Computers in Biology and Medicine 153 (2023) 106544

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

An explainable AI-driven biomarker discovery framework for Non-Small Cell Lung Cancer classification

Kountay Dwivedi [a], Ankit Rajpal [a,*], Sheetal Rajpal [b], Manoj Agarwal [c], Virendra Kumar [d], Naveen Kumar [a]

[a] Department of Computer Science, University of Delhi, Delhi, India
[b] Dyal Singh College, University of Delhi, Delhi, India
[c] Hansraj College, University of Delhi, Delhi, India
[d] Department of Nuclear Magnetic Resonance Imaging, All India Institute of Medical Sciences, New Delhi, India

ARTICLE INFO

Keywords:
Explainable AI
Non-Small Cell Lung Cancer
Biomarkers
Classification
Neural network
Machine learning

ABSTRACT

Non-Small Cell Lung Cancer (NSCLC) exhibits intrinsic heterogeneity at the molecular level that aids in distinguishing between its two prominent subtypes — Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC). This paper proposes a novel explainable AI (XAI)-based deep learning framework to discover a small set of NSCLC biomarkers. The proposed framework comprises three modules — an autoencoder to shrink the input feature space, a feed-forward neural network to classify NSCLC instances into LUAD and LUSC, and a biomarker discovery module that leverages the combined network comprising the autoencoder and the feed-forward neural network. In the biomarker discovery module, XAI methods uncovered a set of 52 relevant biomarkers for NSCLC subtype classification. To evaluate the classification performance of the discovered biomarkers, multiple machine-learning models are constructed using these biomarkers. Using 10-Fold cross-validation, Multilayer Perceptron achieved an accuracy of 95.74% (±1.27) at 95% confidence interval. Further, using Drug-Gene Interaction Database, we observe that 14 of the discovered biomarkers are druggable. In addition, 28 biomarkers aid the prediction of the survivability of the patients. Out of 52 discovered biomarkers, we find that 45 biomarkers have been reported in previous studies on distinguishing between the two NSCLC subtypes. To the best of our knowledge, the remaining seven biomarkers have not yet been reported for NSCLC subtyping and could be further explored for their contribution to targeted therapy of lung cancer.

1. Introduction

Lung cancer has the highest mortality rate among all cancers [1] (GLOBOCAN2020, GlobalCancerObservatory), with a 5-year survival of about 17.8% [2]. The percentage of patients diagnosed with lung cancer has reached 11.4%, with a death rate of 18% [3].

World Health Organization (WHO) has categorized it into two main classes, Small Cell Lung Cancer (SCLC) covering around 15% of the cases, and Non-Small Cell Lung Cancer (NSCLC) covering approximately 85% of the cases [4,5]. NSCLC is further sub-categorized as Lung Adenocarcinoma (LUAD), accounting for about 40% of all lung cancers and Lung Squamous Cell Carcinoma (LUSC) covering about 30% of all the cases [4,5].

Conventional NSCLC diagnosis and treatment methodologies

Patients with NSCLC are often diagnosed at later stages [6,7], with cough and dyspnea as common symptoms, and cardiovascular disease and chronic obstructive pulmonary disease (COPD) as frequent comorbidities [7]. In order to confirm the cancer subtype at the histological level, a biopsy is needed. The TNM staging may also be required for devising an appropriate treatment methodology [8].

Surgery, chemotherapy, and radiotherapy are the standard-of-care treatment followed for NSCLC. Generally, at the early stages of NSCLC, surgery has shown promising results, although studies have revealed that around 30%–55% cases start showing tumor recurrence even after complete resection [9]. Chemotherapy and radiotherapy are generally followed as adjuvant therapy after surgery, and are preferred in later

* Corresponding author.
E-mail addresses: kountaydwivedi@gmail.com (K. Dwivedi), arajpal@cs.du.ac.in (A. Rajpal), sheetal.rajpal.09@gmail.com (S. Rajpal), agar.manoj@gmail.com (M. Agarwal), virendrakumar@aiims.edu (V. Kumar), nk.cs.du@gmail.com (N. Kumar).

https://doi.org/10.1016/j.compbiomed.2023.106544
Received 8 August 2022; Received in revised form 17 December 2022; Accepted 10 January 2023
Available online 12 January 2023

stages [2,10]. In chemotherapy, the patient is administered with anti-cancer drugs to kill rapidly growing cells. However, chemotherapy drugs may cause adverse effects such as systemic toxicity and drug resistance, as they are unable to differentiate tumor cells from normal cells [10,11]. In radiotherapy, high beams of energy are passed to destroy the DNA of the cancerous cells. However, it has been found that postoperative radiotherapy may adversely affect the survival probability of a patient [12,13].

The emerging approach: Targeted therapy

According to [11], "Targeted therapy aims at delivering drugs to particular genes or proteins that are specific to cancer cells or the tissue environment that promotes cancer growth". Currently, there is progress towards NSCLC treatment due to targeted therapy, assisting in the prolonged survival of the patients [10]. However, to develop a targeted therapy, the intrinsic molecular properties of the tumor should be precisely known [14–16] to facilitate the discovery of particular genes, or biomarkers, for which a drug could be devised. NSCLC is a heterogeneous disease at the molecular level, and patients with the same pathological tumor type may require different treatments [10]. Therefore, accurate classification of the tumor is essential for the discovery of biomarkers associated with cancer to device an effective therapy [2,15–17].

Explainable AI for biomarker discovery

Several machine learning models have been developed to perform the classification of NSCLC into its prominent subtypes and/or identification of essential NSCLC biomarkers differentially expressed across the subtypes [17–20]. Although deep learning models have established their supremacy over traditional machine learning algorithms when provided with sufficient data, the contribution of the different features to the task at hand remains opaque to the users of the deep learning models. Recently, the emergence of the explainable AI (XAI) concept has attempted to bridge the gap in deep models' explainability. In various applications, such as interpretation of a 3D brain tumor segmentation model [21], diagnosis of diabetic retinopathy grading [22], discovering essential breast cancer biomarkers and squamous cell cancer biomarkers [23,24], and discriminating ECG signals of COVID-recovered patients from those of healthy patients [25], XAI methods have shown significant potential in explaining the behavior of the model, thereby building trust over it. Motivated by the aforementioned applications of the XAI methods, we aim to exploit the potential of XAI-based feature selection to discover a small set of biomarkers for distinguishing between the two subtypes of NSCLC. Furthermore, to validate the supremacy of XAI-based feature selection, we perform a comparison (please see Section 4.1) of the XAI-based feature selection method with other competitive feature selection methods, namely *ReliefF* [26], *Mutual Information* [27], *Recursive Feature Elimination* [28], *Random Forest* [29], *Extreme Gradient Boosting* or *XGBoost* [30], and *Least Absolute Shrinkage and Selection Operator* or *LASSO* [31].

Related works

Several studies aim at classifying NSCLC into its prominent subtypes — LUAD and LUSC using PET/CT/MRI images or histopathological data. For example, Wang et al. [32] proposed a seed-detection-guided cell segmentation algorithm to segregate the cells in histopathological images. Subsequently, they extracted textural, geometrical, and pixel intensity-based statistical features for classifying the NSCLC into its two subtypes using adaptive boosting (AdaBoost) and random forest (RF), exhibiting 91.7% and 92% accuracy respectively. Hou et al. [33] sliced gigapixel-level whole slide images (WSI) into patches using a convolutional neural network (CNN) over them. They proposed the expectation–maximization (EM) based method to eliminate non-discriminative patches. Finally, they used a decision-fusion algorithm to aggregate the patch-level results to classify WSIs, achieving an accuracy of 79.8% using a support vector machine (SVM). To deal with large-sized whole slide images, Coudray et al. [20] divided each WSI image into the patches of size 512 × 512, and fine-tuned the InceptionV3 model using transfer learning [34] to obtain a state-of-the-art AU-ROC score of 0.97. Further, they demonstrated that six commonly mutated genes were predictable using only the image data, thus suggesting that deep learning models are capable of assisting oncologists in classifying cancer instances and detecting gene mutations. Han et al. [35] extracted 688 radiomics-based features via pre-obtained region-of-interests on PET/CT images. Ten different feature selection techniques were used to rank the features and the top 50 features were selected for each technique. Each subset of features was separately used to train ten machine learning models and the entire feature set was used to train the VGG16 deep learning model [36] for comparative analysis. Conclusively, VGG16 outperformed the rest, securing an AU-ROC score of 0.903 and an accuracy of 84.1%.

Cancer being a genetically diverse disease caused by several molecular aberrations, the researchers have recently focused on exploiting the molecular data for biomarker discovery and classification. Girard et al. [17] constructed a volcano plot to select 42 topmost overly expressed discriminatory genes (21 genes belonging to each class) from MD Anderson Cancer Center (MDACC) microarray dataset. Using the aforementioned sets of 21 genes, they computed the centroid for each subtype. Once the centroids were computed, a sample was assigned to the class whose centroid yielded a higher value of Pearson correlation coefficient (PCC). To validate the results they used The Cancer Genome Atlas (TCGA) RNA-Seq gene expression dataset and achieved an accuracy of 95%. Charkiewicz et al. [37] performed a statistical analysis over a training set of 98 NSCLC instances. Based on Benjamini–Hochberg's adjustment procedure [38], genes with p-value ≤ 0.05 were selected. To identify a gene signature comprising a set of histotypic genes they carried out prediction analysis of microarray (PAM) [39] based on the nearest shrunken centroid algorithm. Thus, they obtained a gene signature comprising a set of 53 genes that yielded an accuracy of 93% on a validation set. [40] used 90 LUAD and 153 LUSC gene expression instances for classification. To select the important genes, they used *ReliefF* algorithm [41], and the *limma* algorithm [42]. While *ReliefF* is a multivariate algorithm that assigns a rank to each feature based on its relevance for classification, *limma* is a statistical R-package, which uses t-statistics to identify differentially expressed genes. Using each algorithm, they selected a feature set of the top 30 scoring genes. Using a naive Bayes classifier, they achieved an AU-ROC score of 0.89 and 0.90 using the feature sets produced using *ReliefF* and *limma*, respectively. Yuan et al. [43] used a dataset comprising 77 LUAD and 73 LUSC instances for classification. For features assessment, they employed Monte Carlo (MCFS) method [44], to discover sets of features based on their contribution to the classification task. For this purpose, they generated a large number of sets of features of different sizes and evaluated their performance using several decision trees. Finally, they employed Incremental Feature Selection (IFS) algorithm [45] to select an optimal subset of 1100 features from the list of assessed features provided by MCFS. They used the Matthews correlation coefficient (MCC) as the evaluation metric, scoring 93.5% MCC, and an accuracy of 96.7%, using a support vector classifier (SVC). [46] ranked the genes using GeneRank [47], which considers gene expression level as well as gene–gene interaction. Further, to select a set of the most relevant (at most eight) genes, they applied radial coordinate visualization (RadViz) [48] over the ranked list of genes. Thus, they achieved an accuracy of 79.4% using SVM on a dataset consisting of 125 instances of RNA-Seq gene expression. [49] used five feature selection methods — minimum redundancy maximum relevance (mRMR), principal component analysis (PCA), differential gene expression analysis (DGE), XGBoost, and least absolute shrinkage and selection operator (LASSO), to select an optimal set of features. Subsequently, they computed the intersection between the set of features obtained by these five methods. They found 17 biomarkers overlapping in three or more feature selection methods. Using a random forest classifier, they achieved an accuracy of 92.9% on a dataset comprising 529 LUAD and 498 LUSC RNA-Seq gene expression instances.

