Computer Methods and Programs in Biomedicine 243 (2024) 107864

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Enlightening the path to NSCLC biomarkers: Utilizing the power of XAI-guided deep learning

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

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

heterogeneous 异质
cellular 细胞的
adenocarcinoma 腺癌
squamous cell carcinoma 鳞状细胞癌

ARTICLE INFO

Keywords:
Non-small cell lung cancer
DNA methylation
Explainable AI
Biomarkers
Classification
Druggability

ABSTRACT

利用甲基化数据在XAI的帮助下找到了一系列NSCLC相关的biomarker

Background and Objective: The early diagnosis of Non-small cell lung cancer (NSCLC) is of prime importance to improve the patient's survivability and quality of life. Being a heterogeneous disease at the molecular and cellular level, the biomarkers responsible for the heterogeneity aid in distinguishing NSCLC into its prominent subtypes–*adenocarcinoma* and *squamous cell carcinoma*. Moreover, if identified, these biomarkers could pave the path to targeted therapy. Through this work, a novel explainable AI (XAI)-guided deep learning framework is proposed that assists in discovering a set of significant NSCLC-relevant biomarkers using methylation data.
Methods: The proposed framework is divided into two blocks– the first block combines an autoencoder and a neural network to classify NSCLC instances. The second block utilizes various eXplainable AI (XAI) methods, namely *IntegratedGradients*, *GradientSHAP*, and *DeepLIFT*, to discover a set of ***seven*** significant biomarkers.
Results: The classification performance of the biomarkers discovered using the proposed framework is evaluated by employing multiple machine learning algorithms, among which the Multilayer Perceptron (MLP) algorithm-based model outperforms others, yielding a 10-fold cross-validation accuracy of 91.53%. An improved accuracy of 96.37% is achieved by integrating RNA-Seq, CNV, and methylation data. On performing statistical analysis using the Friedman and Nemenyi tests, the MLP model is found to be significantly better than other machine learning-based models. Further, the clinical efficacy of the resultant biomarkers is established based on their potential druggability, the likelihood of predicting NSCLC patients' survival, gene-disease association, and biological pathways targeted by them. While the biomarkers *C18orf18, CCNT2, THOP1,* and *TNPO2*, are found potentially druggable, the biomarkers *CCDC15, SNORA9, THOP1*, and *TNPO2* are found prognostically relevant. On further analysis, some of the discovered biomarkers are found to be associated with around 104 diseases. Moreover, five KEGG, ten Reactome, and three Wiki pathways are found to be triggered by the biomarkers discovered.
Conclusion: In summary, the proposed framework uncovers a set of clinically effective biomarkers that accurately classify NSCLC. As a future course of work, efforts would be made to combine a variety of omics data with histopathological data to unveil more precise biomarkers for devising personalized therapy.

1. Introduction

Lung cancer is responsible for the most cancer-related deaths worldwide, with a 5-year survival rate between 15% and 20% ([73,70]). According to a study conducted by ([80]), the estimated number of lung cancer cases in 2020 was 2,206,771 (11.4% of all sites), with 1,796,144 (18% of all sites) cases resulting in a fatality. Lung cancer is primarily categorized as non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) – NSCLC being the most common form, accounting for approximately 85% of cases while SCLC covering the remaining 15% ([36,83]). NSCLC could be further categorized into two major subtypes–*adenocarcinoma* (LUAD) and *squamous cell carcinoma* (LUSC) ([83]).

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

https://doi.org/10.1016/j.cmpb.2023.107864
Received 29 July 2023; Received in revised form 7 October 2023; Accepted 11 October 2023
Available online 20 October 2023

The conventional treatment regime for early-stage NSCLC is surgery, which may be followed by chemotherapy and radiotherapy as adjuvant therapies ([95,33,17]). Nearly 84% of the patients, when diagnosed, are found at advanced stages of NSCLC ([57,95]), making it critically important to increase their chances of survival ([95]). The recent advancement in next-generation sequencing technologies and the development of therapeutic intervention drugs for targetable mutations like $EGFR$, ALK, $PI3K/AKT/mTOR$, etc. have revolutionized biomarker-driven therapeutic strategies for NSCLC treatment ([95,33,17,94,92]). Since NSCLC is a heterogeneous disease at the cellular and molecular level ([12,83,58,92]), making personalized medicines based on the histopathological features of the patient's tumor requires accurate subtyping of NSCLC ([83,58,17]). Indeed, any new NSCLC subtype-specific biomarker may contribute towards the improvement of the patient's survivability and betterment of life ([83,92,33,17]).

1.1. Literature review

histopathological 组织病理学
aberrant 异常
plausibly 似是而非地

The literature comprises numerous works that aim at subtyping NSCLC or discovering NSCLC-relevant biomarkers by employing a variety of omics data such as transcriptomics, genomics, and epigenomics data ([24,19,50,11,28,13,49,71,27,8]). DNA methylation, an epigenetic modification of DNA ([84]), expresses the aberrant chemical modifications in the DNA that may lead to unfavorable alterations in gene expression, plausibly causing tumor proliferation ([5,46,43,40,35]). Although DNA methylation plays an essential role in various biological activities (genomic imprinting, suppressing parasitic DNA sequences, transcriptional repression, etc.), it has been observed that cancerous cells may consist of genome-wide hypomethylation (a major reason for carcinogenesis) ([84]). Consequently, local hypermethylation of certain CpG islands is also observed, which may contribute to cancer development ([84]). In lung cancer, DNA methylation levels are found dysregulated across its subtypes [34]. Moreover, the DNA methylation value of certain biomarker(s) may contribute towards prognosis ([34]). In light of this discussion, a strong conjecture could be made that DNA methylation biomarkers are promising for targeted therapy, early detection, and prognosis of lung cancer.

Despite the significant role of DNA methylation in the early detection of carcinoma, there has been limited focus on this omic type in NSCLC classification and biomarker discovery. Guo et al. ([27]) identified a set of signature genes of DNA methylation to segregate NSCLC tissues from normal tissues. They combined multiple DNA methylation datasets and eliminated various nonbiological noises (known as batch effect) via an empirical Bayes algorithm called *ComBat* ([42]). Subsequently, they conducted a feature selection process using a support vector machine that resulted in five differentially expressed genes for methylation level, capable of distinguishing tumor tissues from normal tissues. These genes were evaluated based on their diagnostic prediction, using four statistical algorithms (with 5-fold cross-validation) – logistic regression, support vector machine, random forest, and Bayes tree. The Bayes tree model outperformed the other models, achieving a prediction accuracy of 91%. Cai et al. ([8]) proposed an ensemble-based feature selection to classify lung cancer into three major types, NSCLC (LUAD and LUSC) and small cell carcinoma (SCLC). Firstly, they ranked the genes using multi-category receiver operating characteristic (multi-ROC) analysis and hypervolume under the manifold (HUM) ([75]) measure. Next, the authors used the minimum redundancy maximum relevance (mRMR) method ([68]) to rank the genes. Finally, the authors employed the random forest (RF) ([7]) method to compute the relevance scores of the genes. Subsequently, the three lists of ranked genes are intersected to form a set of 45 genes. Further, using incremental feature selection (IFS), they selected the top 16 features (biomarkers), which, when provided to support vector classifier, yielded a model that achieved a LUAD-LUSC classification accuracy of 84.6%.

1.2. Research motivation

The established literature discussed so far has employed either machine learning methods or statistical methods, or both for NSCLC subtype classification or biomarker discovery. The utility of AI algorithms, particularly deep learning-based neural networks, has not been explored for this problem and is a subject of active research ([62]). The deep models, indeed, provide improved performance over conventional algorithms; nevertheless, their "*black-box*" nature hinders the ability to explain their decision-making process to the medical practitioners ([48,87,82,53]). The eXplainable AI (XAI) approach provides a set of tools that assist in "*explaining*" the model's decision-making process ([48,66,74]).

Taking a cue from the above discussion, the present work proposes a novel framework that utilizes deep neural networks to discover NSCLC-relevant DNA methylation biomarkers using various XAI tools (Fig. 1). Initially, the dataset is preprocessed to make it compatible with the proposed framework. Next, the preprocessed dataset (*MethylDataset*) is provided as input to the framework that leverages it to uncover significant biomarkers capable of NSCLC classification. The framework comprises two blocks– an autoencoder combined with a neural network that classifies the input NSCLC instances, and a set of XAI tools that interpret the resultant model from the first block, thereby unveiling a set of biomarkers most significant for classification.

The contribution of the work can be summarized as follows:

- A novel, XAI-guided deep learning framework is proposed for the discovery of significant DNA methylation biomarkers capable of NSCLC classification. The framework is robust enough to perform effectively on any omics data (provided certain hyperparameter tuning is addressed).
- The framework assists in unveiling ***seven*** NSCLC-relevant DNA methylation biomarkers, out of which ***four*** are found to be potentially druggable. The remaining ***three*** could be further explored for their clinical relevance.
- Upon further exploration, ***four*** of the discovered biomarkers are found potentially effective in predicting NSCLC patients' survivability (with $0.026 \leq$ p-value ≤ 0.072).

The remainder of the article is organized as follows: Section 2 provides detailed concept of the XAI-guided deep learning framework; Section 3 covers the dataset details; Section 4 provides the description of the experiment and the environment where it is performed; Section 5 discusses the results observed; and finally, Section 7 covers the conclusions and future scope of the work.

2. XAI-guided deep learning framework

The objective of the experiment is to uncover a set of DNA methylation biomarkers significantly contributing to the classification of NSCLC instances into LUAD and LUSC. However, the DNA methylation dataset comprises a huge number of features (genes), far more than the number of instances (NSCLC patients), which may lead to *curse of dimensionality*. One approach to deal with this issue is to compress the enormous input feature space to a smaller space while keeping the loss due to compression to a minimum. The proposed framework incorporates a neural network-based autoencoder to achieve this task. Subsequently, the framework utilizes the compressed space to classify the input instances into their respective classes via a feed-forward neural network. Finally, the framework leverages a set of XAI tools for biomarkers discovery. The framework comprises two blocks:

Encoder-Neural Classifier (*Block-A*): This block utilizes a combination of a deep learning-based autoencoder and a feed-forward deep neural network to accurately distinguish NSCLC instances. The autoencoders ([72]) are neural networks specifically designed to reconstruct the input

