Deep Learning to Discover Genomic Signatures for Racial Disparity in Lung Cancer

Masrur Sobhan
Computing and Information Sciences
Florida International University
Miami, FL, USA
msobh002@fiu.edu

Abdullah Al Mamun
Computing and Information Sciences
Florida International University
Miami, FL, USA
mmamu009@fiu.edu

Raihanul Bari Tanvir
Computing and Information Sciences
Florida International University
Miami, FL, USA
rtanv003@fiu.edu

Mario Jacas Alfonso
Computing and Information Sciences
Florida International University
Miami, FL, USA
mjaca001@fiu.edu

Pablo Valle
Computing and Information Sciences
Florida International University
Miami, FL, USA
pvall017@fiu.edu

Ananda Mohan Mondal
Computing and Information Sciences
Florida International University
Miami, FL, USA
amondal@fiu.edu

2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 978-1-7281-6215-7/20/$31.00 ©2020 IEEE DOI: 10.1109/BIBM49941.2020.9313426

Abstract— Background: In the United States, African American Males (AAM) have the highest lung cancer incidence and mortality rate compared to European American Males (EAM). Cigarette is considered the major risk factor for lung cancer, but smoking alone fails to interpret the rationale for developing lung cancer between AAM and EAM. The higher rates of lung cancer among AAM occur even though they have lower smoking rates, smoke fewer cigarettes per day, and are less likely to be heavy smokers than EAM. Identifying genomic signatures such as key genes that can differentiate lung cancers between AAM and EAM will be a stepping stone to comprehend the disparity of lung cancer between AAM and EAM.

Method: The gene expression profiles of whole blood samples from AAM and EAM patients were used to identify the key genes that can differentiate the lung cancers between AAM and EAM. Due to the US population's imbalanced nature between AAM and EAM, the distribution of samples for the present study is also highly imbalanced (AAM: 15 and EAM: 153). Here, we developed a computational framework using a deep learning-based unsupervised feature selection approach, concrete autoencoder (CAE), which can select actual features rather than latent features. First, we showed that features such as differentially expressed genes (DEGs) discovered by a supervised statistical approach LIMMA could not differentiate lung cancers between AAM and EAM. Then we showed that the CAE could isolate essential features capable of differentiating lung cancers between AAM and EAM.

Results: The proposed framework using CAE was able to detect 34 key features/genes, which outperforms all sets of DEGs identified using three different thresholds on fold change. Using the selected 34 genes, the Random Forest classifier was able to classify lung cancers among AAM and EAM with 99% accuracy and only one false negative.

Conclusion: The proposed framework using CAE reveals the key genes that can differentiate lung tumors between AAM and EAM. These key genes can be used as biomarkers to understand the difference in lung cancer development between AAM and EAM. This study also showed that the CAE is capable of extracting relevant features from a highly imbalanced dataset.

Keywords—Concrete Autoencoder, Feature Selection, Lung Cancer Disparity

I. Introduction

Lung cancer is considered the second most prevalent type of cancer [1] and the leading cause of death in the United States [2]. Lung cancer represents approximately 12.7% of all cancer cases in the United States, and the African American Males (AAM) have the most lung cancer incidence and higher mortality rate than European American Males (EAM) [1]. Cigarette smoking is regarded as one of the major risk factors for lung cancer, but it is noticed that AAM has a lower smoking rate than the EAM [3],[4]. So, some other factors need to be considered, such as geographic regions, the origin of birth, diet, occupations, etc. [3]. Gene expression profiles for the tumor cells or whole blood samples from cancer patients can help identify the disparities among AAM and EAM.