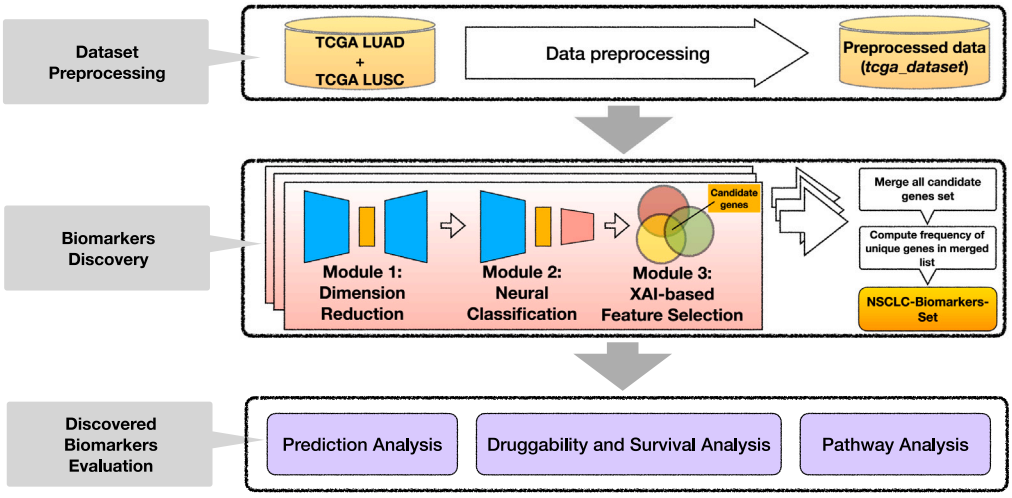


Fig. 1. The outline view of the proposed experiment. Initially, the preprocessing of the dataset is performed. Next, with the help of the developed framework, a set of biomarkers capable of segregating NSCLC into its subtypes is discovered. Finally, the evaluation of the discovered biomarkers is performed.

Research motivation and contribution

Discovering a small set of clinically relevant NSCLC biomarkers is a crucial step in personalized medication [17] as it not only enables us to distinguish between the NSCLC subtypes but also serves as the basis for personalized therapeutic intervention. Recent studies have found a correlation between gene expression profiles and tumor histological subtypes, and thus assert that analysis of gene expression may help capture comprehensive molecular characteristics, aiding in tumor classification [50,51].

The main contributions of this paper are as follows:

1. A deep learning framework is proposed, which utilizes XAI methods to discover a small set of biomarkers that can be used for NSCLC classification.
2. A set of 52 potential biomarkers is discovered. A significant number (45 out of 52) of these biomarkers are in conformity with the established literature. The remaining seven biomarkers could be the subject of further clinical research.
3. The proposed framework utilizes XAI-based feature selection for biomarkers discovery. The XAI-based feature selection method outperformed other competitive feature selection methods in terms of classification performance.
4. The druggability of the discovered biomarkers along with their role in predicting survival is explored. Out of 52 biomarkers, 14 are found to be potentially druggable, and 28 biomarkers are found capable of predicting survival (p-value ≤ 0.05). Also, the biological pathways enriched by the discovered biomarkers are also reported.

Fig. 1 outlines the proposed experiment. Initially, we perform the preprocessing of the dataset. Next, we utilize the framework to discover a set of biomarkers that could segregate NSCLC into its appropriate subtypes — LUAD and LUSC. Lastly, the discovered biomarkers are evaluated on the grounds of their classification accuracy and their clinical relevance.

The rest of the paper is organized as follows: Section 2 provides a brief description of the XAI methods incorporated; Section 3 describes the dataset and the methodology proposed for the experimentation; Section 4 presents the obtained results and the discussion upon them; and finally, Section 5 presents the conclusion and provides a brief scope of future work.

2. Preliminaries

Explainable AI (XAI), is a set of tools or methods that help developers interpret their machine/deep learning model's inherent processing, thus unveiling its "black-box" nature, and gaining the trust of the users. The more a model is interpretable, the more trustworthy it is [52].

Integrated Gradients

Integrated Gradients (IG) is an attribution method that attributes the prediction of a deep neural network to its inputs [53]. The motive is to understand the input/output behavior of a deep network, and thus, assist in its improvisation. The method of attributing the prediction of a deep neural network, as described by Sundararajan et al. [53], is as follows:

Let there be a function $F : R^n \rightarrow [0,1]$, that represents a deep network. Let there be an input $x = (x_1, x_2, \ldots, x_n) \in R^n$. Then, an attribution of the prediction at input x relative to a *baseline* input x' is:

$$A_F(x, x') = (a_1, a_2, \ldots, a_n) \in R^n \tag{1}$$

where $A_F(x, x')$ is a vector, and a_i is the **contribution** of x_i to the prediction $F(x)$. The contribution measures the deviation of an attribute value from a *baseline* counterfactual input.

The integrated gradient along the **ith** dimension for an input x and baseline x' is:

$$IntegratedGrads_i(x) = (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha \tag{2}$$

In practice, Eq. (2) is approximated by using Riemann approximation of the integral with m number of steps, as:

$$IntegratedGrads_i^{approx}(x) := (x_i - x'_i) \left(\sum_{k=1}^{m} \frac{\partial F(x' + \frac{k}{m} \times (x - x'))}{\partial x_i} \times \frac{1}{m} \right) \tag{3}$$

GradientSHAP

The implementation of GradientSHAP is based on the implementation of SHapley Additive exPlanation (SHAP) [54], which is based on the cooperative game theory, Shapley Values, coined by Lloyd Shapley in 1953 [55].

- Each feature of an instance is considered as a *"player"* in a cooperative game.
- The prediction is considered as the *"payout"* or the reward generated by the coalition of each player in the game.
- Shapley values are the *"fair"* distribution of the payout among each player in the coalition.
- The Shapley value of a player is computed as the **average marginal contribution** of that player across all possible coalitions.

To compute the Shapley values, the model needs to retrain on all the feature subsets $S \subseteq F$, where F is the entire feature set. The purpose of this step is to assign importance to a feature that represents the effect of that feature on the prediction of the model by the inclusion of that feature. To compute this effect, two models are trained — one model $f_{S\cup\{i\}}$ that includes the feature, and another model f_S where the feature is withheld. Subsequently, the predictions of both models are compared on the current input:

$$f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S) \tag{4}$$

where x_S is the input feature values in the set S. This effect is computed for all possible subsets $S \subseteq F \setminus \{i\}$. The Shapley values are then computed as the weighted average of all possible differences:

$$\phi_i = \sum_{S\subseteq F\setminus\{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)] \tag{5}$$

The $GradientSHAP$ method adapts the concept of $IntegratedGradients$ to compute the relevance score of a feature i. The SHAP values are approximated by computing the expected gradients from a random sample of baseline distribution [56]. Initially, a set ($\tilde{X}$) of $nPert$ random perturbations (where $nPert$ is a hyperparameter of the method) is generated by adding white Gaussian noise to the input instance x, such that:

$$\tilde{X} := \left\{\tilde{x}_k : \tilde{x}_k \leftarrow whiteGaussianNoise(x)\right\}; \quad k \in \{1,\ldots,nPert\} \tag{6}$$

Subsequently, a set X' is generated, comprising randomly generated $nPert$ instances from a baseline distribution:

$$X' := \{x'_k : x'_k \leftarrow random(baselineDistribution)\}; \quad k \in \{1,\ldots,nPert\} \tag{7}$$

A random point pt_k is selected along the path between each generated $\tilde{x}_k$ and x'_k ($k \in \{1,\ldots,nPert\}$), and the gradient Δ_{i_k} of the feature i with respect to the selected points is computed:

$$\Delta_{i_k} := \frac{\partial M(\tilde{x}_{i_k})}{\partial pt_k}; \quad k \in \{1,\ldots,nPert\} \tag{8}$$

where M is the underlying model. The final SHAP value, or the relevance score, assigned to the feature i is the product of the expectation of the computed gradients and the difference between each $\tilde{x_k}$ (input perturbation) and x'_k (baseline):

$$\phi_i := \left(\frac{1}{nPert}\sum_k \Delta_{i_k}\right) \times (\tilde{x}_k - x'_k); \quad k \in \{1,\ldots,nPert\} \tag{9}$$

DeepLIFT

Shrikumar et al. [57] presented a method, namely Deep Learning Important FeaTures, or DeepLIFT, which is a model interpretation method that computes the importance score using a backpropagation-like algorithm. It tries to explain the difference in output from some "reference output", in terms of the difference of the input from some "reference input". The "reference", in the case of a neuron, is the activation of that neuron when some "reference" input (depends on the domain knowledge, conceptually similar to a *baseline* in $IntegratedGradients$) is passed to the network. Formally, for a neural network F, let there be a target output neuron t. Let $\{x_1,\ldots,x_n\}$ be some set of neurons in an intermediary layer (or input layer) or set of layers that are necessary and sufficient to compute t:

$$t = F(x_1,\ldots,x_n) \tag{10}$$

Let t^0 be the reference activation of t, such that:

$$t^0 = F(x_1^0,\ldots,x_n^0) \tag{11}$$

where $\{x_1^0,\ldots,x_n^0\}$ are the reference activations of input $\{x_1,\ldots,x_n\}$. Then, the difference-from-reference, Δt is computed as:

$$\Delta t = t - t^0 \tag{12}$$

Table 1
Summary of the utilized dataset. After the removal of genes having no effect due to cancer, a total of 20,258 genes remained in the processed dataset. As there were no spurious instances found, hence total instances in the unprocessed as well as processed data remain the same.

Dataset	No. of genes	No. of instances (LUAD/LUSC)
tcga_dataset (unprocessed)	20,530	1129 (576/553)
tcga_dataset (processed)	20,258	

and the contribution scores (or relevance score) assigned by DeepLIFT to Δx_i is $C_{\Delta x_i \Delta t}$, such that:

$$\Sigma_{i=1}^{n} C_{\Delta x_i \Delta t} = \Delta t \tag{13}$$

In Eq. (13), $C_{\Delta x_i \Delta t}$ is considered as the amount of difference-from-reference in t that is attributed to or "blamed" on the difference-from-reference of x_i.

3. Materials and methods

This section provides a detailed description of the dataset, the proposed framework, and the hardware and programming environment in which the experiment was conducted.

