Explainable Machine Learning to Identify Patient-specific Biomarkers for Lung Cancer

Masrur Sobhan
Knight Foundation School of Computing and Information Science
Florida International University
Miami, Florida, USA
msobh002@fiu.edu

Ananda Mohan Mondal
Knight Foundation School of Computing and Information Science
Florida International University
Miami, Florida, USA
amondal@fiu.edu

***Abstract*— *Background:* Lung cancer is the leading cause of cancer compared to other cancers in the USA despite being the most commonly diagnosed. The overall survival rate of lung cancer is not satisfactory even though having cutting edge treatment methods for cancers. Genomic profiling and biomarker gene identification of lung cancer patients due to genomic alteration may play role in the therapeutics of lung cancer patients. The biomarker genes identified by most of the existing methods (statistical and machine learning based) belong to the whole cohort or population. That is why different people having the same disease get the same kind of treatment which causes different outcomes in terms of success and side effects. So, the identification of biomarker genes for individual patients is very crucial to find efficacious therapeutics. *Methods*: In this study we propose a pipeline to identify lung cancer class-specific and patient-specific key genes which may help to formulate effective therapies for the lung cancer patients. We have used two subtypes of lung cancer- lung adenocarcinoma and lung squamous cell carcinoma to identify subtype-specific (class-specific) and patient-specific key genes using an explainable machine learning approach, SHAP. This approach provides scores to each of the genes for individual patients which tell us the attribution of each feature (gene) for each sample (patients). *Result*: In this study, we applied two variations of SHAP- tree explainer and gradient explainer for which tree-based classifier, XGBoost and deep learning-based classifier, convolutional neural network (CNN) was used as classification algorithms, respectively. The classification accuracy of XGBoost and CNN models 96.3% and 92.6% respectively. Both the SHAP explainers provided attribution scores for each of the genes of individual samples. The class-specific identified top 100 genes were compared with the differentially expressed genes (DEGs) as both of them represent the population-based biomarkers. We also showed that there are minimal number of common genes among the class-specific top-100 genes which validates that the genes are truly class-specific. Similarly, we identified the patient specific top-100 genes based on SHAP score. We found that there are very few genes common among the patients which implicates that the identified patient-specific genes belong to the individual lung cancer patients. Our results also show that there lots of common genes when we identified top-100 genes for the healthy individuals which is due to the less mutation and genomic alteration. *Conclusion*: The proposed approach is capable of identifying key biomarker genes using SHAP. In this study demonstrated the result of SHAP by comparing two different explainers in the context of tree-based and neural network-based classifiers trained on the lung cancer RNA-seq data. We also exhibited that the class-specific SHAP genes are biologically significant by comparing with the output of a statistical approach, DGE analysis. This study also determined patient-specific genes from SHAP scores which may provide many biological insights and aid personalized medicine. The proposed pipeline can be used to determine class-specific and patient-specific biologically relevant genes which can be used to for cancer patients' diagnosis.**

Keywords—lung cancer, explainable machine learning, precision medicine, patient-specific biomarkers

I. Introduction (*Heading 1*)

Cancer is a disease in which some cells of the body grow uncontrollably and spread to other organs of the body. Cancer is a genetic disease that is caused by the changes in the genes which control the cell's function, especially the growth and division of cells [1]. Three different kinds of genes are responsible for cancer - proto-oncogenes, tumor suppressor genes, and DNA repair genes [2], [3]. There are more than 100 types of cancers, but carcinomas are the most common type of cancer [1].

Lung cancer is the second most prevalent type of cancer [4], and it is the leading cause of death related to cancer in the United States [5]. There are mainly two types of lung cancer - non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) [6]. Two subtypes of NSCLC are lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC). There have been many studies where lung cancer biomarkers were identified. Some studies identified race-related biomarkers. [7]–[12] A recent study used a deep learning algorithm to find race-specific important genes in lung cancer [13]. There has been much work on classifying lung cancers using gene expression values. The researchers also applied various well-known and novel machine learning and deep learning techniques for the feature selection and classification of lung cancers and other cancer types[14]–[20]. But they have mostly used machine learning and deep learning models as "black boxes." Recently people have been using various approaches to explain the black box model. Some methods have been proposed to make the machine learning and deep learning models explainable, including Shapley Sampling [21], Relevance Propagation [22], LIME [23], ANCHOR [24], and DeepLIFT [25]. In 2017, a game theoretic approach named SHAP (SHapley Additive ExPlanation) [26] was introduced to explain the black box models. Recently there has been adequate work to explain the machine learning models using explainable machine learning algorithms, such as SHAP. Levy et al. used SHAP to discover important methylation states in different cell types and cancer subtypes [27]. Also, in a more recent study, SHAP was used to explain the deep learning model which classified the cancer tissues using RNA-sequence data [28]. Most of the studies identified the global features using SHAP values which represents the average impacts of the genes to that model [29]. Along with various machine learning approaches, researchers used various statistical tools, such as DESeq2 [30], edgeR [31] or LIMMA [32]to identify differentially expressed genes (DEGs) [33]–[38] . It was showed that DEGs help to identify potential genes associated with

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

pathogenesis and prognosis of lung cancer [39] which implicates that DEGs are able to identify important genes across the population. The study [39] identified different set of DEGs for different cohorts which means that DEGs are cohort dependent biomarker genes and they fail to identify patient-specific biomarkers. Study says that the same treatment provided to the patients with the same cancer type may show different kinds of outcome [40]. So, it is important to identify the patient-specific biomarkers which can be achieved by identifying local interpretable features by explaining the machine learning models.

