2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

An Interpretable Approach for Lung Cancer Prediction and Subtype Classification using Gene Expression

Bernardo Ramos, Tania Pereira, João Moranguinho, Joana Morgado, José Luis Costa, and Hélder P. Oliveira *(Member, IEEE)*

***Abstract*—Lung cancer is the deadliest form of cancer, accounting for 20% of total cancer deaths. It represents a group of histologically and molecularly heterogeneous diseases even within the same histological subtype. Moreover, accurate histological subtype diagnosis influences the specific subtype's target genes, which will help define the treatment plan to target those genes in therapy. Deep learning (DL) models seem to set the benchmarks for the tasks of cancer prediction and subtype classification when using gene expression data; however, these methods do not provide interpretability, which is great concern from the perspective of cancer biology since the identification of the cancer driver genes in an individual provides essential information for treatment and prognosis. In this work, we identify some limitations of previous work that showed efforts to build algorithms to extract feature weights from DL models, and we propose using tree-based learning algorithms that address these limitations. Preliminary results show that our methods outperform those of related research while providing model interpretability.**

***Clinical Relevance*: The machine learning methods used in this work are interpretable and provide biological insight. Two sets of genes were extracted: a set that differentiates normal tissue from cancerous tissue (cancer prediction), and a set of genes that distinguishes LUAD from LUSC samples (subtype classification).**

I. Introduction

Cancer is a genetic disease caused by changes to genes that control the way cells operate, especially how they grow and divide. According to the European Lung Foundation (ELF), lung cancer is the largest cancer killer in Europe, accounting for approximately 20% of total cancer deaths [1]. Lung cancer is categorized into two main histological groups: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). NSCLCs are generally subcategorized into adenocarcinoma (LUAD), squamous cell carcinoma (LUSC), and large cell carcinoma (LCC). Accumulating evidence suggests that lung cancer represents a group of histologically and molecularly heterogeneous diseases even within the same histological subtype [2]. Cancer prediction relates to differentiating cancerous tissue from normal tissue, whereas histologic subtype classification differentiates groups within the same type of cancer, based on certain characteristics of the cancer cells. An early diagnosis of cancer is essential for a good prognosis, and accurate histological subtype diagnosis influences the target genes for the specific subtype, which helps defining the treatment plan that can target those genes in therapy.

RNA-sequencing (RNA-seq) is a technique that can examine the quantity and sequences of ribonucleic acid (RNA) in a sample using next-generation sequencing (NGS). It analyzes the transcriptome of gene expression patterns encoded within RNA [3]. RNA-seq tells us which genes are turned on in a cell, what their level of expression is, and at what times they are activated or shut off [4], which can enable towards a deeper understanding of molecular changes that might lead to disease. In Xiao et al. [5], the authors used a multi-model deep learning (DL) based ensemble strategy to distinguish between cancer and non-cancerous samples for stomach adenocarcinoma (STAD), breast invasive carcinoma (BRCA) and LUAD. The best performing model was decision trees (DT) with an accuracy of 0.968±0.023, and the ensemble model provided a boost in accuracy to 0.988±0.018 for the LUAD dataset. In Ahn et al. [6], the authors use a six hidden-layer deep feed-forward network for cancer prediction using gene expression data for 24 different cancer types. The DNN showed an overall accuracy of 0.979, and an algorithm to calculate the individual gene contribution was designed as an effort to provide interpretability on the DNN classifier. The algorithm used to extract feature weights from the DNN bases itself on feature selection techniques, inputting a range of expression values of a gene of interest for the given sample and observing the change in the DNN outcome [6] so that a single weight of gene contribution to the output can be calculated. In De Guia et al. [7], a convolutional neural network (CNN) was built to classify subtypes of 33 cohorts of cancer types using multiclass label classification. The proposed model achieved an overall accuracy of 0.957, and accuracy of 0.950 and 0.910 for the LUAD and LUSC classes, respectively. In Ye et al. [8], the authors use unsupervised learning techniques to identify gene signatures for accurate NSCLC subtype classification. A set of 17 genes was isolated, and multiple classifiers were used for prediction, DTs performed best with an accuracy of 0.922.