3.1. Dataset

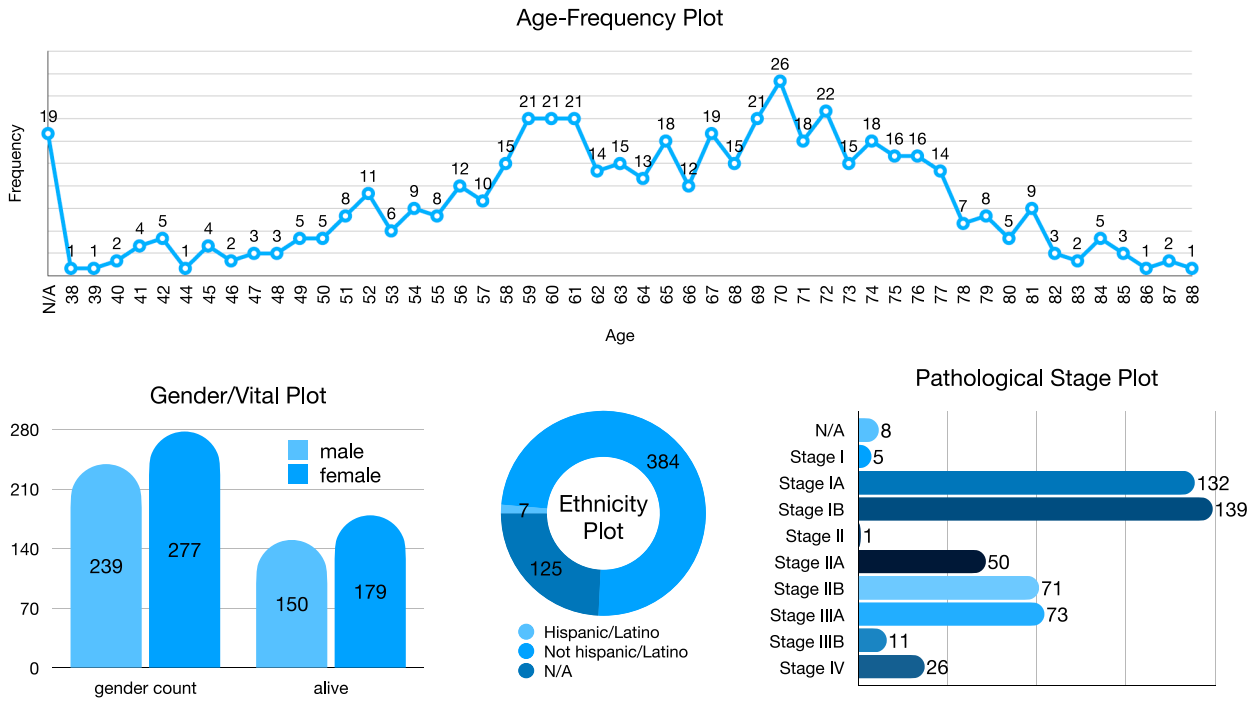
The population under our study is the collection of gene expressions of patients that belong to either of the two subtypes of NSCLC: LUAD and LUSC. For the purpose of experimentation, the publicly available database generated by The Cancer Genome Atlas program (TCGA, funded by the National Institutes of Health (NIH)), is utilized. The $\log_2$ normalized RNA-Seq LUAD and LUSC cohorts from the UCSC Xena repository [58] were downloaded on June 2021 (linktotherepository) for experimentation. The dataset was nearly balanced, with 576 instances of LUAD (51.01% of total instances), and 553 instances of LUSC (48.99% of total instances). There were 20,530 genes present in both cohorts. The demographics of the dataset are presented in Fig. 2. There were 239 males and 277 females in the LUAD cohort, and 371 males and 130 females in the LUSC cohort. In the rest of the paper, we refer to our dataset as ***tcga_dataset***.

Dataset preprocessing

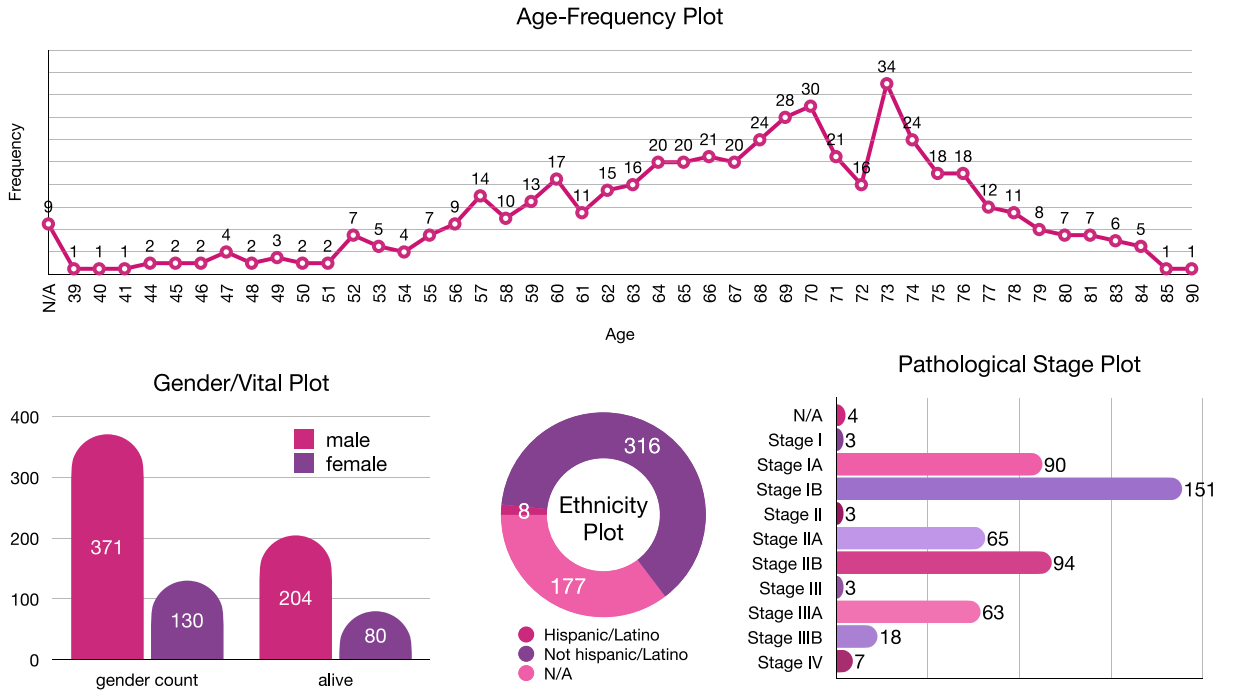
The *tcga_dataset* is first examined for missing values (instances containing NaN or no value for any gene). However, no such records are found. Next, the genes with zero variance across all the patients (indicating no effect due to cancer), are removed. 272 such genes are found and removed, resulting in 20,258 genes as the final input feature space. Finally, the z-score normalization is applied to the dataset gene-wise, since the standardized data could be directly incorporated in the calculation of significant change in gene expressions between different instances and conditions [59]. Table 1 summarizes the *tcga_dataset*.

3.2. Biomarker discovery framework

In this section, the proposed framework is used to discover a small set of biomarkers for discriminating NSCLC subtypes. The Biomarker Discovery Framework 1 comprises three modules — dimension reduction, neural classification and XAI-based feature selection. The *tcga_dataset* (D), comprising 1129 instances (nSamples) and 20,258 genes (nGenes), is passed as an input to the first module (module 1) of the framework. The first module utilizes an autoencoder to generate an embedded vector (*embedded_vector*) of size 512 (nEmbeddings). Subsequently, the *embedded_vector* and D are provided as input to the second module (module 2) of the framework. It comprises a feed-forward neural network (*classifier*) that distinguishes the instances of the dataset into their appropriate classes, namely LUAD and LUSC. Thereafter, the



(a) Demographic details of LUAD cohort



(b) Demographic details of LUSC cohort

Fig. 2. Demographic details of *tcga_dataset*. Fig. 2(a) shows the demographic details of the instances in the LUAD cohort. Fig. 2(b) shows the demographic details of the instances in the LUSC cohort. The mean age of instances in the LUAD cohort was 65.3 years, while in the LUSC cohort, it was 67.2 years.

combined networks obtained from the first and the second modules are passed to the third module (module 3) for interpretation. Module 3 utilizes three XAI methods, $IntegratedGradients$, $GradientSHAP$, and $DeepLIFT$ (XAIMETHODS) to select a set of *candidate_genes* deemed most relevant for classification by the neural network. The aforementioned modules are executed ten times (NITERS) with different seed values. The intent of this repetition is motivated by the fact that the stochastic weight initialization could lead to a slightly different set of results for

different seed values, which would help in capturing the variability. It is to be noted that the output of a single run is a set of candidate genes. On completion of ten runs, each set of candidate genes is merged to form a single list. Subsequently, the frequency of each unique gene in the merged list is computed. Finally, the ***NSCLC-Biomarkers-Set*** is formed with the genes having frequency ≥ 5.

Thus, the proposed approach successfully handles the high dimensional low sample sized ***tcga_dataset*** by capturing the complex non-linear nature of the feature space (here, genes). The XAI methods employed successfully exploit the neural network module to identify the most relevant biomarkers for the classification of NSCLC subtypes. The details of the modules of the proposed framework are as under:

Module 1: Dimension reduction

To handle the high-dimensional preprocessed dataset, a deep learning-based autoencoder is utilized. An autoencoder comprises two consecutive parts — an encoder that shrinks the input feature space to a concise embedded space, and a decoder, which tries to reconstruct the original input from the concise feature space.

An autoencoder is implemented to shrink the set of 20,258 genes to an embedded set of size 512. The encoder component of the autoencoder comprises three layers of successively reducing sizes, i.e., 4096, 2048, and 512. Similarly, the decoder component of the autoencoder comprises three layers of successively increasing sizes, i.e., 512, 2048, and 4096. The optimizer function used is AdamW [60] with weight decay as $1e^{-3}$, and the loss function used is Mean Squared Error (MSE). The autoencoder is trained for 150 epochs, with a learning rate of $1e^{-4}$, and a batch size of 64. After training the autoencoder, a similarity check is performed between the true input and the reconstructed input using Pearson's correlation coefficient (PCC). A mean PCC score of 0.993 is achieved.

Module 2: Neural classification

The second module involves a feed-forward neural network that utilizes the embedded feature space generated by the autoencoder in module 1 (dimension reduction), to classify NSCLC instances into LUAD and LUSC subtypes. The neural network has two *tanh* hidden layers, each comprising 1024 neurons, and a Sigmoid output layer. The network is trained and validated using 5-fold cross-validation, with a learning rate of $1e^{-4}$, batch size of 64, number of epochs equal to 100, and AdamW optimizer.

Module 3: XAI-based feature selection

The third module uses the combined network comprising the autoencoder (module 1) and the feed-forward neural network (module 2). A set of XAI methods is utilized to interpret the feed-forward neural network obtained from module 2 (neural classification) and select a set of *candidate genes* capable of classifying NSCLC instances into their appropriate subtypes. Three XAI methods are utilized — $IntegratedGradients$ [53], $GradientSHAP$ [54] and $DeepLIFT$ [57]. These methods compute the relevance score of each gene, indicating the contribution of that gene towards classifying an instance to a class. Assume there is an instance $i \in \mathbb{R}^p$ with a corresponding set of genes: $\{g_1, g_2, \ldots, g_p\}$, and the relevance score of gene g needs to be computed. If i belongs to a class C, then the relevance score of the gene g is $\phi_g(i)$, where:

$\phi_g(i)$ =contribution of gene g towards classification of instance i into class C

A relevance score of a gene could either be a positive or negative value, signifying its positive or negative contribution towards the prediction. The magnitude of the score states the strength of the contribution.

To compute the mean relevance score, each XAI method ($IntegratedGradients$, $GradientSHAP$, and $DeepLIFT$) is provided with — the feed-forward neural network ($classifier$), the *tcga_dataset* (D), a *baseline_dataset* (here, a data matrix with all zeros of the same size as the D), and the respective class label (LUAD or LUSC), as inputs. The computation of the relevance score for each of the XAI methods is described below:

- $IntegratedGradients$: It uses Eq. (3), where x, i, and F correspond to the input instance, the target gene, and the $classifier$, respectively; and $x' \in baseline\_dataset$.

