Addressing Majority biased prediction of Cancer from Gene Expression

Kushankur Ghosh, Mehrnoosh Bazrafkan, Ramon Diaz Ramos

I. Background

Genes carry the instructions to make proteins, which do much of the work in our cells. Many proteins are switched on or off (gene activation or gene deactivation) that intensely change the whole activity of the cell. Moreover, a gene not typically represented in that cell can be switched on and expressed at high levels. This can result from gene mutation or changes in gene regulation (epigenetics, transcription, post-transcription, translation, or post-translation). So we can define cancer as a disease of transformed gene expression. Many machine learning techniques predict and model gene expression events, including mRNA and protein levels. One of these techniques is deep neural networks. The neural network automatically learns informative sequence representations, and interpreting them enables us to improve our understanding of gene expression. Many deep neural networks, structured and unstructured techniques, have recently been applied on RNA-seq datasets for cancer prediction and classification.

In this study, we inspire an unsupervised variational auto-encoder (VAE) which is proposed in a recent paper by Way and Greene [1]. The proposed auto-encoder has two direct benefits of modelling cancer gene expression data: they can automatically extract non-linear features and learn the reduced dimension manifold of cancer expression space. VAEs are different from deterministic autoencoders because of the added constraint of normally distributed feature activations per sample. This constraint not only regularizes the model but also provides the interpretable manifold.

Additionally, the state-of-the-art classification models are constructed based on the assumption that the training data will maintain an equal amount of data in each class. In real-world data, equal distribution among classes is a perfect scenario for training machine learning models, and it is rare to find. Imbalance among the classes in a dataset is a critical problem in machine learning which results in a biased prediction towards the classes with more amount of training observations [2]. This results in an inferior performance of any classification model on classes with fewer training instances. This problem of majority biased prediction is popularly called the Class-Imbalance. In the recent years, the problem has been explored in variety of real-world domains such as financial inconsistencies [3], social media-based sarcasm detection [4], lesion detection in image analysis [5], and fault diagnosis [6]. For this study, we plan to compare several machine learning classification algorithms with different class-imbalance techniques to gene data to improve the classification performance of the algorithms and VAEs for dimensionality reduction.

II. Methods

In recent years, various attempts have been made to predict cancer from gene expression data by utilizing machine learning techniques [7]–[9]. Through our comprehensive experiments, we will highlight the effects of imbalance data and synthetic resampling strategies on state-of-the-art machine learning models in the context of cancer prediction from gene expression data using VAEs for dimensionality reduction. The framework of our experiment can be segregated into three parts: (1) The first section will concentrate on establishing the detrimental effects of data imbalance in the current context, (2) the second section will apply a VAE approach for feature reduction, and (3) the final step will provide a comparative analysis of existing strategies to mitigate the data-imbalance problem. The final step can be further divided based on techniques like minority up-sampling, majority down-sampling, hybrid resampling, and ensemble learning.

Variational autoencoders (VAEs) are an unsupervised deep neural network approach that generates meaningful latent spaces for image and text data. In Way and Greene's [1] study, they used a VAE to capture the most critical biologically-relevant features. This study will apply the TCGA dataset in the Maximum Mean Discrepancy Variational Autoencoder (MMD-VAE) to identify the most relevant features. MMD-VAE is a type of the InfoVAE family that maximizes Mutual Information between the Isotropic Gaussian Prior (as the latent space) and the Data Distribution. MMD-VAE is an alternative to traditional variational autoencoders that is fast to train, stable, easy to implement, and leads to improved unsupervised features [10].

III. Evaluation

For our experiments, we will primarily follow two sets of testing approaches: (1) 10-fold stratified cross-validation and (2) Balanced Testing [11]. The cross-validation approach is a popular testing regimen constructed by splitting the dataset into k different segments (taking $k = 10$ is a standard approach) followed by an iterative process of assigning one data segment as the holdout set. This approach is beneficial as it gives them the freedom to assign every dataset instance as training and testing data at least once. However, in some cases of extreme imbalance, the cross-validation approach is seen to generate test sets with limited instances belonging to the minority class, leading to a very inferior assessment of the performance of any machine learning model [11]. This motivated us to perform balanced testing by creating independent testing set with equal instances in each class. This will provide an unbiased testing framework by introducing a wide range of data for validating a model with certainty. We will evaluate different imbalance ratios and the degree of

oversampling and undersampling. To measure the performance of the machine learning models, we will be using geometric mean (G-mean), which is a popular measure for imbalanced learning as it is designed to give equal importance to each class [12].