The work done by Mitchell et al. [5] is relevant to our work. But some subtle differences make our research unique. They used expression profiles of tumor and normal tissues, whereas we used expression profiles of blood samples from cancer patients only. They identified key genes using statistical methods (ANOVA and t-test), where they considered tumor tissue as case and normal tissue as control. Besides, they considered both males and females together for their research. Here, we used male patients only, and the EAM patient was the control, while the AAM patient was the case. We used both a statistical approach (identifying DEGs) and a deep learning technique to identify key features from the dataset. Since the dataset was highly imbalanced, 15 AAM versus 153 EAM, a supervised algorithm will be biased to the larger group. Here, we developed a computational framework using a deep learning-based unsupervised algorithm, concrete autoencoder (CAE) [6], to identify the signature genes that can differentiate the lung cancers between AAM and EAM.

In this paper, first, we used LIMMA, an R package, which uses a statistical approach to identify differentially expressed genes (DEGs). Second, CAE was used to isolate the key genes from the original feature space. Finally, three state-of-the-art machine learning algorithms, Support Vector

Machine (SVM), Random Forest (RF), Logistic Regression (LR), were used to check the capabilities of the discovered features to differentiate lung cancers between AAM and EAM.

II. Materials and Methods

The whole blood gene expression profiles of lung cancer patients were obtained from the NCBI GEO database with accession ID, GSE135304 [7]. The dataset contains 712 human whole blood samples (311 males and 401 females) with the information of their demography, disease types, and nodule statuses. As our current hypothesis focuses on the male gender, we categorized 311 male samples based on lung cancers. We found that only 168 patients (15 AAM versus 153 EAM) have lung cancer information for male patients. It is noticeable that the dataset to be analyzed is highly imbalanced. Two approaches were used to isolate the key features that can explain the disparity in lung cancer development between AAM and EAM: (a) LIMMA, a statistical approach, and (b) Concrete Autoencoder, a deep learning-based unsupervised approach.

To check the capability of features selected above in differentiating lung cancers between AAM and EAM, three state-of-the-art classification algorithms including Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR) were used. We used Scikit-Learn, an open-source machine learning library in Python. 5-fold cross-validation was used for measuring the classification performance.

III. Results and Discussion

A. Feature selected by LIMMA

Table-I shows the number of DEGs for lung cancer in AAM compared to EAM. Three sets of DEGs (6, 45, and 317 genes) were found using three different thresholds on fold change, |logFC| ≥ 2.0, 1.0, and 0.5, respectively, with a P-Value ≤ 0.05. These DEGs were used as features for finding the disparity between AAM and EAM, applying three classification algorithms, SVM, RF, and LR.

Table I: The DEGs for AAM compared to EAM. P-Value ≤ 0.05. (↑: upregulated DEGs; ↓: downregulated DEGs)

	|logFC| ≥ 2	**|logFC| ≥ 1**	**|logFC| ≥ 0.5**
DEGs	↑0; ↓6	↑8; ↓37	↑67; ↓250

B. Feature selected by Concrete Autoencoder

Table II shows the distribution of the number of features obtained using two sets of runs, (a) 20 runs to select 317 features, and (b) 140 runs to select 20 features. For k = 317, 20 runs selected a total of 6340 features, of which, 5733 were unique. Similarly, for k = 20, 140 runs selected 2800 features, of which, 2265 were unique. Finally, the intersection of features selected by two sets of runs resulted in 34 features, which were considered as significant features.

Table II: Distribution of the number of features from CAE. First set of runs: k = 317 features and i = 20 runs; Second set of runs: k = 20 features and i = 140 runs.

# of features per run (k)	**# of runs (i)**	**# of total features**	**# of unique features**
317	20	6340	5733
20	140	2800	2265

C. Classification results using the DEGs

Figure 1 shows the performance of three classification algorithms (SVM, RF, and LR) using three sets of DEGs (317, 45, and 6 genes). All sets produced the same level of accuracy ranging from 94% to 97%. But the results are highly biased to the larger group of 153 EAM patients. Of 153 EAM, most of them were predicted correctly, and at most, 2 patients were predicted wrong. On the other hand, of 15 AAM, 5 to 8 patients were predicted wrong.

