TYPE Systematic Review
PUBLISHED 15 February 2023
DOI 10.3389/fbioe.2023.1081950

Artificial intelligence-based non-small cell lung cancer transcriptome RNA-sequence analysis technology selection guide

Min Soo Joo[1], Kyoung-Ho Pyo[2,3,4,5], Jong-Moon Chung[1,6]* and Byoung Chul Cho[5]*

[1]School of Electrical and Electronic Engineering, College of Engineering, Yonsei University, Seoul, Republic of Korea, [2]Department of Oncology, Severance Hospital, College of Medicine, Yonsei University, Seoul, Republic of Korea, [3]Severance Biomedical Science Institute, Yonsei University College of Medicine, Seoul, Republic of Korea, [4]Yonsei New Il Han Institute for Integrative Lung Cancer Research, Yonsei University College of Medicine, Seoul, Republic of Korea, [5]Division of Medical Oncology, Department of Internal Medicine and Yonsei Cancer Center, Severance Hospital, Yonsei University College of Medicine, Seoul, Republic of Korea, [6]Department of Emergency Medicine, College of Medicine, Yonsei University, Seoul, Republic of Korea

OPEN ACCESS

EDITED BY
Mengxing Li,
University of Nebraska-Lincoln,
United States

REVIEWED BY
Jue Zhou,
Experian, United Kingdom
Andrea Qiu,
Merck & Co., Inc., United States
Qiyue Wang,
Amazon, United States
Wei Huang,
Northwestern University, United States

*CORRESPONDENCE
Jong-Moon Chung,
jmc@yonsei.ac.kr
Byoung Chul Cho,
CBC1971@yuhs.ac

SPECIALTY SECTION
This article was submitted to Bioprocess Engineering,
a section of the journal
Frontiers in Bioengineering and Biotechnology

RECEIVED 27 October 2022
ACCEPTED 24 January 2023
PUBLISHED 15 February 2023

CITATION
Joo MS, Pyo K-H, Chung J-M and Cho BC (2023), Artificial intelligence-based nonsmall cell lung cancer transcriptome RNA-sequence analysis technology selection guide.
Front. Bioeng. Biotechnol. 11:1081950.
doi: 10.3389/fbioe.2023.1081950

The incidence and mortality rates of lung cancer are high worldwide, where non-small cell lung cancer (NSCLC) accounts for more than 85% of lung cancer cases. Recent non-small cell lung cancer research has been focused on analyzing patient prognosis after surgery and identifying mechanisms in connection with clinical cohort and ribonucleic acid (RNA) sequencing data, including single-cell ribonucleic acid (scRNA) sequencing data. This paper investigates statistical techniques and artificial intelligence (AI) based non-small cell lung cancer transcriptome data analysis methods divided into target and analysis technology groups. The methodologies of transcriptome data were schematically categorized so researchers can easily match analysis methods according to their goals. The most widely known and frequently utilized transcriptome analysis goal is to find essential biomarkers and classify carcinomas and cluster NSCLC subtypes. Transcriptome analysis methods are divided into three major categories: Statistical analysis, machine learning, and deep learning. Specific models and ensemble techniques typically used in NSCLC analysis are summarized in this paper, with the intent to lay a foundation for advanced research by converging and linking the various analysis methods available.

KEYWORDS

non-small cell lung cancer, transcriptome, RNA, sequence, statistical analysis, machine learning, deep learning

1 Introduction

According to data published in the Cancer Journal for Clinicians (Sung et al., 2021), the number of cancer patients worldwide has increased from 10 million in 2000 to 19.3 million in 2020. During this 20 years period, the number of incidences of lung cancer was the undisputed number one type of cancer. The number of lung cancer patients in 2021 ranked second with 22 million, or 11.4% of all cancer patients. However, as lung cancer ranks first in mortality, it is a disease that has a great impact on human society. Accordingly, various methods of research are being conducted worldwide to elucidate growth mechanisms of lung cancer and develop effective therapeutic agents. Research on non-small cell lung cancer (NSCLC) is predominant among lung cancer types, as it accounts for 80 ~85% of all lung cancer incidents.

In this paper, various artificial intelligence (AI) based transcriptome analysis methods and predictive models are investigated to provide future guidelines for more effective NSCLC treatment development. NSCLC can be subdivided into squamous cell carcinoma, adenocarcinoma, and large cell carcinoma, among which squamous cell carcinoma and adenocarcinoma are most common. Mutations are another important element in the study of NSCLC and being able to accurately classify lung cancer types is essential in selecting treatment options and identifying oncologic mechanisms. This is because the targeted therapy or chemotherapy applied to lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) patients are different (Coudray et al., 2018). In addition to the analysis method of classifying lung cancer subtypes, detailed classification of lung cancer mutations can also elucidate mechanisms and help derive the most effective biomarkers (Zhao et al., 2015; Cohen et al., 2020).

These studies are based on next-generation sequencing (NGS) analysis (De Luca et al., 2021). As research to elucidate these molecular biological mechanisms has been actively conducted, various diagnostic methods and new targeted therapies have become possible. With the development of NGS technology, various genetic mutations have been reported in patients with NSCLC. However, in actual clinical practice, the test methods to determine the treatment policy for NSCLC patients are limited to epidermal growth factor receptor (EGFR) gene mutations and anaplastic lymphoma kinase (ALK) fusion genes (Soda et al., 2007; Hida et al., 2017). Targeted therapy has a high need for whole exome sequencing (WES) and whole genome sequencing (WGS), which can explain the mechanism of resistance for targeted therapies (Kruglyak et al., 2016). However, in the case of immune-therapeutic agents that have been developed, analysis of the expression and pattern of immune-related genes has become more important than mutations secured through genome-wide analysis, so research on translation analysis has emerged.

In recent years, research that enables direct clinical application by applying a cancer prognosis (Han et al., 2019; Volckmar et al., 2019), metastasis (Qi et al., 2017; Kamer et al., 2020; Kim et al., 2020; Tao et al., 2020), and/or treatment response (Jiang et al., 2018) prediction model has been in the spotlight. A representative example is a research method in which a specific overexpressed RNA is discovered and selected to be used as a targeted therapeutic agent in reference to the nature of cancer, in which gene mutation is the main etiology. Although the details will be described later, most studies aim at discovering biomarkers to determine the overall survival (OS) as the output (Yuan et al., 2017; Givechian et al., 2019; Xiong et al., 2020). Although there are limits to accurate analysis of tens of thousands of features per patient using various techniques, this is the most widely used method in modern RNA-seq analysis research.

To help the RNA-seqeuncing (RNA-seq) analysis process, this paper focuses on providing a guideline on various AI and mathematical and statistical methods that can be used in extracting effective RNA information and discovering biomarkers for NSCLC patients. Various RNA-seq analysis methods, such as gene expression, gene fusion, and mutation, have been applied in the past. In addition, gene set enrichment analysis (GSEA) and pathway analysis have been used in various medical fields to extract gene expressions. The accuracy of the prognosis or symptom to be predicted varies depending on the AI model used, and many approaches already exist, which are difficult to distinguish the pros and cons. Therefore, this paper investigates various RNA-seq analysis methods using AI technologies in hope to help future clinical research. For example, a binary prediction model that classifies and clusters a patient's tumor mutation from raw data of the RNA-seq can be constructed, or a regression model that predicts whether a patient will recur or not can be designed.