We hypothesize that biomarker genes may express differently in different patients due to the variability of mutations of genes for which cohort-based gene therapy may not be beneficial to most of the patients. To solve this issue, identifying patient-specific important biomarker genes is very crucial and it may aid to precision medicine or personalized medicine. In this study, we developed a pipeline to discover both global and local NSCLC-associated genes using an explainable machine learning tool, SHAP. This study identified both class-specific and patient-specific genes based on SHAP scores by calculating global and local SHAP score respectively. To our knowledge, there has not been any study of identifying lung cancer patient-specific genes using SHAP.

The later part of this paper is ordered as follows. 'Materials and Methods' section includes the cohort analysis, preparation of the dataset, and the methods used for the research. 'Experimental Results' section provides the outcome of the research and analysis of the results. Finally, conclusions and the future scope will be discussed in the 'Conclusion' section.

II. Materials and Methods

A. Workflow of the study

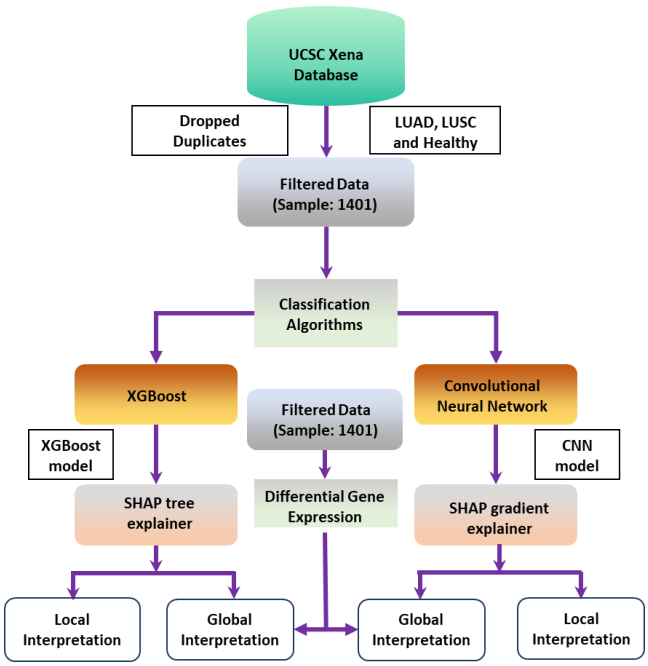

Fig. 1. Workflow of the study to identify patient-specific and class-specific genes.

The overall workflow of this study is shown in Fig 1.

The overall workflow of this study is as follows. At first the lung cancer tumor (LUAD and LUSC), and healthy tissue samples were downloaded from UCSC Xena database. Next the dataset was filtered by dropping the duplicate records of the same patients having the same tumor type. The filtered data was used to classify three different classes (LUAD, LUSC and healthy) using two different algorithms- XGBoost and CNN. Hyperparameters were tuned to achieve a higher classification accuracy. 5-fold cross validation was performed to measure the performance of the two algorithms. Then the two models from two different genre of classification algorithms – XGBoost from tree-based and CNN from deep learning-based classifiers - were used for interpretation using SHAP. As such, we used tree explainer technique for XGBoost and gradient explainer for CNN model for interpretation. Next, we analyzed the two different interpretation techniques to get class-specific genes and patient-specific genes. We also used a statistical tool DESEq2 to get the important genes across the populations.

B. Data Collection and Cohort Analysis

To characterize the lung-cancer associated mRNA, the expression profiles and clinical data associated with lung cancer were collected from UCSC Xena database [40]. The normal tissue samples were downloaded from Xena database and mskcc GitHub repository [41], [42]. There are 1415 samples, including 503 LUAD, 489 LUSC and 423 healthy, as shown in Table I.

TABLE I. Sample distribution and cohort analysis of lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC) and healthy samples

LUAD Tumor Samples	LUSC Tumor Samples	Healthy Samples
503	**489**	**423**
Total= 1415		

C. Data Preparation

Fourteen of 1415 samples were duplicates. We kept only one record of the same patients. So, the final cohort size for this analysis was 1401 with 492 LUAD tumor, 486 LUSC tumor, and 423 healthy samples, respectively. We used the dataset with FPKM values which were already log-normalized. Raw gene count dataset was also considered in this study. The data distribution of the three categories are well distributed and there are little chance of biasness towards the larger group. The final dataset consists of 1401 samples with 19,648 mRNA expression values. Then we used this dataset to classify LUAD, LUSC and healthy using a tree-based machine learning algorithm and a deep learning algorithm.

