Machine learning-based prediction of survival in cancer using multi-omics data

Ayşe Nur Çoruh[1,2], Tunca Doğan[1,2,*]

[1]Biological Data Science Lab, Department of Computer Engineering, Hacettepe University, Ankara, Turkey
[2]Department of Bioinformatics, Graduate School of Health Sciences, Hacettepe University, Ankara, Turkey
*To whom correspondence should be addressed: tuncadogan@gmail.com

Today, cancer is one of the leading causes of death worldwide, according to the World Health Organization. The high lethality of some of the sub-types of cancer increases the importance of correct diagnosis, complete follow-up and effective treatment. Survivability in cancer can be defined as the length of time that patients live after the diagnosis and/or the administration of a certain treatment. The estimation of survival, which is a critical topic in biomedicine, is possible using relevant indicators and historical patient data. For this, computational methods such as machine learning techniques and statistical approaches has been utilized. Until lately, researchers mainly used clinical and demographic data of patient to model survivability, which generally resulted in low success, due to ignoring patient-specific molecular properties that affect both the response given to a treatment and the progression of the disease in general. To solve this problem, personalized medicine-based approaches have been developed and used in cancer research in recent years. As different type of “omics” data is getting easier to be produced/obtained by the ordinary lab, more data has been accumulated on public servers, which allowed computational scientists to build more successful prediction models. There are studies in the literature that use single-omic data to predict survivability in cancer; however, the utilization of multi-omics data is still understudied in this context. Here, the research question is that, “would it be possible to construct more successful survival prediction models by diversifying the input data used during modeling?”.

In this study, we proposed a new computational method to predict the survival of cancer patients. For this purpose, we utilized multi-omics data of patients diagnosed with 1 of the 13 different types of cancer, which are obtained from Genomic Data Commons (GDC) data portal. GDC contain data from different cancer projects, carried out within the scope of The Cancer Genome Atlas (TCGA) Program. In our study, we used mutation, copy number variation (CNV), gene expression, and miRNA expression as our input omic data types. In addition, we incorporated the clinical data and administered drug information of the patients in our dataset, to our input features. We divided patients into two survival groups via determining a specific survival time-based threshold for each tissue/cancer type. In terms of the input genes, we employed (i) the whole human genome, and (ii) genes in the L1000 (landmark) set, in two different settings, to be able to evaluate the better choice in terms of dimensionality and the computational cost. We utilized the random forest algorithm and trained 13 tissue/cancer specific binary classification models (i.e., classes are 0: patient dies before reaching the threshold duration, such as 3 years, 1: patient lives longer than the threshold duration). We employed leave one out cross validation (LOOCV) strategy to calculate performance scores and Kaplan–Meier plots for the evaluation of prediction output. Figure 1 summarizes the overall workflow of the study.

According to our results, models that use multiple types of omic data achieved better prediction performances, compared to the models using a single-omic. We also found that the use of clinical data and drug information of patients, employed in addition to the multi-omics data, further increased the prediction performance, in most tissues. Among different types of omics data, models that utilize mutation and gene expression features obtained the highest prediction performance, in the majority of the tissues. Breast and kidney tissue models performed better than the models of other tissues, probably due to lower heterogeneity among patient signatures. Utilizing the L1000 genes reduced both the noise and the curse of dimensionality, and provided a better performance, as opposed to using the whole genome at the input level. We also observed that applying feature selection during the data preprocessing step did not significantly improve the prediction performance.

These results confirmed the idea that multiple omics data would be successful in determining cancer patients' survival. For future studies, we plan to incorporate additional types of data such as proteomics, lipidomics, glycomics, and etc. where available. We’ll also be trying novel machine/deep learning algorithms, especially in the framework of multi-modal learning.

Keywords: Survival prediction, machine learning, cancer research, multi-omics-based analysis.

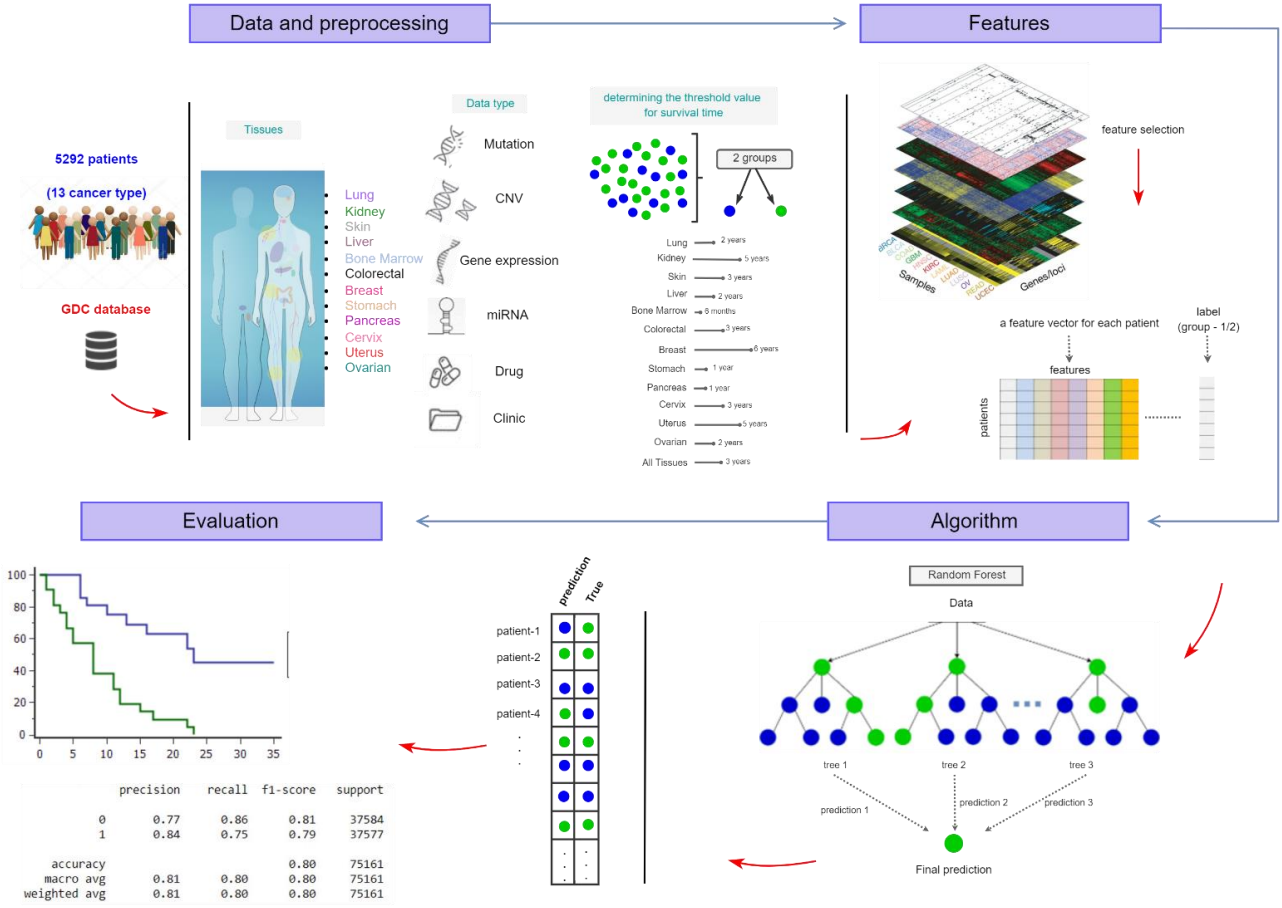

Figure 1. Workflow of the proposed multi-omics-based cancer survival prediction system.