RNA sequencing-based histological subtyping of non-small cell lung cancer with generative adversarial data imputation

Ralph Saber[1,2], Bertrand Routy[2], Simon Turcotte[2], and Samuel Kadoury[1,2]

Abstract— **Non small cell lung cancer (NSCLC) is the most common type of lung cancer and is classified into two main histological subtypes: adenocarcinoma and squamous cell carcinoma. The identification of the histological subtype is a crucial step in the diagnosis of NSCLC. RNA sequencing data hold valuable biological information but may contain missing gene expression counts, which limit their potential exploitation in practice. In this work, we address the issue of missing gene expression data in NSCLC histological subtype prediction from RNA sequencing. To this end, we propose a pipeline based on the generative adversarial imputation network (GAIN) for the generation of plausible imputations of missing data and tree-based ensemble models for NSCLC histological subtype prediction. We adopted a nested cross validation scheme for the evaluation of the classification models. The proposed pipeline exhibited an outstanding performance with an area under the receiver operating characteristic curve of 0.98 ± 0.03 and an accuracy of 0.96 ± 0.05 obtained with the Light Gradient Boosting Machine. Experimental results showed that GAIN-derived imputations are useful to boost classification performance. Finally, we used the Shapley Additive Explanations technique and found a set of genes that were the most relevant for NSCLC subtyping across different models.**

Keywords— **transcriptomics, NSCLC subtyping, generative missing data imputation, interpretable machine learning.**

I. INTRODUCTION

Lung cancer is the leading cause of cancer-related deaths in the world [1], with more than two million diagnoses and 1.7 million deaths in 2018 alone. The two types of lung cancer are small cell lung cancer and non-small cell lung cancer (NSCLC). While the former type proliferates faster, the latter is much more prevalent and accounts for 85% of all lung cancer diagnoses [2]. NSCLC is further subclassified into two main histological subtypes: adenocarcinoma and squamous cell carcinoma [3]. These two subtypes exhibit distinct biological characteristics and may have disparate prognoses [4] and response to treatment [5]. Furthermore, recent evidence suggests that molecular subtypes of lung adenocarcima and squamous cell carcinoma present different immune properties and should be evaluated as a possible biomarker for immunotherapy [6]. The identification of the histological subtype of NSCLC is performed in clinical practice as a crucial step to establish prognosis. Typically, the evaluation of the histological subtype requires the analysis of tissue samples on whole slide images by experienced pathologists. Nevertheless, the manual assessment of histopathological tissue is time consuming, labor intensive and requires technical expertise.

In the last decade, deep learning approaches have been proposed in the biomedical field to perform several tasks based on different types of data, including imaging, clinical and genomics data. These include cancer classification, disease subtyping, immune profiling and the prediction of clinical outcomes [7]. One of the recently investigated medical data types by machine learning experts is transcriptomics data. RNA sequencing allows to evaluate the quantity of ribonucleic acid in a sample and gives insight into the cellular transcriptome. It allows to quantify gene expression using high throughput sequencing methods.

Nevertheless, one of the major challenges in the analysis of medical datasets, particularly RNA sequencing, is handling missing data. Missing gene expression counts in RNA sequencing may occur due to disproportionate polymerase chain reaction amplifications or the breakdown of RNA during library preparation. Discarding missing values is often detrimental to the performance of downstream tasks as it leads to the loss of valuable information.

In this work, we propose a pipeline that tackles the problem of missing expression counts in RNA sequencing for the prediction of NSCLC histological subtypes. To this end, we leverage generative deep learning to provide plausible imputations of the missing RNA sequencing reads and evaluate its benefit in NSCLC subtyping. The proposed pipeline requires no manual feature selection, allowing to holistically evaluate gene interactions. We finally interpret the model' predictions by analyzing the most impactful genes for the prediction of the histological subtype.

The remainder of the article is structured as follows: Section II explains the generative imputation-based pipeline for NSCLC subtyping. Section III describes the results of the proposed pipeline and provides an analysis of the most salient genes in subtype prediction. Finally, section IV summarizes the main contributions of the study.

II. MATERIALS AND METHODS

A. Dataset

This study was performed on the public NSCLC Radiogenomics dataset [8]. The dataset comprises 211 NSCLC patients in total. Patients were included if: (1) RNA sequencing was available, (2) clinicopathological data, including the histological subtype, was provided. For RNA sequencing, reads alignment to the human genome was performed and expression calls were determined in each sample using Fragments Per Kilobase of transcript per Million mapped reads (FPKM). Amongst the 22,126 genes included in the dataset, expression values of 19,342 genes were available.