D. Classification Algorithms

We used two algorithms in our analysis - Extreme Gradient Boosting (XGBoost) [43] and Convolutional Neural Network (CNN) [44]. XGBoost is a decision tree-based machine learning algorithm which uses a process called boosting to help improve performance. It is an optimized gradient boosting algorithm through parallel processing, tree pruning, handling missing values and regularization to avoid bias or overfitting.

CNN is a type of deep neural network primarily used in image classification or computer vision applications. But it has also wide applications in analyzing tabular data. The convolution layers extract features from the samples. A small filter or kernel scans through the samples and extracts features from the samples. The following layer is the pooling layer which downsamples the feature map extracted by the convolution layer. The pooling layer runs a filter across the feature map and takes the specific information from that filter.

It translates the features' exact spatial information to latent information. The final pooling layer is then flattened out and transformed into a one-dimensional array, and fed it to the fully connected layers that predict the output.

The samples of each class were divided into 80/20 split in a stratified manner for training and testing respectively. 5-fold cross validation was used for measuring the classification performance. For the stratification of the samples, we used StratifiedKFold from scikit-learn library. Hyperparameters were also tuned to get optimized results from both XGBoost and CNN classifier.

Next, the contribution of all the features of individual samples for the two classifiers were determined. We wanted to identify the reasons of the models' success or accuracy. Feature contributions, both globally and locally, can decipher the models' accomplishment. That is why we applied an explainable machine learning tool which can identify the feature contribution that caused the models' success.

E. Global and Local Feature Interpretation

Global Feature Interpretation

The global features are a set of features that reflect the average behavior of a cohort of samples or patients. For global feature interpretation, we used two techniques: (a) DESeq2, a statistical tool and (b) SHAP (sHapley Additive exPlanations) a game theoretic approach. DESeq2 is a tool for differential gene expression analysis of RNA-seq data. It provides a list of important genes for a cohort of patients, which reflect the average or global impact of genes across the cohort.

SHAP is a game theoretic approach to explain the output of any machine learning model. It takes the machine learning or deep learning algorithms into account and then calculate score of each feature. First step to calculate the SHAP score is taking the differences of the model's prediction with and without a feature from all the coalition sets. Then by taking the average of all the values from each of the coalition set provides the SHAP score. In short, the average marginal contribution of a feature value across all possible coalition is the SHAP score. The collective SHAP values can show how much each predictor or feature contributes, either positively or negatively, to the target variable or output of the model. The collective SHAP values refer to the global features of the dataset.

Local Feature Interpretation

Local features are a set of features that reflect the characteristics or behavior of an individual sample or patient. Along with identifying global important features, SHAP identifies local important features as well. Each sample for each feature or predictor gets its own SHAP value. It increases the transparency by calculating the contributions of the predictors. Traditional feature importance or selection algorithms produce results across the entire population, not on each individual. The idea of local interpretability of SHAP was used for identifying patient-specific genes which may help devise the strategy for personalized treatment.

III. Experimental Result

A. Classification Accuracy

The dataset was divided into 80/20 split for training and testing. Also, a 5-fold cross validation was performed to measure the performance of the models. The testing accuracy of algorithms from 5 folds was measured and then the average was calculated to finalize the accuracy. Table II summarizes the results of 5-fold cross-validation. The testing accuracy of XGBoost and CNN were 96.3% and 92.6%, respectively.

TABLE II. Results of 5-Fold Cross-Validation. First Row: Distribtuiion of Actual Labeled data; Second row: Distribtuiion of correctly predicted data; Third row: average classification accuracy of 5-fold cross-validation.

	XGBoost			CNN		
	LUAD	LUSC	Healthy	LUAD	LUSC	Healthy
Actual Data	492	486	423	492	486	423
Correct Prediction	473	453	423	468	450	379
Accuracy	96.3%			92.6%		

B. Differential Gene Expression Analysis

Differential gene expression (DGE) analysis was performed using the statistical tool DESeq2. Raw counts of gene expression value were used in this analysis. We used 492 LUAD tumor samples, 486 LUSC tumor samples, 59 healthy tissues from LUAD patients and 51 healthy tissues from LUSC patients. We ran the DGE analysis tool on LUAD and LUSC samples separately to get the most important lung cancer subtype (LUAD and LUSC) specific-genes. These genes represent the average behavior of the population related to the subtypes (LUAD and LUSC). These genes can also be named as global features as they represented the average importance of the cohorts. We identified LUAD-specific differentially expressed genes (LUAD-DEGs) and LUSC-specific differentially expressed genes (LUSC_DEGs) based on the thresholds $|\log_2 \text{Fold-change}| > 3$ and adjusted p-value < 0.001, which provided us 1,037 and 1,773 genes, respectively.