In this paper, RNA-seq analysis using AI technologies are divided into three categories, which are statistical analysis, machine learning, and deep learning. Statistical analysis is effective in finding out which gene groups have significant values. Statistical techniques help to visualize the analyzed result as a volcano plot or heatmap to identify the tendency. Various test techniques are included by applying existing statistical analysis methods used in other fields (Sun et al., 2018; Passiglia et al., 2019). The regression method is represented by the Cox Proportional Hazard model (Cox-ph model) approach (Sun et al., 2022). Survival analysis based on the Cox-ph model are predominant. Tools such as TIDE is a representative example (Jiang et al., 2018). The cox-ph model is a method extensively and generously used in NSCLC RNA-seq analysis and is classified as a statistical analysis category because it is standardized with a specific formula. Machine learning techniques are mainly used for classification and prediction analysis. For example, subtypes of NSCLC can be distinguished using SVM, or tumor mutations can be clustered using logistic regression. Among advanced machine learning techniques, representative schemes include supervised learning and AutoEncoder with Cox regression network (AECOX) (Huang et al., 2020), which discovers specific expression genes by combining models (e.g., SVM and universal classification tools) to construct an ensembled-model, and methods combining two or more machine learning tools are under development. Deep learning-based analysis is mainly used as a method of substituting a neural network into a pipeline. For example, Cox-nnet (Travers et al., 2018) inserts a neural network in front of the Cox-ph model, and DeepSurv (Katzman et al., 2018) uses a neural network in the survival analysis. DCNet (Wang et al., 2022) is a good model to classify subtypes of lung cancer.

2 RNA-sequencing analysis identified by target

2.1 Preprocessing data for analysis

To properly use RNA-seq data, preprocessing is indispensable, as the range of values varies according to the type of RNA-seq and because gene expressions are different depending on the patient and patient group. The order of the preprocessing RNA-seq data method is slightly different depending on the nature of the RNA-seq data. First, in the case of bulk RNA-seq data, differentially expressed genes are derived through trimming, counting, and normalization. An alignment process may be added after trimming, which is an additional way to align data. Second, in the case of single-cell RNA-seq data, Quality Control and Normalization are performed, and the results are grouped through Clustering. After this, differential expressed genes were obtained in each group. For more details on the process, please refer to Figure 2B and Figure 6. Gene selection plays a role in reducing the absolute amount of RNA-seq data to be analyzed by extracting a specifically expressed gene mainly through machine learning. Gene selection creates an environment in which bio-marker discovery can be made easier by reducing the analysis time and deriving more effective factors. In particular, genes are selected through a classifier, such as a decision tree or support vector

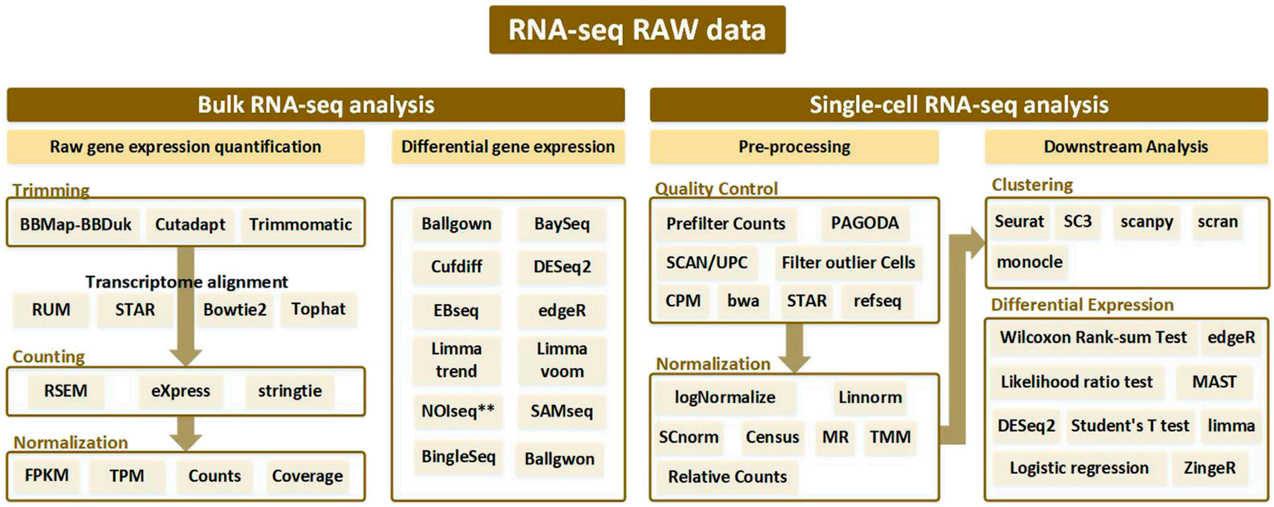

FIGURE 1
Analysis pipeline according to bulk RNA-seq and single-cell RNA-seq from RNA-seq raw data. Bulk RNA-seq analysis is divided into two pipelines: raw gene expression quantification (which leads to trimming, counting, and normalization) and differential gene expression analysis. Alignment is performed after trimming to remove contamination from the sequenced data, and various tools used at this time are generally bowtie-based. Afterwards, the counting process is performed to remove the biased sequencing data, and after normalization, the processed data goes through the final step of discriminating the expressed gene through differential gene expression analysis. In single-cell RNA-seq analysis, processes such as trimming and mapping are batch-processed in the quality control stage, and normalization is performed. After that, a clustering procedure to differentiate between similar cell types is essential. Then the single-cell RNA-seq analysis workflow is completed with differential expression analysis.

machine (SVM). As shown in Figure 1, by using the normalization technique, which is often used in statistics, it is possible to unify the data units and reduce the influence of the size factor. In addition, using the ranking expression technique, it is possible to determine the criteria for an appropriate data set by obtaining a differential expression. Through this, the number of genes can be adjusted according to the analysis goal and used for further detailed analysis. Survival analysis is commonly conducted using traditional statistical analysis methods rather than AI techniques. Prognostic analysis uses various learning methods from machine learning and deep learning, where recently, tumor microenvironment analysis applying convolutional neural network (CNN) technology using image data is being actively conducted.

2.2 Categorization of RNA-sequencing analysis techniques

Research using RNA-seq can be subdivided according to various purposes. Depending on the target, it is broadly classified into classification and prediction in a wide range, but it can be subdivided into biomarker, detection, survival analysis, *etc.* In the predictive biomarker category, studies were also conducted to identify immune checkpoint inhibitors (ICB) (Wiesweg et al., 2019) or to identify the mechanisms of biomarkers that affect responsiveness to immunotherapy (Jiang et al., 2018). Beyond simple elucidation of biomarkers, technologies can be further subdivided based on research that analyzes the prognosis of responsiveness to immunotherapy (Auslander et al., 2018; Jiang et al., 2018; Kapil et al., 2018; Althammer et al., 2019; Nikolas Kather et al., 2019; Fu et al., 2020; He et al., 2020). There are also analysis methods that predict metastasis (Qi et al., 2017; Kamer et al., 2020; Kim et al., 2020; Tao et al., 2020) or identifies indicators related to recurrence after cure or treatment (Lu et al., 2012; Galvez et al., 2020). Analysis of mutation patterns in lung cancer (Zhao et al., 2015; Cohen et al., 2020) and prognostic prediction (Han et al., 2019; Volckmar et al., 2019), which predicts the prognosis of lung cancer patients differently from the previous prediction category, are analyzed as targets, and survival analysis (Yuan et al., 2017; Givechian et al., 2019; Xiong et al., 2020) can be classified as a prognostic biomarker category. Subtypes of lung cancer can be classified by applying ensemble machine learning tools with multi-class classification capability. The process starts with an analysis (Huang et al., 2017; Hsu and Dong, 2018) that classifies malignant and benign based binary classification using various machine learning techniques (Cai et al., 2015; Tian, 2017; Su et al., 2020; Huang et al., 2021; Wang et al., 2022).