IV. Data

To conduct our experiments, we will use the cancer genome atlas (TCGA) pan-cancer RNA sequential gene expression data [1] which was made publicly available in 2013 [13], [14]. The dataset covers over 10,000 tumour samples with 33 categories, which we can further segregate into different stages of cancer. Five thousand gene samples document the potential of the tumours. We have found that there is an unequal class distribution and class imbalance that could impact the class prediction performance (i.e., breast cancer patients (1246), rectum adenocarcinoma (105), kidney chromophobe (91), mesothelioma(87), uveal melanoma (80), adrenocortical carcinoma(79), uterine carcinosarcoma (57), lymphoid neoplasm diffuse large B-cell lymphoma (48), cholangiocarcinoma (45)).

V. Task Distribution and Schedule

We will divide the project as follows:

	ACTIVITY	PLAN START	PLAN DURATION	PERIODS 1	2	3	4	5	6	7
All	Literature Review	1	2							
All	Write Introduction, Related Work and Methods Section	2	2							
MB	Data Exploration with VAE and MMD-VAE	2	2							
RD	Data Visualization	3	1							
RD, KG	Apply techniques for data imbalance	4	2							
MB	Run experiments. Model training and testing.	4	2							
KG, RD	Write Discussion and Results	6	2							
All	Preliminary Project Demo	6	1							
All	Final Project Demo	7	1							

Fig. (1) Task distribution and schedule of project.

VI. Significance

Our study will add value to the scientific literature in cancer classification with gene data of imbalance classes in high-dimensional data. We have detected few studies that focus on data imbalance in gene expressions for cancer classification. Hence, we aim to improve classification performance by applying different class imbalance techniques to aid physicians in their diagnoses.

References

[1] G. P. Way and C. S. Greene, "Evaluating deep variational autoencoders trained on pan-cancer gene expression," *arXiv preprint arXiv:1711.04828*, 2017.

[2] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent data analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[3] D. Veganzones and E. Séverin, "An investigation of bankruptcy prediction in imbalanced datasets," *Decision Support Systems*, vol. 112, pp. 111–124, 2018.

[4] A. Banerjee, M. Bhattacharjee, K. Ghosh, and S. Chatterjee, "Synthetic minority oversampling in addressing imbalanced sarcasm detection in social media," *Multimedia Tools and Applications*, vol. 79, no. 47, pp. 35 995–36 031, 2020.

[5] A. Bria, C. Marrocco, and F. Tortorella, "Addressing class imbalance in deep learning for small lesion detection on medical images," *Computers in biology and medicine*, vol. 120, p. 103735, 2020.

[6] C. Wang, C. Xin, and Z. Xu, "A novel deep metric learning model for imbalanced fault diagnosis and toward open-set classification," *Knowledge-Based Systems*, vol. 220, p. 106925, 2021.

[7] J. Pati, "Gene expression analysis for early lung cancer prediction using machine learning techniques: An eco-genomics approach," *IEEE Access*, vol. 7, pp. 4232–4238, 2018.

[8] A. Gumaei, R. Sammouda, M. Al-Rakhami, H. AlSalman, and A. El-Zaart, "Feature selection with ensemble learning for prostate cancer diagnosis from microarray gene expression," *Health Informatics Journal*, vol. 27, no. 1, p. 1460458221989402, 2021.

[9] K. F. Mahin, M. Robiuddin, M. Islam, S. Ashraf, F. Yeasmin, and S. Shatabda, "Panclassif: Improving pan cancer classification of single cell rna-seq gene expression data using machine learning," *Genomics*, 2022.

[10] S. Zhao, J. Song, and S. Ermon, "Infovae: Information maximizing variational autoencoders," *arXiv preprint arXiv:1706.02262*, 2017.

[11] K. Ghosh, C. Bellinger, R. Corizzo, B. Krawczyk, and N. Japkowicz, "On the combined effect of class imbalance and concept complexity in deep learning," in *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 2021, pp. 4859–4868.

[12] H. Guo, H. Liu, C. Wu, W. Zhi, Y. Xiao, and W. She, "Logistic discrimination based on g-mean and f-measure for imbalanced problem," *Journal of Intelligent & Fuzzy Systems*, vol. 31, no. 3, pp. 1155–1166, 2016.

[13] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart, "The cancer genome atlas pan-cancer analysis project," *Nature genetics*, vol. 45, no. 10, pp. 1113–1120, 2013.

[14] G. P. Way and C. S. Greene, "Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders," in *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2018: Proceedings of the Pacific Symposium*. World Scientific, 2018, pp. 80–91.

[1] https://github.com/greenelab/tybalt