C. Global Interpretability using SHAP

We used the explainable machine learning tool, SHAP, to identify the important genes by leveraging XGBoost and CNN classifier models. The important genes were compared with the differential gene expression genes derived from DESeq2 tool discussed in section 'B'. SHAP and DESeq2 tools both were used to identify the important genes across the population.

In our analysis total number of features (genes) used for XGBoost and CNN algorithms were 19,648. SHAP generates shapely score for each gene for each patient. The scores were then averaged across the samples of correctly classified classes. Thus, we got three sets of genes (LUAD-specific, LUSC-specific, and healthy-specific) with scores. We sorted the genes of each class based on the shapely values. Both XGBoost and CNN generated 5 different models because of five-fold validation. For global interpretation, we considered the average of the five models' output (five set of test data from 5-fold) from XGBoost and CNN. Next, we took the top 1037 genes from LUAD and 1773 genes from LUSC class each, same as the number of DEGs. The top genes of LUAD and LUSC classes were compared with LUAD-DEGs and LUSC-DEGs, respectively. From the analysis we noticed that tree explainer leveraging XGBoost model and gradient explainer for CNN classifier model were able to identify significant number of global genes for both LUAD and LUSC classes which is shown in Fig. 2. From the figure it is clear that XGBoost model identified 89 LUAD and 214 LUSC common genes with LUAD DEGs and LUSC DEGs

respectively. Whereas CNN only identifies 68 LUAD common and 218 LUSC common genes.

Next, we tried to see whether the sorted genes based on shapely values were truly class-specific genes. If these genes are truly class-specific genes then there must be zero or few overlap among the three groups (LUAD, LUSC, and Healthy). This scenario is shown in Fig. 3. We considered the top 100 genes from LUAD, LUSC and Healthy samples seperately. We found that there is no common gene among the three sets derived from both XGBoost and CNN. There are very few or no common genes when considered two of the three classes. Thus, we can conclude that the identified top-100 genes for three classes are truly class-specific.

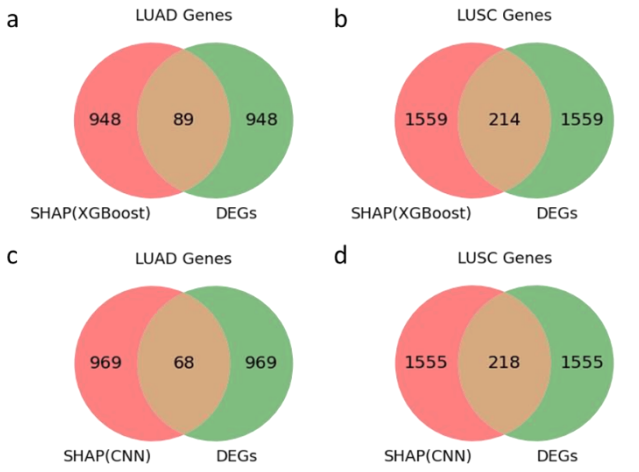

Fig. 2. Venn diagram of SHAP genes and DEGs. (a) and (c) represent the SHAP genes and DEGs for LUAD tumor. (b) and (d) represent the SHAP genes and DEGs for LUSC tumor.

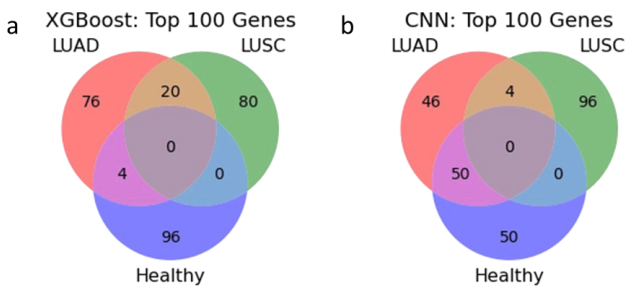

Fig. 3. Venn diagram of class-specific genes. (a) Top 100 genes of LUAD, LUSC and Healthy from XGBoost. (b) Top 100 genes of LUAD, LUSC and Healthy from CNN.

SHAP also provides us the information of important genes that contributed most to the model along with its shapely scores and class-impact of the genes. Fig. 4 shows the top 10 genes that contributed most to the XGBoost model (CNN is not shown). It also provides the information about the class-specific impact of the genes. For example, ACVRL1 (gene) contributed most to both healthy class and model output. TP63 contributed most to the LUSC class and slightly contributed to LUAD class. This means that TP63 gene could be an important biomarker for LUSC. Similarly, we can say that GOLM1 is an important biomarker gene for LUAD.

