DOPE: a predictive multi-omic model for cancer prognosis

Cancan Huang, Reetam Ganguli, Laura McCallion, Anessa Petteruti
{cancan_huang, reetam_ganguli,
laura_mccallion, anessa_petteruti}@brown.edu

Brown University — October 15, 2021

Literature Review

Cancer has been ranked as the second leading cause of death with about 1 out of 10 adults in the United States been diagnosed with cancer (Aruleba *et al.*, 2020; Siegel *et al.*, 2019). Gynecologic malignancies account for approximately 12% of all new cancer cases and 15% of all female cancer survivors (Salani *et al.*, 2017). In the United States, approximately 84,000 new cases of gynecologic malignancies are diagnosed resulting in about 2,800 deaths annually (Stewart *et al.*, 2013). The scale of this disease has motivated global efforts to save patients from the hand of devil.

One important goal for current cancer prognosis is to provide guidance for treatment to achieve better survival rate on the basis of patients' clinical profile. RP *et al.*, 1950 firstly computed the Life Table to obtain the enhanced frequency distribution of patients' survival times, followed by a nonparametric method proposed by Kaplan and Meier, 1958 for survival curve as well as the Proportional Hazards (PH) regression model by Cox, 1972. While those statistical methods have been widely used, most of them are mainly focuses only on specific clinical data, such as cancer types, cancer diagnosis and etc.

In recent years, researchers aiming at early diagnosis and curing these diseases have been fueled by many advances in experimental techniques, which output high throughput and high dimensional multi-omics data of patient samples (Wooster and Weber, 2003; Goossens *et al.*, 2015). With the development of experimental techniques and the emerging abundant data, for example, genomics data (i.e., whole genome data), expression data (i.e., mRNA data) and epigenetic data (i.e., chromosomal modifications), it is necessary and demanding for developing efficient and effective computational methods capable of understanding and handling those multi-omics data for more accurate cancer prognosis(Zhu *et al.*, 2020).

To solve these problems and take advantage of the massive data, other methods, including machine learning techniques are applied. Alexe *et al.* (Alexe *et al.*, 2007) combined principal components analysis (PCA) and k-clustering for breast cancer progression analysis. Xu *et al.*, 2012 adopted support vector machine (SVM) for breast cancer prognosis on gene expression dataset. Graf *et al.*, 2021 use genetic association study find the strong correlation between the copy number variation (CNV) signatures and the ovarian cancer. Other machine learning approaches, such as Bayesian networks, decision trees and semi-supervised learning, have also been applied in cancer prognosis prediction and shown good performance (Kourou *et al.*, 2015).

Deep learning, a branch of machine learning, shows promise in excellent performance in recent years due to the establishment of public accessible large-scale cancer databases as well as breakthroughs in model architectures (Zhu *et al.*, 2020). For example, The Cancer Genome Atlas (TCGA) database, containing both clinical and molecular data from over 11,000 tumor patients covering 33 different cancer types, has been widely used for different tasks (Tomczak *et al.*, 2015). Ching *et al.*, 2018 developed fully-connected neural network model for predicting survival time. The model is trained on TCGA gene expression data, clinical data as well as survival data and achieved better performance than Cox methods and random forest. Guo *et al.*, 2020 proposed a pipeline to identify ovarian cancer subtypes based upon multi-omics ovarian cancer features (mRNA, miRNA, and CNV). A denoising autoencoder is used to generate low dimensional representation from the multi-omics ovarian cancer features. K-means clustering is then labeled reconstructed features for ovarian cancer subtypes. After obtaining the labels, a simple logistic regression model with only mRNA as input is used for the final identification. Sun *et al.*, 2019 collected point mutations from healthy tissues and tumor tissues as input for their deep neural network (DNN) model, which exhibited comparable performance in classification of 12 types of cancers.

The successes of deep learning model in utilizing the massive data of multiple types are delightful. **Therefore, in our project, we aim at implementing a deep learning model trained with multi-omic**

data for predicting some medical outcomes, such as survival time, recurrence and etc, for ovarian and breast cancer patients.

References

Alexe, G., Dalgin, G., Ganesan, S., Delisi, C., and Bhanot, G. (2007). Analysis of breast cancer progression using principal component analysis and clustering. *Journal of biosciences*, **32**(1), 1027–1039.

Aruleba, K., Obaido, G., Ogbuokiri, B., Fadaka, A. O., Klein, A., Adekiya, T. A., and Aruleba, R. T. (2020). Applications of computational methods in biomedical breast cancer imaging diagnostics: A review. *Journal of Imaging*, **6**(10), 105.

Ching, T., Zhu, X., and Garmire, L. X. (2018). Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS computational biology*, **14**(4), e1006076.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, **34**(2), 187–202.

Goossens, N., Nakagawa, S., Sun, X., and Hoshida, Y. (2015). Cancer biomarker discovery and validation. *Translational cancer research*, **4**(3), 256.

Graf, R. P., Eskander, R., Brueggeman, L., and Stupack, D. G. (2021). Association of copy number variation signature and survival in patients with serous ovarian cancer. *JAMA Network Open*, **4**(6), e2114162–e2114162.

Guo, L.-Y., Wu, A.-H., Wang, Y.-x., Zhang, L.-p., Chai, H., and Liang, X.-F. (2020). Deep learning-based ovarian cancer subtypes identification using multi-omics data. *BioData Mining*, **13**(1), 1–12.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, **53**(282), 457–481.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, **13**, 8–17.

RP, G. *et al.* (1950). Calculation of survival rates for cancer. In *Proceedings of the staff meetings. Mayo Clinic*, volume 25, pages 270–286.

Salani, R., Khanna, N., Frimer, M., Bristow, R. E., and Chen, L.-m. (2017). An update on post-treatment surveillance and diagnosis of recurrence in women with gynecologic malignancies: Society of gynecologic oncology (sgo) recommendations. *Gynecologic oncology*, **146**(1), 3–10.

Siegel, R. L., Miller, K. D., and Jemal, A. (2019). Cancer statistics, 2019. *CA: a cancer journal for clinicians*, **69**(1), 7–34.

Stewart, S. L., Lakhani, N., Brown, P. M., Larkin, O. A., Moore, A. R., and Hayes, N. S. (2013). Gynecologic cancer prevention and control in the national comprehensive cancer control program: progress, current activities, and future directions. *Journal of Women's Health*, **22**(8), 651–657.

Sun, Y., Zhu, S., Ma, K., Liu, W., Yue, Y., Hu, G., Lu, H., and Chen, W. (2019). Identification of 12 cancer types through genome deep learning. *Scientific reports*, **9**(1), 1–9.

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary oncology*, **19**(1A), A68.

Wooster, R. and Weber, B. L. (2003). Breast and ovarian cancer. *New England Journal of Medicine*, **348**(23), 2339–2347.

Xu, X., Zhang, Y., Zou, L., Wang, M., and Li, A. (2012). A gene signature for breast cancer prognosis using support vector machine. In *2012 5th International conference on biomedical engineering and informatics*, pages 928–931. IEEE.

Zhu, W., Xie, L., Han, J., and Guo, X. (2020). The application of deep learning in cancer prognosis prediction. *Cancers*, **12**(3), 603.