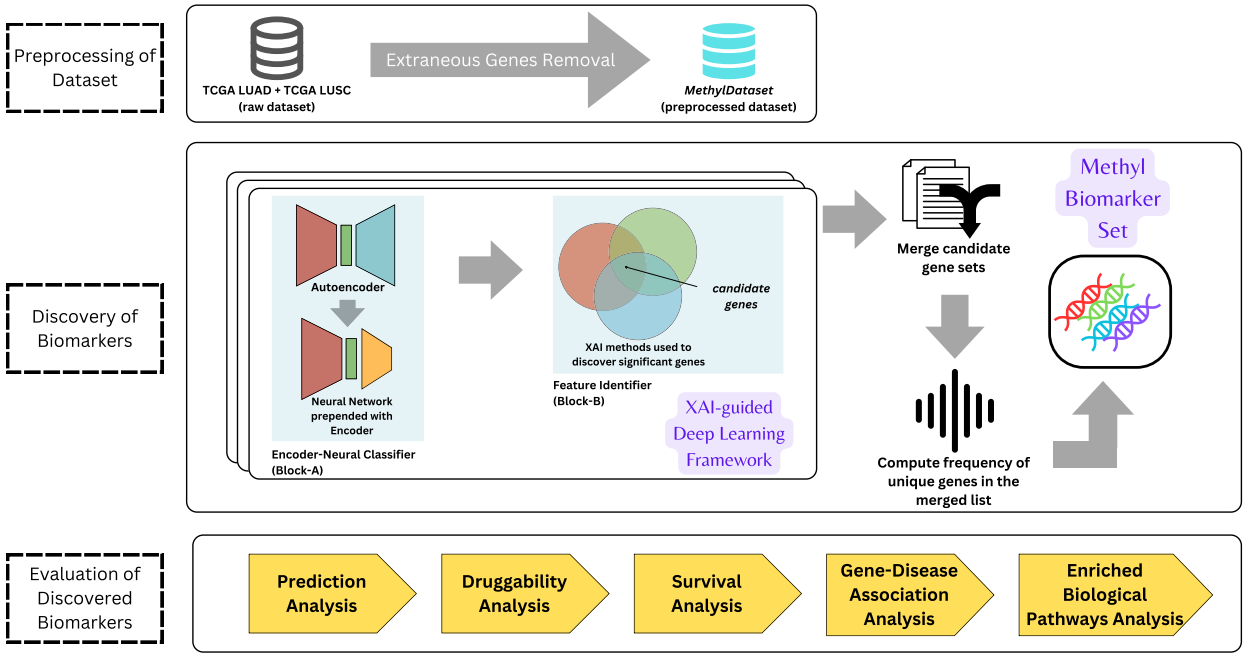

Fig. 1. Illustration of the proposed work. At first, the dataset is preprocessed, followed by discovering the DNA methylation biomarkers (***Methyl Biomarker Set***) via XAI-guided Deep Learning Framework. Conclusively, the discovered biomarkers are analyzed on various grounds.

fed to them ([1,2]). Assume an input vector $v \in \mathbb{R}^d$, the task of the autoencoder is to learn a function $f_{in} : \mathbb{R}^d \to \mathbb{R}^p$ and $f_{out} : \mathbb{R}^p \to \mathbb{R}^d$, such that:

$$\underset{f_{in}, f_{out}}{argmin} E[\Delta(v, f_{out} \circ f_{in}(v))] \tag{1}$$

where E: expectation over D, which is the distribution followed by input vector v, and Δ is the function computing the loss incurred during the reconstruction of v. The function $f_{in} : \mathbb{R}^d \to \mathbb{R}^p$ is called the *encoder*, while $f_{out} : \mathbb{R}^p \to \mathbb{R}^d$ is referred to as the *decoder*. When $p < n$, it states that the autoencoder is performing some kind of compression or feature extraction ([1]).

The autoencoder is utilized to compress the enormous feature (gene) space of the input dataset. Following the compression task, the *encoder* function, $f_{in} : \mathbb{R}^d \to \mathbb{R}^p$ is prepended to a feed-forward neural network, F, that ultimately predicts the class $\tilde{c}$ for the input vector v. Let layer $[l]$ be the first and only hidden layer in F with a single neuron. The inputs to this neuron will be the output vector $\tilde{v} \in \mathbb{R}^p$ of the *encoder* function f_{in}. The elements $\tilde{v}_1, \ldots, \tilde{v}_p$ of vector $\tilde{v}$ are weighted by corresponding weights $w_1^{[l]}, \ldots, w_p^{[l]}$ ([30]). The neuron comprises a bias b that is summed with the weighted input to form a net input $\tilde{v}_{net}$ ([30]):

$$\begin{aligned} \tilde{v}_{net} &= (w_1^{[l]}\tilde{v}_1 + w_2^{[l]}\tilde{v}_2 + \ldots + w_p^{[l]}\tilde{v}_p) + b \\ &= \mathbf{W}\tilde{\mathbf{v}} + b \end{aligned} \tag{2}$$

The output of the neuron is then computed as:

$$a = \sigma(\tilde{v}_{net}) \tag{3}$$

where σ is (generally) a non-linear "activation" function. Typically, for a binary class prediction problem:

$$\sigma(\tilde{v}_{net}) = \frac{1}{1 + \exp^{-\tilde{v}_{net}}} \tag{4}$$

As mentioned, the *encoder* ($f_{in} : \mathbb{R}^d \to \mathbb{R}^p$) outputs the compressed dataset, which is then passed as input to the neural network. Subsequently, the network is trained on the compressed dataset, resulting in a trained model F capable of accurately classifying NSCLC instances.

Feature Identifier *(Block-B)*: This block utilizes a set of XAI tools to interpret the trained model F resulting from *Block-A* and identifies a set of biomarkers significantly contributing towards the classification process. The XAI tools utilized, namely *IntegratedGradients (IG)* ([79]), *GradientSHAP (GS)* ([20]) and *DeepLIFT (DL)* ([76]), are model-specific methods that require the trained model F and an input vector $v \in \mathbb{R}^d$ with features $(v_1, \ldots, v_d)$, interpret the latent processing of F w.r.t. v and quantize the contribution of each feature towards predicting the class of v.

The ***IntegratedGradients (IG)*** quantizes the prediction of F w.r.t. v relative to a *baseline* ([79]) input $v' \in \mathbb{R}^d$:

$$A_F(v, v') = (a_1, a_2, .., a_d) \in R^d \tag{5}$$

where $A_F(v, v')$ is a vector, and a_i is the quantized contribution of v_i towards the prediction. Considering a straight line path from $v_i' \to v_i$, the gradients along all the points are computed, and the integrated gradients for v_i are the cumulative sum of these gradients ([79]).

$$IG(v_i) = a_i = (v_i - v_i') \times \int_{\alpha=0}^{1} \frac{\partial F(v' + \alpha \times (v - v'))}{\partial v_i} d\alpha \tag{6}$$

The ***GradientSHAP (GS)*** ([20]) is an enhanced version of *IntegratedGradients*, where the contribution score of each feature v_i of input vector v is computed by taking into account multiple *baselines* (unlike *IntegratedGradients* with a single *baseline*) randomly sampled from a distribution B and averaging the gradients over them ([20]):

$$GS(v_i) = \int_{v'} \Big(IG(v_i) \Big) p_B(v') dv' \tag{7}$$

where v' is the *baseline* sample drawn from the distribution B and $IG(v_i)$ is computed according to Equation (6).

The XAI method ***DeepLIFT*** was introduced by [76]. Importantly, it computes the difference in output from some "reference output" in terms of the difference of the input from some "reference input" ([76]). Here, the "reference" signifies the activation a of a neuron when the *baseline* v' is

Table 1
Demographic details of the respective TCGA-LUAD and TCGA-LUSC datasets.

Characteristic	TCGA-LUAD	TCGA-LUSC
Mean Age	65.02	67.54
Male/Female	214/244	274/96
Alive/Dead/Not Reported	157/281/0	146/204/20
Not Hispanic or Not Latino	343	239
Hispanic or Latino	7	6
Not Reported	108	125

inputted to the model F. More formally, for a target neuron t with its reference activation t^0, the difference-from-reference, Δt is computed as ([76]):

$$t = F(n_1, \ldots, n_r) \tag{8}$$

$$t^0 = F(n_1^0, \ldots, n_r^0) \tag{9}$$

$$\Delta t = t - t^0 \tag{10}$$

where $(n_1, \ldots, n_r)$ and $(n_1^0, \ldots, n_r^0)$ are necessary and sufficient neurons to compute t and t^0, respectively ([76]). The quantized contribution scores $C_{\Delta n_i \Delta t}$ for neurons $(n_1, \ldots, n_r)$ is assigned such that:

$$\Sigma_{i=1}^{r} C_{\Delta n_i \Delta t} = \Delta t \tag{11}$$

that is, $C_{\Delta n_i \Delta t}$ is equal to the difference-from-reference in t that is *blamed* upon the difference-from-reference of n_i ([76]).

3. Dataset details and preprocessing

Dataset Description

In this work, the dataset used for experimentation is generated by the TCGA Research Network. They provide two datasets for NSCLC, one for *adenocarcinoma* (TCGA-LUAD) and the other for *squamous cell carcinoma* (TCGA-LUSC), comprising 458 and 370 instances, respectively. The instances in the TCGA-LUAD and TCGA-LUSC datasets are described by the methylation intensities of sets of 20,115 and 20,116 genes, respectively. These datasets are constructed using Illumina Infinium assay that measures the methylated and unmethylated probe intensities of the alleles of the CpG probe under investigation and computes the beta values as the ratio of the aforementioned probe intensities, as follows ([18]):

$$\beta := \frac{max(probe_{methyl}, 0)}{max(probe_{methyl}, 0) + max(probe_{unmethyl}, 0) + \mathfrak{c}} \tag{12}$$

where $\mathfrak{c}$ represents a constant and $probe_{methyl}$ and $probe_{unmethyl}$ represent the methylation intensity values of the methylated and unmethylated CpG probes, respectively. The range of β, the methylation value, is between 0 and 1. While $\beta = 0$ indicates that all CpG copies are unmethylated, $\beta = 1$ indicates that copies are methylated ([18]). Thus, the lower β value refers to *hypomethylation*, whereas its higher value indicates *hypermethylation* ([67]). In the TCGA-LUAD and TCGA-LUSC datasets, given a data instance, the beta values are mapped over the human genome to obtain a single value for each gene. The TCGA-LUAD and TCGA-LUSC datasets are downloaded from LinkedOmics ([86]) portal, last updated on June 16, 2021. The demographic details of these datasets is provided in Table 1.

Preprocessing

The TCGA-LUSC and TCGA-LUAD datasets comprise identical sets of genes except for the $MBD3L2$ gene, which is available only in the TCGA-LUSC dataset. Therefore, $MBD3L2$ is removed from the TCGA-LUSC dataset to make the two datasets compatible. The TCGA-LUAD dataset comprises 92,12,670 (458 × 20,115) entries, including 5,608 missing entries. The TCGA-LUSC dataset comprises 74,42,920 (370 × 20,116) entries, including 4,500 missing entries. It is to be noted that approximately 0.06% of the entries were missing in each dataset. To

Table 2
Although the Mean imputer and the KNN-imputer yield similar accuracy, the margin of error is quite small for the *mean imputer* compared to the *KNN imputer*.

Imputation methodology	10-fold cross-validation Accuracy (95% confidence interval)	Coefficient of Variance
mean imputer	**92.39 (± 3.01)**	**1.08**
median imputer	90.56 (± 3.84)	1.41
most-frequent-value imputer	91.05 (± 3.25)	1.18
constant imputer	88.90 (± 4.27)	1.60
KNN imputer	92.47 (± 3.91)	1.40

Table 3
List of hyperparameter values selected for the training of autoencoder and feed-forward neural network.

Network	Hyperparameter Values
Autoencoder	epochs: 20 optimizer: AdamW ([54]) criterion: mean squared error loss learning rate: $5e^{-7}$ batch size: 32
Feed-forward Neural Network	epochs: 20 optimizer: AdamW criterion: binary cross entropy loss learning rate: $5e^{-7}$ batch size: 32

cater to this issue, multiple imputation methods are employed (please see Table 2) to fill the missing values in datasets. Although the *mean imputer* and the *KNN imputer* yield similar accuracy, the margin of error is quite small for the *mean imputer* compared to the *KNN imputer*. Further, the *mean imputer* is computationally more efficient than the *KNN imputer*. Therefore, we preferred the *mean imputer* to fill in the missing values over the *KNN imputer*.

As the TCGA-LUAD and TCGA-LUSC datasets comprised the patients of *adenocarcinoma* and *squamous cell carcinoma*, respectively, the two datasets are merged to form a single dataset comprising patients of both the NSCLC subtypes. The combined dataset will be referred to as *MethylDataset* in the remainder of the article.