In general, previous works show us that DL models are prime candidates for the tasks of cancer classification and subtype prediction, and although efforts have been made

B. Ramos and J. Moranguinho are with the INESC TEC - Institute for Systems and Computer Engineering, Technology and Science, Portugal and FEUP - Faculty of Engineering, University of Porto, Portugal.

T. Pereira is with the INESC TEC, Portugal.

J. L. Costa is with the i3S - Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Portugal and IPATIMUP - Institute of Molecular Pathology and Immunology of the University of Porto, Portugal.

J. Morgado and H. P. Oliveira are with the INESC TEC and FCUP - Faculty of Science, University of Porto, Portugal.

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project UIDB/50014/2020.

to provide interpretability from the DL learners, we can pinpoint two downfalls of previous approaches: first, the DNN's performed better when performing feature selection a priori using variance selection techniques, which might leave out important genes that are not selected by variance or other techniques; secondly, the algorithm used to extract weights from the network bases itself on leave one out technique, that might fail to capture correlations between variables post feature selection. In this work, we use a model that requires no a priori feature selection so that all features are considered as input to the model. The proposed work aims to provide a method for cancer prediction and subtype classification, which outperforms state-of-the-art DL methods while providing interpretability, which is a great concern from a pathologist's perspective since identifying the cancer driver genes in an individual provides essential information for treatment and prognosis.

II. Materials and Methods

A. Dataset

Using R BioConductor framework with the TCGABiolinks package, we queried for gene expression quantification (RNA-seq) data from The Cancer Genome Atlas (TCGA) project. The data was retrieved from the Genomic Data Commons (GDC) legacy database and entailed tier 3 post-normalized data of the GDC workflow. We queried for tissue types primary solid tumour (TP) and solid tissue normal (NT) for the LUAD and LUSC projects. Using GDCprepare, we proceed to add clinical information for the patients and remove duplicate patient records. A total of 598 LUAD and 553 LUSC samples were retrieved, with expression for 20,531 genes, out of which 59 and 51 were normal tissue samples, respectively.

B. Preprocessing

The nomenclature for columns identifiers in BioConductor is the Human Genome Organisation (HUGO) symbol by default. Some errors were detected, namely duplicated gene names corresponding to different entries on distinct databases. Therefore, according to the gene metadata file, all identifiers were renamed as a combination of their HUGO symbol and Entrez gene. Upon analysis, we identified two different kinds of "duplicate" samples belonging to the same patient: 1) samples with the same vial and portion but different plate; 2) samples with a different vial. To our understanding, case one represents duplicate samples tested on different plates for reproducibility, while two refers to samples from different regions of the tumour. Therefore, for case one, we averaged the samples' values of expression, and for two, we maintained all the samples.

C. Experiment Design

The workflow for the experiment can be divided into two stages. In the first stage, we use gene expression normalized data and employ two binary classifiers. One to distinguish between tumour and tissue normal samples and another to distinguish LUAD from LUSC samples. For the cancer classification problem two approaches were experimented to tackle the 9:1 ratio class imbalance of positive samples: balancing the weights of the labels; use Adaptive Synthetic (ADASYN) to oversample the negative class in the train and validation sets. Furthermore, we conduct a data analysis stage to analyze two sets of features extracted from the interpreted models. The first feature set represents expressed genes that better differentiate TP from NT samples, and the second feature set represents features that distinguish between LUAD and LUSC subtypes.

D. Classification

Gradient boosting decision trees (GBDT) are a family of ensemble models of decision trees with various implementations such as XGBoost or pGBRT. Although these are popular machine learning algorithms, the efficiency and scalability are still unsatisfactory when the feature dimension is high, and data size is large. The classifier used in this work is the light gradient boosting machine (LightGBM), which attempts to fix this problem by implementing two innovative techniques: Gradient-based One-Side Sampling (GOSS) which excludes a significant proportion of data instances with small gradients, and Exclusive Feature Bundling (EFB) that bundles mutually exclusive features, therefore, reducing the number of features [9].