[1]RS and SK are with Polytechnique Montréal, Montreal, Qc, Canada

[2]RS, BR, ST and SK are with the Centre de recherche du centre hospitalier de l'Université de Montréal, Montreal, Qc, Canada

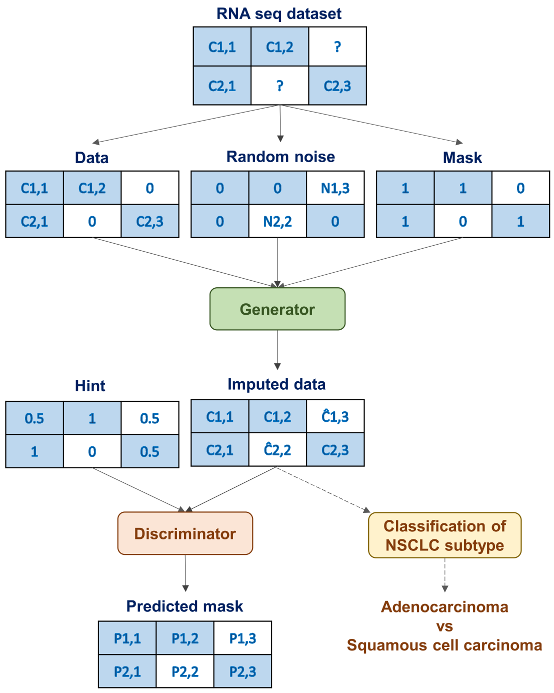

Fig. 1. Proposed pipeline for NSCLC subtyping from RNA sequencing with GAIN-generated imputations of missing gene expression counts. For simplicity, three gene counts of two patients are represented in the figure.

Moreover, missing data existed in 14,074 of the available genes.

After applying the inclusion criteria, 127 patients were included in this work. The ensuing cohort comprised 31 lung squamous cell carcinoma lesions and 96 lung adenocarcinoma lesions.

B. Generative imputation of missing gene expression values

In this work, we propose to generatively impute missing values in the RNA sequencing dataset by training a Generative Adversarial Imputation Network (GAIN) [9] (Figure 1). GAIN is a variation of the Generative Adversarial Network (GAN) that was conceived for generative imputation of missing data. As in the traditional GAN model, the two major components of GAIN are the generator and the discriminator networks. The generator is trained to produce reasonable imputations of missing values within the dataset whereas the discriminator is trained to distinguish between original and imputed values. The generator takes as input the original data, a random noise and a binary mask. The mask enables the generator to distinguish between observed (1) and missing (0) data and allows it to generate plausible imputations of the missing data, while keeping the original observations unaltered. The discriminator attempts to predict the binary mask, taking as inputs the imputed data produced by the generator and a hint matrix. The latter provides partial information on the mask to the discriminator and ensures that the imputations produced by the generator follow the actual data distribution. GAIN is trained iteratively: at each iteration, the discriminator weights are updated to maximize the probability that it predicts the mask and the generator is adversarially trained to minimize this probability.

The objective function is therefore represented by a min-max problem as follows:

$$min_G max_D \mathbb{E}[\mathcal{L}(M, \widehat{M})] \tag{1}$$

where G is the generator, D the discriminator, $\mathcal{L}$ the cross-entropy loss function, M the actual mask and $\widehat{M}$ the mask predicted by the discriminator.

In this work, the generator and discriminator are fully connected neural networks with 3 layers each, of respective dimensions $N \times 2$, N and N, where N is the number of features. Due to memory constraints, the dataset features were divided into 3 subsets and generative data imputation was performed on each subset separately.

C. Imputed RNA-seq data preprocessing

Given the wide range of expression values across different genes, standardization was performed prior to training the classification models so as to obtain a null mean and unit standard deviation. Moreover, the Synthetic Minority Over-sampling Technique (SMOTE) [10] was applied prior to training in order to account for the class imbalance. Instances from the minority class were hence oversampled by a factor of three using SMOTE technique.

D. Histological subtype classification

In this work, we trained and evaluated six different machine learning models: decision trees (DT), random forests (RF), and four gradient boosting ensemble models namely the extreme gradient boosting (XGBoost), the adaptive boosting (Adaboost), categorical boosting (CatBoost) and light gradient-boosting (LightGBM) machines. All models were trained with and without the GAIN-imputed RNA sequencing dataset. For comparison purposes, we also discarded genes with missing expression values yielding a feature set of 5,268 genes. The models were subsequently trained using the same strategy and their performance was compared.

