Assessing Reproducibility and Veracity across Machine Learning Techniques in Biomedicine: A Case Study using TCGA Data

Ahyoung Amy Kim[a*], Samir Rachid Zaim[a-c], Vignesh Subbian[d-e]

a. Graduate Interdisciplinary Program in Statistics, The University of Arizona
b. Center for Biomedical Informatics & Biostatistics, University of Arizona Health Sciences
c. Department of Medicine, College of Medicine-Tucson, The University of Arizona
d. Department of Systems and Industrial Engineering, The University of Arizona
e. Department of Biomedical Engineering, The University of Arizona

* Corresponding author at: The University of Arizona, 617 N. Santa Rita Ave., P.O. Box 210089, Tucson, Arizona 85721, United States.
E-mail address: akim127@email.arizona.edu (A.A. Kim)

Abstract

Background: Many studies that aim to identify gene biomarkers using statistical methods and translate them into FDA-approved drugs have faced challenges due to lack of clinical validity and methodological reproducibility. Since genomic data analysis relies heavily on these statistical learning tools more than before, it is vital to address the limitations of these computational techniques.

Methods: Our study demonstrates these methodological gaps among most common statistical learning techniques used in gene expression analysis. To assess the classification ability and reproducibility of statistical learning tools for gene biomarker detection, six state-of-the-art machine learning models were trained on four different cancer data retrieved from The Cancer Genome Atlas (TCGA). Standard performance metrics including specificity, sensitivity, precision, and F1 score were evaluated to investigate the classification ability. For analysis of reproducibility, the identifiability of gene classifiers was examined by quantifying the consistency of the chosen classifier genes.

Results: Among the six state-of-the-art machine learning methods, the random forest had the best classification ability overall. Very few genes were selected by multiple methods, which suggests poor identifiability and reproducibility of statistical learning methods for gene expression data. Our results demonstrated the challenges of reproducing discoveries from gene expression analysis due to the inherent differences that exist in statistical machine learning methods.

Conclusion: Since statistical machine learning models can have large variations in high-dimensional settings such as analysis of gene expression data, transparent analysis procedures including data preprocessing, model parameterization, and evaluation and choice of interpretable models are required for clinical validity and utility.

Keywords: Reproducibility; Classification; Neoplasm; Machine Learning; TCGA

1. Introduction

Over the last decade, the goal of identifying gene biomarkers for precision medicine and clinical decision making and translating them into FDA-approved drugs has yielded limited results with significant federal investment in data-driven biomarker related research grants [1,2]. Despite these major investments, only a handful of new single-gene product biomarkers have been translated to FDA-approved clinical practice [3]. Among various reasons, the major challenges in gene biomarker detection may be attributed to the assumption of homogeneity in highly heterogenous human populations [4,5] as well as reproducibility of results from biomarker studies [6]. In this study, we focus on a major obstacle in translating gene biomarkers into clinical practice: a lack of methodological reproducibility induced by the large number of variabilities in statistical learning methods used in gene biomarker detection. We note that inconsistent gene expression classifiers is an emerging and important issue in bioinformatics [7]. However, in this data science-anchored age of genomic data analysis, it is crucial that we continue to address the limitations of these computational techniques as our research relies on them more than before. The goal of this work is to demonstrate the methodological gaps of the most common statistical learning techniques used in gene expression analysis using publicly-available data from The Cancer Genome Atlas (TCGA) and show the inherent variability that is unaccounted for in selecting one computational method over another.

Given that finding molecular abnormalities that cause cancer is critical for developing and implementing treatments, TCGA [8] provides genomic data such as RNA-Seq, miRNA-Seq, and methylation on 33 cancer types collected from more than 11,000 patients' tissue samples over 12 years since 2006. Many studies have utilized TCGA data to understand biological processes associated with cancer, to identify differentially expressed genes, to characterize molecular and genomic features of cancer, and finally to provide insights into potential treatments for cancer. In general, RNA-Seq transcriptome data are better suited for identifying transcriptomic changes associated with human cancers [9]. However, analyses of RNA-seq suffer from high dimensionality challenges, where there are more than 10,000 genes associated with one tissue sample from a subject, but far few samples. To address the high dimensionality and advance our understanding of cancer, including finding potential therapeutic targets, statistical and machine learning techniques and the accessible nature of TCGA data has been particularly useful in various studies.

Among those studies that apply machine learning algorithms to TCGA datasets, some studies introduced new methods to identify key driver genes more effectively, while others applied existing machine learning approaches. For example, a new method called DriverML that integrated Rao's score and supervised machine learning approach to identify cancer driver genes outperformed some of the existing methods [10]. Another study demonstrated a method to efficiently combine different types of molecular data and identify dominant biological processes active in tumor [11]. On the other hand, studies also used existing machine learning models such as the support vector machine (SVM)-recursive feature elimination and forward-SVM to screen differentially expressed genes and predict overall survival [12], to infer gene-interactions of a subset of genes extracted from TCGA dataset using graphical models [13], or to compare performances of machine learning techniques in predicting survival and metastasis outcomes in cancer [14]. The SVM method was also used to successfully classify adjacent normal and cancer samples in seven different cancer types using a subset of differentially expressed genes [9]. Moreover, TCGA data was used to validate differentially expressed genes in stage I papillary

thyroid carcinoma and normal adjacent tissues sequenced by RNA-Seq [15]. Some studies tried to resolve prevailing problems within gene expression data including interpretability and missing data by identifying minimal set of biomarkers to classify cancer status and different subtypes of cancer [16] or by applying imputation techniques to address missing values in important clinical features and dimensionality reduction techniques to reduce computational cost while maintaining good performance [17].