In connection with cancer classification analysis, AI techniques are used to classify the reactive prognosis according to the type of surgery or the patient group that distinguishes between malignant and benign status. In this case, since there are two final analysis targets, machine learning tools such as SVM or decision tree are suitable to use as they are very effective in binary classification (Han et al., 2014; Peng et al., 2015; Huang et al., 2017; Hsu and Dong, 2018; Reynders et al., 2018). In addition, analysis based on gender, age, overall survival (OS), *etc.*, is also needed. In general, classifying cancer subtypes of NSCLC is treated as a multi-group classification problem because there are more than three groups to be distinguished (Cai et al., 2015; Tian, 2017; Su et al., 2020).

Prediction techniques are also very important as they help estimate future prognosis based on surgery or treatment method (Zhou et al., 2016; Ahmed et al., 2018; Han et al., 2019; Volckmar et al., 2019; He et al., 2020). In addition, progression-free survival

TABLE 1 Analysis methodology identified by target. Three categories of analysis methodology are identified based on by the target: analysis methods that identifies a predictive biomarker according to the target of the analysis, identify biomarkers related to the prognosis of the patient, and classify the type of cancer.

Category	Target	Author and year of publication
Predictive Biomarker	ICB	Wiesweg et al. (2019)
	Response to Immunotherapy	Jiang et al. (2018)
	Metastasis	Qi et al. (2017)
		Kim et al. (2020)
		Kamer et al. (2020)
		Tao et al. (2020)
	Recurrence	Lu et al. (2012)
		Galvez et al. (2020)
Prognostic Biomarker	Mutation	Cohen et al. (2020)
		Zhao et al. (2015)
	Prognostic prediction	Han et al. (2019)
		Volckmar et al. (2019)
	Survival Analysis	Yuan et al. (2017)
		Givechian et al. (2019)
		Xiong et al. (2020)
Cancer Classification	Malignant or Benign	Hsu and Dong (2018)
		Huang et al. (2017)
	Cancer subtype	Tian (2017)
		Su et al. (2020)
		Cai et al. (2015)
		Huang et al. (2021)
		Wang et al. (2022)

(PFS) analysis has been used since the 2010s to predict the recurrence probability (Lu et al., 2012; Galvez et al., 2020). It is also possible to rank treated patients by extracting expressed genes and analyzing RNA based on how highly a specific treatment was responsive (Han et al., 2019). Studies that have focused on survival analysis use various criteria ranging from simply distinguishing between dead and alive (Jefferson et al., 1997) to predicting the survival rate (Yuan et al., 2017; Givechian et al., 2019; Xiong et al., 2020). Among these studies, RNA-seq data has been used in survival analysis in many ways, which are based on the various targets of classification. Analysis using only numerical values of the RNA from a statistical point of view has been performed by various researchers (Li et al., 2014; Conesa et al., 2016; Afonso et al., 2019; Sharma et al., 2019). In (Byron et al., 2016), the authors only use RNA-seq data in making survival predictions, where the data was transformed to fit the format of clinical data.

2.3 Process of RNA-sequencing data analysis

After setting the analysis target, the overall process flow is explained in the following. First, it is critical to pre-process the data in accordance with the analysis module and decide if clinical data will be combined. If the goal is to analyze a patient's prognosis, adding clinical data increases the accuracy. On the other hand, a simple classification process that distinguishes between malignant and benign NSCLC patient predictions can be conducted without clinical data. Therefore, it is important to add or subtract data according to the analysis target. Once the data is ready, it should go through a normalization process. This process facilitates gene expression analysis because each RNA has a different scale. For example, the RNA normalization process includes fragments per kilobase of transcript per million mapped reads (FPKM), reads per kilobase per millions mapped reads (RPKM), and transcripts per million (TPM). However, if the extracted RNA-seq data does not show a significant difference in scale, the raw counts of RNA data itself can be directly used.

2.3.1 Analysis pipeline of bulk RNA-sequencing

This procedure is a preparation step for analysis using FASTQ from raw RNA-seq data. FASTQ is a text-based format to store all quality (Phred) scores for each nucleotide expressed in ASCII code in the adenine, guanine, cytosine, thymine (AGCT) biological sequence. It can be divided into a technique to quantify and analyze gene expression from RNA-seq raw data and a technique to analyze differential gene expressions. First, considering the gene expression quantification technique, it mainly proceeds to 1) Trimming, 2)

TABLE 2 Analysis methodology identified by AI technique. The statistical part includes analysis using simple mathematical models including test techniques and regression represented by the Cox-ph model. AI techniques using big data can be divided into machine learning and deep learning. In the machine learning part, SVM is mainly used, and the ensembled model with Decision Tree added is also increasing in frequency recently. In the Deep Learning part, research is actively underway by adding neural networks to various models.

Category	Model & algorithm	Author and year of publication
Statistical Analysis	Copy Number Variant (CNV)	Zhao et al. (2015)
	Kaplan-Meier, Log-rank Test	Sun et al. (2018)
	Gene-Expression	Lu et al. (2012)
	Cox-ph model	Sun et al. (2022)
	TIDE	Jiang et al. (2018)
Machine Learning	Ensembled-model	Wiesweg et al. (2019)
	SVM	Zhou et al. (2016)
		Cai et al. (2015)
		Su et al. (2020)
	LR	Tian (2017)
		Cai et al. (2015)
	WGRFE	Su et al. (2020)
	AECOX	Huang et al. (2020)
Deep Learning	Neural Network	Faraggi and Simon (1995)
		Travers et al. (2018)
	DeepSurv	Katzman et al. (2018)
		Jiang et al. (2020)
		Huang et al. (2021)
		Wang et al. (2022)

Counting, and 3) Normalization. Trimming is the procedure of removing contaminated or low-quality data from raw data. It is an optional step, but the quality of the data increases when it is executed, and the final analysis accuracy increases. Differentially expressed genes (DEG) analysis is a process of extracting meaningful RNAs. For this purpose, various test techniques can be used based on the *p*-value and the fold change (FC) value. DEG analysis can be performed with the Kaplan-Meier test using the log-rank test, or DEG analysis can be performed through gene pathway analysis. As shown in Figure 2A, the input data to be applied to the AI module needs to be prepared through this process. The next step is to build a learning module to apply the input data based on the analysis target, which can be assisted by the methods presented in Table 1 and Table 2. In AI analysis, the ratio of dividing the data for training, testing, and validation are important. In general, the analysis performed is good when the data is divided in a balanced form among the training and test sets, and it is useful to use validation techniques such as cross-validation. However, in addition to various validation methods and test set split methods, genetic algorithm-based approaches also exist, which are effective when the data is unbalanced. Genetic algorithms help in selecting several solutions in advance and obtains the most suitable solution over generations. If a genetic algorithm is applied to spilt the data, the genetic algorithm can be used to appropriately divide the training data set and the test data set so artificial intelligence technologies can be effectively applied. When there is abundant patient data, this is less of a problem, but in most cases, patient data is limited, and the ratio of dividing the training, validation, and test set is dependent on the data features and AI analysis algorithm.

After completing the previous three steps of Figure 2B, in the final performance evaluation process, the performance of the area under the curve (AUC) and receiver operating characteristic (ROC) curves need to be analyzed. Or the ranking of RNA expressions through the Sharpe value needs to be analyzed. These processes are summarized in Figure 2 and Table 1. Through the analysis that actively utilizes Table 2, the entire process of RNA-seq data analysis of NSCLC patients can be confirmed.