E. Evaluation strategy and experimental setup

For all classification models, we applied a nested cross-validation (CV) scheme for simultaneous hyperparameter tuning and model evaluation. An outer five-fold CV was conducted by first dividing the dataset into five folds: four folds were used for training and the fifth for testing. Moreover, an inner three-fold CV was performed on the training folds in order to find the optimal hyperparameters of each model. A grid search was hence conducted on the models' hyperparameters then the models were retrained on the four training folds using the selected hyperparameters before testing on the fifth fold. The process was repeated five times so as to cover the entire dataset. All splits performed were stratified in order to maintain the same proportion of each class in the training and testing subsets.

The GAIN model was trained for 1000 epochs with a batch size of 64 and Adam optimizer. Training was performed on an NVIDIA GeForce GTX TITAN Xp 12GB. The performance metrics adopted were the area under the curve (AUC), accuracy, precision, recall and the F1-score.

F. Model interpretability analysis

In a subsequent step, we attempted to interpret the model's behavior by running a post hoc analysis on the model's predictions. We used the Shapley Additive Explanations (SHAP) technique [11] for this purpose. For each feature (gene), the algorithm computes a Shapley value that reflects the impact of that feature on the model's final decisions. The higher the absolute Shapley value of a given gene, the more significant its contribution is to the final predictions. A positive Shapley value encodes a positive contribution whereas a negative Shapley value represents a negative impact. In this study, we utilized the Tree SHAP method, an optimized implementation of SHAP specifically devised for decision tree-based models, which takes into account the number of subsets flowing into each node of the decision trees and has hence a much lower computational complexity.

III. Results and discussion

A. Prediction of NSCLC histological subtype

Table I summarizes the nested CV performance of the different models trained for the prediction of the NSCLC histological subtype: adenocarcinoma vs squamous cell carcinoma. The results show a consistent improvement in the performance of all the models when trained on the GAIN-imputed version of the RNA sequencing dataset. In fact, the decision tree witnessed an important increase in the AUC from 0.77 ± 0.11 (when discarding genes with missing values) to 0.91 ± 0.08 (when generating imputations of the missing values using GAIN model). Furthermore, models trained on the GAIN-imputed RNA sequencing data exhibited a much higher precision-recall balance mirrored by the high F1-scores ($\geq$ 0.95). As expected, ensemble models outperformed the DT with AUCs greater than 0.97 when using GAIN-generated imputations. An improved performance was observed with gradient boosting ensemble models, especially the LightGBM, XGBoost and CatBoost models. The LightGBM model achieved the highest F1-score (F1-score = 0.98 ± 0.03, AUC = 0.98 ± 0.03, accuracy = 0.96 ± 0.05). The best time complexity was also obtained with the LightGBM as training was faster compared to the other ensemble models.

In order to demonstrate the importance of the holistic information acquired when training models on the entire RNA sequencing data, we also trained the LightGBM model on a subset of the 100 genes that exhibited the highest variance across patients. Figure 2 depicts the receiver operating characteristic curves using the three approaches: (1) when training the model on the 100 most variant genes, (2) when discarding genes with missing values and training the model on the remaining ones and (3) when training the model on the entire RNA sequencing dataset and using GAIN-generated imputations. The results show that the LightGBM had a higher AUC when forgoing the feature selection step (0.95 ± 0.03 without feature selection vs 0.90 ± 0.04 when manually selecting genes). The AUC was even higher when using GAIN imputations. Figure 3 shows that a similar performance boost was observed in terms of accuracy (from 0.86 ± 0.06 to 0.96 ± 0.05) and F1 score (from 0.67 ± 0.17 to 0.98 ± 0.03).

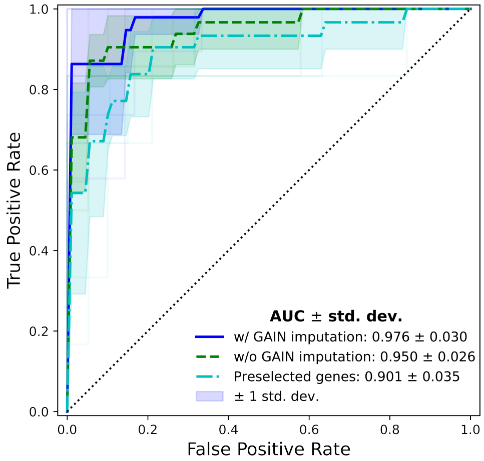

Fig. 2. Receiver operating characteristic curve of the LightGBM model when trained on the GAIN-imputed RNA sequencing dataset, without GAIN imputation and on manually selected genes.