In this study, we conducted a comprehensive analysis of state-of-the-art supervised statistical and machine learning models, including deep learning models, and quantified variability and reproducibility of results from these models. By doing so, we demonstrate these machine learning approaches from the perspectives of interpretability, clinical relevance, performance, reproducibility, as well as general utility.

2. Materials and methods

2.1 TCGA data

TCGA contains 33 different cancer types and several different technical platforms. Of the 33 cancer types, we selected bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), liver hepatocellular carcinoma (LIHC), and lung adenocarcinoma (LUAD) to explore data with different sample sizes and to consider the extent to which data are separable between the normal and tumor cases. Considering different sample sizes and separability allows for examining how these factors affect the performance of various machine learning methods. To determine the separability of normal and tumor cases, we performed principal component analyses (PCAs) on all available datasets (See Appendix A Figure A1) in the TCGA2STAT [18] library in R environment. Although TCGA data can be obtained from another source such as the Genomic Data Commons (GDC) which may include more samples, the data can be more easily accessed via the TCGA2STAT library in the R environment and are already preprocessed and formatted to be ready for statistical analyses [18]. Each subject in the dataset is labeled as tumor or normal, allowing for a binary classification problem. Among different data platforms built in the TCGA2STAT library, we chose RNA-Seq data with reads per kilobase of transcript per million mapped reads (RPKM) to acquire normalized counts (See Appendix B Table B1). Genes that had all zero values across all samples or had less than 10 RPKM reads across all samples were removed. Even after removing those genes, the number of genes in the data was still more than 18,000 maintaining high dimensionality (19,518 genes for the BLCA, 20,404 genes for the BRCA, 18,274 genes for LIHC, and 19,870 genes for the LUAD).

2.2 Techniques

Various statistical and machine learning techniques used as part of our reproducibility analyses are listed in Table 1 along with their respective R packages and references. To demonstrate the extent of reproducibility attained in using the selected machine learning techniques, the following same tasks was performed using each model: (1) predict the label of interest and (2) produce a classifier (i.e., conduct feature selection to select a final model). The models were evaluated and assessed based on these two tasks, but rather than comparing and contrasting each to find the optimal classifiers – a task that has been extensively conducted in the

past – we focused more on the consistency and reproducibility of each of them. The overall workflow is summarized in Figure 1.

The ability of whether a model can perform feature selection is also noted in Table 1. Features selection is of particular importance since we want to determine which gene is important to predict the status of each patient and to measure consistencies of machine learning models. Although random forests, deep neural networks, and gradient boosting machine do not perform signed feature selection, we used functions in randomForest package [19] and h2o package [20] to compute the variable importance and sort the importance measures to see which genes are selected frequently or considered important in constructing the model.

Table 1.
Techniques, references, package interface, and feature selection ability.

Method	Ref.	Package Interface	Feature Selection
Lasso	[21]	glmnet [22]	Yes
Elastic Net	[23]	glmnet [22]	Yes
Support Vector Machine (SVM)	[24]	e1071 [25]	No
Random Forests (RF)	[26]	randomForest [19]	No
Artificial Neural Networks (ANN)	[27]	h2o [20]	No
Gradient Boosting Machines (GBM)	[28]	h2o [20]	No