2.3.2 Analysis pipeline of single-cell RNA-sequencing

Until now, the transcriptome of a cell population has been studied, but the precise results of the study are limited because the expression patterns of each cell are different. Therefore, there is a limit to the analysis to understand the cancer microenvironment in which various types of cells exist only by bulk RNA-seq. A method for deconvolution of the cell type ratio was developed from the bulk RNA-seq results (Sun et al., 2022). But since this method requires significant reference data, this increases the complexity even more, and therefore may need to be verified in another way, which may be difficult in reality. Accordingly, there was a demand to study cell interactions and tissue functions through cell-level analysis, and as a part of this, research on single-cell RNA-seq emerged. Single-cell RNA-seq analysis can be largely divided into a pre-processing part and a

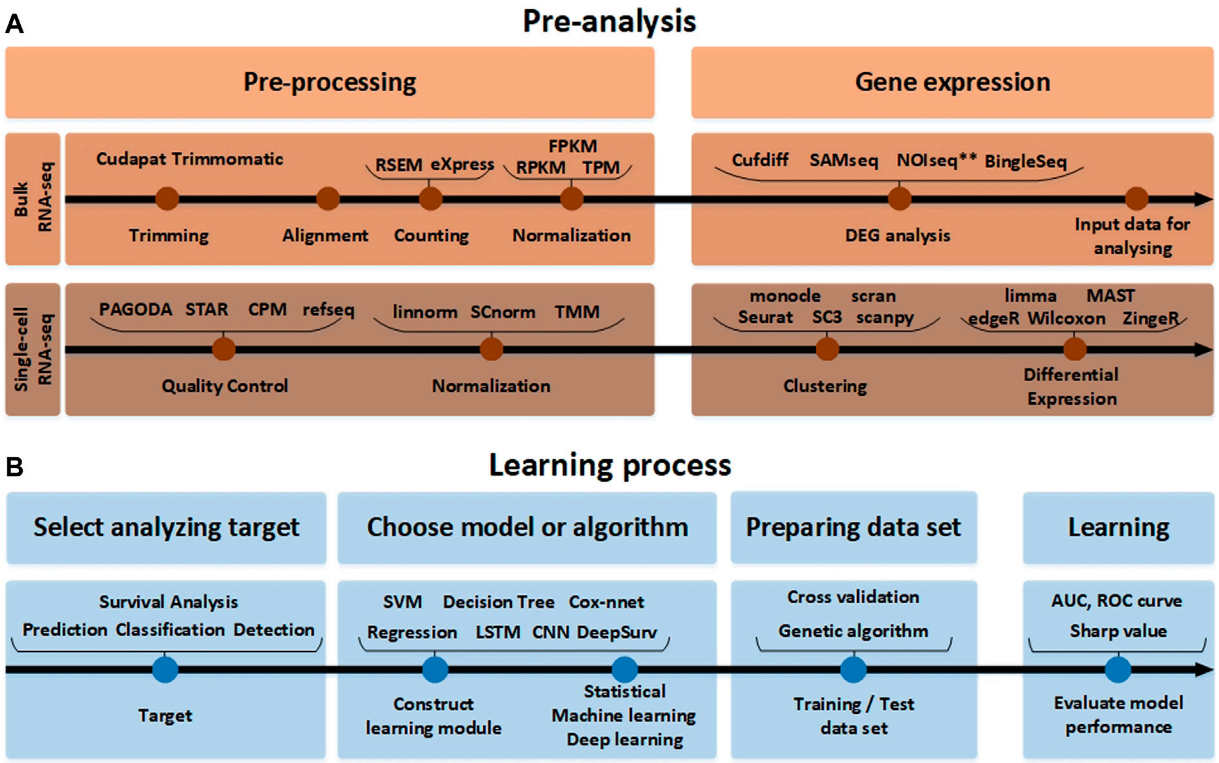

FIGURE 2
The RNA-seq data analysis process. **(A)** Before the analysis of RNA-seq data. The pre-analysis part includes the conversion process to form the input data into an appropriate format that can help in analyzing the RNA-seq data. Since the detailed tools used for processing and gene expression analysis are different in single-cell RNA-seq and bulk RNA-seq analysis methods, they are indicated separately. **(B)** The learning process part after the analysis of RNA-seq data. In accordance with the target to analyze the input data obtained in this way. It it beneficial to pre-process the data using various normalization tools again and learn through the AI model.

downstream analysis part. In the pre-processing part, quality control is performed first. Quality control in raw read is tested through various methods to confirm sequencing fish, PCR artifacts, and contamination. This is an essential step to remove outliers. In simple terms, it can be said that it is the task of selecting data for easy analysis. Data that has undergone quality control goes through normalization, and then the pre-processing (for analysis) is finished. Then by clustering similar references, differential expression analysis can be performed. The data after this process is now ready to be applied to machine learning tools.

3 RNA-sequencing analysis using artificial intelligence techniques

RNA-seq analysis methods can be classified based on the data type to be analyzed and the AI technique to be applied. Analysis methods using AI can be divided into three categories: statistical analysis, machine learning, and deep learning. First, statistical analysis methods can be divided into test technique, proportional hazard model, and regression. Test techniques can be subdivided into the Kaplan-Meier test (Fu et al., 2020) and log-Rank test (Sun et al., 2018), which are frequently used in survival prediction and deriving gene expressions (Lu et al., 2012). A more advanced analysis method is the regression method represented by the Cox-ph model. Survival analysis based on the Cox-ph model has shown a predominant performance, in which TIDE is a representative example (Jiang et al., 2018). Statistical analysis techniques are advantageous when the data set is small or when the target to be analyzed is clear. For example, it is effective to use statistical analysis when identifying the tendency in the presence or absence of cancer recurrence. Second, in machine learning, supervised learning and AECOX (Huang et al., 2020), which discovers specific expression genes by combining models (e.g., SVM), universal classification tools (to construct ensembled-models), and methods combining two or more machine learning tools have been developed. Machine learning techniques such as SVM, random forest and decision trees are mainly used for classification and prediction, which help distinguish the subtype of the NSCLC or the normal and abnormal status. Machine learning algorithms show a good performance in binary classification (Han et al., 2014; Peng et al., 2015; Huang et al., 2017; Hsu and Dong, 2018; Reynders et al., 2018) and multi-class classification (Tian, 2017; Su et al., 2020), where the analysis method differs depending on the nature of the input data. Depending on the raw-data and target classification domain, there are various methods that can be applied from basic machine learning techniques, which include SVM, logistic regression, artificial neural network (ANN) (Khan et al., 2001), and AECOX (Huang et al., 2020), which combines neural networks. Third, deep learning is a category of machine learning that uses ANNs with multiple hidden layers (and each hidden layer consists of more artificial neurons) to provide higher levels of precision, classification, and estimations. Deep learning has been used in AECOX models and regression models like Cox-nnet