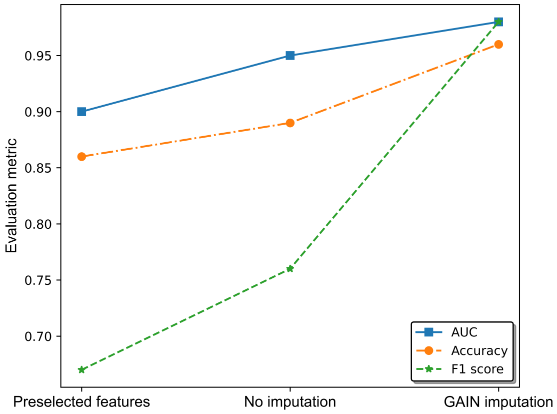

Fig. 3. Evolution of the performance of the LightGBM model with different preprocessing approaches.

B. Feature contribution analysis

Figure 4 depicts the summary plot of the SHAP technique for the LightGBM model when trained on the GAIN-imputed dataset. The top 20 genes with the highest Shapley values are included in the plot. Genes are listed in a decreasing order of importance from top to bottom (mirrored by their absolute Shapley values). As shown in the figure, the gene with the highest importance was the Sphingosine-1-Phosphate Receptor 5 gene or S1PR5. Sphingosine 1-phosphate is involved in cell proliferation and S1PR receptors are being investigated as potential targets for lung diseases and cancer [12]. It is to be noted that a high concordance in the SHAP results was found between LightGBM and XGBoost models (not shown) with S1PR5 found as top contributing gene

Model	GAIN imputation	AUC	Accuracy	Precision	Recall	F1-score
DT	✗	0.77 ± 0.11	0.79 ± 0.11	0.57 ± 0.14	0.75 ± 0.17	0.64 ± 0.13
	✓	0.91 ± 0.08	0.94 ± 0.04	**0.95 ± 0.06**	0.97 ± 0.05	0.96 ± 0.03
RF	✗	0.97 ± 0.04	0.88 ± 0.05	0.92 ± 0.11	0.59 ± 0.21	0.69 ± 0.17
	✓	0.97 ± 0.05	0.94 ± 0.05	0.93 ± 0.07	0.99 ± 0.03	0.96 ± 0.03
AdaBoost	✗	0.95 ± 0.05	0.90 ± 0.05	0.86 ± 0.16	0.75 ± 0.17	0.78 ± 0.09
	✓	0.97 ± 0.04	0.92 ± 0.07	0.93 ± 0.06	0.97 ± 0.05	0.95 ± 0.04
XGBoost	✗	0.96 ± 0.04	0.87 ± 0.05	0.80 ± 0.19	0.75 ± 0.22	0.74 ± 0.10
	✓	**0.98 ± 0.03**	0.94 ± 0.06	**0.95 ± 0.07**	0.98 ± 0.03	0.96 ± 0.04
CatBoost	✗	0.96 ± 0.03	0.87 ± 0.06	0.76 ± 0.17	0.82 ± 0.23	0.75 ± 0.12
	✓	**0.98 ± 0.04**	0.95 ± 0.05	**0.95 ± 0.07**	0.99 ± 0.02	0.97 ± 0.03
LightGBM	✗	0.95 ± 0.03	0.89 ± 0.05	0.84 ± 0.17	0.75 ± 0.22	0.76 ± 0.12
	✓	**0.98 ± 0.03**	**0.96 ± 0.05**	**0.95 ± 0.07**	**1.00 ± 0.00**	**0.98 ± 0.03**

TABLE I

PERFORMANCE COMPARISON OF THE DIFFERENT MODELS TRAINED FOR THE PREDICTION OF THE NSCLC SUBTYPES WITH AND WITHOUT GAIN-BASED MISSING DATA IMPUTATION. RESULTS REPRESENT MEAN ± STANDARD DEVIATION IN NESTED CV. TOP RESULTS ARE SHOWN IN BOLD.

and many genes appearing among the most important in the two analyses. For instance, BNC1 and LPCAT1 which was shown to be upregulated in NSCLC, appeared among the top 5 impactful genes in both models. Other common genes were HN1B, JAG1, ALOX15B, NDUFA4L2 and C1orf116. This demonstrates the robustness of the predictions and the reproducibility of the results with different models.