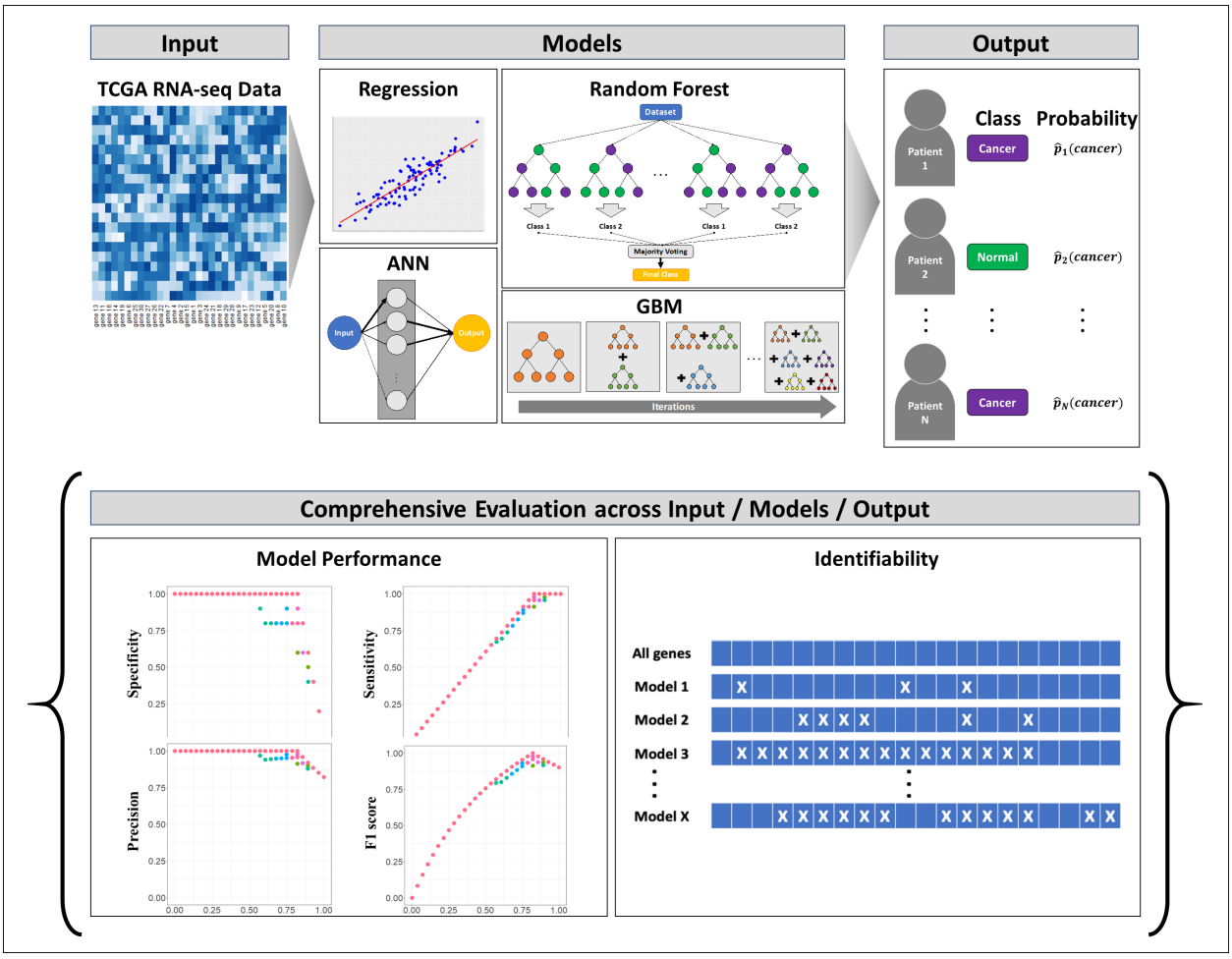

Figure 1.
The overall workflow of the analysis in this study. TCGA RNA-Seq data were used as input, and three regression models – the lasso, elastic net, and SVM – random forest, ANN, and GBM were trained. Then, comprehensive evaluations across input, models, and output were performed. Each model was evaluated based on 4 different performance metrics, and its identifiability was assessed based on the gene classifiers or important genes.

2.3 Study design

To evaluate classification ability of the models, training and validation were evaluated using accuracy metrics including test errors, sensitivity, specificity, precision, and F1 score on the final, cross-validated model for each algorithm (See Table 2). The identifiability assessment is based on the consistency throughout classifiers and variables selected by some important measures. We iterated the 'sampling scheme' to determine how stable the selected genes in the final gene classifiers were depending on what portion of the data was touched in training the model. For each iteration, the training set was created by random sampling of 60% from each group – tumor and normal, and the remaining was used as the test set. The training and test data were different at each iteration, but they were the same across different methods by setting the same seed number for random sampling. For deep learning models, instead of repeating the sampling scheme, models were trained with different configurations (See Table 3). Genes with high importance measures were examined for model identifiability.

Table 2.
Summary of Evaluation Metrics

Evaluation Metrics	**Calculation**
Test error	$\frac{1}{n}\sum_{i\in\{Test\}}(\widehat{y}_i \neq y_i)$
Specificity	$\frac{TN}{TN+FP}$
Sensitivity	$\frac{TP}{TP+FN}$
Precision	$\frac{TP}{TP+FP}$
F1 score	$\frac{2TP}{2TP+TN+FP+FN}$

Note: n is a test sample size, $\widehat{y}_i$ is a predicted label (normal or tumor), and y_i is a true label. TP stands for true positive, TN true negative, FP false positive, and FN false negative.

Table 3.
ANN and GBM models with different configurations.

Method	**# of Nodes**	**# of Trees**	**Learning Rate**	**Bag Fraction**
ANN1	200	-	-	-
ANN2	500	-	-	-

GBM1	-	1,000	0.1	0.5
GBM2	-		0.01	
GBM3	-		0.001	

A three-fold cross-validation (CV) was performed to tune parameters in each model. The number of trees in the random forests was chosen to be 500 after examining error versus the number of trees plot generated by initial fitting of the random forests model. Also, more details about prespecified parameter values for the random forest model as well as the hyperparameter values used by regression models (the lasso, elastic net, and SVM) are provided (See Appendix C Table C1).

2.4 Computing Environment

All analyses were conducted using the R statistical software (R version 3.5.3) [29] and the University of Arizona's High Performance Computing (HPC) environment. All packages have been listed with their specific R package implementation, and for reproducibility of results from this study, the source code for all the analyses has been deposited on GitHub: aykim127/Assess_Reproducibility_ML_TCGA.

3. Results

3.1 Separability using PCA

Among available TCGA RNA-Seq data retrieved from the TCGA2STAT library in R, 11 datasets had both normal and tumor cases. In order to select appropriate datasets for this study, we performed PCA on all of these 11 datasets (See Appendix A Figure A1 for the PCA results of these 11 datasets). Based on 2-dimensional visualizations of the first two components from PCA results, the normal and tumor groups were moderately separable in the BLCA and BRCA data, not separable in the LIHC data, and separable in the LUAD data (Fig. 2).