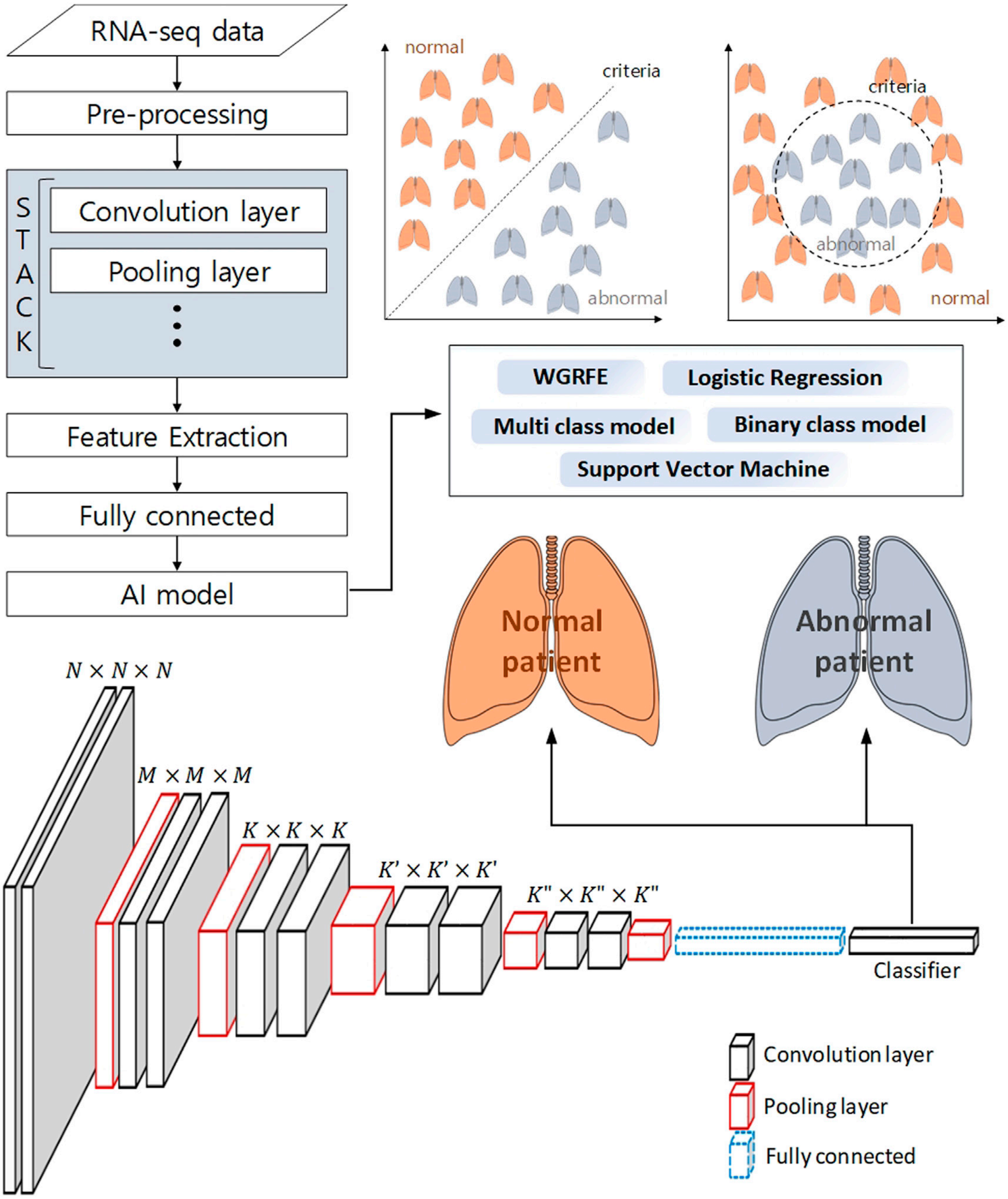

FIGURE 3
Schematic workflow of analysis to classify a normal patient from an abnormal patient using deep learning and machine learning. The pre-processed RNA-seq data is subjected to feature extraction using a neural network, and machine learning-based binary classification is performed using these features.

(Travers et al., 2018). In case the data set is small, deep transfer learning can be applied, or a pre-trained model can be imported and used in the analysis (Jiang et al., 2020). In the following, an overview of AI models used in NSCLC research is provided.

3.1 Statistical analysis and simple regression

Medical practitioners who study cancer patients as well as NSCLC mainly use the survival model. It is a statistical analysis method that estimates the survival time from the start of treatment to death of a patient (Kapil et al., 2018; He et al., 2020). However, since it focuses on simple 'death,' it is an analytical method that has limitations in being able to accurately analyze causal relationships. Regression techniques include the Kaplan-Meier test, log-Rank test, and the Cox's proportional hazard (Cox-ph) model. The Kaplan-Meier test has a disadvantage in that it cannot control its variables, whereas the Cox-ph model enables variable control and is therefore most frequently used. Among regression methods, Kaplan-Meier (Fu et al., 2020), log-rank test (Auslander et al., 2018; Fu et al., 2020), and Cox-ph are most

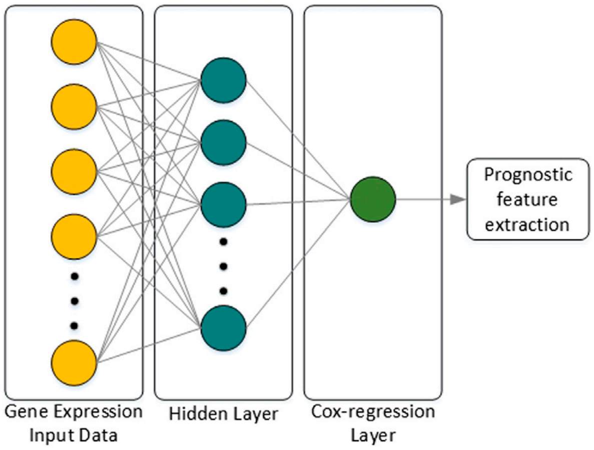

FIGURE 4
Architecture of Cox-nnet. The neural network structure is composed of the input layer, one fully connected hidden layer, and an output Cox regression layer. Finally, among many features of the patient, features related to prognosis are extracted.

frequently used, so they are briefly described in the following. Most regression analysis algorithms use standard methods (e.g., *p*-value), where it is possible to cluster patients based on a major classification criterion. For example, regression is commonly used to divide the overall survival rate, disease-free survival rate, and median survival value into groups of patients who have undergone different treatment methods and determine the significance of the survival rate between the two groups.

3.1.1 Kaplan-Meier

In (Qi et al., 2017), the Kaplan-Meier method was used based on the time of lung cancer surgery as the starting point to analyze the prognosis of patients who underwent surgery for metastatic lung cancer. The Kalman-Meier method (which is also called the product limit method) is effective in calculating the interval survival rate (at each event point during the entire study or analysis period) and finally calculates the cumulative survival rate. Applying this method to the RNA-seq data set, the survival rate can be calculated based on the occurrence of cancer. After that, the data can be arranged in the order in which cancer patients were observed, and the interval survival rate $P(k)$ can be calculated based on the ratio of the number of survivors in each interval. For example, if one person dies during the observation period, the interval survival rate is $(n-1)/n$, where n is the number of patients under observation. Finally, the cumulative survival rate $S(k)$ required for the Kalman-Meier test can be obtained by sequentially multiplying the interval survival rate according to Eq. 1, where N_s is the number of survivors up to period k and N_o represents the number of observations up to period k

$$P(k) = \frac{N_S}{N_O} \tag{1}$$

3.1.2 Log-rank test

The Kaplan-Meier test is an effective method for estimating the interval survival rate and cumulative survival rate, the log-rank test is effective in checking whether the difference in the survival rate between two groups to be discriminated is significant. For example, after using the Kaplan-Meier method, the log-rank test (applying the criteria of *p*-value less than 0.05) is used to compare the survival rates according to the disease-free survival period in (Qi et al., 2017). After the survival curve is drawn using the Kaplan-Meier method, the log-rank test plays a role in determining whether the groups have a significant statistical difference. The log-rank test is most effective when it is simply not possible to visually distinguish differences in survival curves. For example, in some cases a group that appears to have a relatively low survival rate may actually have a high survival rate, which can be accurately confirmed using the log-rank statistical test method.

