Towards Self-Supervised Learning for Prediction of Vital Status of Colorectal Cancer Patients

Girivinay Padegal, Murali Krishna, Om Amitesh B. R, Sathwik Acharya, Gowri Srinivasa
PES Center for Pattern Recognition and the Department of Computer Science and Engineering,
PES University, Bengaluru, India
pes1ug19cs315, muralikrishnam, omamiteshbr, sathwikacharya@pesu.pes.edu, gsrinivasa@pes.edu

Abstract

RNA sequencing (RNA-Seq) is a technique that utilises the capabilities of next-generation sequencing to study cellular transcriptome *i.e.*, to determine the amount of RNA at a given time for a given biological sample. The advancement of RNA-Seq technology has resulted in a large volume of gene expression data for analysis. In this study, we have used a gene expression dataset pertaining to patients diagnosed with colorectal cancer towards predicting the vital status of a given patient. We further show that our computational model (built on top of TabNet), which is pretrained with a self-supervised learning objective function, outperforms popular classic machine learning algorithms such as XGBoost and Decision Trees. To the best of our knowledge, TabNet, which is first pretrained on an unlabelled dataset of multiple types of adenomas and adenocarcinomas, and later fine-tuned on the labelled dataset, shows state-of-the-art results in the context of the estimation of the vital status of colorectal cancer patients. We conclude this study with an understanding of the genes that are important to the prediction task through interpreting the model and corroborate our results with pathological evidence that exists in current literature.

1 Introduction

Recent advancements in sequencing techniques in bio-informatics have gained traction and have given rise to studies pertaining to the identification of cancer bio-markers and quicker diagnosis of the disease. However, not as much research has been done on estimating the vital status of patients diagnosed with cancer. In this study, we analyse the gene expression dataset obtained from the TCGA website, to estimate the vital status of patients in the context of colorectal cancer [1]. The vital status property is a binary variable which indicates whether the patient survived the cancer. A vital status value of 0 represents a patient that survived; a vital status value of 1 represents a patient that succumbed to the cancer. The outcome of this study could not only help determine a suitable course of treatment but also help in a better understanding of the disease through a study of the bio-markers that are identified as important predictors by the computational models.

Due to the tabular nature of the gene expression dataset, applications of self-supervised learning techniques on gene expression have not made it to the limelight as much as they have in the domains of Computer Vision and Natural Language Processing. This is primarily because tabular models must be able to accommodate features from different discrete and continuous distributions and uncover correlations without relying on positional information.

So why is self-supervised learning worth exploring in the context of gene expression data? Apart from the increase in performance that comes with deep learning models, it also enables fusion with multiple modalities (say gene expression and clinical data) – eliminating the need for feature engineering, representation learning and end-to-end compositional multi-task models.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Our work can be summarised as follows:

- We have used feature selection techniques to mitigate the effects of the curse of dimensionality observed due to the high dimensionality of the data. The features chosen are used to train classical machine learning models (such as XGBoost) without the self-supervised learning objective.
- This is followed by filtering the same set of genes obtained from the feature selection technique in the unlabelled dataset and pretraining the TabNet model with a masked feature reconstruction loss as a self-supervised learning objective function.
- The pretrained model is then fine-tuned on the labelled dataset whose performance is compared with those models trained without the self-supervised learning objective function.
- This is concluded by demonstrating the explainability of the TabNet model by plotting its feature importance mapping and providing pathological evidence for the same.

2 Related work

The earliest work on the estimation of vital status using RNA-seq data can be traced to [2], which serves as a solid baseline model. They use a LASSO model and two variants of neural networks to infer that simplistic models like LASSO are computationally less expensive and outperform neural networks for this problem. Further, ensemble learning through the random forest algorithm to avoid the inherent bias present in decision tree classification demonstrated promising success [3]. Given the curse of dimensionality inherent in the data, there is a useful analysis of feature selection techniques for this problem [4]. On the self-supervised learning front, SSL methods have predominantly been focused on natural language processing [5, 6] and computer vision [7, 8, 9], where the concept of correlation between sequential features is at the fore. Recently, these self-supervised learning approaches have been introduced to work with tabular data using attention-based mechanisms [10] and random feature corruption [11], with these methods performing better than common industry favourite supervised learning methods such as XGBoost [12] and LightGBM [13]. [14] proposes a new self-supervised learning model called TabNet that works on tabular data and is heavily referred to in this paper.