4. Proposed methodology

The proposed XAI-guided deep learning framework comprises two blocks– *Encoder-Neural Classifier (Block-A)* and *Feature Identifier (Block-B)*. While Block-A encapsulates an autoencoder and a feed-forward neural network, *Block-B* encapsulates three XAI tools for interpretation– *IntegratedGradients*, *GradientSHAP*, and *DeepLIFT*. The architecture of the autoencoder and the neural network is shown in Fig. 2a and Fig. 2b. The hyperparameter values of both networks are listed in Table 3. The autoencoder of *Block-A* consists of an input layer of size 20,115, four hidden layers of sizes 4,096, 2,048, 2,048, and 4,096, respectively, and an output layer of size 20,115. The autoencoder's first three layers represent the encoder, and the last three represent the decoder. The batch normalization and the rectified linear unit (ReLU) activation are used in conjunction between each pair of autoencoder layers, except the bottleneck layers (please see Fig. 2a). The batch normalization is used to reduce the internal covariate shift ([37]), and the rectified linear unit (ReLU) activation is leveraged to introduce the non-linearity ([64]).

Fig. 2b depicts the feed-forward neural network of *Block-A* that utilizes the output of the autoencoder's encoder, followed by a concatenation of two hidden layers of sizes 128 and 64, respectively, and an output layer of size 2. Each pair of these layers is exposed to a combination of batch normalization and ReLU activation. Finally, the output layer uses a sigmoid activation function to cater to the binary classification problem.

Algorithm 1 gives step-by-step operations performed in the XAI-guided Deep Learning Framework. The *MethylDataset* ($D : \mathbb{R}^{n \times d}; n =$

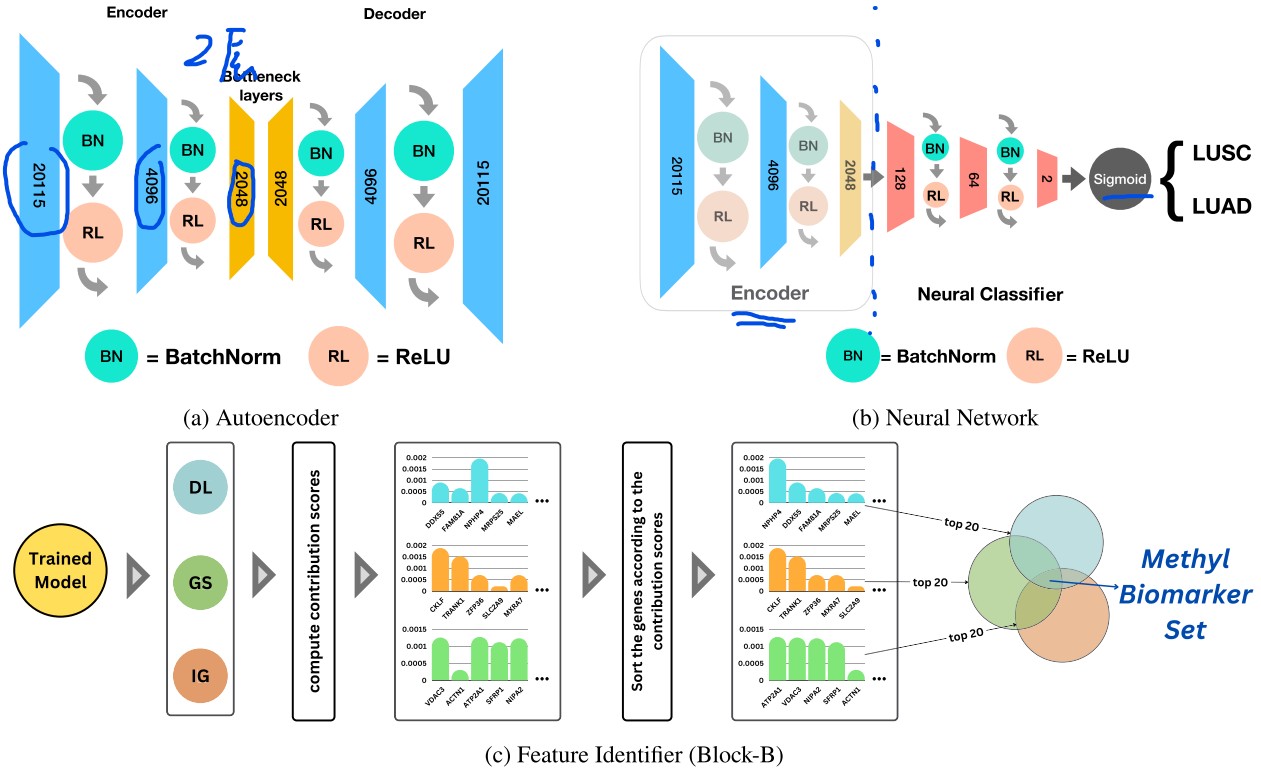

Fig. 2. The diagrammatic representation of the XAI-guided deep learning framework. Fig. 2a shows the architecture of the autoencoder of *Encoder-Neural Classifier (Block-A)*. The weights of the bottleneck layer represent the learned compressed feature space. Fig. 2b shows the architecture of the feed-forward neural network. Notice how the *encoder* function is detached from the autoencoder and prepended to the neural network. This step aids the neural network in utilizing the learned compressed space by the *encoder* function to perform classification. Fig. 2c illustrates the flowchart of the *Feature Identifier (Block-B)* of the XAI-guided deep learning framework.

$828, d = 20,115$), the size of compressed space ($p = 2,048$), the number of iterations of the experiment ($nIter$) (discussed at the end of this section), the list of XAI methods ($xMethods$), and the number of significant genes to be selected from each XAI tool ($nSelect$), are provided as input to the framework.

Execution of Encoder-Neural Classifier (Block-A): Initially, D and p are passed as inputs to the autoencoder AE of Block-A which compresses the feature space $D : \mathbb{R}^{n \times d} \rightarrow \tilde{D} : \mathbb{R}^{n \times p}$ (line: 7). Following the compression, the *encoder* function f_{in} of AE is prepended to the feed-forward neural network $FFNN$, where the compressed space $\tilde{D}$ is utilized to accurately classify the instances in D (line: 9). This ends the execution of Block-A, and the resultant model F is forwarded to Block-B for further processing.

Execution of Feature Identifier (Block-B): The flowchart of the Feature Identifier (Block-B) is illustrated in Fig. 2c. The resultant trained model F from Block-A is consecutively leveraged by the three XAI methods enlisted in $xMethods$ for the discovery of biomarkers (lines: 10–18). Each XAI method is provided with F, D, and $nSelect$ as inputs:

- *IntegratedGradients (IG)*: IG employs Equation (6) for the contribution score computation of each gene. Here, (v, i, F, v') corresponds to a tuple (input vector, target gene, model, *baseline*).
- *GradientSHAP*: GS utilizes Equation (7) to calculate the contribution of each gene. Here, (v, B, i, F) are associated with the tuple (input instance, *baseline* probability distribution, target gene, model). The *baseline* v' is sampled from its distribution B.
- *DeepLIFT (DL)*: DL employs Equation (11) to compute the contribution score. The values (Δn_i and Δt) are the *difference-from-reference* corresponding to neurons n_1 to n_r and target neuron t, respectively.

Fig. 3 illustrates a sample output of the execution of individual XAI methods. A set of twenty genes are randomly selected from each method and their quantized contribution scores are plotted. Post computation of the contribution score of each gene in D, the Feature Identifier block exits. Consequently, for each XAI method, the genes are arranged in decrementing order of their contribution value, and $nSelect$ most significant genes (genes with maximum contribution value) from each sorted list are retained; the rest are discarded. Finally, these three lists are intersected to obtain a set of $candidate\_genes$ (line: 19).

Significance of $nIter$: It is to be noted that the lines: 7–20 of Algorithm 1 represent a single complete iteration of the proposed experiment. The entire experiment is repeated for $nIter$ times (line: 4) with random seeding. The rationale behind this step is to capture the inherent stochasticity of the system, leading to a different initialization of the models' parameters in each iteration. As mentioned earlier, the output of each iteration is the $candidate\_genes$ corresponding to the random seed value generated for that iteration. Each $\{seed, candidate\_genes\}$ pair is appended to a collection named $candidate\_genes\_dict$. After running the experiment for $nIter$ times, the frequency of each distinct gene in the $candidate\_genes\_dict$ is calculated (line: 22). Those genes that have a frequency $\geq 0.5 \times nIter$ are considered as a discovered set of NSCLC-relevant methylation biomarkers named ***MethylBiomarkerSet*** (line: 23). This concludes the experimentation.

Experimental Setup

The experiment is conducted *in-silico*. The details of the system are provided in Table 4. In addition, the algebraic operations are conducted via Numpy v1.19.2, the dataset manipulation is performed via Pandas v1.0.5, and the visualizations (graphs and plots) are conducted over Matplotlib v3.2.2.

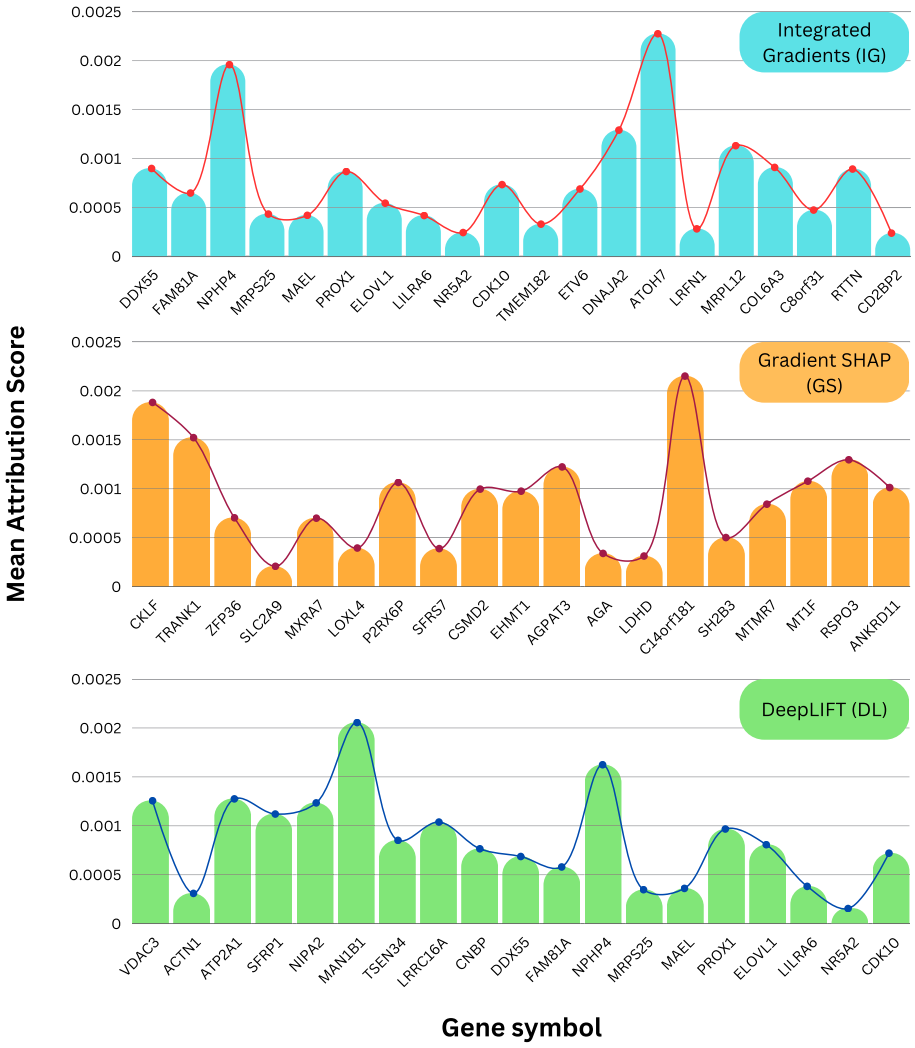

Fig. 3. Illustration of the genes with their contribution score as computed by individual XAI methods. A set of twenty genes are randomly selected from the entire input gene space.