The train, validation and test splits were obtained using stratified sampling with a (70,15,15)% split for cancer prediction and (80,10,10)% for the subtype classification problem. The larger size of the independent test size on the first problem is due to the labels' imbalance; therefore we need to guarantee a minimum amount of negative samples for support in the test and validation sets. Hyper-parameters were tuned with a Bayesian optimizer using 5-fold cross-validation, and when performing data augmentation, the oversampling was done for each fold to avoid data leakage. The train and test splits were executed 100 times, randomizing the split seed to reduce variability and overcome skewness caused by the short sample size and class imbalance. The evaluation metrics' binary cross-entropy (logloss) and area under the ROC Curve (AUC) were used to assess training performance and control overfitting by early stoppage. Logloss metric captures the extent to which predicted probabilities diverge from class labels. Both these metrics evaluate the model's degree of separability. To evaluate the test set's performance, we used AUC, accuracy, precision, recall and f1-score metrics to better assess our results against similar research.

E. Optimization

The Bayesian optimization, in which a learning algorithm's generalization performance is modelled as a sample from a Gaussian process (GP), tries to find the minimum of a function $f(x)$ on some bounded set X. The difference from traditional methods such as randomized search is that it constructs a probabilistic model for $f(x)$ and then exploits this model to make decisions about where in X to evaluate

TABLE I
HYPER-PARAMETERS FOR CANCER AND SUBTYPE CLASSIFICATION.

Hyper-Parameters	Cancer Classification	Subtype Classification
n_estimators	256	154
max_depth	6	8
learning_rate	0.1048	0.9173
feature_fraction	0.2673	0.7457
bagging_fraction	0.1067	0.4718
min_split_gain	0.0002	0.0141
min_child_weight	0.0057	0.0053
min_child_samples	5	17
reg_alpha	0.0140	0.0103
reg_lambda	0.1100	0.1368
scale_pos_weight	4.9604	2.3732

the function next [10], which decreases the cost of finding the solutions when the black-box function f is complex, which is the case of more elaborate machine learning models. A critical step of this optimizer is deciding on an acquisition function, expected improvement was chosen, and the meta-parameter ξ that defines the exploitation-exploration trade-off was optimized by trial and error over five steps in the [$1e^{-4}$, $1e^{-1}$] search space. We optimize the hyper-parameters by minimizing logloss. Logloss takes the "certainty" of classification into account, and this is especially relevant when designing a model to diagnose deadly disease, as we want to penalize bad decisions and not necessarily only improve performance.

III. RESULTS

A. Best Hyperparameters

For each problem, we ran approximately 10,0000 steps of Bayesian optimization with 5-fold cross-validation. The optimal value for the meta-parameter ξ was fixed at $1e^{-2}$, which prioritizes exploration over exploitation. In TABLE I, we present the optimal values for the hyper-parameters of the LGBM classifiers.

The Bayesian optimization's optimal step showed a cross-validation loss of 0.00007 for cancer classification and 0.00048 for subtype classification. Higher depth values showed better results for both problems, and the number of leaves was fixed at $2^{max\_depth} - 1$.
L1 (*reg_alpha*) and L2 (*reg_lambda*) regularization were added to force sparsity and to diminish the value of the weights, which conferred a reduction in standard deviation across runs. The *scale_pos_weight* parameter value was higher in the cancer classification problem considering the 9:1 ratio class imbalance of positive samples.

B. Classification Results

In TABLE II, we present the cancer prediction and subtype classification models' performance. The cancer classification model showed an average AUC of 0.983. The average number of independent test samples' for support was 154.55 for the positive class and for the negative 17.67. The model showed higher precision for the majority class and higher standard deviations for the negative class; this is to be expected because of the low sample count of the negative class which

TABLE II
PERFORMANCE OVER 100 RUNS OF THE LGBM MODEL FOR CANCER AND SUBTYPE CLASSIFICATION.