Biomarker Discovery Framework

Input:
D: Dataset of size (nSamples × nGenes)
nEmbeddings: Size of the embeddings vector
ClassLabels: List of class labels
XaiMethods: List of XAI methods incorporated
nIters: Number of iterations
nSelectGenes: Number of most relevant genes (positively and negatively) to be selected for each class from each XAI method in a single iteration

Output: NSCLC-Biomarkers-Set

begin
  for $i \leftarrow 1$ *to* nIters **do**
    // Module 1: Use autoencoder to generate an embedded vector of size (nEmbeddings) from the input gene set of size (nGenes)
    $embedded\_vector \leftarrow$ ***Autoencoder***(D, nEmbeddings)
    // Module 2: Input $embedded\_vector$ and D to a feed-forward neural network ($classifier$) for classification
    $classifier \leftarrow$ ***FFNN***($embedded\_vector$, D)
    // Module 3: Perform $candidate\_genes$ selection:
    foreach $m \in$ XaiMethods **do**
      foreach $l \in$ ClassLabels **do**
        // Compute mean relevance score of input genes set
        $genes\_scores \leftarrow$ ***MeanRelevanceScore***(D, $classifier, m, l$)
        // Sort $genes\_scores$ in decreasing order of the mean relevance score of the genes
        $sorted\_scores \leftarrow$ ***Sort***($genes\_scores$)
        // Use $sorted\_scores$ to select nSelectGenes genes
        $top\_genes \leftarrow$ ***RelevantGenesSelection***($sorted\_scores$, nSelectGenes)
      end
      // Take the union of top_genes belonging to each class in ClassLables
      $combined\_genes \leftarrow \bigcup_{top\_genes_l}; \forall l \in$ ClassLabels
    end
    // Intersect $combined\_genes$ belonging to each method in XaiMethods
    $candidate\_genes_i \leftarrow \bigcap_{combined\_genes_m}; \forall m \in$ XaiMethods
    $merged\_list := merged\_list.append(candidate\_genes_i)$
  end
  // Compute frequency of each unique candidate gene in $merged\_list$
  $candidate\_genes\_frequency \leftarrow$ ***ComputeFrequency***($merged\_list$)
  // From $candidate\_genes\_frequency$, select the genes with frequency ≥ 5 as NSCLC biomarkers
  foreach $gene \in candidate\_genes\_frequency$ **do**
    if $frequency_{gene} \geq 5$ **then**
      NSCLC-Biomarkers-Set := NSCLC-Biomarkers-Set.$append(gene)$
    end
  end
  return NSCLC-Biomarkers-Set
end

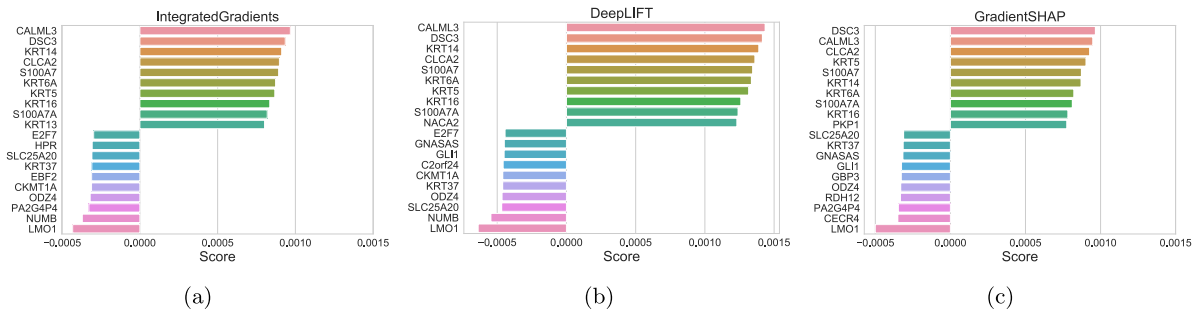


Fig. 3. Mean relevance scores for 20 most relevant genes (positive as well as negative) computed by each XAI method for LUAD.

- $GradientSHAP$: It employs Eq. (9), with $\tilde{X}$ being a set of perturbations for an input instance x; X' denotes the set of instances selected randomly from $baseline\_dataset$; and Δ_i refers to the gradient of target gene (denoted by i) computed by $classifier$ (denoted by M).
- $DeepLIFT$: The Δx_i and Δt in Eq. (13) denotes the difference-from-reference corresponding to a gene i and target label t (= LUAD or LUSC), respectively; and $C_{\Delta x_i \Delta t}$ is the contribution or the relevance score of gene i.

Thereafter, the mean relevance score is computed for each gene for the two classes separately, resulting in two lists of scores computed by each XAI method.

Module 3 of the Biomarker Discovery Framework 1 provides a detailed description of the candidate genes selection process by utilizing the three XAI methods. The step-wise details are as follows:

1. For each class, each XAI method interprets the feed-forward neural network and computes the mean relevance score of each gene across all the instances in the dataset.
2. The two lists of mean relevance scores (each belonging to one of two classes — LUAD and LUSC) are sorted in decreasing order of the mean relevance score of the genes. Fig. 3 shows the mean relevance score of the 20 most relevant LUAD genes computed by the individual XAI methods.
3. From each sorted list, a number ($nSelectGenes$) of the most relevant genes (positively as well as negatively) are selected. This number ($nSelectGenes$) was empirically found to be 150. Thus, for each class, a set of 300 most relevant genes is selected.
4. A union is performed on both the sets of selected genes (each set belonging to one of the classes), resulting in a combined set of the most relevant LUAD and LUSC genes.
5. Steps 1 to 4 are repeated for each XAI method, resulting in three sets of combined genes. Finally, an intersection is performed over them to obtain a set of genes that are selected as most relevant by all three methods. This set of obtained genes is named *candidate genes set*.

3.3. Hardware and programming environment

The entire experiment is performed on Acer Predator Helios 300 (PH317-53) system with a Core i7-9750H CPU clocked at 2.60 GHz. The primary memory is 16 GB and the operating system is Windows 10 Home edition. The system has a dedicated CUDA-enabled NVIDIA GeForce GTX 1660-Ti GPU with 6 GB memory, the CUDA version being 10.2.

The implementation is done on Python v3.7.7 programming language, utilizing the PyTorch v1.8.1 library [61]. The XAI methods are utilized from the PyTorch-based Captum v0.4.0 library [62]. We use Numpy v1.19.2 for algebraic operation, Pandas v1.0.5 for dataset operations, Matplotlib v3.2.2 and Seaborn v0.10.1 for graphs and plots visualization.

4. Results and discussions

The objective of the proposed study is to discover a small set of clinically relevant NSCLC biomarkers for their potential application in targeted therapy. A deep learning framework is developed, using the XAI methods *GradientSHAP, IntegratedGradients*, and *DeepLIFT* and a set of 52 biomarkers (***NSCLC-Biomarkers-Set***) are discovered.

Out of 52 biomarkers that are discovered using the proposed framework, 45 have already been reported in earlier studies. In Fig. 4, we have included 12 studies that show significant overlap with the discovered biomarkers, cumulatively accounting for 32 (61.5%) of the 45 overlapping genes. In addition, [63,64] identified $PTTG3P$ as an NSCLC biomarker whose high expression value may lead to shorter survival of the patients. [65] investigated immunogenomic patterns of LUSC patients based on 11 immune-related genes, including $RNASE7$, to improve the prognosis of LUSC. [66] discovered a novel therapeutic biomarker of NSCLC, namely, $RPL7$ which is deregulated by platinum-based chemotherapy. [67] studied RNA and protein levels of $S100A7$, and concluded that while its specific expression was found in LUSC, adenosquamous carcinoma, and large cell lung carcinoma, it was not detected in LUAD and small cell lung carcinoma. They also observed that an elevated expression of $S100A7$ found in the serum of LUSC patients makes it a potential lung cancer biomarker. [68] studied the mRNA and methylation status of $KLK10$ (alias: $PRSSL1$), and found its epigenetic inactivation a common event contributing to NSCLC pathogenesis. They stated $KLK10$ as a tumor suppressor gene in NSCLC and may be used as a potential biomarker. [69] performed genomic pan-cancer classification using TCGA gene expression data and identified numerous sets of 20 genes capable of classifying 31 types of cancers, including LUAD and LUSC. Three genes — $NACA2$, $PA2G4P4$, and $C14orf19/IGBP1P1$ of our discovered genes overlapped with their 20 most frequently selected genes. Moreover, $BNC1$ which was found to differentiate between genders by [69] is in our discovered set of genes. [70] identified eight genes, including $A2ML1$, that were related to TRIM58/cg26157385 methylation, and thus may be considered as a potential biomarker for LUSC treatment. [71] studied the protein expression and DNA methylation levels of $S100A7A$, an alias for $S100A15$, and concluded that its increased gene expression and decreased methylation of its gene promotor region was associated with potentially high metastasis and poor outcome in LUAD. [72] examined the mRNA expression of Desmogleins 1–3 ($DSG1$, $DSG2$, $DSG3$), and DNA methylation levels of Desmogleins 1–2 ($DSG1$, $DSG2$), concluding DSG2 and DSG3 as potential diagnostic biomarkers for LUSC, and DSG3 as a potential biomarker for lung cancer differentiation. [73] studied the expression levels of lncRNA $FTH1P3$ (alias: $FTHL3$), and found them to be highly expressed in NSCLC tissues as compared to matched normal tissues, suggesting it to be a promising biomarker for NSCLC. [74] showed $HEY1$ as a Notch3-dependent gene that lies in the Notch receptor pathway. They also showed that Notch3 peptides could help in the apoptosis process leading to tumor suppression in lung cancer. To the best of our knowledge, the remaining seven genes, namely, *AP2M1, C9orf69, FLJ44635, ID2B, CEL, LOC442308* and *LOC728758*,

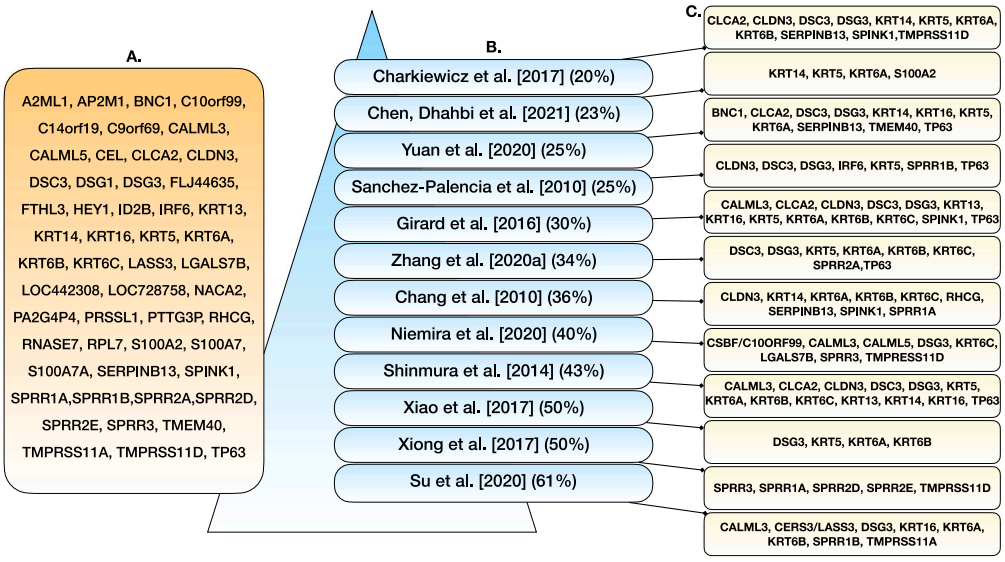


Fig. 4. Published works that show significant overlap with the discovered biomarkers [17,37,43,49,75–82]. **A.** lists out the discovered biomarkers in the *NSCLC-Biomarkers-Set*. **B.** presents a pyramid that identifies the specific works by the first author's name and year of publication, along with the percentage of overlapping genes (in increasing order) with the *NSCLC-Biomarkers-Set* discovered using the proposed framework. In each case, an arrow points to the list of overlapping genes shown in **C**.

Table 2
Hyperparameter values of each model used in the evaluation.

Model	Hyperparameter values
Multilayered Perceptron (MLP)	solver='adam', epochs=100, hidden layers=(512, 256, 128), activation='relu', learning rate=$1e^{-5}$, batch size=32, alpha=0.02
Logistic Regression (LR)	l1_ratio=0.04, penalty='elasticnet', epochs=100, solver='saga'
XGBoost (XGB)	eta=0.1, max depth=10, booster='gbtree', alpha=0.5
Support Vector Classifier (SVC)	C=1.0, Kernel='rbf', gamma='scale', probability=False

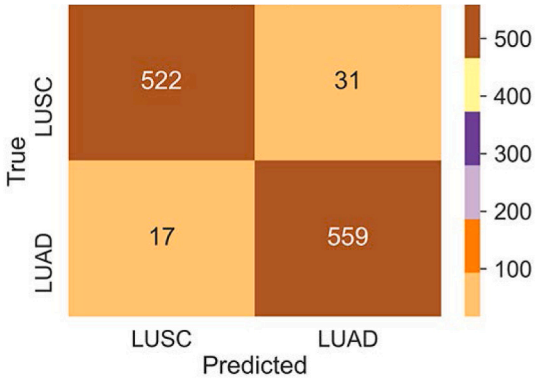


Fig. 5. Confusion Matrix of Multilayer Perceptron (MLP) model using 10-Fold CV.

have not been reported so far for subtyping NSCLC and could be the subject of further investigation by the clinicians for devising a targeted therapy for the non-small cell lung cancer patients.