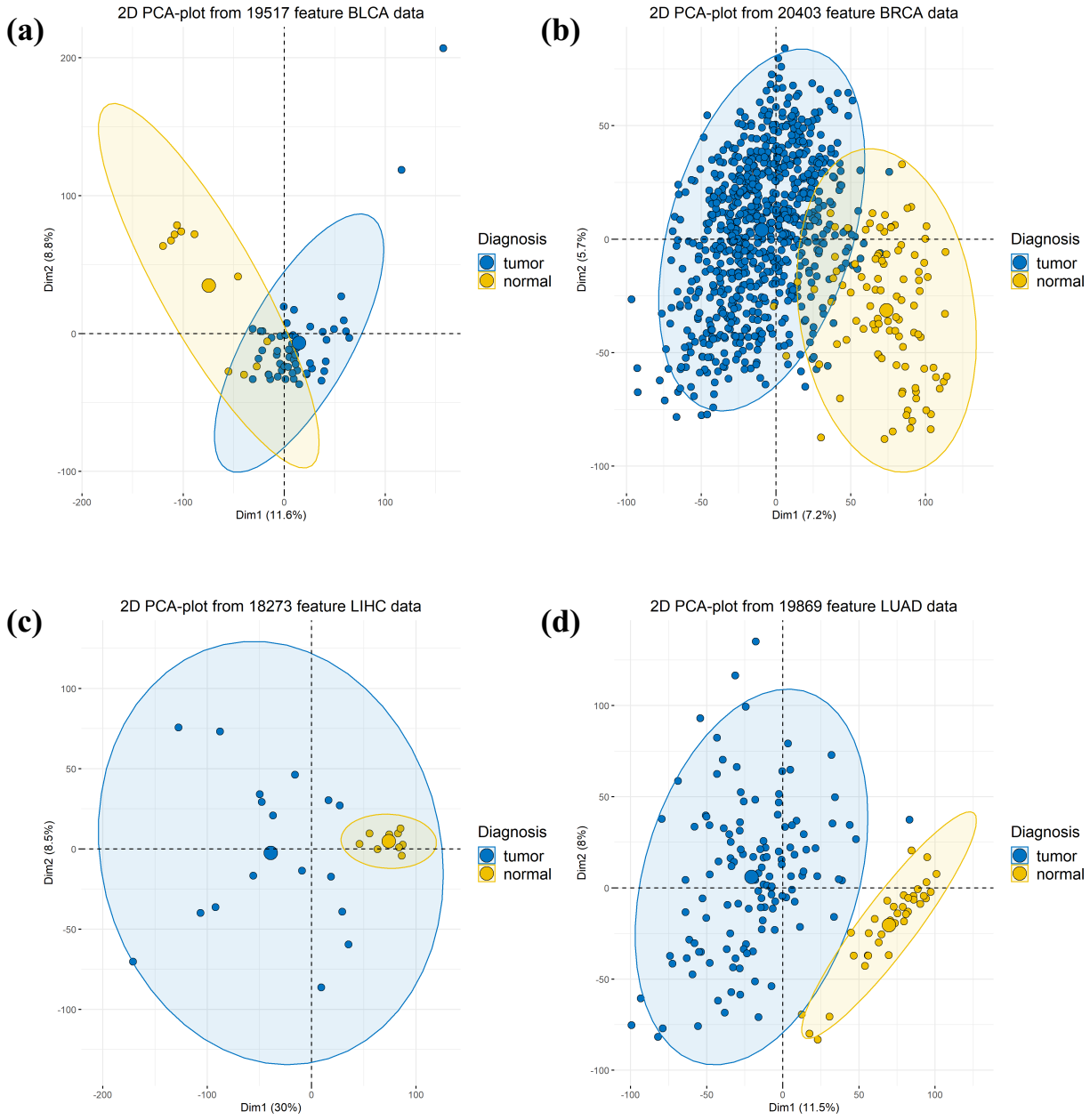

Figure 2.
The 2D plots of the first two components from the principal component analysis results. (a) There is a large overlap between the normal and tumor groups for the BLCA data, but there are also some cases that are separable. (b) Two groups are moderately separable for the BRCA data, and there are many cases outside of the overlapping region. (c) All normal cases are inside the elliptical region of the tumor group for the LIHC data. Based on this PCA result, the LIHC data is considered as a non-separable case. (d) The LUAD data is a separable case, since there is only one normal case that touches the boundary of the elliptical region for the tumor group.

3.2 Classification ability

Classification abilities were computed based on the test error, specificity, sensitivity, precision, and F1 score. The results from the best model among the lasso, elastic net, SVM, and random forest, and the best model among various deep learning models were visualized for each dataset. For the BLCA data, the random forest model was the best (test error = 0.083, specificity = 0.885, sensitivity = 0.584, precision = 0.978, and F1 score = 0.663) among the lasso, elastic net, SVM, and random forest, and for deep learning models, ANN2 was the best (Fig. 3). All datasets had the similar patterns of specificity, sensitivity, precision, and F1 score. Specificity and precision were relatively high, whereas sensitivity and F1 score were relatively low suggesting that the proportions of correctly classified groups in all methods were relatively low. Except for the BLCA data, little variation in different model fits from 30 iterations was observed (See Appendix D for detailed results) and was confirmed by overlapping lines in the figures.