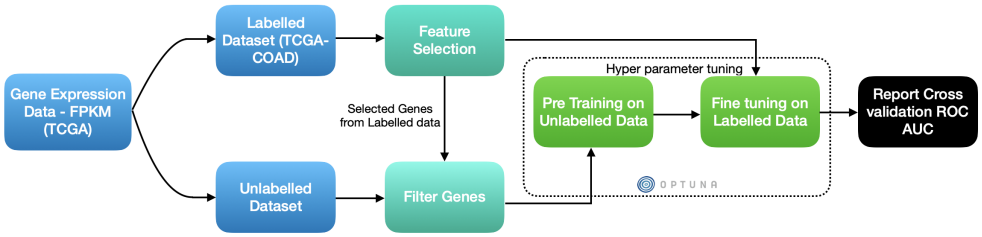

Figure 1: Workflow used for the self supervised approaches in this study.

3 Methodology

The workflow shown in Figure 1. consists of data collation, feature selection, pretraining and finetuning of the TabNet model followed by interpreting the model, outlined in the sections below[1].

3.1 Procurement of data

The data used is obtained from TCGA [1]. The labelled dataset is from the COAD project focusing on Colon Adenocarcinoma. This data was downloaded from the GDC data portal, with filters set to obtain the vital status annotation. We obtained 519 samples, each having gene expression data for $> 60,000$ genes. The raw gene expression data from the labs has been pre-normalised by the experimental team accounting for sequencing depth and gene length. Additionally, an unlabelled dataset consisting of the same gene IDs, with a total size of 4801 samples from various types of adenomas and adenocarcinomas is also procured by the same means for self-supervised learning purposes.

[1]The code pertaining to this paper can be found at `https://bit.ly/3SljHSO` for reproducibility purposes

3.2 Feature selection

A filter of protein-coding genes (obtained from BioMart) is first used to reduce this to 19559 genes from the provided 60000 genes. We then propose different feature selection techniques, namely, PCA [15], T-test [16] and Lasso [17] for further downstream analysis. Other statistical-based feature selection techniques such as Laplacian score and UMAP [18] were used but none of these methods performed as well as the above-mentioned ones.

3.3 TabNet pretraining and fine-tuning

The TabNet architecture [14] has an encoder and decoder module wherein the former is inspired by top-down attention using sparse instance-wise feature selection constructed on top of a sequential multi-step architecture. For the self-supervised learning objective, a decoder architecture has been proposed for reconstructing the masked features from the encoded representation. To do this a binary mask S has been chosen as $\mathbf{S} \in \{0,1\}^{B \times D}$ wherein the fully connected layer and feature transformers have to predict the masked feature at each decision step based on the reconstruction loss as follows:

$$\sum_{b=1}^{B}\sum_{j=1}^{D}\left|\frac{(\hat{f}_{b,j}-f_{b,j})\cdot S_{b,j}}{\sqrt{\sum_{b=1}^{B}(f_{b,j}-1/B\sum_{b=1}^{B}f_{b,j})^2}}\right|^2 \tag{1}$$

Here $f_{b,j}$ and $\hat{f}_{b,j}$ correspond to the feature importance score and the expected feature importance score of the j^{th} feature in the b^{th} sample.

The same feature set that has been chosen by the corresponding feature selection method has been used in the unlabelled dataset for pretraining purpose. Once the embedding space has been learnt via pretraining, the model is then fine tuned on the labelled dataset.

4 Experiments and results

In line with the theme of this study, that is, showing the performance gain using a self-supervised learning approach for vital status prediction, the results can be seen in Table 1. Despite trying a host of computational models, we have only enlisted those that show the highest scores. With respect to the TabNet model, the Adam [19] optimiser with a ReduceLROnPlateau scheduler was used. The metrics reported in this study pertain to the average AUC Score of the model's performance on a 5-fold cross-validation performed on the labelled (TCGA-COAD) dataset [2].

A pretrained TabNet model with the employment of T-test as a feature selection technique shows the highest score of **0.8132**. Optuna [20] was used to tune hyperparameters such as weight decay, learning rate, patience (for early stopping) and gamma (feature reuse in masks). A set of 5 trials is performed by Optuna to find the best set of hyperparameters that maximise the model's AUC Score. All mentioned experiments were carried out in Google Colab using a NVIDIA Tesla K80 GPU.