4.1. Classification results of NSCLC-Biomarkers-Set

The classification performance of the *NSCLC-Biomarkers-Set* is evaluated in terms of accuracy, balanced accuracy, and AU-ROC score.

Four machine learning models are developed using Multilayer Perceptron (MLP), Logistic Regression (LR), Extreme Gradient Boosting (XGB), and Support Vector Classifier (SVC) algorithms. The models are validated by employing Leave-one-out-cross-validation (LOOCV) and 10-Fold cross-validation at 95% confidence interval (C.I.). The hyperparameter values for each model are selected based on experimentation. Table 2 shows the various models and values of the hyperparameters that are found by experimentation.

Table 3 shows the performance of all the classification models using LOOCV and 10-fold CV over *tcga_dataset*. The overall best performance was achieved by the MLP model — 95.75% accuracy on LOOCV, 95.74%±1.27 accuracy on 10-fold CV (95% C.I.), 95.71% balanced accuracy, and 98.89 ± 0.64 AU-ROC score. Fig. 5 presents the confusion matrix of MLP (computed over 10-Fold CV).

The XAI-based feature selection methodology is compared with various competitive feature selection methods, namely *Support Vector Machine with Recursive Feature Elimination* (SVM-RFE), *ReliefF*, *Mutual Information* (MI), *Least Absolute Shrinkage and Selection Operator* (LASSO), *Random Forest* (RF), and *Extreme Gradient Boosting* (XGB). The comparison is made with respect to 10-fold cross-validation classification accuracy achieved using the set of 52 highest-ranking biomarkers obtained from each of the aforementioned methods. Table 4 shows the comparison between classification accuracy obtained using the competing feature selection methods vis-a-vis the XAI-based feature selection method. It is evident that the XAI-based feature selection method outperforms the other feature selection methods.

4.2. NSCLC-Biomarkers-Set druggability

The potential druggability of the *NSCLC-Biomarkers-Set* is reported with the help of **Drug-Gene Interaction Database** or **DGIdb** [83]. It is an online resource (www.dgidb.org) that could be utilized for exploring drug-gene interaction or the potential druggability of a gene. Out of 52 discovered biomarkers, 14 are included in the DGIdb. Table 5 shows the relevant category for each of the aforementioned genes. To the best of our knowledge, we are reporting the genes $AP2M1$ and CEL as NSCLC biomarkers for the first time. However, these genes need to be evaluated clinically for their therapeutic value in the treatment of NSCLC.

$A2ML1$ is a protease inhibitor gene, recently added to RAS-related pathway regulators that are activated in lung adenocarcinoma (LUAD)

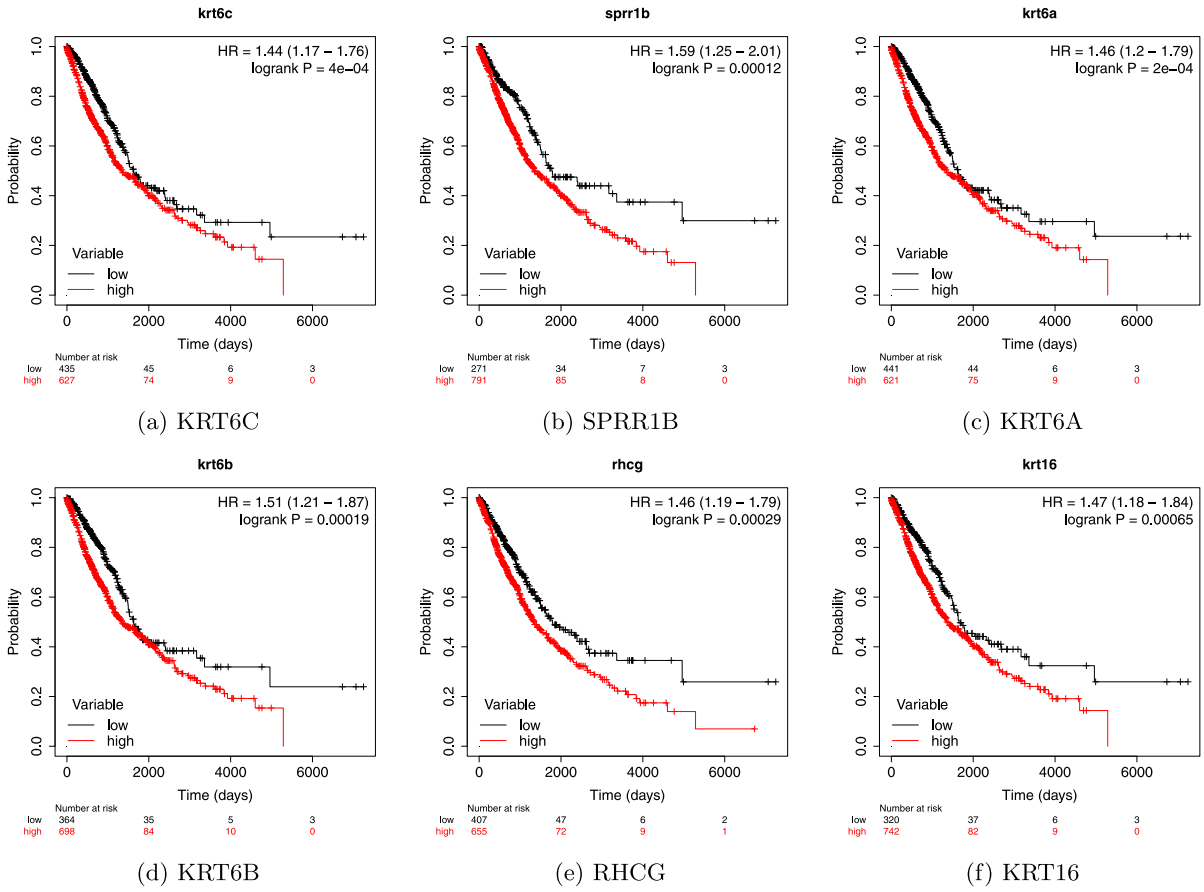


Fig. 6. Kaplan–Meier curve of six genes out of 28 genes with least *p*-value. The survival period (in the number of days) and the probability of survival are indicated along the horizontal and vertical axes, respectively. The curves with black and orange color represent the group of instances having low and high expression values respectively. The HR ratio depicts the survival probability of the first group (black) over the second group (red). It was observed that the low expression value of these genes contributed to a higher survival probability.

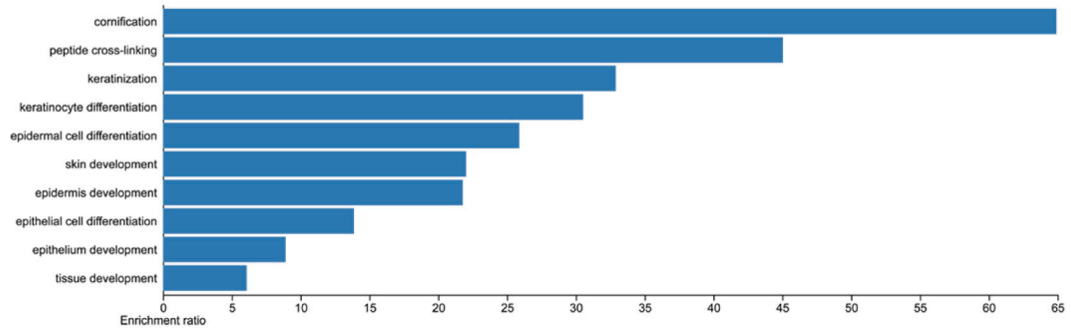
Table 3
Classification accuracy and AU-ROC score obtained by employing MLP, LR, XGB, and SVC models.

	LOOCV accuracy (%)	10 Fold CV accuracy (%) (95% C.I.)	AU-ROC (95% C.I.)
Multilayer Perceptron (MLP)	95.75	95.74 ± 1.27	98.89 ± 0.64
Logistic Regression (LR)	95.04	94.95 ± 1.51	98.67 ± 0.73
XGBoost (XGB)	95.48	95.21 ± 1.39	98.47 ± 0.73
Support Vector Classifier (SVC)	95.48	95.57 ± 1.43	98.76 ± 0.74

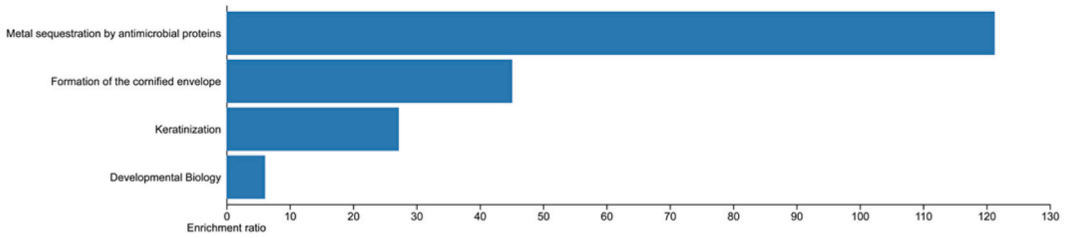
Table 4
Comparison between various feature selection methods and XAI-based feature selection. It is observed that XAI-based feature selection outperforms the other competitive methods, yielding maximum classification accuracy.

Feature selection method	Accuracy (%) (95% C.I)
SVM-RFE	93.62 ± 1.18
MI	93.80 ± 1.10
ReliefF	93.89 ± 0.98
LASSO	91.76 ± 2.59
XGB	92.47 ± 1.60
RF	92.20 ± 1.97
XAI-based	**95.74 ± 1.27**

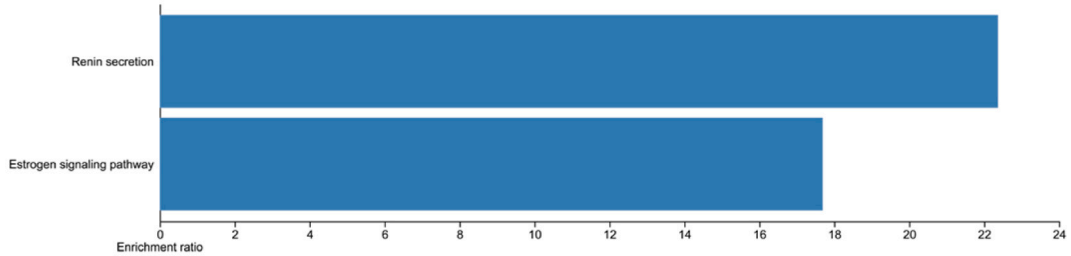
[84]. *KRAS* is a form of *RAS* isoform which is 33% mutated in lung carcinoma [84]. *AP2M1* is regarded as a universal host protein that is exploited by various viral infections, including COVID-19 [85]. These viruses affect the host protein, which eventually catalyzes the gene expression and signaling pathways, such as PI3K/AKT pathway, which is very prominent in various carcinomas, including the lung. [79] found *CLCA2* as a novel potential immunohistochemical biomarker to differentially segregate LUAD and LUSC. Upon further analysis, they reported that loss of *CLCA2* is a poor prognostic factor in female LUSC patients. *CLCA2* is targeted by *p53* and negatively regulates the proliferation, migration, and invasion of cancer cells [79]. Claudin-3 (*CLDN3*) has been identified as a positive regulator of cancer stemness and cancer stem-like cells-mediated chemoresistance in nonsquamous NSCLC, and hence, targeting it may provide a worthy NSCLC therapy [86]. Hairy/Enhancer-Of-Split Related With YRPW Motif Protein 1 (*HEY1*) is a *NOTCH* signaling gene, which is found abnormally active in NSCLC. Various antibody-based biologics that target *NOTCH* ligands and receptors have been devised as investigational drugs [87]. Interferon regulatory factor 6 (*IRF6*) gene is found to be upregulated in both LUAD and LUSC when compared to normal tissues [88], and that miRNA-320 is negatively related to the expression of *IRF6* in lung cancer. [88] concluded that their findings could help explore therapeutic drugs related to the miRNA-320/*IRF6* signaling axis for NSCLC treatment. Serine Peptidase Inhibitor Kazal Type 1 (*SPINK1*) is found to be a promoter of proliferation in several cancers and is highly expressed