3.1.3 Cox Proportional Hazard model

The Cox-ph model is the best-known method for screening prognostic variables that have a significant effect on the survival rate of lung cancer. In particular, the Cox-ph model is commonly used to compare patients with any RNA factor in two patient groups. For example, a group of patients with a high and low TGFB1 ratio has been analyzed as a survival fraction over time in (Cohen et al., 2020). A new biomarker can be estimated by selecting significant RNA groups among multiple RNA groups using the Cox-ph model. Most RNA sequence studies aim to calculate a significant biomarker candidate group by using this proportional hazard model and set a standard using the built-in *p*-value and fold change value. The previous Kaplan-Meier test and log-rank test are non-parametric analysis methods because they do not reflect the characteristics of the data. On the other hand, the Cox-ph model can predict the survival period using a regression model under the assumption that the survival time distribution (e.g., normal distribution) of lung cancer patients exists. In addition, the Cox-ph model utilizes the hazard ratio (HR), where it assumes that the HR is always constant. The Cox-ph model uses survival functions like Kaplan-Meier, where a survival function expressed by $S(t)$ represents 1 at first, and $S(t)$ would converge to 0 when infinite time has elapsed. Using a NSCLC patient as an example, the Cox-ph model will have $S(t) = 1$ at the starting point of observation. Through the survival function, it is possible to derive the lifetime distribution function $F(t)$, which represents the probability that the observed event will occur within a specific time. By differentiating the lifetime distribution function, it is possible to obtain the survival density function $f(t)$, which represents the event rate at a specific time. The relation of $S(t)$, $F(t)$, and $f(t)$ is expressed in Eq. 2.

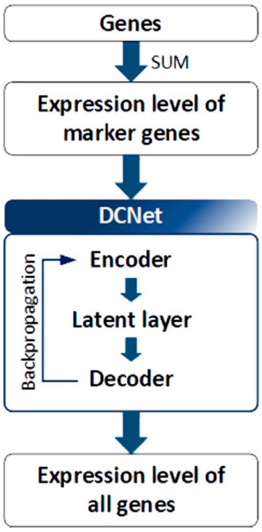
Genes
SUM
Expression level of marker genes
DCNet
Backpropagation
Encoder
Latent layer
Decoder
Expression level of all genes

FIGURE 5
Architecture of DCNet. Gene expression levels were considered to be input and output neurons in DCNet. The expression levels of marker genes and all genes are used as input and output neurons, respectively. Finally, the encoder layer and latent layer of the neural network are transferred, and the activation value of the latent layer indicates the abundance ratio of cells.

$$S(t) = P(T > t) = 1 - F(t) = \int_0^{\infty} f(u)du \tag{2}$$

The hazard function $H(t)$ uses the probability that an event will occur at an arbitrary point in time $h(t)$. The hazard function is based on a conditional probability that represents the probability that an event will occur for a case in which the event has not occurred until a specific time. That is, the hazard function $H(t)$ uses the probability relation of $h(t) = f(t)/s(t)$ and is defined in Eq. 3.

$$H(t) = \int_0^{t} h(u)du = -logS(t) \tag{3}$$

3.2 Machine learning

In the analysis of RNA sequencing of NSCLC using machine learning, supervised learning techniques represented by SVM classify subtypes of cancer, and regression models represented by LR are used for survival analysis. Machine learning is used to assist decision-making in clinical studies of various cancer types (including lung cancer) and can be used with a combination of various other analysis techniques (Huang et al., 2020). NSCLC related candidate RNAs with high rankings can be derived from the RNA-seq data using pre-processing techniques. In addition, by analyzing expression differences, normal and abnormal lung adenocarcinoma (LUAD) patient groups can be comparatively analyzed. For example, in (Cai et al., 2015; Zhou et al., 2016; Su et al., 2020) a classifier for diagnosis of LUAD using SVM is proposed where molecular markers are discovered from this classifier.

Machine learning techniques do not stand alone. Machine learning techniques use a combination of two or more techniques depending on their purpose. In (Afonso et al., 2019), genes were selected using the whole gene recursive feature elimination (WGRFE) technique and then subtypes of NSCLC were classified using SVM. If these ensembled models are properly combined, more accurate results can be obtained. As shown in Figure 3, by applying a classification model after analyzing with statistical analysis or deep learning tools, modules such as WGRFE and SVM can be applied in the last analyzing stage to assist the final criterion decision making. The results show that the performance can be significantly improved when machine learning techniques are used in conjunction with other analysis techniques. After pre-processing the RNA-seq data to extract the features, various techniques (e.g., statistical tools and neural networks) can be applied. The next technique to apply depends on the analysis target and the extracted features, where Figure 3 can be used to help choose the final algorithm needed to distinguish the abnormal and normal status as well as the level of difference.

3.3 Deep learning

Deep learning based RNA-seq analysis has been in the spotlight recently. Deep learning in the medical field has been mostly used for pattern recognition and medical image processing. For example, various imaging methods have been introduced to detect and analyze cell tissues (Volckmar et al., 2019). This is a representative method that can help reveal the origin of lung cancer, which applies imaging the tissue of cells. In these studies, deep learning is used to analyze the tumor's microenvironment through comparative analysis between the cancer microenvironment and surrounding cells. In particular, deep learning CNN analysis on CT image data is widely used (Ma et al., 2013; Su et al., 2020). In addition, non-image data has been converted into image-like data such that CNN analysis can applied (Arbour et al., 2021). Recently, research on survival prognosis of cancer patients using deep learning has been actively conducted. However, deep learning models are rarely used as a stand-alone method in predicting a prognosis. The most widely used NSCLC deep learning models include Cox-nnet, DeepSurv, and AECOX, which are briefly described below.