Table 4
Hardware and software utilized for experimentation.

Hardware properties	
Model	Acer Predator Helios 300
Processor	Core i7 9750H (clocked 2.60 GHz)
Primary Memory	16 GB
Graphics Processing Unit	NVIDIA GeForce GTX 1660-Ti 6 GB dedicated graphics memory
Software properties	
Operating System	Windows 10 Home
Programming Environment	Python v3.7.7
Implementation Library	PyTorch v1.8.1
XAI Library	Captum v0.4.0

Table 5
The discovered set of biomarkers, when utilized as input feature set, give high performance in terms of classification accuracy by various machine learning models.

Machine Learning Models	Accuracy (%) (using 10-fold cross-validation)
Multilayer Perceptron Classifier (MLP)	91.53
Random Forest Classifier (RF)	84.17
Logistic Regression Classifier (LR)	83.48
Support Vector Classifier (SVC)	85.66

类似验证testing

5. Result analysis

The objective of the experiment is to discover a set of the most significant DNA methylation biomarkers capable of accurately classifying NSCLC instances into their appropriate subtypes – LUAD and LUSC. The deep learning-based framework employed in this experiment assists in finding out a set of ***seven*** such biomarkers, named *MethylBiomarkerSet*—$\{C18orf18, CCDC15, CCNT2, EXOC6, SNORA9, THOP1, TNPO2\}$.

The classification performance of *MethylBiomarkerSet* is assessed using four machine learning algorithms — multilayer perception classifier (MLP), random forest classifier (RF), logistic regression classifier (LR), and support vector classifier (SVC). Each of these algorithms is provided with the *MethylDataset* as input with only *MethylBiomarkerSet* as genes; the rest of the genes are discarded. Among the four trained models, the MLP-based model achieves the maximum average 10-fold cross-validation accuracy of 91.53%. The confusion matrix of the MLP model is presented in Fig. 4. Furthermore, the classification performance of each model is enlisted in Table 5.

Statistical Significance of XAI-guided Deep Learning Framework: To observe the statistical significance of the results obtained by the proposed framework over those obtained by the competitive machine learning algorithms (mentioned in Table 5), the Friedman test is utilized

Algorithm 1 XAI-guided Deep Learning Framework.

Require:
$D \leftarrow \{x_i, y_i\}_{i=1}^{n}; x \in \mathbb{R}^d; y \in \{LUSC, LUAD\}$
$p \leftarrow$ size of compressed space
$nIter \leftarrow$ number of times the experiment is repeated
$xMethods \leftarrow$ List of the XAI methods
$nSelect \leftarrow$ number of genes to be selected from each XAI method
Ensure: Return $\{MethylBiomarkerSet\}$

$candidate\_genes\_dict \leftarrow [\,]$
$candidates\_freq \leftarrow [\,]$
$MethylBiomarkerSet \leftarrow [\,]$
for $i = 1$ **to** $nIter$ **do**
  $genes_{IG}, genes_{GS}, genes_{DL} \leftarrow [\,]$
  $candidate\_genes \leftarrow [\,]$
  // **Encoder-Neural Classifier (Block-A)**
  $\tilde{D} \leftarrow AE(D, p)$
  $f_{in} \leftarrow detachEncoder(AE)$
  $F \leftarrow FFNN(f_{in}(D, \tilde{D}))$
  // **Feature Identifier (Block-B)**
  for all $M \in xMethods$ **do**
    if $M == IntegratedGradients$ **then**
      $genes_{IG} := IG(F, D, nSelect)$
    else if $M == GradientSHAP$ **then**
      $genes_{GS} := GS(F, D, nSelect)$
    else
      $genes_{DL} := DL(F, D, nSelect)$
    end if
  end for
  $candidate\_genes \leftarrow \bigcap\{genes_{IG}, genes_{GS}, genes_{DL}\}$
  $candidate\_genes\_dict.append(candidate\_genes)$
end for
// **Frequency computation**
$candidates\_freq := computeFreq(candidate\_genes\_dict)$
// **Biomarkers Discovery**
for all $gene \in candidates\_freq$ **do**
  if $freq(gene) \geq 0.5 \times nIter$ **then**
    $MethylBiomarkerSet.append(gene)$
  end if
end for
return $MethylBiomarkerSet$

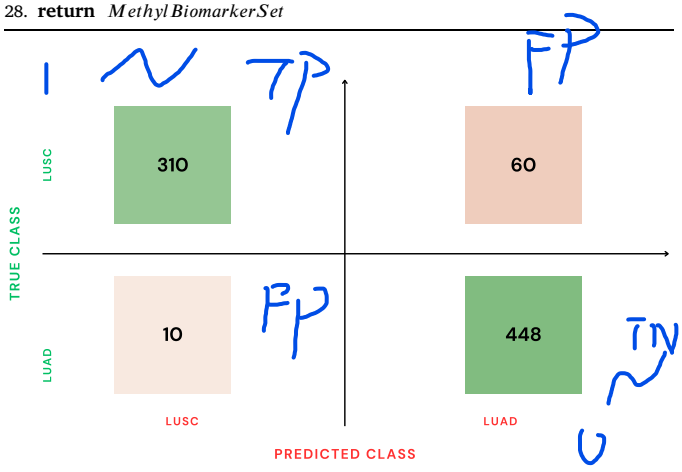

Fig. 4. Confusion matrix of the MLP classifier.

([16]). For 30 different runs, all the classifiers under comparison are 10-fold cross-validated on the *MethylDataset* with the *MethylBiomarkerSet* as the input features. Keeping the level of significance $\alpha = 0.05$, the Friedman test resulted in a χ^2 statistic of 49.31 ($p - val : 1.115e^{-10}$). The expected ranking of all the algorithms is as follows (a lower number indicates better rank): MLP (1.2); SVC (2.6), LR (2.7); RF (3.5). The result thus obtained rejects the null hypothesis that there exists no significant difference between the results yielded by the proposed framework and state-of-the-art machine learning classifiers.

Furthermore, by utilizing the Nemenyi test ([16]), the expected ranking of the algorithms under comparison, unveiled by the Friedman test, is analyzed. The Nemenyi test states that two algorithms differ significantly if the difference between the expected rank of those algorithms is greater than or equal to a critical distance CD, defined as:

$$CD := q_\alpha \times \sqrt{\frac{k \times (k+1)}{6 \times N}} \tag{13}$$

Here, $q_\alpha = 2.85$; number of runs, $N = 30$; and number of classifiers under comparison, $k = 4$. Putting these values in Equation (13) $CD = 0.95$. Fig. 5 shows the expected ranking of the algorithm (a lower number indicates better rank). The red-colored, blue-colored, and green-colored lines denote the difference between the expected rank of MLP and SVC (= 1.4), SVC and RF (= 0.9), and RF and LR (= 0.8), respectively. While no significant difference could be witnessed among SVC, RF, and LR classifiers (as the differences are $\leq CD$), a considerably significant difference in the rank is observed between MLP and the rest of the classifiers.

5.1. Druggability analysis

The clinical relevance of the *MethylBiomarkerSet* is observed and reported by utilizing the online database, **DGIdb** ([26]). Two out of the seven discovered biomarkers– $CCNT2$ and $THOP1$ are included in the DGIdb database. To explore the potential druggability of the remaining discovered biomarkers, the relevant literature is surveyed, revealing two more biomarkers– $C18orf18$ and $TNPO2$, as potentially druggable.

$C18orf18$, an alias of $TP53LC13$, is a $TP53$-inducible putative long non-coding RNA (lncRNA), which is one of the fifteen $TP53$-regulated lncRNA capable of encoding peptides, thus assisting in expanding the $TP53$ network ([91]). $TNPO2$, an alias to $TRN2$, is a nuclear trafficking factor responsible for the nuclear import of an RNA-binding protein HuR, through which it regulates apoptosis ([88]). Thus, an increase/decrease in its level may result in inverse consequential effects on cell death ([88]). Moreover, $TNPO2$ is found to be a promoter of gastric carcinoma cell proliferation along with an inhibitor of apoptosis ([55,25]). Also, [60] have identified the overexpression of $TNPO2$ in malignant pleural mesothelioma (MPM) tissues and cells, thus stating it as a probable novel cancer gene.

As far as we can tell, the following discovered genes – $CCDC15$, $EXOC6$, and $SNORA9$, have not yet been reported in the literature as potential biomarkers and could be considered for further clinical evaluation to check their potency for targeted therapy.

5.2. Survival analysis

To observe the potential of the *MethylBiomarkerSet* to predict the likelihood of survival of NSCLC patients, the Kaplan-Meier (KM) Plotter tool ([47,29]) is utilized. The clinical data of the patients in the *MethylDataset* is downloaded from the LinkedOmics portal and is provided as input to the KM Plotter tool with the input set of genes as *MethylBiomarkerSet*. The univariate Cox regression model is used to compute the hazard ratio and the p-value of the biomarkers. Four out of seven biomarkers are found prognostically significant with a p-value ranging between [0.026 — 0.072].

Fig. 6 shows the KM curve of the four prognostically relevant biomarkers. From the plots, it could be observed that while the lower methylation values of $CCDC15, SNORA9$, and $TNPO2$ contribute towards higher survivability of the patients, the lower methylation values of $THOP1$ contribute towards lower survivability of the patients.

5.3. Gene-disease association analysis

Gene-disease association (GDA) provides an in-depth understanding of a disease's aetiology (the set of causes of a disease) ([65]). Though

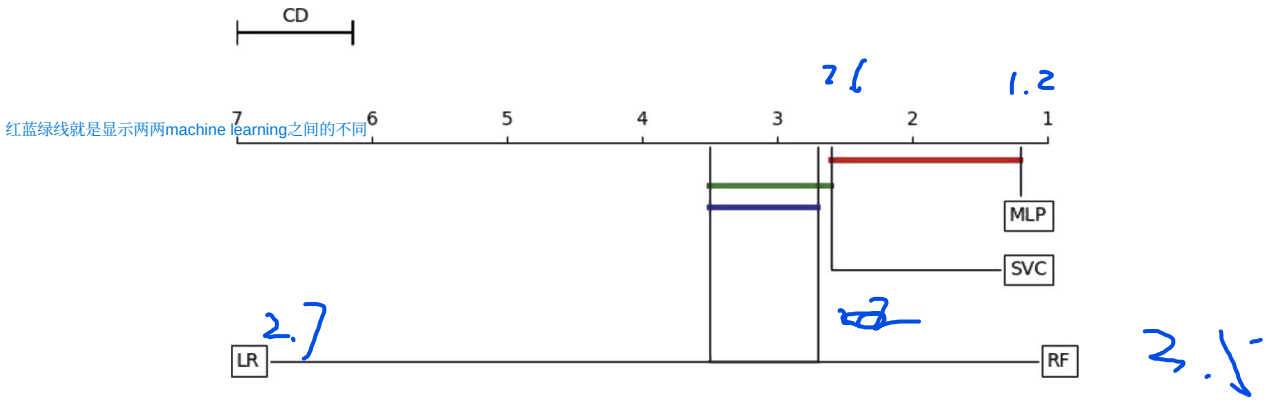

Fig. 5. The expected rank of each algorithm is marked along the axis (a lower number indicating better rank). The MLP model ranks higher in terms of accuracy as compared to SVC, RF, and LR.