(a) GO Biological Processes



(b) Reactome Pathway



(c) KEGG Pathway

Fig. 7. Pathway enrichment analysis of ***NSCLC-Biomarkers-Set***. The Benjamini–Hochberg test was applied to avoid Type-1 errors or false positives, and the p-value was adjusted with False Discovery Rate (FDR) ≤ 0.05. (a) Gene Ontology (GO) Biological Processes (b) Reactome Pathways (c) KEGG Pathways.

in NSCLC. It is studied that $SPINK1$ inhibits apoptosis in NSCLC by maintaining redox homeostasis driven by regulating the nuclear factor erythroid 2-related factor two pathways [89,90]. $TMPRSS11D$ belongs to the largest group of pericellular serine proteases — Type II transmembrane serine protease family [91]. It was found by [91] that $TMPRSS11D$ possesses high mRNA and protein expressions leading to the poor overall survival of NSCLC patients. They concluded that $TMPRSS11D$ could aid tumorigenesis via cell proliferation, invasion and metastasis, and inflammation, and therefore should be targeted to prevent metastasis in NSCLC. Tumor protein p63 ($TP63$) is directly associated with Syntaxin Binding Protein 4 ($STXBP4$), and regulating $STXBP4$ could lead to regulating TP63, which is considered a highly specific LUSC biomarker [92].

4.3. Survival analysis

Survival prediction of lung cancer patients is important for the patients as well as the clinicians [93]. In this work, we have used Kaplan--Meier(KM)Plotter for predicting the survivability of the NSCLC patients by [94,95]. The survival data associated with 1078 patient instances in *tcga_dataset* was downloaded from the Genomic Data Commons (GDC) data portal. A total of 28 genes out of 52 were found capable of predicting the survival probability (p-value ≤ 0.05). Using KM curves, we found that 28 of the discovered biomarkers (having p-value ≤ 0.05) contributed to computing the survival of the NSCLC patients. Fig. 6 depicts the KM curves for six out of 28 genes having the least p-value for the two groups of instances — one with low expression values and another with high expression values of the gene under consideration. The survival period (in the number of days) and the probability of survival are indicated along the horizontal and vertical axes, respectively. The curve in orange color shows the instances with a high expression value of the gene for the specific (survival period in the number of days, survival probability) pair. Similarly, the curve in black color shows the instances with a low expression value of the gene for the specific (survival period in the number of days, survival probability) pair. The Hazard-Ratio (HR) for each of these genes was found to be in the interval [1.44, 1.59]. Thus, these genes established their importance in prognostic evaluation by segregating the high survival probability group from the low survival probability group, based on the differences in the expression level.

4.4. Enriched pathway analysis

An over-representation analysis (ORA) of the ***NSCLC-Biomarkers-Set*** is performed to detect known biological processes that are overly represented or "enriched". For this purpose, we used WEB-based GEne SeT AnaLysis Toolkit (WebGestalt) [96]. To avoid Type-1 error (false-positive), the Benjamini–Hochberg test was employed, with the false discovery rate (FDR) set to 0.05. Fig. 7 shows Gene Ontology (GO) Biological Process, Reactome Pathway, and KEGG Pathway being targeted by ***NSCLC-Biomarkers-Set***.

Figs. 7(a) and 7(b) depict the significance of *Keratinization* in GO Biological Process and Reactome Pathway. [97] found that *Keratinization*

Table 5
14 out of 52 discovered biomarkers were found potentially druggable on DGIdb.

Gene	Categories	Source(s)
A2ML1	DRUGGABLE GENOME PROTEASE INHIBITOR ENZYME	HingoraniCasas GO, dGene Pharos
AP2M1	KINASE	Pharos
CALML5	ENZYME	Pharos
CEL	DRUGGABLE GENOME CELL SURFACE ENZYME	RussLampel, HopkinsGroom, HingoraniCasas GO GuideToPharmacology
CLCA2	DRUGGABLE GENOME	RussLampel, HingoraniCasas
CLDN3	TRANSPORTER	HumanProteinAtlas
HEY1	CLINICALLY ACTIONABLE	FoundationOneGenes, CarisMolecularIntelligence
IRF6	TRANSCRIPTION FACTOR	Pharos
RHCG	DRUGGABLE GENOME TRANSPORTER	RussLampel GuideToPharmacology, Pharos
SERPINB13	DRUGGABLE GENOME PROTEASE INHIBITOR	RussLampel, HopkinsGroom, HingoraniCasas HopkinsGroom, dGene
SPINK1	DRUGGABLE GENOME CLINICALLY ACTIONALBLE PROTEASE INHIBITOR	HingoraniCasas Tempus dGene
TMPRSS11A	PROTEASE DRUGGABLE GENOME	HopkinsGroom, dGene HopkinsGroom, RussLampel
TMPRSS11D	PROTEASE DRUGGABLE GENOME ENZYME	HopkinsGroom, GO, dGene HopkinsGroom, RussLampel, HingoraniCasas GuideToPharmacology
TP63	CLINICALLY ACTIONABLE DRUGGABLE GENOME TRANSCRIPTION FACTOR	MskImpact, FoundationOneGenes, Tempus HopkinsGroom, RussLampel Pharos

is associated with poor prognosis of LUSC. In fact, Keratins are proteins that play an important role in maintaining the structural integrity of cells and may be involved in cell differentiation. Keratinocytes, after their apoptotic death, create a keratin layer that could be considered a marker of well-differentiated LUSC.

Fig. 7(c) shows the importance of *Estrogen signaling* and *Renin secretion* pathway in lung cancer diagnosis and treatment [98,99]. [98] observed that *estrogens and growth factor* act as a promoter of tumor progression in NSCLC. They noted that estrogen receptors (ER) are found in significant proportions of NSCLC specimens. Experimenting on archival NSCLC tumors, they observed that $EGFR$ kinase inhibitor drug Faslodex alone as well as with erlotinib helped in restraining NSCLC growth. [99] observed that *renin–angiotensin system* (RAS) regulates certain functional capabilities, such as sustained angiogenesis and evasion of apoptosis, which are associated with lung cancer tumor progression and malignant transformation. So, they suggested that inhibiting RAS may serve as a significant adjuvant therapy in lung cancer.

4.5. Comparison with state-of-the-art

Table 6 shows a comparison of our findings with the state-of-the-art works [17,46,49]. Since the experiment is performed with TCGA RNA-Seq gene expression data, the results are compared with the studies involving the same. Compared to [49], the proposed framework achieved higher accuracy while using a significantly smaller set of biomarkers (the accuracy achieved by the proposed framework is compared with the highest accuracy achieved by [49]). Although [17,46] worked with smaller sets of genes compared to the proposed work, our method outperformed in terms of accuracy.

Table 6
Comparison of our proposed work with the state-of-the-art works.

Published work	#Genes/Features	Accuracy (%)
Girard et al. [17]	42	95
Chen and Dhahbi [49]	500	94.2
Tian [46]	8	92.48
Proposed Work	**52**	**95.75**

5. Conclusion and scope of future work

In this paper, an XAI-based deep learning framework is proposed to discover a small set of clinically relevant NSCLC biomarkers capable of classifying the NSCLC instances to their respective subtypes. By utilizing the framework, a set of 52 NSCLC biomarkers are discovered, 45 of which are found to be overlapping with the literature. To the best of our knowledge, the remaining seven genes are being reported for the first time for their relevance in NSCLC subtyping and could be further investigated for devising targeted therapy for NSCLC patients. We have demonstrated that the discovered set of biomarkers aids in classifying NSCLC instances accurately. Moreover, the XAI-based feature selection method incorporated in the proposed framework outperformed other feature selection methods in terms of classification accuracy. Further, we found 14 of the discovered biomarkers to be potentially druggable and 28 biomarkers with p-value ≤ 0.05 useful for predicting survival outcome. Pathway analysis using Gene Ontology (GO) Biological Process showed ten biological processes enriched by the discovered biomarkers. Similarly, four and two pathways are enriched by

Table 7
List of 45 biomarkers (out of the discovered set of 52 biomarkers), and the corresponding articles in conformity with them.

Genes	Articles
A2ML1	Zhang et al. [70]
BNC1	Yuan et al. [43]
C10orf99	Niemira et al. [78]
C14orf19	Li et al. [69]
CALML3	Girard et al. [17], Niemira et al. [78], Shinmura et al. [79], Su et al. [82]
CALML5	Niemira et al. [78]
CLCA2	Charkiewicz et al. [37], Yuan et al. [43], Girard et al. [17], Shinmura et al. [79]
CLDN3	Charkiewicz et al. [37], Sanchez-Palencia et al. [75], Girard et al. [17], Chang et al. [77], Shinmura et al. [79]
DSC3	Charkiewicz et al. [37], Yuan et al. [43], Sanchez-Palencia et al. [75], Girard et al. [17], Zhang et al. [76], Shinmura et al. [79]
DSG1	Saaber et al. [72]
DSG3	Charkiewicz et al. [37], Yuan et al. [43], Sanchez-Palencia et al. [75], Girard et al. [17], Zhang et al. [76], Niemira et al. [78], Shinmura et al. [79], Xiao et al. [80], Su et al. [82], Saaber et al. [72]
FTHL3	Li and Wang [73]
HEY1	Lin et al. [74]
IRF6	Sanchez-Palencia et al. [75]
KRT13	Girard et al. [17], Shinmura et al. [79]
KRT14	Charkiewicz et al. [37], Chen and Dhahbi [49], Yuan et al. [43], Chang et al. [77],Shinmura et al. [79]
KRT16	Yuan et al. [43], Girard et al. [17], Shinmura et al. [79], Su et al. [82]
KRT5	Charkiewicz et al. [37], Chen and Dhahbi [49], Yuan et al. [43], Sanchez-Palencia et al. [75], Girard et al. [17], Zhang et al. [76], Shinmura et al. [79], Xiao et al. [80]
KRT6A	Charkiewicz et al. [37], Chen and Dhahbi [49], Yuan et al. [43], Girard et al. [17], Zhang et al. [76], Chang et al. [77], Shinmura et al. [79], Xiao et al. [80], Su et al. [82]
KRT6B	Charkiewicz et al. [37], Girard et al. [17], Zhang et al. [76], Chang et al. [77], Shinmura et al. [79], Xiao et al. [80], Su et al. [82]
KRT6C	Girard et al. [17], Zhang et al. [76], Chang et al. [77], Niemira et al. [78], Shinmura et al. [79]
LASS3	Su et al. [82]
LGALS7B	Niemira et al. [78]
NACA2	Li et al. [69]
PA2G4P4	Li et al. [69]
PRSSL1	Zhang et al. [68]
PTTG3P	Yang et al. [63], Huang et al. [64]
RHCG	Chang et al. [77]
RNASE7	Zhang et al. [65]
RPL7	Ryan et al. [66]
S100A2	Chen and Dhahbi [49]
S100A7	Zhang et al. [67]
S100A7A	Chen et al. [71]
SERPINB13	Charkiewicz et al. [37], Yuan et al. [43], Chang et al. [77]
SPINK1	Charkiewicz et al. [37], Girard et al. [17], Chang et al. [77]
SPRR1A	Chang et al. [77], Xiong et al. [81]
SPRR1B	Sanchez-Palencia et al. [75], Su et al. [82]
SPRR2A	Zhang et al. [76]
SPRR2D	Xiong et al. [81]
SPRR2E	Xiong et al. [81]
SPRR3	Niemira et al. [78], Xiong et al. [81]
TMEM40	Yuan et al. [43]
TMPRSS11A	Su et al. [82]
TMPRSS11D	Charkiewicz et al. [37], Niemira et al. [78], Xiong et al. [81]
TP63	Yuan et al. [43], Sanchez-Palencia et al. [75], Girard et al. [17], Zhang et al. [76], Shinmura et al. [79]

(*continued on next page*)

(*continued*).

the discovered biomarkers while using Reactome and KEGG Pathway databases. However, NSCLC being a genetically diverse disease, a single omics data may not be sufficient enough to capture the heterogeneity of the tumor and we expect the analysis involving multi-omics data, or fusing omics data with histopathological images to throw more light on the heterogeneity of NSCLC.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Kountay Dwivedi would like to thank University Grants Commission, New Delhi, India, for providing Junior Research Fellowship (Reference ID: 190510173202). The team would also like to thank Prof. Shandar Ahmad and his team at the School of Computational and Integrative Sciences, Jawaharlal Nehru University, for his insightful suggestions, and Dr. Debasis Dash, Institute of Genomics and Integrative Biology, Council of Scientific and Industrial Research, New Delhi, India, for providing his eminent guidance. Also, Dr. Ankit Rajpal (PI) and Dr. Manoj Agarwal (Co-PI) would like to thank the Institute of Eminence (IoE), University of Delhi, India to provide a minor research grant under the Faculty Research Programme (Ref. No./IoE/2021/12/FRP).