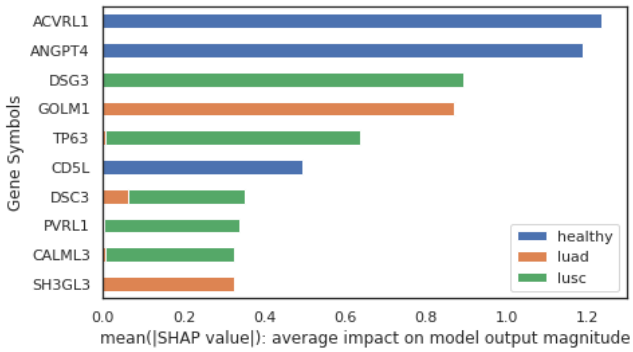

Fig. 4. Barplot of top 10 genes. The X-axis is the mean SHAP values scored by XGBoost. The values indicate the average score of the model output for the genes. Blue, orange and green color represents three different classes-healthy, luad and lusc. The Y-axis represents the top gene symbols determined by tree explainer.

D. *Local Interpretability using SHAP*

We identified the most salient genes from the XGBoost and Convolutional neural network (CNN) model using SHAP for each gene and each sample. This level of local interpretability helped to identify patient-specific biomarkers which may be used as personalized medicine or therapy. To get the scores for each of the genes and samples, we trained both XGBoost and CNN with 80% of the data and tested with the rest 20% of the data. We followed this procedure five times and each of the cases there were new 20% of the data in the testing set, thus providing 100% of data after testing. But XGBoost and CNN have 96.3% and 92.6% accuracy respectively which indicates that there are few false predictions. Next, we discarded the false predicted samples and kept only the true prediction.. Out of 1,401 samples, the number of correctly classified samples were 1,349 and 1,297 for XGBoost and CNN, respectively. So, each of the samples has all the genes scored based on shapely values. Next, we sorted all the genes in a descending order based on the score. Fig. 5 shows the most important genes for a single patient. This figure is a force plot for a particular LUSC tumor patient. The predicted SHAP score of this sample is 6.23 where the base value is 0.8992. This score indicates that the expression values of the genes for this patient has higher influence in the model. The base value is the average of the model output of LUSC class. The red arrow indicates that the genes pushed the model score higher and blue arrow indicates the genes that pushed the model score lower. From the gene expression values, we also see that DSG3 has high expression value and SLC4A4 has a low expression value thus the former is red and the latter is blue in color.

Fig. 5. Force plot of a single LUSC patient. The numerical values along with the genes are the expression values for this patient. This plot shows the most important genes for this particular patient.

Next, we tried to interpret the patient-specific genes of each of the samples. We wanted to make sure whether these genes are really patient-specific or not. To prove it we considered randomly chosen five LUAD and five LUSC samples. For each of the patients we picked the top 100 genes based on the SHAP score (higher SHAP scored genes were chosen). Our hypothesis was that if these genes are really patient-specific then there should be very few overlapping

genes as each individual has different mutations of genes and different expression profiles. To validate this hypothesis, we plotted heatmaps for LUAD samples which is shown in Figure 6. From the figure we see that there are very few overlapping genes from the tree explainer output leveraging XGBoost model. On the other hand, gradient explainer was able to find totally unique genes or almost zero overlapping genes among the five patients. The same scenario was observed with the LUSC patients as well (not shown). Also, the heatmap was plot across all the patients and very few overlapping genes were found. This indicates that, even though these samples are coming from the same class, SHAP was able to score patient-specific genes.

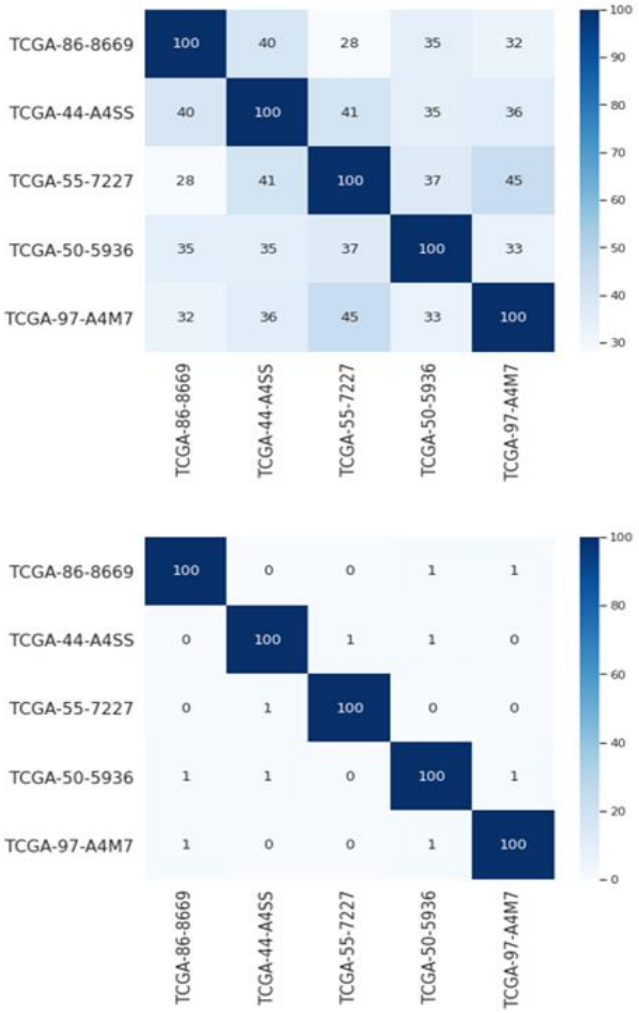

Fig. 6. Heatmap of LUAD patients with corresponding top 100 genes. (a) Heatmap of 100 genes derived from tree explainer (XGBoost model) across 5 LUAD patients. (b) Heatmap of 100 genes derived from gradient explainer (CNN model) across 5 LUAD patients.

