Taewon Kim[1], Bo Kyu Choi[1], Jin Yong Kim[1], and Yu Rang Park[1,*]

[1] *Department of Biomedical system informatics, Yonsei University*
**Corresponding author: yurangpark@yuhs.ac*

Background:

Transfer learning through large datasets can improve the performance of learning in the primary task. In medical field, whether this benefit can be maintained when the target diseases in the pretraining phase and in the primary task is an area of interest. We analyzed mortality data from pediatric cancer patients to validate this concept by using VAECox structure, a deep learning model architecture with two step approach integrating Variational Autoencoders (VAEs) with a classifier.

Method:

Data was extracted from Therapeutically Applicable Research to Generate Effective Treatments (TARGET) and The Cancer Genome Atlas (TCGA). TARGET is a project for pediatric cancer research and TCGA is a project that provides information on a wide range of cancer genomes. From TCGA, 8,754 samples across cancers (e.g., BRCA, BLCA, HNSC) were divided into 7,003 training, 875 validation, and 876 testing groups. From TARGET, 2,870 samples of pediatric including Acute Lymphoblastic Leukemia in both phase II (ALL-P2) and phase III (ALL-P3), Acute Myeloid Leukemia (AML), Osteosarcoma (OS), and Wilms Tumor (WT) were divided into 2,296 training, 287 validation, and 287 testing groups. Model 1 trained both the encoder and classifier by using only the TARGET data and Model 2 trained the encoder using the TCGA data and predicted on the TARGET data. A 5 fold cross validation was done and it was repeated 10 times to evaluate performance.

Result:

In four types of pediatric cancer, Model 2 outperformed Model 1 with a C-index margin of 0.021, 0.007, 0.009, and 0012 for ALL-P3, AML, OS, and WT, respectively. The only exception was for ALL-P2, where Model 1 surpassed Model 2 with a C-index margin of 0.006.

Discussion:

The model pretrained in the TCGA dataset yielded better predictions in pediatric cancer mortality data. TCGA dataset is substantially more extensive compared to the TARGET dataset, but it consists of only adult cancer data and did not include any cancer types for mortality prediction. This suggests that pretraining models on larger datasets can improve effectiveness of predictive modeling even when the source and target datasets are from different diseases.