Appendix

See Table 7.

References

[1] J. Ferlay, M. Ervik, F. Lam, M. Colombet, L. Mery, M. Piñeros, A. Znaor, I. Soerjomataram, F. Bray, Global cancer observatory: Cancer today, Lyon, France: Int. Agency Res. Cancer 3 (20) (2018) 2019.

[2] C. Zappa, S.A. Mousa, Non-small cell lung cancer: Current treatment and future advances, Transl. Lung Cancer Res. 5 (3) (2016) 288.

[3] H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: Cancer J. Clin. 71 (3) (2021) 209–249.

[4] W.D. Travis, E. Brambilla, M. Noguchi, A.G. Nicholson, K.R. Geisinger, Y. Yatabe, D.G. Beer, C.A. Powell, G.J. Riely, P.E. Van Schil, et al., International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung Adenocarcinoma, J. Thoracic Oncol. 6 (2) (2011) 244–285.

[5] K. Inamura, Lung cancer: Understanding its molecular pathology and the 2015 WHO classification, Front. Oncol. 7 (2017) 193.

[6] R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics, 2019, CA: Cancer J. Clin. 69 (1) (2019) 7–34.

[7] F. Kocher, W. Hilbe, A. Seeber, A. Pircher, T. Schmid, R. Greil, J. Auberger, M. Nevinny-Stickel, W. Sterlacci, A. Tzankov, et al., Longitudinal analysis of 2293 NSCLC patients: A comprehensive study from the TYROL registry, Lung Cancer 87 (2) (2015) 193–200.

[8] N. Duma, R. Santana-Davila, J.R. Molina, Non–small cell lung cancer: Epidemiology, screening, diagnosis, and treatment, Mayo Clin. Proc. 94 (8) (2019) 1623–1640.

[9] H. Uramoto, F. Tanaka, Recurrence after surgery in patients with NSCLC, Transl. Lung Cancer Res. 3 (4) (2014) 242.

[10] J. Dong, B. Li, D. Lin, Q. Zhou, D. Huang, Advances in targeted therapy and immunotherapy for non-small cell lung cancer based on accurate molecular typing, Front. Pharmacol. 10 (2019) 230.

[11] V.V. Padma, An overview of targeted cancer therapy, BioMedicine 5 (4) (2015) 1–6.

[12] M. Reck, D.F. Heigener, T. Mok, J.-C. Soria, K.F. Rabe, Management of non-small-cell lung cancer: Recent developments, Lancet 382 (9893) (2013) 709–719.

[13] P.M.-a.T. Group, Postoperative radiotherapy in non-small-cell lung cancer: Systematic review and meta-analysis of individual patient data from nine randomised controlled trials, Lancet 352 (9124) (1998) 257–263.

[14] S. Carnio, S. Novello, M. Papotti, M. Loiacono, G.V. Scagliotti, Prognostic and predictive biomarkers in early stage non-small cell lung cancer: Tumor based approaches including gene signatures, Transl. Lung Cancer Res. 2 (5) (2013) 372.

[15] W.D. Travis, E. Brambilla, G.J. Riely, New pathologic classification of lung cancer: Relevance for clinical practice and clinical trials, J. Clin. Oncol. 31 (8) (2013) 992–1001.

[16] W. Zhao, H. Wang, Y. Peng, B. Tian, L. Peng, D.-C. Zhang, ΔNp63, CK5/6, TTF-1 and napsin A, a reliable panel to subtype non-small cell lung cancer in biopsy specimens, Int. J. Clin. Exper. Pathol. 7 (7) (2014) 4247.

[17] L. Girard, J. Rodriguez-Canales, C. Behrens, D.M. Thompson, I.W. Botros, H. Tang, Y. Xie, N. Rekhtman, W.D. Travis, I.I. Wistuba, et al., An expression signature as an aid to the histologic classification of non–small cell lung cancer, Clin. Cancer Res. 22 (19) (2016) 4880–4889.

[18] S. Huang, J. Yang, S. Fong, Q. Zhao, Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges, Cancer Lett. 471 (2020) 61–71.

[19] S.M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G.S. Corrado, A. Darzi, et al., International evaluation of an AI system for breast cancer screening, Nature 577 (7788) (2020) 89–94.

[20] N. Coudray, P.S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A.L. Moreira, N. Razavian, A. Tsirigos, Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning, Nat. Med. 24 (10) (2018) 1559–1567.

[21] H. Saleem, A.R. Shahid, B. Raza, Visual interpretability in 3D brain tumor segmentation network, Comput. Biol. Med. 133 (2021) 104410.

[22] M. Shorfuzzaman, M.S. Hossain, A. El Saddik, An explainable deep learning ensemble model for robust diagnosis of diabetic retinopathy grading, ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) 17 (3s) (2021) 1–24.

[23] S. Rajpal, M. Agarwal, V. Kumar, A. Gupta, N. Kumar, Triphasic DeepBRCA-A deep learning-based framework for identification of biomarkers for breast cancer stratification, IEEE Access 9 (2021) 103347–103364.

[24] J. Meena, Y. Hasija, Application of explainable artificial intelligence in the identification of Squamous cell Carcinoma biomarkers, Comput. Biol. Med. 146 (2022) 105505.

[25] A. Agrawal, A. Chauhan, M.K. Shetty, M.D. Gupta, A. Gupta, et al., ECG-iCOVIDNet: Interpretable AI model to identify changes in the ECG signals of post-COVID subjects, Comput. Biol. Med. 146 (2022) 105540.

[26] I. Kononenko, Estimating attributes: Analysis and extensions of RELIEF, in: European Conference on Machine Learning, Springer, 1994, pp. 171–182.

[27] R. Battiti, Using mutual information for selecting features in supervised neural net learning, IEEE Trans. Neural Netw. 5 (4) (1994) 537–550.

[28] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (1) (2002) 389–422.

[29] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[30] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[31] R. Tibshirani, Regression shrinkage and selection via the Lasso, J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1) (1996) 267–288.

[32] H. Wang, F. Xing, H. Su, A. Stromberg, L. Yang, Novel image markers for non-small cell lung cancer classification and survival prediction, BMC Bioinformatics 15 (1) (2014) 1–12.

[33] L. Hou, D. Samaras, T.M. Kurc, Y. Gao, J.E. Davis, J.H. Saltz, Patch-based convolutional neural network for whole slide tissue image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2424–2433.

[34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.

[35] Y. Han, Y. Ma, Z. Wu, F. Zhang, D. Zheng, X. Liu, L. Tao, Z. Liang, Z. Yang, X. Li, et al., Histologic subtype classification of non-small cell lung cancer using PET/CT images, Eur. J. Nucl. Med. Mol. Imaging 48 (2) (2021) 350–360.

[36] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.

[37] R. Charkiewicz, J. Niklinski, J. Claesen, A. Sulewska, M. Kozlowski, A. Michalska-Falkowska, J. Reszec, M. Moniuszko, W. Naumnik, W. Niklinska, Gene expression signature differentiates histology but not progression status of early-stage NSCLC, Transl. Oncol. 10 (3) (2017) 450–458.

[38] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing, J. R. Stat. Soc. Ser. B Stat. Methodol. 57 (1) (1995) 289–300.

[39] R. Tibshirani, T. Hastie, B. Narasimhan, G. Chu, Diagnosis of multiple cancer types by Shrunken centroids of gene expression, Proc. Natl. Acad. Sci. 99 (10) (2002) 6567–6572.

[40] A.L. Pineda, H.A. Ogoe, J.B. Balasubramanian, C. Rangel Escareño, S. Visweswaran, J.G. Herman, V. Gopalakrishnan, On predicting lung cancer subtypes using 'omic'data from tumor and tumor-adjacent histologically-normal tissue, BMC Cancer 16 (1) (2016) 1–11.

[41] I. Kononenko, E. Šimec, M. Robnik-Šikonja, Overcoming the myopia of inductive learning algorithms with RELIEFF, Appl. Intell. 7 (1) (1997) 39–55.

[42] G.K. Smyth, Linear models and empirical bayes methods for assessing differential expression in microarray experiments, Stat. Appl. Genet. Mol. Biol. 3 (1) (2004).

[43] F. Yuan, L. Lu, Q. Zou, Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms, Biochimica Et Biophysica Acta (BBA)-Mol. Basis Dis. 1866 (8) (2020) 165822.

[44] M. Dramiński, A. Rada-Iglesias, S. Enroth, C. Wadelius, J. Koronacki, J. Komorowski, Monte Carlo feature selection for supervised classification, Bioinformatics 24 (1) (2008) 110–117.

[45] H. Liu, R. Setiono, Incremental feature selection, Appl. Intell. 9 (3) (1998) 217–230.

[46] S. Tian, Classification and survival prediction for early-stage lung Adenocarcinoma and squamous cell Carcinoma patients, Oncol. Lett. 14 (5) (2017) 5464–5470.

[47] J.L. Morrison, R. Breitling, D.J. Higham, D.R. Gilbert, GeneRank: Using search engine technology for the analysis of microarray experiments, BMC Bioinformatics 6 (1) (2005) 1–14.

[48] P. Hoffman, G. Grinstein, K. Marx, I. Grosse, E. Stanley, DNA visual and analytic data mining, in: Proceedings. Visualization'97 (Cat. No. 97CB36155), IEEE, 1997, pp. 437–441.

[49] J.W. Chen, J. Dhahbi, Lung adenocarcinoma and lung squamous cell carcinoma cancer classification, biomarker identification, and gene expression analysis using overlapping feature selection methods, Sci. Rep. 11 (1) (2021) 1–15.

[50] M.E. Garber, O.G. Troyanskaya, K. Schluens, S. Petersen, Z. Thaesler, M. Pacyna-Gengelbach, M. Van De Rijn, G.D. Rosen, C.M. Perou, R.I. Whyte, et al., Diversity of gene expression in adenocarcinoma of the lung, Proc. Natl. Acad. Sci. 98 (24) (2001) 13784–13789.

[51] D.A. Wigle, I. Jurisica, N. Radulovich, M. Pintilie, J. Rossant, N. Liu, C. Lu, J. Woodgett, I. Seiden, M. Johnston, et al., Molecular profiling of non-small cell lung cancer and correlation with disease-free survival, Cancer Res. 62 (11) (2002) 3005–3008.