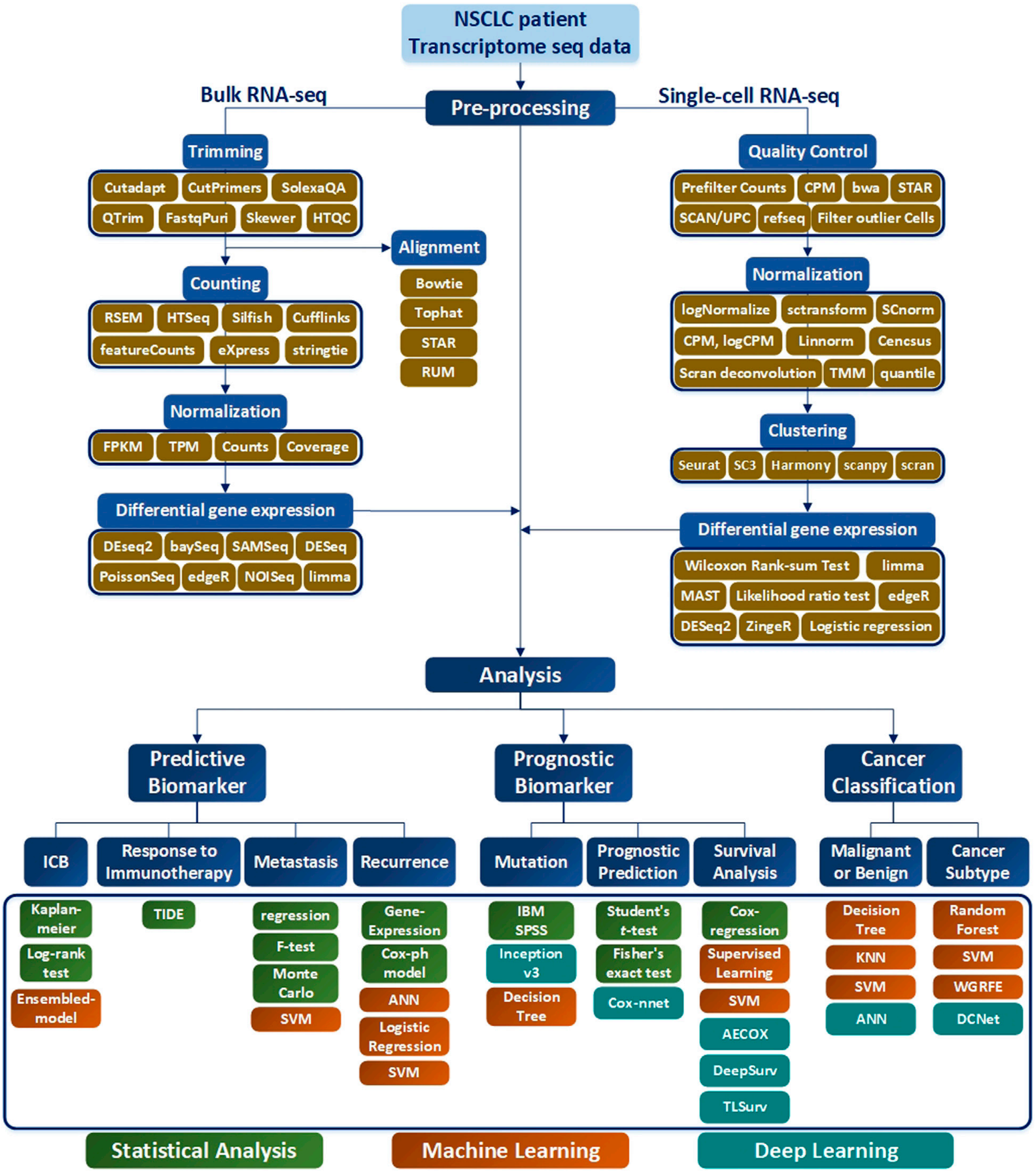

FIGURE 6
RNA-seq workflow of how to preprocess data and conduct analysis. The preprocessing section is divided into bulk RNA-seq and single-cell RNA-seq, where each require quality control. In addition, differential gene expressions are derived to obtain data for analysis. In the Analysis section, the target is classified into Predictive Biomarker, Prognostic Biomarker, and Cancer Classification, and analysis methods are sub-classified according to each purpose. In addition, the necessary tools according to the data type are indicated in green for Statistical Analysis, orange for the Machine Learning techniques, and blue for the me`thods using Deep Learning.

3.3.1 Cox-nnet

Classification or prognostic prediction using the machine learning techniques described above have meaning only in the final stage analysis. Thus, research on combining the characteristics of deep neural networks (DNNs) with regression models was attempted, and an extension of the Cox-regression model using a DNN was proposed in (Travers et al., 2018), which was named Cox-nnet. The characteristics of Cox-nnet are explained in Figure 4. Cox-nnet is used to make predictions, where it was first used in cancer survival predictions. However, its simple model has a disadvantage in that it is difficult to apply when the dimension of the input data (i.e., number of input data types) increases.

3.3.2 DeepSurv

DeepSurv is a multilayer perceptron model which consists of hidden layers consisting of fully-connected non-linear activation functions similar to the Faraggi-simon network consisting of a single hidden layer with two or three nodes (Faraggi and Simon, 1995). DeepSurv uses a non-linear proportional hazard model, which uses a neural network inside a Cox hazard model. DeepSurv includes one or more hidden layers, weight decay regulation, and activation functions such as an exponential linear unit (ELU) or a rectified linear unit. The DeepSurv performance can be improved by adding hidden layers to form a DNN so that the covariates of the first hidden layer of the DNN are used as input to the Cox proportional hazard model. The output of the DNN can be made to be a single node that estimates the hazard function $H_\theta(x)$ parameter based on the DNN weight θ (Katzman et al., 2018). DeepSurv can adjust the spacing of the non-linear model distributions by adjusting the network output nodes and can predict individual non-linear distributions for a single data input. DeepSurv can be used in a variety of survival analysis applications. Examples of this approach can be found in many treatment recommendations, which is a medical application that provides treatment recommendations based on a set of patient observations.

3.3.3 DCNet

DCNet is an autoencoder-based deep learning model that predicts about 400 cell types from a bulk RNA-seq dataset and discovers marker genes (Wang et al., 2022). As presented in Figure 5, the DCNet model consists of a total of three layers: an input layer corresponding to the marker gene, a hidden layer represented by the cell type, and an output layer composed of TCGA gene data. Therefore, it is possible to identify the relationship between a marker gene and a cell and identify a TME-specific biomarker through DCNet. DCNet can also be used for the purpose of classifying cancer subtypes. However, as in other deep learning models, the lack of data greatly affects the final performance, so to prevent this, oversampling techniques were introduced to solve the class label imbalance problem. In order to use the deep learning model effectively, a sufficient amount of training data must be secured. Then the insufficient part can be solved by using fine-tuning techniques to update the weights of the network. The DCNet model is meaningful in that it improves robustness and stability by applying a deep learning-based framework to RNA-seq research that uses simple machine learning or simple statical analysis techniques.

4 Conclusion

Selection of a RNA-seq method to apply to NSCLC data depends on the target of the analysis type, which may not be easy to select. To assist this process, in this paper, various analysis methods depending on the objective of RNA-seq are summarized in Table 1, and a methodological approach guideline is presented in Figure 6. For RNA-seq analysis of NSCLC patients, separate pipelines must be used by dividing bulk RNA-seq and single-cell RNA-seq. There is no big difference between bulk RNA-seq and single-cell RNA-seq in the basic principle of analyzing RNA-seq. However, whereas bulk RNA-seq analyzes the average value of whole cells, single-cell RNA-seq outlines each cell separately and analyzes the average value of each cell type. As shown in Figure 6, trimming to remove contamination or low-quality data and subsequent quantification are performed using various counting tools such as HTSeq. The sequence alignment process is preemptively performed. Then the pre-processing of the bulk RNA-seq can be completed by using DEseq2 or Limma to check differential gene expressions. In single-cell RNA-seq, the quality control process is carried out in units of 1 cell, where the information obtained from several cells is well classified, the similar cells are collected, and the clustering process is added, in which groups of similar cells are grouped and analyzed. Differential gene expression analysis is also performed in single-cell RNA-seq, and various tools such as MAST are used in addition to DESeq2. One effective way to analyze the subtypes of NSCLC patients would be to first cluster the NSCLC patient data corresponding to each subtype and then divide the RNA-seq data into different data sets. Using this as the input data, a classification scheme can be selected based on the analysis objective using Figure 6, and the data can be analyzed according to the number of classes. As a result, RNA that affects the subtype of the tumor can be extracted by sharp value-based ranking.

This paper investigates various AI analysis methods of RNA-seq research and provides guidance on how to apply this in NSCLC analysis and predictions. Although there are many papers on this area, there is no known single dominant method on how to apply systematic AI learning technologies in analyzing NSCLC RNA-seq data. The best way to obtain the most accurate analysis result is to select the research goal and the corresponding model properly, in which Table 1 and Figure 6 can provide some guidance. Although this paper is limited to NSCLC patients, it can be applied to other cancer types, such as breast cancer or colorectal cancer, which will be the focus of future research.

Author contributions

MJ was in charge of the bibliography search and the overall design of the thesis, K-HP provided medical advice, and J-MC and BC conducted the review of the entire thesis.

Funding

This work was supported by the National Research Foundation (NRF) of the Republic of Korea government Ministry of Science and ICT (MSIT) grant no. 2022R1A2C3005817. This research was also supported by the Bio & Medical Technology Development Program of the NRF funded by the Republic of Korea government MSIT grant no. 2022M3E5F3081138. This research was also supported by the MSIT grant D0318-21-1008 and the National Fire Agency grant 20017102 of the Republic of Korea.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Afonso, L. C. S., GustavoRosa, H., Pereira, C. R., Weber, S. A. T., Hook, C., Albuquerque, V. H. C., et al. (2019). A recurrence plot-based approach for Parkinson's disease identification. *Future Gener. Comput. Syst.* 94, 282–292. doi:10.1016/j.future.2018.11.054

Ahmed, H., Parmar, C., Coroller, T. P., Grossmann, P., Roman, Z., Kumar, A., et al. (2018). Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. *PLoS Med.* 15 (11), e1002711. doi:10.1371/journal.pmed.1002711

Althammer, S., Heng Tan, T., Spitzmüller, A., Rognoni, L., Tobias, W., Herz, T., et al. (2019). Automated image analysis of nsclc biopsies to predict response to anti-pd-l1 therapy. *J. Immunother. cancer* 7 (1), 121–212. doi:10.1186/s40425-019-0589-x