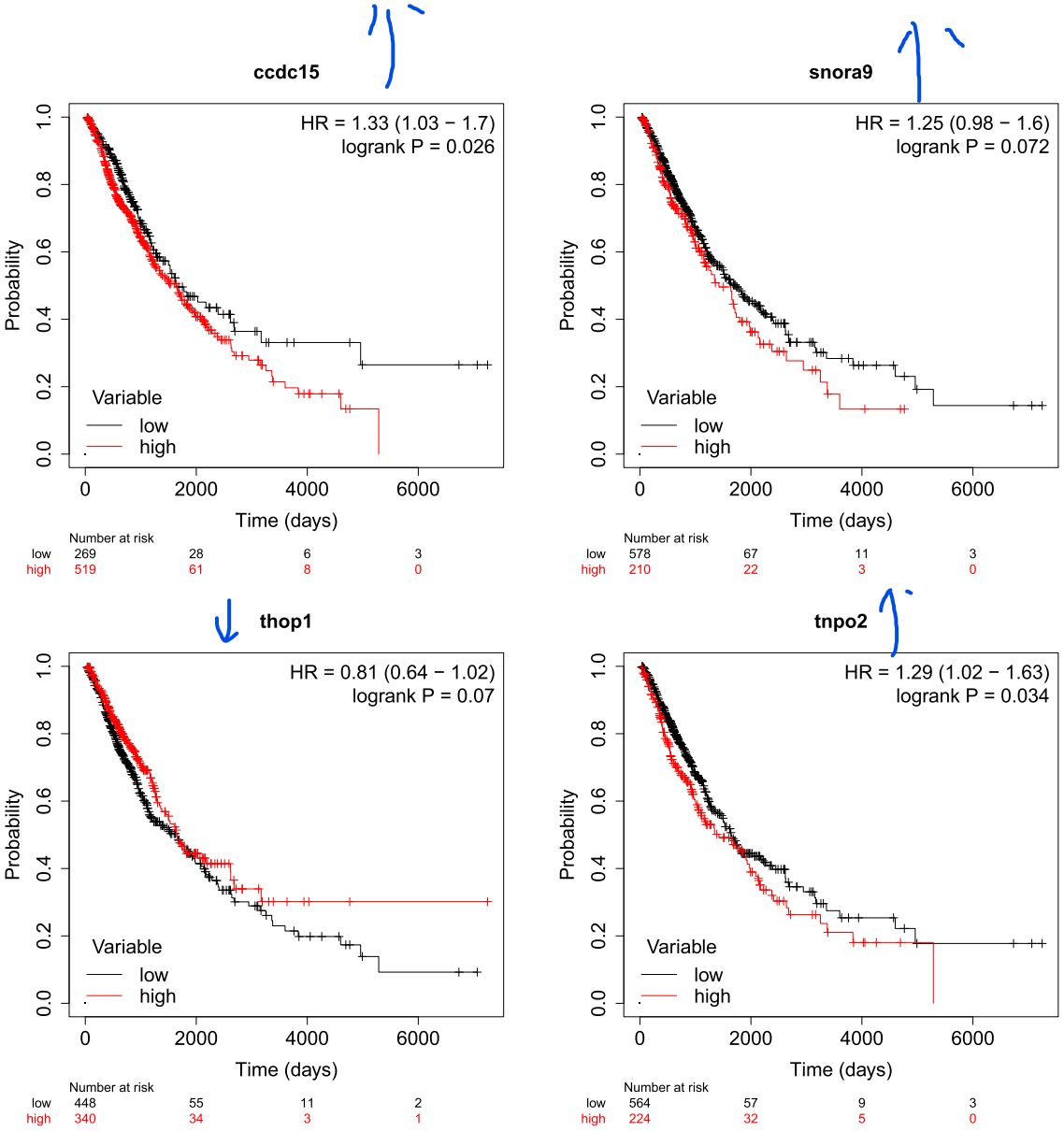

Fig. 6. KM curves w.r.t. the four biomarkers with least p-value. While the x-axis represents the survival period (in the number of days), the y-axis represents the probability of survival. The black curve shows the set of patients with low methylation values, and the orange curve shows the set of patients with high methylation values. The HR value signifies the probability of survival of the first set of patients over the second set of patients. These curves indicate that the low methylation value of the mentioned genes, except $THOP1$, contributes to a higher survival probability.

总共找到相关性疾病有104个 这里罗列的是分数最高的几个

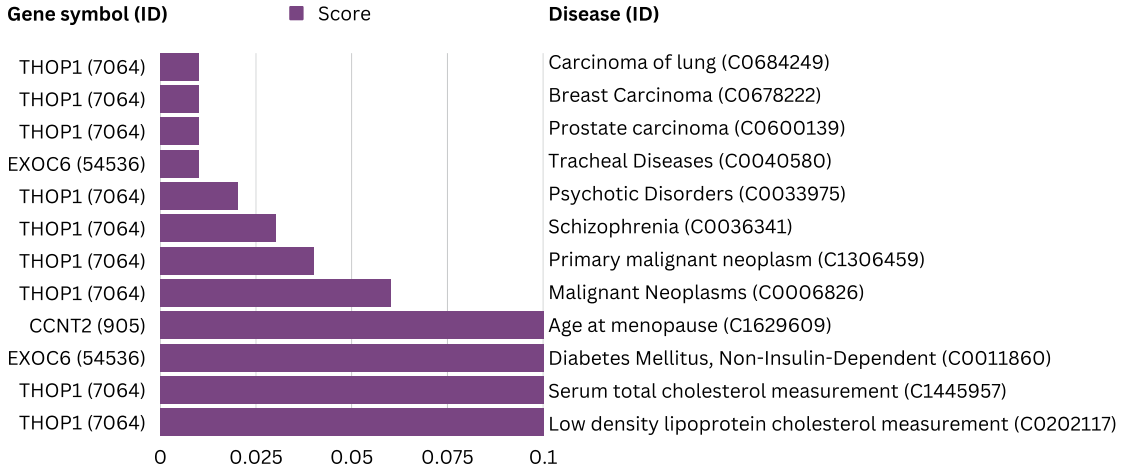

Fig. 7. Gene-disease association graph: plot of twelve GDAs with maximum GDA score. The higher score signifies a better linkage between the gene and the associated disease.

still an active area of research, it provides a better understanding of the linkage between a gene and a disease, thus helping in the development of improved therapeutic strategies ([65]). To explore the GDA of the *MethylBiomarkerSet*, the online database, DisGeNET ([69]) is used that currently contains more than a million of GDAs (source: DisGeNET) and provides a score for each GDA. This GDA score (based on supporting evidence) portrays the strength of the linkage between the disease and the gene, thus helping in the selection of the prioritized GDAs ([69]). Multiple factors play a role in GDA score computation, some of them being the number of sources that report the association, the animal models where the association has been studied, and the number of supporting publications from text-mining-based sources ([69]. The mathematical formulation for the computation of the GDA score is provided in Appendix A.

A total of 104 diseases are found to be associated with three of the discovered biomarkers– $CCNT2$,$THOP1$, and $EXOC6$. A list of diseases with the maximum GDA score is shown in Fig. 7. It is observed that $THOP1$ has associations with schizophrenia, breast and prostate carcinomas, and lung carcinoma. The $EXOC6$ has a higher association with diabetes mellitus (non-insulin-dependent), while $CCNT2$ has a higher association with age at menopause. With this GDA study, it could be affirmed that a few of the discovered biomarkers are highly active in numerous diseases and could be further explored w.r.t. those diseases to develop improved therapy.

5.4. Enriched biological pathways analysis

The pathway enrichment analysis of the ***MethylBiomarkerSet*** is accomplished by utilizing the Reactome, Wikipathways, and KEGG pathways databases. For Reactome and Wikipathways analysis, the online WEB-based GEne SeT AnaLysis Toolkit (WebGestalt) is used ([51]). The Reactome database is an open-source open-access pathway database comprising various pathways such as classical intermediary metabolism, signaling, transcriptional regulation, apoptosis, and disease ([23]). The Wikipathways is an open, collaborative platform dedicated to providing pathways information and is maintained by the biological community. The Kyoto Encyclopedia of Genes and Genomes (KEGG) database provides information about genomic functionalities and association of genes with higher order functional information such as cellular processes ([44]). Table 6, Table 7, and Table 8 provide the enriched pathways being targeted by KEGG, Reactome, and Wikipathways, respectively.

- ***hsa03013***: The nucleocytoplasmic transport is associated with the intracellular localization and regulation of proteins, which is generally found overly-expressed in lung, colon, and breast cancers ([32]).
- ***hsa03250, R-HSA-167287, R-HSA-167290, R-HSA-167152***: HIV-1 or human immunodeficiency virus type-1 is associated with various diseases, such as AIDS-defining carcinoma, Kaposi sarcoma, and cervical carcinoma ([38]). The HIV-1 encodes two gene-regulatory proteins– Tat and Rev among the total 15 proteins that it encodes. The HIV-1 Tat or Trans-Activator of Transcription primarily attempts to elongate the viral transcripts that lead to

Table 6
The KEGG pathways targeted by the discovered biomarkers.

Gene	Pathway Entry	Name
$TNPO2$	hsa03013	Nucleocytoplasmic transport
$CCNT2$	hsa03250	Viral life cycle - HIV-1
	hsa05202	Transcriptional misregulation in cancer
$THOP1$	hsa04614	Renin-angiotensin system
	hsa05143	African trypanosomiasis

Table 7
The list of Reactome pathways triggered by the discovered set of biomarkers.

Gene	Pathway Entry	Name
$EXOC6$	R-HSA-5620916	VxPx cargo-targeting to cilium
	R-HSA-264876	Insulin processing
	R-HSA-5620920	Cargo trafficking to the periciliary membrane
$CCNT2$	R-HSA-2173796	SMAD2/SMAD3:SMAD4 heterotrimer regulates transcription
	R-HSA-167287	HIV elongation arrest and recovery
	R-HSA-167290	Pausing and recovery of HIV elongation
	R-HSA-2173793	Transcriptional activity of SMAD2/SMAD3:SMAD4 heterotrimer
	R-HSA-167152	Formation of HIV elongation complex in the absence of HIV Tat
	R-HSA-112382	Formation of RNA Pol II elongation complex
	R-HSA-75955	RNA Polymerase II Transcription Elongation

Table 8
List of Wikipathways triggered by the discovered set of biomarkers.

Gene	Pathway Entry	Name
$EXOC6$	WP3670	Simplified Interaction Map Between LOXL4 and Oxidative Stress Pathway
	WP4352	Ciliary landscape
$CCNT2$	WP4673	Genes involved in male infertility

hampering the cells' (defected by Tat) capability of replicating effectively in cell culture environment ([59]). In addition, Kirk et al. ([45]) hypothesized that HIV infection (independent of smoking) could lead to lung cancer. In their study, the authors observed that out of 27 lung cancer deaths, 14 patients had HIV infection. However, only one among the 27 patients was identified as a smoker.