Next, we hypothesized that there should be many overlapping genes in the healthy samples. This is because, there should be very few mutations of genes as the tissue samples is not affected by tumor. Again, we plotted heatmap with randomly chosen 5 healthy patients shown in Figure 7. From the heatmap, it is evident that there are lots of overlapping genes which proves our hypothesis.

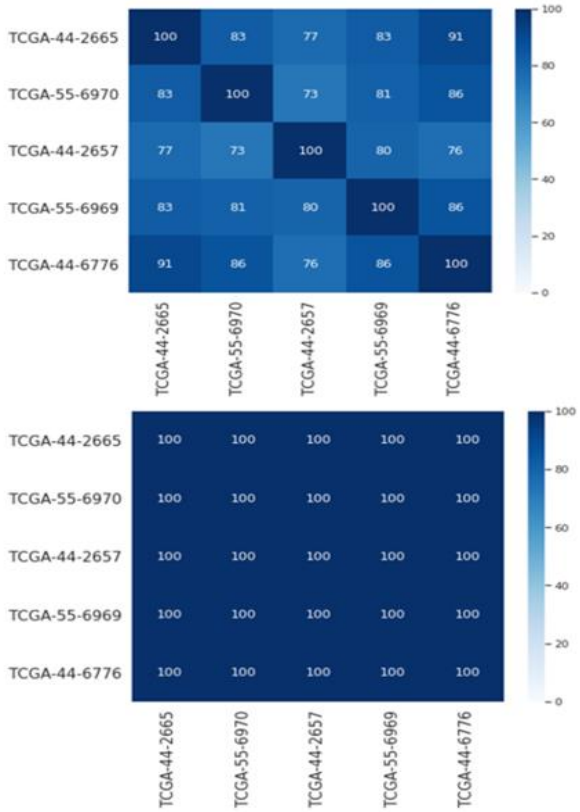

Fig. 7. Heatmap of healthy samples with corresponding top 100 genes. (a) Heatmap of 100 genes derived from tree explainer (XGBoost model) across 5 healthy samples. (b) Heatmap of 100 genes derived from gradient explainer (CNN model) across 5 healthy samples.

IV. Discussion

Most of the prior machine learning and deep learning works were involved in cancer classification and the algorithms were used as a 'black box'. But recently quite a few algorithms like- SHAP, LIME, ANCHOR, DeepLIFT, etc. algorithms have been introduced to explain the black box In this study, we used a tree-based algorithm, XGBoost, and a deep learning classifier CNN to classify the two types of lung cancer (LUAD and LUSC) and Healthy cohorts. Then the models generated by the classifiers were used in SHAP to explain the output of the models. SHAP is a unified approach to explain machine learning models which addresses the limitations of the black box models by explaining local and global features. We used two different explainers- tree explainer for XGBoost model and gradient explainer for CNN model. Tree explainer is a fast and exact method to estimate SHAP values for the tree models. Gradient explainer is another kind of SHAP explainer which can handle neural network models. In this study we tried to address an important task which may play vital role in the field of healthcare, personalized medicine, by adopting the proposed pipeline.

SHAP is able to identify global features that explains the impact of the model output of the whole population. To identify whether the SHAP explainability model were able to identify plausible features, we compared the output of the two explainers with the differential gene expression (DGE) analysis tool DESEq2 output. DGE tool was used as the reference to assess the correctness of the predicted genes from the explainers. Unlike DESEq2, there is no standard approach for selecting SHAP features (genes). That is why we ranked the genes based on the SHAP values and considered only top

ranked genes to compare with differentially expressed genes (DEGs). The outputs of both the explainers had some common genes with the output DGE analysis. We also tried to find out the common genes among the three classes (LUAD, LUSC and healthy) and found very few genes are overlapping among two classes and none of the genes were overlapped among the three classes. It tells us that SHAP was able to identify biologically significant class-specific genes.

One of the greatest challenges in healthcare is to identify patient-specific important biomarker which can aid to personalized medicine. In this study, we addressed this issue by explaining local interpretability of SHAP output. SHAP scores were assigned to every gene of every sample leveraging the modification of the game theoretic approach. So, each of the genes of every sample consists of a SHAP score which is then ranked based on the score. To explain the local interpretability, we considered the top 100 genes of each patient. We tried to find out the common genes among the samples of the same classes and found that tree explainer output has very few common genes across the samples, where gradient explainer has almost zero overlapping genes across the samples. It tells us that SHAP is able to identify patient-specific important genes in the tumor classes (LUAD and LUSC) as the tumor is more likely to work differently in different patients. We also noticed that there are lots of overlapping genes across the healthy samples. It is understandable because there is no mutation or few genomic alterations in the patients.