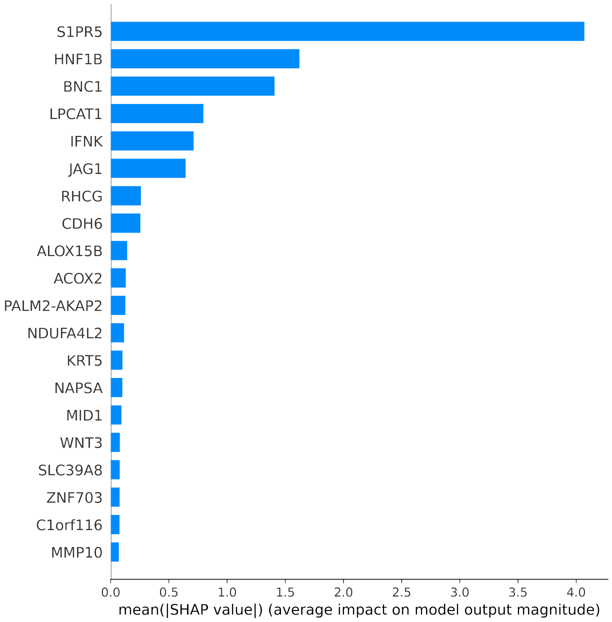

Fig. 4. SHAP summary plots of the LightGBM model trained on the GAIN-imputed RNA sequencing dataset. Top 20 genes are listed in a decreasing order of importance from top to bottom.

IV. CONCLUSIONS

In this study, we proposed a pipeline based on the GAIN network and gradient boosting ensemble models for the prediction of the histological subtypes of NSCLC from RNA sequencing data. We also provided interpretations of the predictions using the SHAP technique. Experimental results show that including all the RNA sequencing genes and leveraging a generative network to impute missing gene expressions improved the predictive models' performance. This pipeline is an automated, rapid and cost effective tool for the identification of the histological subtype of NSCLC. In a future step, we will integratively leverage transcriptomics data with imaging and clinical data in a predictive pipeline trained on a multi-institutional dataset.

ACKNOWLEDGMENT

This work was supported by the Canada Research Chairs, the National Science and Engineering Research Council of Canada and the Université de Montréal Roger Des Groseillers Research Chair in Hepatopancreatobiliary Surgical Oncology. BR, ST and SK are scientists of the Centre de recherche du Centre hospitalier de l'Université de Montréal supported by the Fonds de recherche du Québec - Santé.

REFERENCES

[1] Krishna Chaitanya Thandra et al., "Epidemiology of lung cancer," *Contemp Oncol (Pozn)*, vol. 25, no. 1, pp. 45–52, Feb. 2021.

[2] Meina Wang et al., "Toward personalized treatment approaches for non-small-cell lung cancer," *Nat Med*, vol. 27, no. 8, pp. 1345–1356, Aug. 2021.

[3] Menno Tamminga et al., "Immune microenvironment composition in non-small cell lung cancer and its association with survival," *Clin Transl Immunology*, vol. 9, no. 6, pp. e1142, June 2020.

[4] Bing-Yen Wang et al., "The comparison between adenocarcinoma and squamous cell carcinoma in lung cancer patients," *J Cancer Res Clin Oncol*, vol. 146, no. 1, pp. 43–52, Nov. 2019.

[5] Hiroaki Nomori et al., "Differences between lung adenocarcinoma and squamous cell carcinoma in histological distribution of residual tumor after induction chemoradiotherapy," *Discov Oncol*, vol. 12, no. 1, pp. 36, Sept. 2021.

[6] Hawazin Faruki et al., "Lung adenocarcinoma and squamous cell carcinoma gene expression subtypes demonstrate significant differences in tumor immune landscape," *J Thorac Oncol*, vol. 12, no. 6, pp. 943–953, Mar. 2017.

[7] Ralph Saber et al., "Radiomics using computed tomography to predict CD73 expression and prognosis of colorectal cancer liver metastases," *J Transl Med*, vol. 21, no. 1, pp. 507, July 2023.

[8] Shaimaa Bakr et al., "A radiogenomic dataset of non-small cell lung cancer," *Sci Data*, vol. 5, pp. 180202, Oct. 2018.

[9] Jinsung Yoon et al., "Gain: Missing data imputation using generative adversarial nets," 2018.

[10] Nitesh V. Chawla et al., "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, jun 2002.

[11] Scott Lundberg and Su-In Lee, "A unified approach to interpreting model predictions," 2017.

[12] Hongyu Chen et al., "Sphingosine 1-phosphate receptor, a new therapeutic direction in different diseases," *Biomed Pharmacother*, vol. 153, pp. 113341, July 2022.