Metrics	(Mean ± Standard Deviation)			
	Cancer Classification		***Subtype Classification***	
AUC	0.983±0.017		0.971±0.018	
Accuracy	0.995±0.006		0.971±0.018	
	Positive	*Negative*	*Positive*	*Negative*
Precision	0.997±0.005	0.976±0.036	0.962±0.020	0.980±0.022
Recall	0.997±0.004	0.969±0.046	0.980±0.022	0.961±0.023
F1-score	0.997±0.003	0.972±0.030	0.971±0.014	0.970±0.015

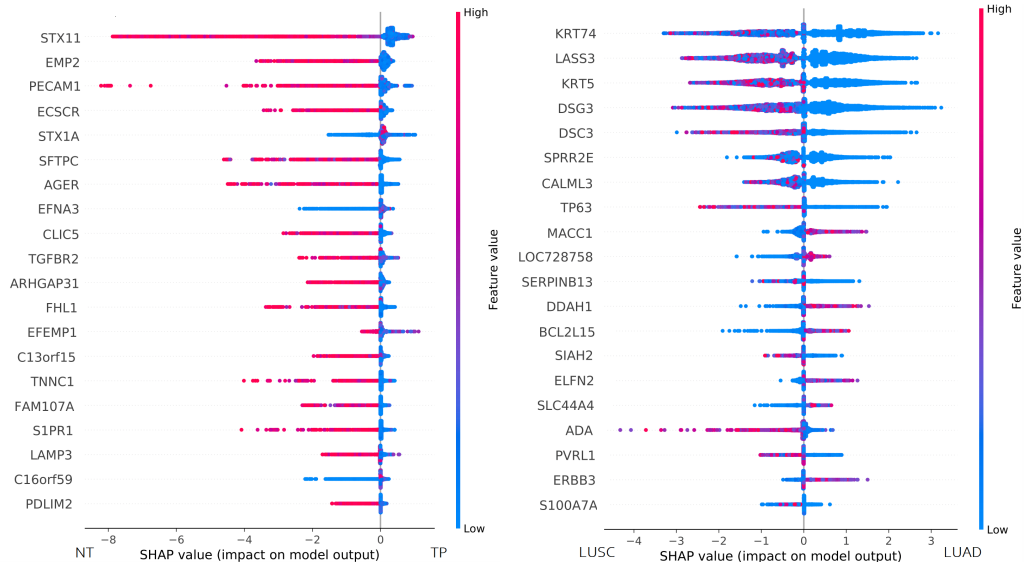
should induce more variance in results. The cancer subtype classification model showed an average AUC and accuracy of 0.971. The average support for the positive class was 52 and 50 for the negative. Precision was better for the negative class (LUSC), and variance was more stable amongst both classes.

C. Most Relevant Gene Signatures

To provide model interpretability, we used SHapley Additive exPlanations (SHAP) technique. This method explains individual predictions by estimating each feature's contribution to the corresponding prediction and, consequently, assigning it a SHAP value. Features with larger absolute SHAP values are more important for prediction and can positively or negatively impact the prediction depending on its sign.

In Figures 1(a), 1(b) we provide the SHAP summary plot for cancer and subtype classification problems across 100 runs, which combines feature importance with feature effects. The plot's y-axis identifies a gene, represented by its HUGO symbol and Entrez Gene and the x-axis the corresponding SHAP values for each data instance. The genes are ordered on the y-axis by overall predictive importance, and the top 20 most important genes were selected for analysis. The colour gives us a visual representation of features' original value distributions, categorized into low or high values of gene expression. This visualization can give us a holistic view of the model's decision as it conjugates the importance of the features with the effect on prediction while showing the value distribution of those features in the original data.