5 Interpretability and conclusion

The TabNet architecture, which consists of an encoder (composed of attentive transformer, feature transformer and feature masking) and the decoder (which consists of the feature transformer only), employs sparse feature selection at each decision step. Unlike SHap [21] which relies on cooperative game theory that averages the marginal contribution of a feature instance over all coalitions, TabNet quantifies the aggregate feature importance at each decision step in the following way:

$$M_{agg-b,j}\sum_{i=1}^{N_{steps}}\eta_b[i]M_{b,j}[i]/\sum_{j=1}^{D}\sum_{i=1}^{N_{steps}}\eta_b[i]M_{b,j}[i]^2 \tag{2}$$

Here, $M_{b,j}[i]$ corresponds to feature importance of $f_{b,j}$ and $\eta_b[i]$ weighs the aggregate contribution of the i^{th} decision step via a ReLU activation function [22].

[2]The best performing models were also trained and tested on a dataset of kidney cancers (KIPAN) to ensure that no overfitting occured, and produced consistent results, code for which is available at https://bit.ly/3SljHS0

Table 1: ROC AUC Score as per feature reduction and Training approach

Model	Training Approach	Feature Reduction	AUC
Logistic Regression	Supervised	Lasso	0.624 ± 0.060
Logistic Regression	Supervised	PCA	0.592 ± 0.052
Logistic Regression	Supervised	T-Test	0.495 ± 0.061
Neural Network	Supervised	Lasso	0.544 ± 0.1
Neural Network	Supervised	PCA	0.509 ± 0.09
Neural Network	Supervised	T-Test	0.662 ± 0.07
XGBoost	Supervised	Lasso	0.571 ± 0.08
XGBoost	Supervised	PCA	0.519 ± 0.08
XGBoost	**Supervised**	**T-Test**	$\mathbf{0.721 \pm 0.061}$
TabNet	Self Supervised	Lasso	0.762 ± 0.018
TabNet	Self Supervised	PCA	0.680 ± 0.040
TabNet	**Self Supervised**	**T-Test**	$\mathbf{0.813 \pm 0.042}$

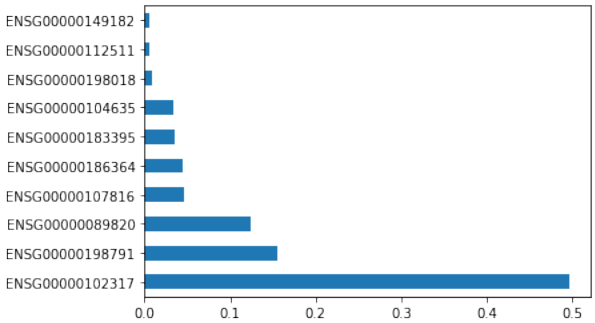

Figure 2: List of features and their Importance as per TabNet

The top contributing genes have been shown in Figure 2. We also surveyed existing literature to see if the top contributing genes in the TabNet model had any links to pathological evidence. We observed that the 4 out of the top 5 genes that contribute to the model's performance have scientific evidences which links it to the prognosis of colorectal cancer. The gene ID ENSG00000102317 (which contributes the highest towards the TabNet model prediction) corresponds to gene RBM3. [23]'s work shows the effect of over-expression of RBM3 and its subsequent effects on epithelial proliferation and stemness in colorectal cancer. Interestingly enough, NUDT17 has not been shown to be marker for colorectal cancer prognosis, but has been reported to show favourable cancer specificity to renal cancer [24]. Therefore, the findings of this paper may perhaps spark an interest in researching the contribution of this gene to the prognosis of colorectal cancer.

A summary of the findings can be found in Table 2. [25] provides factual pathological evidence that corroborates the findings of the top genes provided by the TabNet model (under the self-supervised learning setting) which contribute the most for vital status prediction in colorectal cancer. As a concluding remark, our work shows the state of the art results achieved for vital status prediction and contrasts the same between self supervised learning algorithms and supervised learning algorithms.

Table 2: Gene ID and its pathological evidence towards Colon Cancer

Gene ID	Gene Name	Pathological evidence [25]
ENSG00000102317	RBM3	Yes[23]
ENSG00000198791	CNOT7	Yes[26]
ENSG00000089820	ARHGAP4	Yes[27]
ENSG00000107816	LZTS2	Yes[25]
ENSG00000186364	NUDT17	No Evidence Found

References

[1] S Kirk et al. "Radiology data from the cancer genome atlas colon adenocarcinoma [TCGA-COAD] collection". In: *The Cancer Imaging Archive* 10 (2016), K9.