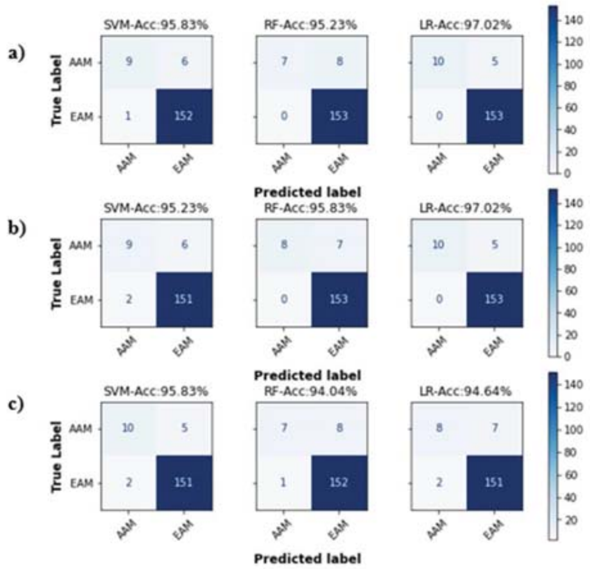

Fig 1. **Confusion matrix and corresponding accuracy using DEGs.** Classification algorithms used: SVM, RF, and LR. a) Results using 317 DEGs, b) Results using 45 DEGs, and c) Results using 6 DEGs.

D. Classification results using the features from CAE

Figure 2 shows the classification accuracy, and confusion matrix using three sets of features selected from two sets of CAE runs, 20 runs with k = 317 and 140 runs with k = 20. The conditions for feature selection are: (i) the number of features should be as few as possible, (ii) the accuracy using the selected features should be > 90%, and (iii) the wrong prediction for the smaller group (AAM) should be as minimum as possible.

Feature selection from 20 runs with k = 317: Counting genes that appeared in more than 2, 3, and 4 runs resulted in 72, 7, and 1 gene, respectively. Figure 2 (a & b) shows the classification performance using 7 and 72 genes, respectively. It is clear from Figure 2a that performance with 7 genes is the worst since all AAM samples are predicted wrong using SVM and RF. On the other hand, 72 genes perform the best, 100% accuracy using SVM, but poor results using RF (5 out of 15 AAM samples are predicted wrong).

Feature selection from 140 runs with k = 20: Figure 2c shows the confusion matrix using 31-gene set appeared in more than 3 runs. It is clear that the 31-gene set performs way better than that of the 7-gene set (2a) and slightly worse than the 72-gene set (2b). Since our goal is to select as few features as possible, 72 gene-set is high to design a wet lab experiment for further investigation.

Common features between two sets of runs: It is clear from Figure 2b and 4c that both sets of runs (20 runs with k = 317 and 140 runs with k = 20) have significant features capable of differentiating lung cancer between AAM and EAM. This observation motivated us to use the common 34 features between the two sets of runs.

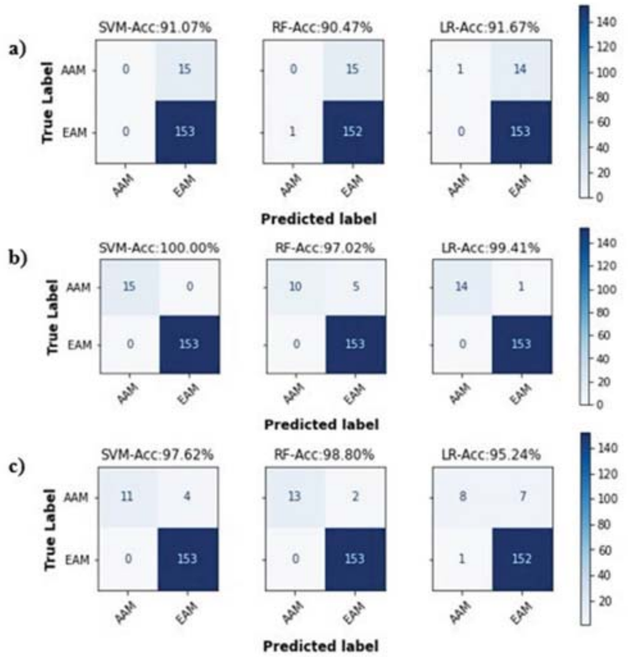

Fig 2. **Confusion matrix and corresponding accuracy for three classifiers (SVM, RF, and LR) using the features isolated by CAE.** a) Results using 7 genes: selected from 20 runs with k = 317, each gene appearing in more than 3 runs; b) Results using 72 genes: selected from 20 runs with k = 317, each gene appearing in more than 2 runs c) Results using 31 genes: selected from 140 runs with k = 20, each gene appearing in more than 3 runs.