For the cancer classification problem, a total of 1,183 genes were selected by the model for prediction. Figure 1(a) shows the genes that more adequately distinguish between cancerous and non-cancerous samples according to the model. The negative weights represent features with a negative effect on prediction, which equates to features whose effect helps to predict NT samples, and the positive weights bind the decision to predict cancerous tissue. Amongst these top 20 genes, we can clearly see a pattern used for prediction: most of the selected genes when over-expressed affect the prediction negatively; and when under-expressed affect the prediction positively. Exceptions to this are STX1A, EFNA3 and C16orf59 genes, which present mostly low expression in both classes and some visible over-expression, which positively impacts the prediction. Generally, the analysis of



(a) Cancer Prediction (b) Subtype Classification

Fig. 1. SHAP values over 100 runs for the LGBM model. The y-axis identifies genes and the x-axis the corresponding SHAP values for each data instance.

the 20 most important gene expression signatures for cancer prediction shows a pattern of selecting signatures with a high expression that are important to predict for normal tissue.

For the cancer subtype classification problem, 2,685 genes presented non-null weights and therefore, were used for prediction. Figure 1(b) shows that the expression value of genes is more balanced across features with a positive and negative effect on model output. Negative weights bias decision to predicting LUSC samples and positive weights bias the decision to predict LUAD samples. We can infer two groups of genes that show identical patterns: MACC1, LOC728759, DDAH1, BCL2L15, ELFN2, SLC44A4 and ERBB3 represent the first group that shows mostly over-expression when binding the decision to predict LUAD, and under-expression when negatively affecting the prediction; the second group containing the remaining 13 genes present a symmetrical pattern with mostly over-expression when important for the model to predict LUSC tissue.

IV. Conclusions and Perspectives

This work proposes a methodology for lung cancer prediction and subtype classification based on gradient boosted trees. A critical difference between our proposed approach and the DL models, covered in state of the art, is that LGBM eradicates the need for a priori feature selection as the model removes redundant features by performing EFB. Two feature sets were extracted using model interpretability that should provide biological insight on differences in gene expression between cancerous and healthy tissue, and LUAD and LUSC subtypes. Preliminary results show that our methods outperform previous work results' that use DL methods for lung cancer prediction and subtype classification. In future work, we intend to further validate these results by extending the learners to other cancer types and performing validation to datasets outside of the TCGA scope.

Acknowledgment

We acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health for the free publicly available TCGA database used in this work. This database ensures that the necessary ethical approvals regarding data access were obtained.

References

[1] Altekruse et al., "SEER Cancer Statistics Review 1975-2007 National Cancer Institute," *Cancer*, pp. 1975–2007, 2010.

[2] K. Inamura, "Lung cancer: understanding its molecular pathology and the 2015 wHO classification," *Frontiers in Oncology*, vol. 7, no. AUG, pp. 1–7, 2017.

[3] Z. Wang et al., "RNA-Seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[4] F. Ozsolak et al., "RNA sequencing: advances, challenges and opportunities," *Nature Reviews Genetics*, vol. 12, no. 2, pp. 87–98, 2011.

[5] Y. Xiao et al., "A deep learning-based multi-model ensemble method for cancer prediction," *Computer Methods and Programs in Biomedicine*, vol. 153, pp. 1–9, 2018.

[6] T. Ahn et al., "Deep Learning-based Identification of Cancer or Normal Tissue using Gene Expression Data," *Proceedings - 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018*, pp. 1748–1752, 2019.

[7] J. M. de Guia et al., "Deepgx: Deep learning using gene expression for cancer classification," in *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Aug 2019, pp. 913–920.

[8] X. Ye, W. Zhang, and T. Sakurai, "Adaptive Unsupervised Feature Learning for Gene Signature Identification in Non-Small-Cell Lung Cancer," *IEEE Access*, vol. 8, pp. 154 354–154 362, 2020.

[9] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems*, vol. 2017-December, pp. 3147–3155, 2017.

[10] J. Snoek et al., "Practical bayesian optimization of machine learning algorithms," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2*, 2012, p. 2951–2959.