- ***hsa05202***: The transcriptional regulation is one of the key processes of $TP53$ tumor suppressor gene– one of the most crucial biomarkers found readily dysregulated in various carcinomas, including NSCLC ([81]).
- ***hsa04614***: The renin-angiotensin system is responsible for sustaining angiogenesis and evading apoptosis– two highly dysregulated functions in lung cancer, catalyzing cancer progression and malignancy transformation ([19]).
- ***hsa05143***: The African trypanosomiasis is a sleeping sickness disorder specifically found in sub-Saharan Africa due to *tsetse* fly's bite that may, in turn, could seriously impact the central nervous system and may even cause death ([56]).
- ***R-HSA-5620916, R-HSA-5620920, WP4352***: The VxPx cargo-targeting to the cilium, the cargo trafficking to the periciliary membrane, and the ciliary landscape pathways and events are centered around *cilium*– a hub of nearly thousands of proteins ([39]). A significant association has been identified between primary cilium and lung cancer via a pathway called *Hedgehog* pathway that is largely responsible for cell growth and differentiation ([31,52]).
- ***R-HSA-264876***: The insulin-like growth factor receptor pathway or the IGF pathway is majorly involved in various processes such as invasiveness, survival, and tumor cell proliferation. Thus, the IGF is considered a potential therapeutic target in NSCLC ([21]).
- ***R-HSA-2173796, R-HSA-2173793***: The *SMADS*, typically $SMAD2/SMAD3$ are receptor-regulated *SMAD* proteins that are significantly involved in the signal transmission of transforming growth factor-β (TGF-β) ([93]). This signaling pathway is included in numerous essential biological processes such as cell growth, angiogenesis, apoptosis, and differentiation. The TGF-β signaling can inhibit or promote tumor growth based on the stage of the tumor. This effect of TGF-β is mediated through the dependent and independent mechanisms of $SMADS$ proteins, thus opening an opportunity for personalized medication for NSCLC ([93,41]).
- ***R-HSA-75955, R-HSA-112382***: The $CKAP2L$ is a promotor of NSCLC through direct binding to RNA polymerase II, thus promoting cell cycle genes, in turn enabling tumor progression ([63]).
- ***WP3670***: Choi et al. ([14]) identified $LOXL4$ knockdown to be directly related to lung cancer metastasis through collagen-dependent extracellular matrix changes in triple-negative breast cancer. In addition, Wang et al. ([89]) found that the upregulation of LOX family genes associated with tumor infiltration could act as a prognostic marker for early-stage LUAD patients and that high protein and mRNA expressions of LOX are observed in NSCLC cells (as compared to normal cells).

5.5. Performance of XAI-guided deep learning framework over multiomics data

To evaluate the robustness and the diversity of the proposed XAI-guided deep learning framework, the transcriptomics (RNA-Seq gene expression) and the genomics (copy number variation) profiles of the NSCLC instances are explored, in addition to the epigenomics profile (DNA methylation). The transcriptomics and the genomics datasets of TCGA-LUAD and LUSC are downloaded from cBioportal ([22,10]). The number of instances and genes for both datasets are shown in Table 9.

Initially, the instances found common in all the omics datasets (transcriptomics, genomics, and epigenomics) are retained (discarding the rest), resulting in a set of 440 LUAD and 370 LUSC instances. Next, the preprocessing of transcriptomics and genomics datasets is carried out in the same manner (wherever applicable) described in Section 3.

Table 9
Dataset details of the RNA-Seq gene expression and CNV datasets.

dataset	#TCGA-LUAD	#TCGA-LUSC	#Genes
RNA-Seq gene expression	510	484	20,531
Copy number variation	511	487	24,777

The transcriptomics and the genomics data are consecutively provided as input to the XAI-guided deep learning framework, uncovering a set of 52 and 39 RNA-Seq and CNV biomarkers, respectively. The classification performance of these discovered RNA-Seq and CNV biomarkers, along with the DNA methylation biomarkers, are individually recorded. Further, the performance on different combinations of these sets of biomarkers ([DNA methylation + CNV], [DNA methylation + RNA-Seq], [RNA-Seq + CNV], and [DNA methylation + CNV + RNA-Seq]) is subsequently recorded. It is to be noted that only the commonly occurring genes between the respective datasets and their different combinations are considered for experimentation, yielding a set of 32, 34, 39, and 63 biomarkers for [DNA methylation + CNV], [DNA methylation + RNA-Seq], [RNA-Seq + CNV], and [DNA methylation + CNV + RNA-Seq]), respectively.

5.5.1. Classification performance of various combinations of omics data

The comparative classification performance analysis of the individually discovered biomarkers, along with the 32 [DNA methylation + CNV], the 34 [DNA methylation + RNA-Seq], the 39 [RNA-Seq + CNV], and the 63 [DNA methylation + CNV + RNA-Seq] biomarkers is provided in Table 10, in addition to the box plot illustrating the classification performance in Fig. 8. It is worth noting that the algorithm utilized to train and validate each of the aforementioned gene sets is the MLP classifier. The columns iter-1 to iter-10 represent the expected 10-fold cross-validation accuracy observed in each iteration of the experiment. The last column shows the overall average classification accuracy achieved by different sets of biomarkers. Indeed, the overall expected accuracy yielded by the set of 63 [DNA methylation + CNV + RNA-Seq] biomarkers relatively outperforms the rest.

5.6. Comparison with the state-of-the-art

Table 11 shows a comparison of the classification performance of the *MethylBiomarkerSet* and the 63 [DNA methylation + CNV + RNA-Seq] biomarkers with the state-of-the-art works.

The performance of the discovered *MethylBiomarkerSet* achieves a significantly higher accuracy of 91.53% as compared to the accuracy of 84.6% and 93.21% achieved by Cai et al. ([8]) and Carillo-Perez et al. ([9]), respectively. Extending the comparison, the discovered set of 63 biomarkers using multiomics data ([DNA methylation + CNV + RNA-Seq]) outperforms Carillo-Perez et al. ([9]) with an accuracy of 96.37%.

6. Discussion

The proposed framework used an autoencoder and a neural network to classify NSCLC instances. Further, using various eXplainable AI (XAI) methods, namely *IntegratedGradients, GradientSHAP*, and *DeepLIFT*, we discovered a set of ***seven*** significant biomarkers. In this section, we discuss the strengths of the proposed framework and the effect of demographic variations on the proposed model's performance. We also discuss the scope of the future work.

6.1. Strength

Concerns have been raised about the trustworthiness of the deep learning models as they remain opaque to the end-user. This issue is now potentially subdued, thanks to the rising applications of the eXplainable AI (xAI) methods that provide a reasonable and human-friendly interpretation of the model's internal processing, bridging the

Table 10
The comparative classification performance analysis of the various sets of discovered biomarkers (individually and when combined). The checkmark indicates which set(s) of biomarkers is/are combined. Clearly, 63 [DNA methylation + CNV + RNA-Seq] biomarkers achieve a remarkable 96.37% accuracy, surpassing the other combinations.

Omics Type			#Genes	Accuracy (using 10-fold cross-validation)										Overall Mean Accuracy (%)
DNA Methyl	RNA-Seq	CNV		iter-1	iter-2	iter-3	iter-4	iter-5	iter-6	iter-7	iter-8	iter-9	iter-10	
✓	–	–	7	0.9277	0.9638	0.8915	0.9156	0.8915	0.9638	0.9036	0.9156	0.8658	0.9146	91.53
–	✓	–	52	0.9468	0.9424	0.9459	0.9486	0.9442	0.9495	0.945	0.9388	0.9415	0.945	94.47
–	–	✓	39	0.8367	0.8357	0.8416	0.8347	0.8367	0.8377	0.8377	0.8327	0.8356	0.8368	83.65
✓	–	✓	32	0.91	0.9411	0.9054	0.882	0.921	0.9878	0.909	0.925	0.8781	0.9236	91.83
✓	✓	–	34	0.9265	0.9553	0.962	0.9343	0.98	0.962	0.9481	0.9195	0.9	0.9779	94.65
–	✓	✓	59	0.998	0.935	0.9632	0.9503	0.8989	0.9621	0.945	0.8988	0.9857	0.9655	95.02
✓	✓	✓	**63**	**0.9698**	**0.9803**	**0.9512**	**0.9563**	**0.9615**	**0.9803**	**0.961**	**0.9445**	**0.9627**	**0.97**	**96.37**

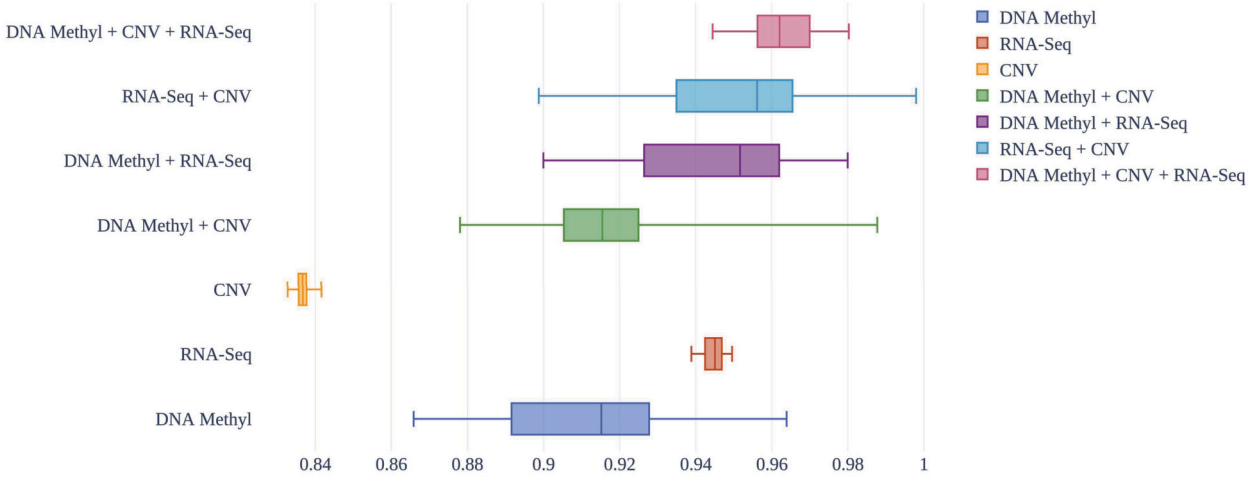

Fig. 8. Boxplot for classification performance analysis of various sets of discovered biomarkers.

Table 11
The classification performance of the proposed work compared with the state-of-the-art works. Indeed the present work outperforms Cai et al. ([8]) in terms of the number of genes as well as accuracy, nevertheless, it achieves competitive results with Carillo-Perez et al. ([9]).

	#Genes	Accuracy (%)
Cai et al. ([8])	16 [DNA methyl]	84.6
Carillo-Perez et al. ([9])	6 [DNA methyl]	93.21
	24 [DNA methyl + RNA-Seq + CNV]	94.11
Proposed Work	**7 [DNA methyl]**	**91.53**
	63 [DNA methyl + RNA-Seq + CNV]	**96.37**

gap between the decision-making criteria and the end-users' comprehension. However, the proposed framework addresses these concerns by deploying the XAI methods to discover biomarkers that would assist oncologists in determining the subtype of an NSCLC patient.

6.2. Potential biases

While the TCGA program covers a broad range of demographics, some researchers have pointed out certain biases in the TCGA datasets. For example, [77] noted that the TCGA datasets comprise a significant number of samples based on the U.S. population, while the Asian and Hispanic populations are underrepresented. Similarly, [90] observed that although TCGA datasets comprise a similar distribution of samples based on sex, the distribution based on race, ethnicity, and diagnostic stage of cancers were dissimilar in proportion. Further, it is noteworthy that even within the U.S. subjects, a large portion of TCGA instances are *white* (as compared to *black* cases). Moreover, the Hispanic and other (not reported) instances are relatively low (please see Table 1).