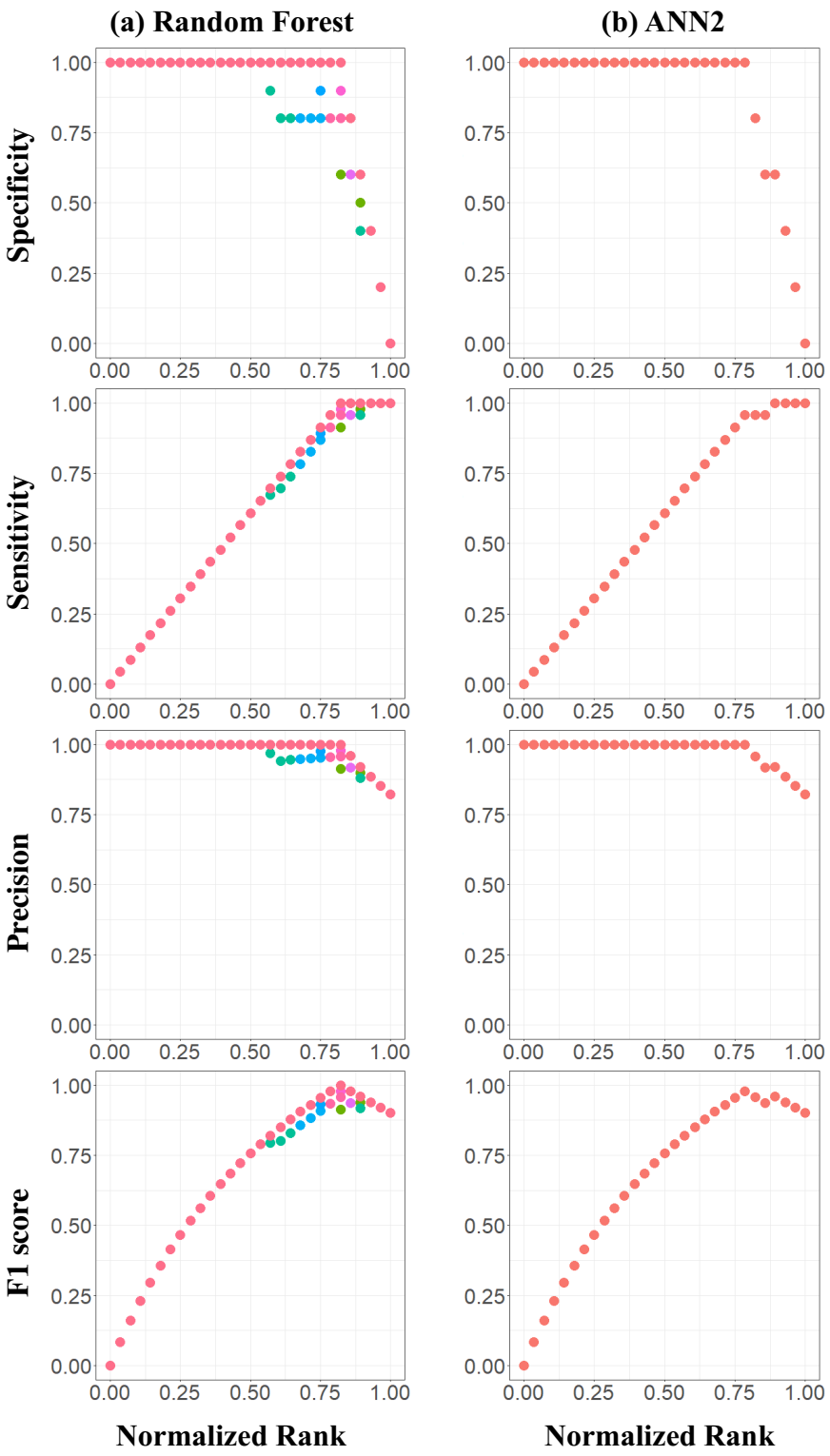

Figure 3.
Classification abilities of selected models measured by specificity, sensitivity, precision, and F1 score for the BLCA data: (a) random forest and (b) ANN2. For the random forest model, different colors represent different model fits from 30 iterations of random sampling.

3.2 Model identifiability

To examine the model identifiability, we selected a list of genes that had non-zero coefficients in the lasso and the elastic net models or higher importance measures in the random forests, ANN, and GBM (Table 4) and investigated the consistency in gene selection across all models. Genes selected by the lasso and elastic net were almost identical. For ANN models, even though only 10 genes were listed in Table 4, these genes had scaled importance of all above 0.9, and furthermore, all genes had scaled importance close to 0.6 or above. On the other hand, GBM models commonly selected C13orf36 gene. SFTPC gene was the only gene that was commonly selected by multiple methods including the lasso, elastic net, and random forests for the LUAD data (See Appendix E Tables E1-E3 for the results of the BLCA, BRCA, and LIHC data).

Table 4.
Model identifiability results for the LUAD data. The numbers next to gene names indicate the number of times that a gene had non-zero coefficient values among 30 iterations for the lasso and elastic net and the number of times that a gene was one of genes with top Gini index measure among 30 iterations for the random forest. For deep learning models, the listed genes were top genes according to their scaled importance measures.