V. Conclusion

Majority of previous studies identified only cohort-based important genes or population-based important genes. But it was observed that different patients require different kind of treatment for the same disease due to the different genomic alterations and mutations. In this study we addressed two important issues of therapeutics- identification of subtype-specific (class-specific) and patient-specific genes. To solve these issues, we developed a pipeline which can identify both subtype-specific and patient-specific genes leveraging SHAP scores. For this analysis, we used rna-seq data of lung cancer to show that SHAP was able to identify both class-specific and patient-specific genes. This study shows that SHAP can be used to find many biological insights by identifying local (patient-specific) and global (class-specific) genes which may help to develop better therapeutics for individual patients.

All the output shown in this analysis is totally machine learning and deep learning-based computational outcome. These outcomes should be verified in the wet lab to strongly validate our result. If they can be verified in the wet lab, the pipeline can be used to identify important genes for any types of disease.

References

[1] "What Is Cancer? - National Cancer Institute." .

[2] G. Mendiratta, E. Ke, M. Aziz, D. Liarakos, M. Tong, and E. C. Stites, "Cancer gene mutation frequencies for the U.S. population," Nat. Commun. 2021 121, vol. 12, no. 1, pp. 1–11, Oct. 2021.

[3] P. A. Futreal et al., "A CENSUS OF HUMAN CANCER GENES," Nat. Rev. Cancer, vol. 4, no. 3, p. 177, 2004.

[4] "Lung and Bronchus Cancer — Cancer Stat Facts." https://seer.cancer.gov/statfacts/html/lungb.html (accessed Dec. 24, 2020).

[5] "Common Cancer Sites — Cancer Stat Facts." .

[6] K. Inamura, "Lung Cancer: Understanding Its Molecular Pathology and the 2015 WHO Classification," Front. Oncol., vol. 7, no. AUG, p. 193, Aug. 2017.

[7] S. D. Stellman et al., "Lung cancer risk in white and black Americans," Ann. Epidemiol., vol. 13, no. 4, pp. 294–302, 2003.

[8] V. L. Shavers and M. L. Brown, "Racial and ethnic disparities in the receipt of cancer treatment," J. Natl. Cancer Inst., vol. 94, no. 5, pp. 334–357, 2002.

[9] V. A. Zavala et al., "Cancer health disparities in racial/ethnic minorities in the United States," Br. J. Cancer, vol. 124, no. 2, pp. 315–332, 2021.

[10] C. Zeng, W. Wen, A. K. Morgans, W. Pao, X. O. Shu, and W. Zheng, "Disparities by race, age, and sex in the improvement of survival for major cancers: Results from the National Cancer Institute Surveillance, Epidemiology, and End Results (SEER) program in the United States, 1990 to 2010," JAMA Oncol., vol. 1, no. 1, pp. 88–96, 2015.

[11] D. B. White, L. M. Koehly, A. Omogbehin, and C. M. McBride, "African Americans' responses to genetic explanations of lung cancer disparities and their willingness to participate in clinical genetics research," Genet. Med., vol. 12, no. 8, pp. 496–502, Aug. 2010.

[12] J. D. Campbell et al., "Comparison of Prevalence and Types of Mutations in Lung Cancers Among Black and White Populations," JAMA Oncol., vol. 3, no. 6, pp. 801–809, Jun. 2017.

[13] M. Sobhan, A. Al Mamun, R. B. Tanvir, M. J. Alfonso, P. Valle, and A. M. Mondal, "Deep Learning to Discover Genomic Signatures for Racial Disparity in Lung Cancer," in 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Dec. 2020, pp. 2990–2992.

[14] I. H. Lee, G. H. Lushington, and M. Visvanathan, "A filter-based feature selection approach for identifying potential biomarkers for lung cancer," J. Clin. Bioinforma., vol. 1, no. 1, pp. 1–8, 2011.

[15] Z. Cai, D. Xu, Q. Zhang, J. Zhang, S. M. Ngai, and J. Shao, "Classification of lung cancer using ensemble-based feature selection and machine learning methods," Mol. Biosyst., vol. 11, no. 3, pp. 791–800, 2015.

[16] S. Tian, H. H. Chang, and C. Wang, "Weighted-SAMGSR: Combining significance analysis of microarray-gene set reduction algorithm with pathway topology-based weights to select relevant genes," Biol. Direct, vol. 11, no. 1, pp. 1–15, 2016.

[17] L. Zhang, L. Wang, B. Du, T. Wang, P. Tian, and S. Tian, "Classification of Non-Small Cell Lung Cancer Using Significance Analysis of Microarray-Gene Set Reduction Algorithm," Biomed Res. Int., vol. 2016, 2016.