To evaluate the effect of the aforementioned biases in the present study, we trained our model on a set of 582 instances (343 LUAD and 239 LUSC) of patients from the *MethylDataset* who were not ***Hispanic*** or ***Latino***. Subsequently, on evaluating the performance of the trained model on the set of 246 instances (115 LUAD and 131 LUSC) from the *MethylDataset* whose ethnicity was ***Hispanic*** or ***Latino*** or ***not reported***, we found 10-fold cross-validation accuracy of 82.59% (± 1.65), which is about 10% lower than the accuracy achieved on the entire *MethylDataset*. Thus, we conclude that the demography variations impact the model's performance on the population that was underrepresented in the dataset.

6.3. Future scope

The proposed framework indeed demonstrates its prowess in the discovery of NSCLC-relevant biomarkers. Moreover, the framework is versatile enough to process any omics data for biomarkers discovery. However, the effectiveness of the proposed methodology while dealing with a combination of multiomics data with histopathological data is yet to be evaluated and will be part of future work. It would further assist medical practitioners in devising enhanced therapeutic interventions for NSCLC patients.

7. Conclusion

The work presents a novel XAI-guided deep learning framework capable of discovering a set of significant methylation biomarkers that can aid in NSCLC subtype classification and prognosis. The framework

assists in unveiling *seven* NSCLC-relevant biomarkers. These biomarkers are evaluated for classification performance over multiple machine learning algorithms, with the multilayer perceptron (MLP)-based model outperforming the others. Integrating RNA-Seq, CNV, and methylation data yielded an improved accuracy of 96.37%, outperforming the state-of-the-art works. The machine learning classifiers employed in the experiment are examined for their statistical significance– the MLP-based classifier was ranked the highest among the others.

The clinical efficacy of the discovered set of seven NSCLC-relevant biomarkers is observed by exploring their potential druggability, their likelihood of predicting the survival of NSCLC patients, the other diseases associated with the discovered biomarkers, and the various biological pathways that are being targeted by them. The literature and DGIdb database reveal that *four* out of seven discovered biomarkers are potentially druggable. The Kaplan-Meier curves depict four discovered biomarkers capable of predicting NSCLC patients' survival. More than 100 diseases are found to be associated with three of the biomarkers in the *MethylBiomarkerSet*. Moreover, five, ten, and three major KEGG, Reactome, and Wiki pathways, respectively, are found to be triggered by the discovered biomarkers.

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.

Data availability

The datasets used in the experiment are publicly available in the LinkedOmics and cBioportal portals.

Acknowledgements

Kountay Dwivedi would like to thank University Grants Commission, New Delhi, India, for providing Junior Research Fellowship (Reference ID: 190510173202). The authors would also like to acknowledge Professor Ranjita Pandey, Department of Statistics, Faculty of Mathematical Sciences, University of Delhi, for useful discussions regarding applying statistical techniques. The team would also like to thank Gorseet Paul Singh and Vaibhav Maheshwari at the Department of Computer Science, University of Delhi, India, for assisting in validating the statistical significance of the classifiers.

Appendix A. Calculation of gene-disease association score

The gene-disease score S_{GDA} ranges from 0 to 1 and is computed as:

$$S_{GDA} := W_{uniprot} + W_{ctd} + W_{mouse} + W_{rat} + W_{lit}$$

where [85,15,78,4]:

$$W_{uniprot} = \begin{cases} 0.3 & \text{if association reported in [85]} \\ 0, & \text{otherwise} \end{cases}$$

$$W_{ctd} = \begin{cases} 0.3 & \text{if association reported in [15] human dataset} \\ 0, & \text{otherwise} \end{cases}$$

$$W_{rat} = \begin{cases} 0.1 & \text{if association reported in [78] or CTD rat dataset} \\ 0, & \text{otherwise} \end{cases}$$

$$W_{mouse} = \begin{cases} 0.1 & \text{if association reported in [4] or CTD mouse dataset} \\ 0, & \text{otherwise} \end{cases}$$

$$W_{lit} = \begin{cases} max & \text{if } \frac{n_{gd}\times 100}{N_{lit}} \geq max \\ \frac{n_{gd}\times 100}{N_{lit}} & \text{if } \frac{n_{gd}\times 100}{N_{lit}} < max \quad \text{otherwise} \end{cases}$$

Here, lit can be Genetics Association Database (GAD) ([3]), a literature-derived human gene-disease network (LHGDN) ([61]), and/or BeFree database ([6]); the n_{gd} represents the number of publications that have reported a GDA in the source; and N_{lit} represents the total number of publications in the source ([69]). The value max is computed as:

$$max = \begin{cases} 0.08 & \text{if } lit == \text{GAD} \\ 0.06 & \text{if } lit == \text{LHGDN} \vee \text{BeFree} \end{cases}$$

References

[1] P. Baldi, Autoencoders, unsupervised learning, and deep architectures, in: Proceedings of ICML Workshop on Unsupervised and Transfer Learning, 2012, pp. 37–49.
[2] D. Bank, N. Koenigstein, R. Giryes, Autoencoders, arXiv preprint, arXiv:2003.05991, 2020.
[3] K.G. Becker, K.C. Barnes, T.J. Bright, S.A. Wang, The genetic association database, Nat. Genet. 36 (5) (2004) 431–432.
[4] J.A. Blake, R. Baldarelli, J.A. Kadin, J.E. Richardson, C.L. Smith, C.J. Bult, Mouse genome database (mgd): knowledgebase for mouse–human comparative biology, Nucleic Acids Res. 49 (D1) (2021) D981–D987.
[5] C. Bock, Analysing and interpreting dna methylation data, Nat. Rev. Genet. 13 (10) (2012) 705–719.
[6] À. Bravo, J. Piñero, N. Queralt-Rosinach, M. Rautschka, L.I. Furlong, Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research, BMC Bioinform. 16 (2015) 1–17.
[7] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[8] Z. Cai, D. Xu, Q. Zhang, J. Zhang, S.-M. Ngai, J. Shao, Classification of lung cancer using ensemble-based feature selection and machine learning methods, Mol. BioSyst. 11 (3) (2015) 791–800.
[9] F. Carrillo-Perez, J.C. Morales, D. Castillo-Secilla, O. Gevaert, I. Rojas, L.J. Herrera, Machine-learning-based late fusion on multi-omics and multi-scale data for non-small-cell lung cancer diagnosis, J. Pers. Med. 12 (4) (2022) 601.
[10] E. Cerami, J. Gao, U. Dogrusoz, B.E. Gross, S.O. Sumer, B.A. Aksoy, et al., The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data, Cancer Discov. 2 (5) (2012) 401–404.
[11] 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.
[12] Z. Chen, C.M. Fillmore, P.S. Hammerman, C.F. Kim, K.-K. Wong, Non-small-cell lung cancers: a heterogeneous set of diseases, Nat. Rev. Cancer 14 (8) (2014) 535–546.
[13] C.H.Y. Cheung, H.-F. Juan, Quantitative proteomics in lung cancer, J. Biomed. Sci. 24 (1) (2017) 1–11.
[14] S.K. Choi, H.S. Kim, T. Jin, W.K. Moon, Loxl4 knockdown enhances tumor growth and lung metastasis through collagen-dependent extracellular matrix changes in triple-negative breast cancer, Oncotarget 8 (7) (2017) 11977.
[15] A.P. Davis, T.C. Wiegers, R.J. Johnson, D. Sciaky, J. Wiegers, C.J. Mattingly, Comparative toxicogenomics database (ctd): update 2023, Nucleic Acids Res. 51 (D1) (2023) D1257–D1262.
[16] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.
[17] 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.
[18] P. Du, X. Zhang, C.-C. Huang, N. Jafari, W.A. Kibbe, L. Hou, S.M. Lin, Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis, BMC Bioinform. 11 (2010) 1–9.
[19] K. Dwivedi, A. Rajpal, S. Rajpal, M. Agarwal, V. Kumar, N. Kumar, An explainable ai-driven biomarker discovery framework for non-small cell lung cancer classification, Comput. Biol. Med. 106544 (2023).
[20] G. Erion, J.D. Janizek, P. Sturmfels, S.M. Lundberg, S.-I. Lee, Improving performance of deep learning models with axiomatic attribution priors and expected gradients, Nat. Mach. Intell. 3 (7) (2021) 620–631.
[21] M.J. Fidler, D.D. Shersher, J.A. Borgia, P. Bonomi, Targeting the insulin-like growth factor receptor pathway in lung cancer: problems and pitfalls, Ther. Adv. Med. Oncol. 4 (2) (2012) 51–60.
[22] J. Gao, B.A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S.O. Sumer, et al., Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal, Sci. Signal. 6 (269) (2013) pl1.
[23] M. Gillespie, B. Jassal, R. Stephan, M. Milacic, K. Rothfels, A. Senff-Ribeiro, et al., The reactome pathway knowledgebase 2022, Nucleic Acids Res. 50 (D1) (2022) D687–D692.
[24] L. Girard, J. Rodriguez-Canales, C. Behrens, D.M. Thompson, I.W. Botros, H. Tang, 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.
[25] L. Gong, T. Wen, Z. Li, Y. Wang, J. Wang, X. Che, et al., Tnpo2 operates downstream of dync1i1 and promotes gastric cancer cell proliferation and inhibits apoptosis, Cancer Med. 8 (17) (2019) 7299–7312.
[26] M. Griffith, O.L. Griffith, A.C. Coffman, J.V. Weible, J.F. McMichael, N.C. Spies, et al., Dgidb: mining the druggable genome, Nat. Methods 10 (12) (2013) 1209–1210.