Methods		**Selected Genes**
Lasso		SFTPC(30), SCGB1A1(12), SFTPA1(4)
Elastic Net		SFTPC(30), SCGB1A1(12), SFTPA1(4)
RF		SFTPC(13), LIMS2(10), EMP2(9), CLIC5(9), ITLN2(7), FAM189A2(7), STX11(6), PYCR1(6), CAV1(6)S
ANN	ANN1	C9orf4, ELMO1, SF3A1, TP53INP2, COQ5, WDHD1, NCRNA00176, ADRA1A, CPLX4, POMT2
	ANN2	C7orf23, HCP5, EBF3, MAPK12, MKS1, NFE2L1, ANAPC1, HES7, SLC37A1, CYP46A1
GBM	GBM1	C13orf36
	GBM2	C13orf36
	GBM3	C13orf36

4. Discussion

There has been a growing interest in biomarker research for precision medicine [30], and consequently, the volume of literature devoted to biomarker identification and characterization has been expanding in parallel [2]. Although gene expression microarray data has been the major tool in traditional biomarker research, RNA-Seq data has emerged as a popular source for biomarker identification, given their reproducible nature and high-resolution expression [31] For example, RNA-Seq data has been shown to be capable of successfully identifying biomarker signatures for different types of cancer using TCGA as one source of data [32]. Despite the capability and advantages of the RNA-Seq data, one challenging aspect of biomarker identification using this type of data is that among more than 10,000 genes in these RNA-Seq data, only a small subset of DNA biomarkers will be related to specific diseases [33]. To identify

this subset of biomarkers, various statistical and machine learning algorithms [9,32], including deep learning methods [34], that are capable of feature selection or measuring importance of these features have been widely adopted. However, these models tend to have low stability of feature selection due to high dimensionality of gene expression data and low sample size of clinical datasets [33]. In addition, deep learning models suffer from overfitting and high variance for such high dimensional and low sample size data [35], and they are usually more difficult to interpret and to understand the algorithms built inside the model, thereby limiting their effectiveness especially in the medical domain [36], where human experts' understanding of results from these models can be critical. Furthermore, even if a model identifies a marker as important, it may not be clinically relevant as one study has shown that only a handful of 150 identified biomarkers associated with tongue squamous cell carcinoma are found to be clinically valid [37].

Although there are many studies aimed at identifying a set of genes that are crucial to cancer prevention and early diagnosis, many of the preclinical cancer studies, including those published in top-tier journals could not be reproduced [38]. One major reason for non-reproducibility was due to inappropriate use of statistics [38]. Since machine learning algorithms, are popular in gene identification studies, assessing the reproducibility of these models calls for significant attention among the scientific community. In this study, the classification ability, model identifiability, and reproducibility of state-of-the-art statistical machine learning and deep learning models were assessed.

The low stability of feature selection and the problem of model identifiability were apparent from our results. As we iterated the random sampling process, the input data to train models were changed, selecting different genes and resulting inconsistency in gene classifiers. Depending on the portion of data used as the training set, different subsets of genes were selected, or none of genes was selected more than the half of the total iteration times based on their importance measures. For the ANN models, genes were not differentiated by the importance measures. There was no overlap among the gene classifier across all methods for all datasets included in this study, and this suggests low stability of feature selection in machine learning and deep learning models in applications to high-dimensional data. The high dimensionality of our data also has an impact on the classification ability suggested by overall low sensitivity. The overall classification ability did not improve even for the BRCA data, which is almost 30 times and 10 times larger than the LIHC data and the BLCA data, respectively. When data is high dimensional, the model predictability and classification accuracy can be lowered by overfitting, which occurs when the model is over trained on the training dataset and properly, and even perfectly classifies the training labels but not the test dataset due to lack of generalization capability [39].

Overall, the classification ability was roughly the same and most of selected models poorly performed for the chosen dataset. The ratio of the number of events to the number of predictor variables knowns as Events Per Variable (EPV) is a factor that can affect the performance of binary logistic regression analysis, and 10 EPV rule has been widely adopted in many studies [40]. Although 10 EPV rule has not been fully supported by many findings, the data used in this study have EPV much below 10 in addition to small sample sizes compared to the number of predictor variables. The low EPV together with the overall small sample sizes could have led to overall poor performances of chosen models.

For all four datasets, most of deep learning models did not perform well. Considering the fact that the BRCA data alone took more than 3 hours to train 3 different GBM models, these

deep learning models were not efficient and did not provide performance gains (See Appendix F). Deep learning models have been popular choices for studies related to classification or prediction problems in biomedicine, but the larger question is whether such models are appropriate for various sample sizes. Despite high prediction accuracy, most deep learning models are black-box models and are not as interpretable as penalized regression models with explicit mathematical formulations. In order for deep learning models to have more effectiveness in the medicine, models that have the ability to interpret the decisions and underlying algorithms explicitly [36], while maintaining a high-level of prediction accuracy are needed.

This study is limited by the naïve assumption of gene independence, since genes form complex mechanistic networks that represent their roles and biological functions that they regulate. However, if techniques do not agree in the simple scenario where genes are treated as identically and independently distributed, then it will be highly unlikely to attain unity with much more difficulty in the complex task of inferring a gene network. Another limitation of this study is the class imbalance within the dataset, which can make the model prediction biased by allowing models choose the majority class to achieve higher prediction accuracy in that majority class. We could match tumor and normal pairs, but it would have greatly decreased the sample sizes of all four datasets. Furthermore, to increase stability and performance of our models by reducing the size of data, we could have selected a subset of genes and trained our models. However, if we were to subset genes arbitrarily, clinical relevance of those selected genes may not be justified, and therefore, we chose to limit the parameter search space instead by setting specific values for some parameters.

5. Conclusion

This study assessed the reproducibility and veracity of the state-of-the-art statistical machine learning models by quantifying classification ability and model identifiability using high dimensional TCGA data. Model selection and feature selection are open-ended questions in statistics and machine learning, as no one model can guarantee to recover the true signal nor guarantee universal optimal performance. Our results showed that since various machine learning techniques use different underlying algorithms to make prediction and perform feature selection, it is difficult for the end-users to reproduce discoveries from the genomic data analysis and to choose proper models. Hence, it is important to conduct transparent analyses in the research community, especially in the field of biomedicine, to provide end-users with the means to select the most appropriate tools for their task. Furthermore, more transparency-driven reproducibility studies must be conducted to improve reliability and provide a more honest evaluation of the capabilities of machine learning models especially in the task of biomarker discovery.