[18] B. Shin et al., "Cascaded Wx: A novel prognosis-related feature selection framework in human lung adenocarcinoma transcriptomes," Front. Genet., vol. 10, no. JUN, pp. 1–11, 2019.

[19] R. Alanni, J. Hou, H. Azzawi, and Y. Xiang, "Deep gene selection method to select genes from microarray datasets for cancer classification," BMC Bioinformatics, vol. 20, no. 1, pp. 1–15, 2019.

[20] A. Al Mamun et al., "Multi-Run Concrete Autoencoder to Identify Prognostic lncRNAs for 12 Cancers," Int. J. Mol. Sci., vol. 22, no. 21, p. 11919, Nov. 2021.

[21] E. Štrumbelj and I. Kononenko, “Explaining prediction models and individual predictions with feature contributions,” Knowl. Inf. Syst., vol. 41, no. 3, pp. 647–665, Nov. 2014.

[22] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. R. Müller, and W. Samek, “On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation,” PLoS One, vol. 10, no. 7, p. e0130140, Jul. 2015.

[23] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier,” NAACL-HLT 2016 - 2016 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Demonstr. Sess., pp. 97–101, Feb. 2016.

[24] M. T. Ribeiro, S. Singh, and C. Guestrin, “Anchors: High-Precision Model-Agnostic Explanations.”

[25] A. Shrikumar, P. Greenside, and A. Kundaje, “Learning Important Features Through Propagating Activation Differences,” 34th Int. Conf. Mach. Learn. ICML 2017, vol. 7, pp. 4844–4866, Apr. 2017.

[26] S. Lundberg and S. Lee, “A unified approach to interpreting model predictions,” 2017.

[27] J. J. Levy, A. J. Titus, C. L. Petersen, Y. Chen, L. A. Salas, and B. C. Christensen, “MethylNet: An automated and modular deep learning approach for DNA methylation analysis,” BMC Bioinformatics, vol. 21, no. 1, pp. 1–15, Mar. 2020.

[28] M. Yap et al., “Verifying explainability of a deep learning tissue classifier trained on RNA-seq data,” Sci. Reports 2021 111, vol. 11, no. 1, pp. 1–12, Jan. 2021.

[29] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, “Explainable AI: A Review of Machine Learning Interpretability Methods,” Entropy 2021, Vol. 23, Page 18, vol. 23, no. 1, p. 18, Dec. 2020.

[30] E. Porcu et al., “Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome,” Nat. Commun. 2021 121, vol. 12, no. 1, pp. 1–9, Sep. 2021.

[31] Q. Sun et al., “Differential Expression and Bioinformatics Analysis of circRNA in Non-small Cell Lung Cancer,” Front. Genet., vol. 11, p. 1454, Nov. 2020.

[32] N. Shriwash, P. Singh, S. Arora, S. M. Ali, S. Ali, and R. Dohare, “Identification of differentially expressed genes in small and non-small cell lung cancer based on meta-analysis of mRNA,” Heliyon, vol. 5, no. 6, p. e01707, 2019.

[33] S. Wen et al., “Four differentially expressed genes can predict prognosis and microenvironment immune infiltration in lung cancer: a study based on data from the GEO,” BMC Cancer, vol. 22, no. 1, pp. 1–12, Dec. 2022.

[34] J. M. Xue, Y. Liu, L. H. Wan, and Y. X. Zhu, “Comprehensive Analysis of Differential Gene Expression to Identify Common Gene Signatures in Multiple Cancers,” Med. Sci. Monit., vol. 26, pp. e919953-1, Feb. 2020.

[35] S. Udhaya Kumar et al., “Analysis of differentially expressed genes and molecular pathways in familial hypercholesterolemia involved in atherosclerosis: A systematic and bioinformatics approach,” Front. Genet., vol. 11, pp. 1–16, Jul. 2020.

[36] M. I. Love, W. Huber, and S. Anders, “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2,” Genome Biol., vol. 15, no. 12, p. 550, Dec. 2014.

[37] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data,” Bioinformatics, vol. 26, no. 1, p. 139, Nov. 2010.

[38] G. K. Smyth, “Limma: Linear Models for Microarray Data.”

[39] M. Ni et al., “Identification of candidate biomarkers correlated with the pathogenesis and prognosis of non-small cell lung cancer via integrated bioinformatics analysis,” Front. Genet., vol. 9, no. OCT, pp. 1–14, 2018.

[40] “UCSC Xena.” .

[41] Q. Wang et al., “Unifying cancer and normal RNA sequencing data from different sources,” Sci. Data 2018 51, vol. 5, no. 1, pp. 1–8, Apr. 2018.

[42] “GitHub - mskcc/RNAseqDB.” .

[43] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., vol. 13-17-August-2016, pp. 785–794, Mar. 2016.

[44] K. O’Shea and R. Nash, “An Introduction to Convolutional Neural Networks,” Nov. 2015.