[27] S. Guo, F. Yan, J. Xu, Y. Bao, J. Zhu, X. Wang, et al., Identification and validation of the methylation biomarkers of non-small cell lung cancer (nsclc), Clin. Epigenet. 7 (1) (2015) 1–10.
[28] R. Guttapadu, T. Katte, D. Sayeeram, S. Bhatia, A.R. Abraham, K. Rajeev, et al., Identification of novel biomarkers for lung squamous cell carcinoma, 3 Biotech 13 (2) (2023) 72.
[29] 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.
[30] M.T. Hagan, H.B. Demuth, M. Beale, Neural Network Design, PWS Publishing Co., 1997.
[31] M. Higgins, I. Obaidi, T. McMorrow, Primary cilia and their role in cancer, Oncol. Lett. 17 (3) (2019) 3041–3047.
[32] R. Hill, B. Cautain, N. De Pedro, W. Link, Targeting nucleocytoplasmic transport in cancer therapy, Oncotarget 5 (1) (2014) 11.
[33] F.R. Hirsch, G.V. Scagliotti, J.L. Mulshine, R. Kwon, W.J. Curran, Y.-L. Wu, L. Paz-Ares, Lung cancer: current therapies and new targeted treatments, Lancet 389 (10066) (2017) 299–311.
[34] P.H. Hoang, M.T. Landi, Dna methylation in lung cancer: mechanisms and associations with histological subtypes, molecular alterations, and major epidemiological factors, Cancers 14 (4) (2022) 961.
[35] K. Holm, C. Hegardt, J. Staaf, J. Vallon-Christersson, G. Jönsson, H. Olsson, et al., Molecular subtypes of breast cancer are associated with characteristic dna methylation patterns, Breast Cancer Res. 12 (3) (2010) 1–16.
[36] K. Inamura, Lung cancer: understanding its molecular pathology and the 2015 who classification, Front. Oncol. 7 (2017) 193.
[37] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in: International Conference on Machine Learning, 2015, pp. 448–456.
[38] M. Isaguliants, E. Bayurova, D. Avdoshina, A. Kondrashova, F. Chiodi, J.M. Palefsky, Oncogenic effects of hiv-1 proteins, mechanisms behind, Cancers 13 (2) (2021) 305.
[39] H. Ishikawa, J. Thompson, J.R. Yates, W.F. Marshall, Proteomic analysis of mammalian primary cilia, Curr. Biol. 22 (5) (2012) 414–419.
[40] R. Jaenisch, A. Bird, Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals, Nat. Genet. 33 (3) (2003) 245–254.
[41] H.-S. Jeon, J. Jen, Tgf-^{2} signaling and the role of inhibitory smads in non-small cell lung cancer, J. Thorac. Oncol. 5 (4) (2010) 417–419.
[42] W.E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods, Biostatistics 8 (1) (2007) 118–127.
[43] P.A. Jones, J.D. Buckley, The role of dna methylation in cancer, Adv. Cancer Res. 54 (1990) 1–23.
[44] M. Kanehisa, M. Furumichi, Y. Sato, M. Kawashima, M. Ishiguro-Watanabe, Kegg for taxonomy-based analysis of pathways and genomes, Nucleic Acids Res. 51 (D1) (2023) D587–D592.
[45] G.D. Kirk, C. Merlo, P. O'Driscoll, S.H. Mehta, N. Galai, D. Vlahov, et al., Hiv infection is associated with an increased risk for lung cancer, independent of smoking, Clin. Infect. Dis. 45 (1) (2007) 103–110.
[46] R. Lakshminarasimhan, G. Liang, The role of dna methylation in cancer, in: DNA Methyltransferases-Role and Function, 2016, pp. 151–172.
[47] 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.
[48] S.M. Lauritsen, M. Kristensen, M.V. Olsen, M.S. Larsen, K.M. Lauritsen, M.J. Jørgensen, et al., Explainable artificial intelligence model to predict acute critical illness from electronic health records, Nat. Commun. 11 (1) (2020) 3852.
[49] B.-Q. Li, J. You, T. Huang, Y.-D. Cai, Classification of non-small cell lung cancer based on copy number alterations, PLoS ONE 9 (2) (2014) e88300.
[50] J. Liao, J. Shen, Q. Leng, M. Qin, M. Zhan, F. Jiang, Microrna-based biomarkers for diagnosis of non-small cell lung cancer (nsclc), Thorac. Cancer 11 (3) (2020) 762–768.
[51] 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.
[52] H. Liu, A.A. Kiseleva, E.A. Golemis, Ciliary signalling in cancer, Nat. Rev. Cancer 18 (8) (2018) 511–524.
[53] H.W. Loh, C.P. Ooi, S. Seoni, P.D. Barua, F. Molinari, U.R. Acharya, Application of explainable artificial intelligence for healthcare: a systematic review of the last decade (2011–2022), Comput. Methods Programs Biomed. (2022) 107161.
[54] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint, arXiv:1711.05101, 2017.
[55] X. Lv, Y. Zhao, L. Zhang, S. Zhou, B. Zhang, Q. Zhang, et al., Development of a novel gene signature in patients without helicobacter pylori infection gastric cancer, J. Cell. Biochem. 121 (2) (2020) 1842–1854.
[56] D. Mabille, L. Dirkx, S. Thys, M. Vermeersch, D. Montenye, M. Govaerts, et al., Impact of pulmonary African trypanosomes on the immunology and function of the lung, Nat. Commun. 13 (1) (2022) 7083.
[57] U. Majeed, R. Manochakian, Y. Zhao, Y. Lou, Targeted therapy in advanced non-small cell lung cancer: current advances and future trends, J. Hematol. Oncol. 14 (1) (2021) 1–20.
[58] F.Z. Marino, R. Bianco, M. Accardo, A. Ronchi, I. Cozzolino, F. Morgillo, et al., Molecular heterogeneity in lung cancer: from mechanisms of origin to clinical implications, Int. J. Med. Sci. 16 (7) (2019) 981.
[59] Z. Mbita, R. Hull, Z. Dlamini, Human immunodeficiency virus-1 (hiv-1)-mediated apoptosis: new therapeutic targets, Viruses 6 (8) (2014) 3181–3227.
[60] O. Melaiu, E. Melissari, L. Mutti, E. Bracci, C. De Santi, C. Iofrida, et al., Expression status of candidate genes in mesothelioma tissues and cell lines, Mutat. Res. 771 (2015) 6–12.
[61] J.A. Mitchell, A.R. Aronson, J.G. Mork, L.C. Folk, S.M. Humphrey, J.M. Ward, Gene indexing: characterization and analysis of nlm's generifs, in: Amia Annual Symposium Proceedings, vol. 2003, 2003, p. 460.
[62] B. Mohanta, P. Das, S. Patnaik, Healthcare 5.0: a paradigm shift in digital healthcare system using artificial intelligence, iot and 5g communication, in: 2019 International Conference on Applied Machine Learning (ICAML), 2019, pp. 191–196.
[63] T. Monteverde, S. Sahoo, M. La Montagna, P. Magee, L. Shi, D. Lee, et al., Ckap2l promotes non–small cell lung cancer progression through regulation of transcription elongation, Cancer Res. 81 (7) (2021) 1719–1731.
[64] V. Nair, G.E. Hinton, Rectified linear units improve restricted Boltzmann machines, in: Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010, pp. 807–814.
[65] K. Opap, N. Mulder, Recent advances in predicting gene–disease associations, F1000Res. 6 (2017).
[66] U. Pawar, D. O'Shea, S. Rea, R. O'Reilly, Incorporating explainable artificial intelligence (xai) to aid the understanding of machine learning in the healthcare domain, in: Aics, 2020, pp. 169–180.
[67] M.A. Peinado, Hypomethylation of dna, in: M. Schwab (Ed.), Encyclopedia of Cancer, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 1791–1792.
[68] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.
[69] J. Piñero, N. Queralt-Rosinach, A. Bravo, J. Deu-Pons, A. Bauer-Mehren, M. Baron, et al., Disgenet: a discovery platform for the dynamical exploration of human diseases and their genes, Database 2015 (2015) bav028.
[70] R. Pirker, Conquering lung cancer: current status and prospects for the future, Pulmonology 26 (5) (2020) 283–290.
[71] Z.-W. Qiu, J.-H. Bi, A.F. Gazdar, K. Song, Genome-wide copy number variation pattern analysis and a classification signature for non-small cell lung cancer, Genes Chromosomes Cancer 56 (7) (2017) 559–569.
[72] James L. McClelland, David E. Rumelhart, Geoffrey E. Hinton, The appeal of parallel distributed processing, vol. 3, MIT Press, Cambridge MA, 1986, p. 44.
[73] J. Sandoval, J. Méndez González, E. Nadal, G. Chen, F.J. Carmona, S. Sayols, et al., A prognostic dna methylation signature for stage i non-small-cell lung cancer, J. Clin. Oncol. 31 (2013) 4140–4147.
[74] D. Saraswat, P. Bhattacharya, A. Verma, V.K. Prasad, S. Tanwar, G. Sharma, et al., Explainable ai for healthcare 5.0: opportunities and challenges, IEEE Access (2022).
[75] B.K. Scurfield, Multiple-event forced-choice tasks in the theory of signal detectability, J. Math. Psychol. 40 (3) (1996) 253–269.
[76] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: International Conference on Machine Learning, 2017, pp. 3145–3153.
[77] D.E. Spratt, T. Chan, L. Waldron, C. Speers, F.Y. Feng, O.O. Ogunwobi, J.R. Osborne, Racial/ethnic disparities in genomic sequencing, JAMA Oncol. 2 (8) (2016) 1070–1074.
[78] R.G. Steen, A.E. Kwitek-Black, C. Glenn, J. Gullings-Handley, W. Van Etten, O.S. Atkinson, et al., A high-density integrated genetic linkage and radiation hybrid map of the laboratory rat, Genome Res. 9 (6) (1999) AP1–AP8.
[79] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International Conference on Machine Learning, 2017, pp. 3319–3328.
[80] 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.
[81] K. Suzuki, H. Matsubara, et al., Recent advances in p53 research and cancer treatment, BioMed Res. Int. 2011 (2011).
[82] Q. Teng, Z. Liu, Y. Song, K. Han, Y. Lu, A survey on the interpretability of deep learning in medical diagnosis, Multimed. Syst. (2022) 1–21.
[83] W.D. Travis, E. Brambilla, A.G. Nicholson, Y. Yatabe, J.H. Austin, M.B. Beasley, et al., The 2015 world health organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification, J. Thorac. Oncol. 10 (9) (2015) 1243–1260.
[84] J.A. Tsou, J.A. Hagen, C.L. Carpenter, I.A. Laird-Offringa, Dna methylation analysis: a powerful new tool for lung cancer diagnosis, Oncogene 21 (35) (2002) 5450–5461.
[85] Uniprot: the universal protein knowledgebase in 2023, Nucleic Acids Res. 51 (D1) (2023) D523–D531.
[86] S.V. Vasaikar, P. Straub, J. Wang, B. Zhang, Linkedomics: analyzing multi-omics data within and across 32 cancer types, Nucleic Acids Res. 46 (D1) (2018) D956–D963.
[87] A. Vellido, The importance of interpretability and visualization in machine learning for applications in medicine and health care, Neural Comput. Appl. 32 (24) (2020) 18069–18083.
[88] C. von Roretz, A.M. Macri, I.-E. Gallouzi, Transportin 2 regulates apoptosis through the rna-binding protein hur, J. Biol. Chem. 286 (29) (2011) 25983–25991.

[89] W. Wang, X. Wang, F. Yao, C. Huang, Lysyl oxidase family proteins: prospective therapeutic targets in cancer, Int. J. Mol. Sci. 23 (20) (2022) 12270.
[90] X. Wang, J.T. Steensma, M.H. Bailey, Q. Feng, H. Padda, K.J. Johnson, Characteristics of the cancer genome atlas cases relative to us general population cancer cases, Br. J. Cancer 119 (7) (2018) 885–892.
[91] W. Xu, C. Liu, B. Deng, P. Lin, Z. Sun, A. Liu, et al., Tp53-inducible putative long non-coding rnas encode functional polypeptides that suppress cell proliferation, Genome Res. (2022), gr–275831.
[92] S.-R. Yang, A.M. Schultheis, H. Yu, D. Mandelker, M. Ladanyi, R. Büttner, Precision medicine in non-small cell lung cancer: current applications and future directions, in: Seminars in Cancer Biology, vol. 84, 2022, pp. 184–198.
[93] T. Yokoyama, T. Kuga, Y. Itoh, S. Otake, C. Omata, M. Saitoh, K. Miyazawa, Smad2δexon3 and smad3 have distinct properties in signal transmission leading to tgf-β–induced cell motility, J. Biol. Chem. 299 (2) (2023).
[94] M. Yuan, L.-L. Huang, J.-H. Chen, J. Wu, Q. Xu, The emerging treatment landscape of targeted therapy in non-small-cell lung cancer, Signal Transduct. Targeted Ther. 4 (1) (2019) 61.
[95] C. Zappa, S.A. Mousa, Non-small cell lung cancer: current treatment and future advances, Transl. Lung Cancer Res. 5 (3) (2016) 288.