Funding sources

- VS was supported, in part, by the National Science Foundation under grant #1838745.

Summary Table

What was already known on the topic:

- Non-reproducibility of cancer studies is prevalent, and often connected to inappropriate use of statistical methods.
- Different types of machine learning algorithms have been adopted to identify gene classifiers for various types of cancer, but no one algorithm can guarantee to recover the true signal or universal optimal performance.

What this study added to our knowledge:

- The consistency and reliability of state-of-the-art machine learning techniques were assessed by a case study using TCGA data, and the limitation of these machine learning algorithms was demonstrated by evaluations of model performance and identifiability.
- Genomics data analysis relies heavily on these statistical learning methods more than before, so it is important to acknowledge and address the limitations of these computational techniques.

Acknowledgements

- The results published or shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.
- The authors would like to acknowledge the University of Arizona's High-Performance Computing (HPC) for providing the space and computing hours to conduct our simulation studies and analyses.

References

[1] S.F. Terry, Obama's precision medicine initiative, Genet. Test. Mol. Biomarkers. 19 (2015) 113–114. doi:10.1089/gtmb.2015.1563.

[2] A.S. Ptolemy, N. Rifai, What is a biomarker? Research investments and lack of clinical integration necessitate a review of biomarker terminology and validation schema, Scand. J. Clin. Lab. Invest. 70 (2010) 6–14. doi:10.3109/00365513.2010.493354.

[3] S.R. Zaim, Q. Li, A.G. Schissler, Y.A. Lussier, Emergence of pathway-level composite biomarkers from converging gene set signals of heterogeneous transcriptomic responses, in: Pacific Symp. Biocomput., World Scientific Publishing Co. Pte Ltd, 2018: pp. 484–495. doi:10.1142/9789813235533_0044.

[4] D.C. Wang, X. Wang, Systems heterogeneity: An integrative way to understand cancer heterogeneity., Semin. Cell Dev. Biol. 64 (2017) 1–4. doi:10.1016/j.semcdb.2016.08.016.

[5] E.A. Mroz, J.W. Rocco, The challenges of tumor genetic diversity., Cancer. 123 (2017) 917–927. doi:10.1002/cncr.30430.

[6] L.M. McShane, In Pursuit of Greater Reproducibility and Credibility of Early Clinical Biomarker Research, Clin. Transl. Sci. 10 (2017) 58–60. doi:10.1111/cts.12449.

[7] J. Massagué, Sorting Out Breast-Cancer Gene Signatures, N. Engl. J. Med. 356 (2007) 294–297. doi:10.1056/NEJMe068292.

[8] J.N. Weinstein, E.A. Collisson, G.B. Mills, K.R. Mills Shaw, B.A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J.M. Stuart, The Cancer Genome Atlas Pan-Cancer analysis project, 2013. doi:10.1038/ng.2764.

[9] L. Peng, X.W. Bian, D.K. Li, C. Xu, G.M. Wang, Q.Y. Xia, Q. Xiong, Large-scale RNA-

Seq Transcriptome Analysis of 4043 Cancers and 548 Normal Tissue Controls across 12 TCGA Cancer Types, Sci. Rep. 5 (2015) 13413. doi:10.1038/srep13413.
[10] Y. Han, J. Yang, X. Qian, W.C. Cheng, S.H. Liu, X. Hua, L. Zhou, Y. Yang, Q. Wu, P. Liu, Y. Lu, DriverML: a machine learning algorithm for identifying driver genes in cancer sequencing studies, Nucleic Acids Res. 47 (2019) e45. doi:10.1093/nar/gkz096.
[11] T. Bismeijer, S. Canisius, L.F.A. Wessels, Molecular characterization of breast and lung tumors by integration of multiple data types with functional sparse-factor analysis, PLoS Comput. Biol. 14 (2018) 1–28. doi:10.1371/journal.pcbi.1006520.
[12] R.Z. Dong, X. Yang, X.Y. Zhang, P.T. Gao, A.W. Ke, H. chuan Sun, J. Zhou, J. Fan, J. bin Cai, G.M. Shi, Predicting overall survival of patients with hepatocellular carcinoma using a three-category method based on DNA methylation and machine learning, J. Cell. Mol. Med. 23 (2019) 3369–3374. doi:10.1111/jcmm.14231.
[13] H. Zhao, Z.-H. Duan, Cancer Genetic Network Inference Using Gaussian Graphical Models, Bioinform. Biol. Insights. 13 (2019) 117793221983940. doi:10.1177/1177932219839402.
[14] L. Tapak, N. Shirmohammadi-Khorram, P. Amini, B. Alafchi, O. Hamidi, J. Poorolajal, Prediction of survival and metastasis in breast cancer patients using machine learning classifiers, Clin. Epidemiol. Glob. Heal. (2018) 1–7. doi:10.1016/j.cegh.2018.10.003.
[15] J. Han, M. Chen, Y. Wang, B. Gong, T. Zhuang, L. Liang, H. Qiao, Identification of Biomarkers Based on Differentially Expressed Genes in Papillary Thyroid Carcinoma, Sci. Rep. 8 (2018) 9912. doi:10.1038/s41598-018-28299-9.
[16] M. Sherafatian, Tree-based machine learning algorithms identified minimal set of miRNA biomarkers for breast cancer diagnosis and molecular subtyping, Gene. 677 (2018) 111–118. doi:10.1016/j.gene.2018.07.057.
[17] M.C. Rendleman, J.M. Buatti, T.A. Braun, B.J. Smith, C. Nwakama, R.R. Beichel, B. Brown, T.L. Casavant, Machine learning with the TCGA-HNSC dataset: Improving usability by addressing inconsistency, sparsity, and high-dimensionality, BMC Bioinformatics. 20 (2019) 1–9. doi:10.1186/s12859-019-2929-8.
[18] Y.W. Wan, G.I. Allen, Z. Liu, TCGA2STAT: Simple TCGA data access for integrated statistical analysis in R, Bioinformatics. 32 (2016) 952–954. doi:10.1093/bioinformatics/btv677.
[19] A. Liaw, M. Wiener, Classification and Regression by randomForest, 2002. http://www.stat.berkeley.edu/.
[20] A. Candel, E. Ledell, A. Bartz, Deep Learning with H2O, 2018. http://h2o.ai/resources/ (accessed September 10, 2019).
[21] R. Tibshirani, Regression Shrinkage and Selection Via the Lasso, J. R. Stat. Soc. Ser. B. 58 (1996) 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x.
[22] T. Hastie, J.Q. Stanford, Glmnet Vignette, 2016. http://cran.us.r-project.org (accessed September 10, 2019).
[23] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (2005) 301–320. doi:10.1111/j.1467-9868.2005.00503.x.
[24] J.A.K. Suykens, J. Vandewalle, Least squares support vector machine classifiers, Neural Process. Lett. 9 (1999) 293–300. doi:10.1023/A:1018628609742.
[25] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, A.W. Maintainer, The e1071 Package, 2005.
[26] Breiman L., Machine Learning, 45(1), 5–32., Stat. Dep. Univ. California, Berkeley, CA