[52] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38.

[53] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 3319–3328.

[54] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).

[55] L. Shapley, in: E. Artin, M. Morse (Eds.), Quota solutions op n-person games1, 1953, p. 343.

[56] Captum, Captum: Model interpretability for Pytorch, 2019, https://captum.ai/api/_modules/captum/attr/_core/gradient_shap.html#GradientShap.

[57] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: International Conference on Machine Learning, PMLR, 2017, pp. 3145–3153.

[58] M.J. Goldman, B. Craft, M. Hastie, K. Repečka, F. McDade, A. Kamath, A. Banerjee, Y. Luo, D. Rogers, A.N. Brooks, et al., Visualizing and interpreting cancer genomics data via the Xena platform, Nature Biotechnol. 38 (6) (2020) 675–678.

[59] C. Cheadle, M.P. Vawter, W.J. Freed, K.G. Becker, Analysis of microarray data using Z score transformation, J. Mol. Diagn. 5 (2) (2003) 73–81.

[60] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, 2017, arXiv preprint arXiv:1711.05101.
[61] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Adv. Neural Inf. Process. Syst. 32 (2019).
[62] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, et al., Captum: A unified and generic model interpretability library for pytorch, 2020, arXiv preprint arXiv:2009.07896.
[63] S. Yang, X. Wang, J. Liu, B. Ding, K. Shi, J. Chen, W. Lou, Distinct expression pattern and prognostic values of pituitary Tumor transforming gene family genes in non-small cell lung cancer, Oncol. Lett. 18 (5) (2019) 4481–4494.
[64] H.-T. Huang, Y.-M. Xu, S.-G. Ding, X.-Q. Yu, F. Wang, H.-F. Wang, X. Tian, C.-J. Zhong, The novel lncRNA PTTG3P is downregulated and predicts poor prognosis in non-small cell lung cancer, Arch. Med. Sci. AMS 16 (4) (2020) 931.
[65] J. Zhang, J. Zhang, C. Yuan, Y. Luo, Y. Li, P. Dai, W. Sun, N. Zhang, J. Ren, J. Zhang, et al., Establishment of the prognostic index of lung squamous cell Carcinoma based on immunogenomic landscape analysis, Cancer Cell Int. 20 (1) (2020) 1–16.
[66] S.-L. Ryan, K.A. Dave, S. Beard, M. Gyimesi, M. McTaggart, K.B. Sahin, C. Molloy, N.S. Gandhi, E. Boittier, C.G. O'Leary, et al., Identification of proteins deregulated by platinum-based chemotherapy as novel biomarkers and therapeutic targets in non-small cell lung cancer, Front. Oncol. 11 (2021) 241.
[67] H. Zhang, Q. Zhao, Y. Chen, Y. Wang, S. Gao, Y. Mao, M. Li, A. Peng, D. He, X. Xiao, Selective expression of S100A7 in lung squamous cell Carcinomas and large cell Carcinomas but not in Adenocarcinomas and small cell Carcinomas, Thorax 63 (4) (2008) 352–359.
[68] Y. Zhang, H. Song, Y. Miao, R. Wang, L. Chen, Frequent transcriptional inactivation of Kallikrein 10 gene by CpG Island hypermethylation in non-small cell lung cancer, Cancer Sci. 101 (4) (2010) 934–940.
[69] Y. Li, K. Kang, J.M. Krahn, N. Croutwater, K. Lee, D.M. Umbach, L. Li, A comprehensive genomic pan-cancer classification using the cancer genome Atlas gene expression data, BMC Genomics 18 (1) (2017) 1–13.
[70] W. Zhang, Q. Cui, W. Qu, X. Ding, D. Jiang, H. Liu, TRIM58/cg26157385 methylation is associated with eight prognostic genes in lung squamous cell Carcinoma, Oncol. Rep. 40 (1) (2018) 206–216.
[71] Y.-C. Chen, M.-C. Lin, C.-C. Hsiao, Y.-X. Zheng, K.-D. Chen, M.-T. Sung, C.-J. Chen, T.-Y. Wang, Y.-Y. Lin, H.-C. Chang, et al., Increased S100a15 expression and decreased DNA methylation of its gene promoter are involved in high metastasis potential and poor outcome of lung adenocarcinoma, Oncotarget 8 (28) (2017) 45710.
[72] F. Saaber, Y. Chen, T. Cui, L. Yang, M. Mireskandari, I. Petersen, Expression of desmogleins 1–3 and their clinical impacts on human lung cancer, Pathol. Res. Prac. 211 (3) (2015) 208–213.
[73] Z. Li, Y. Wang, Long non-coding RNA FTH1P3 promotes the metastasis and aggressiveness of non-small cell lung carcinoma by inducing epithelial-mesenchymal transition, Int. J. Clin. Exper. Pathol. 12 (10) (2019) 3782.
[74] L. Lin, R. Mernaugh, F. Yi, D. Blum, D.P. Carbone, T.P. Dang, Targeting specific regions of the Notch3 ligand-binding domain induces apoptosis and inhibits tumor growth in lung cancer, Cancer Res. 70 (2) (2010) 632–638.
[75] A. Sanchez-Palencia, M. Gomez-Morales, J.A. Gomez-Capilla, V. Pedraza, L. Boyero, R. Rosell, M.E. Fárez-Vidal, Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer, Int. J. Cancer 129 (2) (2011) 355–364.
[76] H. Zhang, Z. Jin, L. Cheng, B. Zhang, Integrative analysis of methylation and gene expression in lung adenocarcinoma and squamous cell lung Carcinoma, Front. Bioeng. Biotechnol. 8 (2020) 3.
[77] H.-H. Chang, J.M. Dreyfuss, M.F. Ramoni, A transcriptional network signature characterizes lung cancer subtypes, Cancer 117 (2) (2011) 353–360.
[78] M. Niemira, F. Collin, A. Szalkowska, A. Bielska, K. Chwialkowska, J. Reszec, J. Niklinski, M. Kwasniewski, A. Kretowski, Molecular signature of subtypes of non-small-cell lung cancer by large-scale transcriptional profiling: Identification of key modules and genes by weighted gene co-expression network analysis (WGCNA), Cancers 12 (1) (2020) 37.
[79] K. Shinmura, H. Igarashi, H. Kato, Y. Kawanishi, Y. Inoue, S. Nakamura, H. Ogawa, T. Yamashita, A. Kawase, K. Funai, et al., CLCA2 as a novel immunohistochemical marker for differential diagnosis of squamous cell carcinoma from adenocarcinoma of the lung, Dis. Markers 2014 (2014).
[80] J. Xiao, X. Lu, X. Chen, Y. Zou, A. Liu, W. Li, B. He, S. He, Q. Chen, Eight potential biomarkers for distinguishing between lung adenocarcinoma and squamous cell carcinoma, Oncotarget 8 (42) (2017) 71759.
[81] Y. Xiong, L. Mingzhen, P. Zhang, L. Zhang, Y. Yue, Study on genetype in lung squamous carcinoma by high-throughput of transcriptome sequence, Zhongguo Fei Ai Za Zhi 20 (11) (2017).
[82] R. Su, J. Zhang, X. Liu, L. Wei, Identification of expression signatures for non-small-cell lung carcinoma subtype classification, Bioinformatics 36 (2) (2020) 339–346.
[83] M. Griffith, O.L. Griffith, A.C. Coffman, J.V. Weible, J.F. McMichael, N.C. Spies, J. Koval, I. Das, M.B. Callaway, J.M. Eldred, et al., DGIdb: mining the druggable genome, Nature Methods 10 (12) (2013) 1209–1210.
[84] D.K. Simanshu, D.V. Nissley, F. McCormick, RAS proteins and their regulators in human disease, Cell 170 (1) (2017) 17–33.
[85] S. Yuan, H. Chu, J. Huang, X. Zhao, Z.-W. Ye, P.-M. Lai, L. Wen, J.-P. Cai, Y. Mo, J. Cao, et al., Viruses harness YxxØ motif to interact with host AP2M1 for replication: A vulnerable broad-spectrum antiviral target, Sci. Adv. 6 (35) (2020) eaba7910.
[86] L. Ma, W. Yin, H. Ma, I. Elshoura, L. Wang, Targeting claudin-3 suppresses stem cell-like phenotype in nonsquamous non-small-cell lung carcinoma, Lung Cancer Manag. 8 (1) (2019) LMT04.
[87] M. Katoh, M. Katoh, Precision medicine for human cancers with Notch signaling dysregulation, Int. J. Mol. Med. 45 (2) (2020) 279–297.
[88] Y. Liu, G. Shao, Z. Yang, X. Lin, X. Liu, B. Qian, Z. Liu, Interferon regulatory factor 6 correlates with the progression of non-small cell lung cancer and can be regulated by miR-320, J. Pharm. Pharmacol. 73 (5) (2021) 682–691.
[89] T.-C. Lin, Functional roles of SPINK1 in cancers, Int. J. Mol. Sci. 22 (8) (2021) 3814.
[90] M. Guo, X. Zhou, X. Han, Y. Zhang, L. Jiang, SPINK1 is a prognosis predicting factor of non-small cell lung cancer and regulates redox homeostasis, Oncol. Lett. 18 (6) (2019) 6899–6908.
[91] X. Cao, Z. Tang, F. Huang, Q. Jin, X. Zhou, J. Shi, High TMPRSS11D protein expression predicts poor overall survival in non-small cell lung cancer, Oncotarget 8 (8) (2017) 12812.
[92] E.-O. Bilguun, K. Kaira, R. Kawabata-Iwakawa, S. Rokudai, K. Shimizu, T. Yokobori, T. Oyama, K. Shirabe, M. Nishiyama, Distinctive roles of syntaxin binding protein 4 and its action target, TP63, in lung squamous cell carcinoma: A theranostic study for the precision medicine, BMC Cancer 20 (1) (2020) 1–14.
[93] C.M. Lynch, B. Abdollahi, J.D. Fuqua, A.R. de Carlo, J.A. Bartholomai, R.N. Balgemann, V.H. van Berkel, H.B. Frieboes, Prediction of lung cancer patient survival via supervised machine learning classification techniques, Int. J. Med. Inform. 108 (2017) 1–8, http://dx.doi.org/10.1016/j.ijmedinf.2017.09.013, URL https://www.sciencedirect.com/science/article/pii/S1386505617302368.
[94] A. Lánczky, B. Győrffy, et al., Web-based survival analysis tool tailored for medical research (KMplot): Development and implementation, J. Med. Internet Res. 23 (7) (2021) e27633.
[95] B. Győrffy, A. Lanczky, A.C. Eklund, C. Denkert, J. Budczies, Q. Li, Z. Szallasi, An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients, Breast Cancer Res. Treat. 123 (3) (2010) 725–731.
[96] Y. Liao, J. Wang, E.J. Jaehnig, Z. Shi, B. Zhang, WebGestalt 2019: Gene set analysis toolkit with revamped UIs and APIs, Nucleic Acids Res. 47 (W1) (2019) W199–W205.
[97] H.J. Park, Y.-J. Cha, S.H. Kim, A. Kim, E.Y. Kim, Y.S. Chang, Keratinization of lung squamous cell carcinoma is associated with poor clinical outcome, Tuberculosis Respiratory Dis. 80 (2) (2017) 179–186.
[98] D.C. Márquez-Garbán, H.-W. Chen, M.C. Fishbein, L. Goodglick, R.J. Pietras, Estrogen receptor signaling pathways in human non-small cell lung cancer, Steroids 72 (2) (2007) 135–143.
[99] M.J. Catarata, R. Ribeiro, M.J. Oliveira, C. Robalo Cordeiro, R. Medeiros, Renin-angiotensin system in lung tumor and microenvironment interactions, Cancers 12 (6) (2020) 1457.