[2] Daniel Urda et al. "Deep learning to analyze RNA-seq gene expression data". In: *International work-conference on artificial neural networks*. Springer. 2017, pp. 50–59.

[3] Thomas Liuksiala. *Predicting cancer survival with machine learning*. https://geneviatechnologies.com/blog/predicting-cancer-survival-with-machine-learning/. Accessed: 2022-09-20.

[4] Ruizhi Xiang et al. "A comparison for dimensionality reduction methods of single-cell RNA-seq data". In: *Frontiers in genetics* 12 (2021), p. 646936.

[5] Kaitao Song et al. "Mpnet: Masked and permuted pre-training for language understanding". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 16857–16867.

[6] Colin Raffel et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." In: *J. Mach. Learn. Res.* 21.140 (2020), pp. 1–67.

[7] Jean-Bastien Grill et al. "Bootstrap your own latent-a new approach to self-supervised learning". In: *Advances in neural information processing systems* 33 (2020), pp. 21271–21284.

[8] Kaiming He et al. "Momentum contrast for unsupervised visual representation learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9729–9738.

[9] Yonglong Tian, Dilip Krishnan, and Phillip Isola. "Contrastive multiview coding". In: *European conference on computer vision*. Springer. 2020, pp. 776–794.

[10] Gowthami Somepalli et al. "Saint: Improved neural networks for tabular data via row attention and contrastive pretraining". In: *arXiv preprint arXiv:2106.01342* (2021).

[11] Dara Bahri et al. "Scarf: Self-supervised contrastive learning using random feature corruption". In: *arXiv preprint arXiv:2106.15147* (2021).

[12] Tianqi Chen et al. "Xgboost: extreme gradient boosting". In: *R package version 0.4-2* 1.4 (2015), pp. 1–4.

[13] Guolin Ke et al. "Lightgbm: A highly efficient gradient boosting decision tree". In: *Advances in neural information processing systems* 30 (2017).

[14] Sercan Ö Arik and Tomas Pfister. "Tabnet: Attentive interpretable tabular learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 8. 2021, pp. 6679–6687.

[15] Hervé Abdi and Lynne J Williams. "Principal component analysis". In: *Wiley interdisciplinary reviews: computational statistics* 2.4 (2010), pp. 433–459.

[16] Tae Kyun Kim. "T test as a parametric statistic". In: *Korean journal of anesthesiology* 68.6 (2015), pp. 540–546.

[17] J Ranstam and JA Cook. "LASSO regression". In: *Journal of British Surgery* 105.10 (2018), pp. 1348–1348.

[18] Leland McInnes, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426* (2018).

[19] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[20] Takuya Akiba et al. "Optuna: A Next-generation Hyperparameter Optimization Framework". In: *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2019.

[21] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017).

[22] Abien Fred Agarap. "Deep learning using rectified linear units (relu)". In: *arXiv preprint arXiv:1803.08375* (2018).

[23] Anand Venugopal et al. "RNA binding protein RBM3 increases β-catenin signaling to increase stem cell characteristics in colorectal cancer cells". In: *Molecular carcinogenesis* 55.11 (2016), pp. 1503–1516.

[24] Yue Wang et al. "NUDT expression is predictive of prognosis in patients with clear cell renal cell carcinoma". In: *Oncology letters* 14.5 (2017), pp. 6121–6128.

[25] Mathias Uhlen et al. "A pathology atlas of the human cancer transcriptome". In: *Science* 357.6352 (2017), eaan2507.

[26] James Flanagan et al. "Analysis of the transcription regulator, CNOT7, as a candidate chromosome 8 tumor suppressor gene in colorectal cancer". In: *International journal of cancer* 106.4 (2003), pp. 505–509.

[27] Ming-sheng Fu et al. "Analysis of ARHGAP4 Expression With Colorectal Cancer Clinical Characteristics and Prognosis". In: *Frontiers in Oncology* 12 (2022).

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] One of the main limitations of this study is that it is still a proof-of-concept with promising results that paves way for a better understanding of the factors that underlie the disease. But to be useful as a clinical decision support system, the model would have to be trained and tested on a larger volume of data, that is currently not available in the public domain.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] TCGA curates data from thousands of patients after obtaining informed consent from each one; we do not feel it necessary to further detail this, as the process is made transparent via the TCGA website.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] Due to data being obtained from TCGA, a government effort, all personally identifiable details are obscuted as per standard procedure.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]