94720. (2001). doi:10.1023/A:1010933404324.
[27] J. Zurada, Introduction to artificial neural systems, 1992. http://www.jaicobooks.com/j/pdf hed/j-878 artificial neural systems.pdf (accessed September 10, 2019).
[28] J.H. Friedman, Greedy Function Approximation: A Gradient Boosting Machine, 2001.
[29] R. Ihaka, R. Gentleman, R: A Language for Data Analysis and Graphics, J. Comput. Graph. Stat. 5 (1996) 299–314. doi:10.1080/10618600.1996.10474713.
[30] F.S. Collins, H. Varmus, A new initiative on precision medicine, N. Engl. J. Med. 372 (2015) 793–795. doi:10.1056/NEJMp1500523.
[31] H. Han, X. Jiang, Disease Biomarker Query from RNA-Seq Data, Cancer Inform. 13s1 (2014) CIN.S13876. doi:10.4137/CIN.S13876.
[32] I.H. Wei, Y. Shi, H. Jiang, C. Kumar-Sinha, A.M. Chinnaiyan, RNA-Seq accurately identifies cancer biomarker signatures to distinguish tissue of origin, Neoplasia. 16 (2014) 918–927. doi:10.1016/j.neo.2014.09.007.
[33] W. Awada, T.M. Khoshgoftaar, D. Dittman, R. Wald, A. Napolitano, A review of the stability of feature selection techniques for bioinformatics data, Proc. 2012 IEEE 13th Int. Conf. Inf. Reuse Integr. IRI 2012. (2012) 356–363. doi:10.1109/IRI.2012.6303031.
[34] C.A. Targonski, C.A. Shearer, B.T. Shealy, M.C. Smith, F.A. Feltus, Uncovering biomarker genes with enriched classification potential from Hallmark gene sets, Sci. Rep. 9 (2019) 9747. doi:10.1038/s41598-019-46059-1.
[35] B. Liu, Y. Wei, Y. Zhang, Q. Yang, Deep Neural Networks for High Dimension, Low Sample Size Data, in: Proc. Twenty-Sixth Int. Jt. Conf. Artif. Intell. {IJCAI-17}, 2017: pp. 2287–2293. doi:10.24963/ijcai.2017/318.
[36] A. Holzinger, C. Biemann, C.S. Pattichis, D.B. Kell, What do we need to build explainable AI systems for the medical domain?, (2017) 1–28. http://arxiv.org/abs/1712.09923.
[37] A.A. Hussein, T. Forouzanfar, E. Bloemena, J. de Visscher, R.H. Brakenhoff, C.R. Leemans, M.N. Helder, A review of the most promising biomarkers for early diagnosis and prognosis prediction of tongue squamous cell carcinoma, Br. J. Cancer. 119 (2018) 724–736. doi:10.1038/s41416-018-0233-4.
[38] C.G. Begley, Six red flags for suspect work., Nature. 497 (2013) 433–434. doi:10.1038/497433a.
[39] M. Daoud, M. Mayo, A survey of neural network-based cancer prediction models from microarray data, Artif. Intell. Med. 97 (2019) 204–214. doi:10.1016/j.artmed.2019.01.006.
[40] M. van Smeden, J.A.H. de Groot, K.G.M. Moons, G.S. Collins, D.G. Altman, M.J.C. Eijkemans, J.B. Reitsma, No rationale for 1 variable per 10 events criterion for binary logistic regression analysis, BMC Med. Res. Methodol. 16 (2016) 163. doi:10.1186/s12874-016-0267-3.