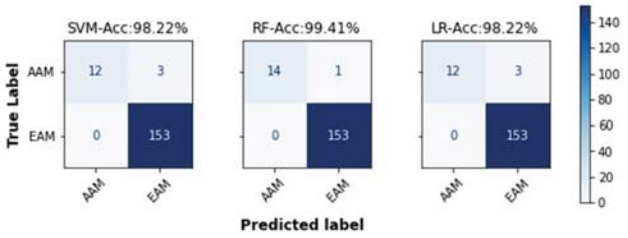

Fig 3. **Confusion matrix and corresponding accuracy using 34 common features from two sets of CAE run.** 20 runs with k = 317 and 140 runs with k = 20.

Figure 3 shows the classification performance using 34-gene set. It is clear that 34 gene-set was able to classify the lung cancers among AAM and EAM very well. The whole set of 153 EAM samples were predicted correctly by three classifiers. Of 15 EAM samples, RF made 14, and SVM and LR made 12 correct predictions. Comparing 31-gene set performance in Figure 2c, 34-gene set in Figure 3 produced better results. Thus, the signature of the 34-gene set can be used to develop a wet lab experiment to find the disparity of lung cancer between AAM and EAM. The list of 34 genes is provided below.

ACKR1, AIRE, ATP6V0D1, CAMP, CASP1, CCDC125, DEFA1B, DYSF, FAM210B, FCGR3B, FRAT2, GNAS, ILMN_1693262, ILMN_1762189, ILMN_1827887, ILMN_2338997, ILMN_3246805, ILMN_3278879, ITPRIP, LINC00173, LOC644936, MUC6, MXD1, NLRP12, POLR3C, RNA28S5, RNA28S5, S100P, SERPINA13P, SLC6A15, TDP1, TNPO3, UBA52, WARS.

IV. Conclusion and Future Work

We developed a computational framework using a deep learning-based unsupervised feature selection algorithm, Concrete Autoencoder (CAE), to identify the key genes related to the disparity in lung cancers between AAM and EAM. This study shows that whole blood samples carry the signature of health disparity in lung cancer between AAM and EAM.

ACKNOWLEDGMENT

This research was partially funded by NSF CAREER award #1651917 (transferred to #1901628) and NSF RAPID award #2037374.

References

[1] "Lung and Bronchus Cancer — Cancer Stat Facts." .
[2] "Common Cancer Sites — Cancer Stat Facts." .
[3] S. D. Stellman et al., "Lung cancer risk in white and black Americans," Ann. Epidemiol., vol. 13, no. 4, pp. 294–302, 2003.
[4] "Lung Cancer Disparities." .
[5] K. A. Mitchell, A. Zingone, L. Toulabi, J. Boeckelman, and B. M. Ryan, "Comparative transcriptome profiling reveals coding and noncoding RNA differences in NSCLC from African Americans and European Americans," Clin. Cancer Res., vol. 23, no. 23, pp. 7412–7425, Dec. 2017.
[6] A. Abid, M. F. Balin, and J. Zou, "Concrete autoencoders: Differentiable feature selection and reconstruction," 36th Int. Conf. Mach. Learn. ICML 2019, vol. 2019-June, pp. 694–711, 2019.
[7] "GEO Accession viewer." .
[8] B. Phipson, S. Lee, I. J. Majewski, W. S. Alexander, and G. K. Smyth, "Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression," Ann. Appl. Stat., vol. 10, no. 2, pp. 946–963, 2016.