Arbour, K. C., Luu, A. T., Jia, L., Rizvi, H., Plodkowski, A. J., Sakhi, M., et al. (2021). Deep learning to estimate RECIST in patients with NSCLC treated with PD-1 blockade. *Cancer Discov.* 11 (1), 59–67. doi:10.1158/2159-8290.cd-20-0419

Auslander, N., Zhang, G., Lee, J. S., Frederick, D. T., Miao, B., Moll, T., et al. (2018). Robust prediction of response to immune checkpoint blockade therapy in metastatic melanoma. *Nat. Med.* 24 (10), 1545–1549. doi:10.1038/s41591-018-0157-9

Byron, S. A., Van Keuren-Jensen, R. K., Engelthaler, D. M., Carpten, J. D., and Craig, D. W. (2016). Translating rna sequencing into clinical diagnostics: Opportunities and challenges. *Nat. Rev. Genet.* 17 (5), 257–271. doi:10.1038/nrg.2016.10

Cai, Z., Dong, X., Zhang, Q., Zhang, J., Ngai, S. M., and Shao, J. (2015). Classification of lung cancer using ensemble-based feature selection and machine learning methods. *Mol. Biosyst.* 11 (3), 791–800. doi:10.1039/c4mb00659c

Cohen, D., Hondelink, L. M., Solleveld-Westerink, N., Uljee, S. M., Ruano, D., Cleton-Jansen, A. M., et al. (2020). Optimizing mutation and fusion detection in nsclc by sequential dna and rna sequencing. *J. Thorac. Oncol.* 15 (6), 1000–1014. doi:10.1016/j.jtho.2020.01.019

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., et al. (2016). A survey of best practices for rna-seq data analysis. *Genome Biol.* 17 (1), 13–19. doi:10.1186/s13059-016-0881-8

Coudray, N., Santiago Ocampo, P., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., et al. (2018). Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat. Med.* 24 (10), 1559–1567. doi:10.1038/s41591-018-0177-5

De Luca, C., Pepe, F., Iaccarino, A., Pisapia, P., Righi, L., Listì, A., et al. (2021). Rna-based assay for next-generation sequencing of clinically relevant gene fusions in non-small cell lung cancer. *Cancers* 13 (1), 139. doi:10.3390/cancers13010139

Faraggi, D., and Simon, R. (1995). A neural network model for survival data. *Statistics Med.* 14 (1), 73–82. doi:10.1002/sim.4780140108

Fu, J., Li, K., Zhang, W., Wan, C., Zhang, J., Jiang, P, et al. (2020). Large-scale public data reuse to model immunotherapy response and resistance. *Genome Med.* 12 (1), 21–28. doi:10.1186/s13073-020-0721-z

Galvez, C., Jacob, S., Finkelman, B. S., Zhao, J., Tegtmeyer, K., YoungChae, K., et al. (2020). The role of egfr mutations in predicting recurrence in early and locally advanced lung adenocarcinoma following definitive therapy. *Oncotarget* 11 (21), 1953–1960. doi:10.18632/oncotarget.27602

Givechian, K. B., Garner, C., Benz, S., Song, B., Rabizadeh, S., and Soon-shiong, P. (2019). An immunogenic nsclc microenvironment is associated with favorable survival in lung adenocarcinoma. *Oncotarget* 10, 1840–1849. doi:10.18632/oncotarget.26748

Han, S. S., Kim, W. J., Hong, Y., Hong, S. H., Lee, S. J., DongRyu, R., et al. (2014). Rna sequencing identifies novel markers of non-small cell lung cancer. *Lung Cancer* 84 (3), 229–235. doi:10.1016/j.lungcan.2014.03.018

Han, X., Tan, Q., Yang, S., Li, J., Xu, J., Hao, X., et al. (2019). Comprehensive profiling of gene copy number alterations predicts patient prognosis in resected stages i–iii lung adenocarcinoma. *Front. Oncol.* 9, 556. doi:10.3389/fonc.2019.00556

He, B., Dong, D., She, Y., Zhou, C., Fang, M., Zhu, Y., et al. (2020). Predicting response to immunotherapy in advanced non-small-cell lung cancer using tumor mutational burden radiomic biomarker. *J. Immunother. cancer* 8 (2), e000550. doi:10.1136/jitc-2020-000550

Hida, T., Nokihara, H., Kondo, M., Kim, Y. H., Azuma, K., Seto, T., et al. (2017). Alectinib versus crizotinib in patients with alk-positive non-small-cell lung cancer (j-alex): An open-label, randomised phase 3 trial. *Lancet* 390 (10089), 29–39. doi:10.1016/S0140-6736(17)30565-2

Hsu, Y. H., and Dong, S. (2018). Cancer type prediction and classification based on rna-sequencing data. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* 2018, 5374–5377. doi:10.1109/EMBC.2018.8513521

Huang, G., Zhan, H., Yu, H., Chen, L., You, L., Huang, T., et al. (2021). Identifying lung cancer cell markers with machine learning methods and single-cell rna-seq data. *Life* 11 (9), 940–949. doi:10.3390/life11090940

Huang, Z., Chen, L., and Wang, C. (2017). Classifying lung adenocarcinoma and squamous cell carcinoma using rna-seq data. *Cancer Stud Mol Med Open J.* 3 (2), 27–31. doi:10.17140/csmmoj-3-120

Huang, Z., Travis, S., Johnson, Z. H., Helm, B., Cao, S., Zhang, C., et al. (2020). Deep learning-based cancer survival prognosis from rna-seq data: Approaches and evaluations. *BMC Med. genomics* 13 (5), 41–12. doi:10.1186/s12920-020-0686-1

Jefferson, M. F., Pendleton, N., Lucas, S. B., and Horan, M. A. (1997). Comparison of a genetic algorithm neural network with logistic regression for predicting outcome after surgery for patients with nonsmall cell lung carcinoma. *Cancer Interdiscip. Int. J. Am. Cancer Soc.* 79 (7), 1338–1342. doi:10.1002/(sici)1097-0142(19970401)79:7<1338::aid-cncr10>3.0.co;2-0

Jiang, P., Gu, S., Deng, P., Fu, J., Sahu, A., Hu, X., et al. (2018). Signatures of t cell dysfunction and exclusion predict cancer immunotherapy response. *Nat. Med.* 24 (10), 1550–1558. doi:10.1038/s41591-018-0136-1

Jiang, Y., Alford, K., Frank, K., Tong, L., and Wang, M. D. (2020). "Tlsurv: Integrating multi-omics data by multi-stage transfer learning for cancer survival prediction," in Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, 1–10.

Kamer, I., Steuerman, Y., Daniel-Meshulam, I., Perry, G., Izraeli, S., Perelman, M., et al. (2020). Predicting brain metastasis in early stage non-small cell lung cancer patients by gene expression profiling. *Transl. Lung Cancer Res.* 9 (3), 682–692. doi:10.21037/tlcr-19-477

Kapil, A., Meier, A., Zuraw, A., Steele, K. E., Rebelatto, M. C., Schmidt, G., et al. (2018). Deep semi supervised generative learning for automated tumor proportion scoring on nsclc tissue needle biopsies. *Sci. Rep.* 8 (1), 17343–17410. doi:10.1038/s41598-018-35501-5

Katzman, J. L., Shaham, U., Alexander, C., Bates, J., Jiang, T., and Kluger, Y. (2018). Deepsurv: Personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* 18 (1), 24–12. doi:10.1186/s12874-018-0482-1

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Frank, W., et al. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.* 7 (6), 673–679. doi:10.1038/89044

Kim, N., Kim, H. K., Lee, K., Hong, Y., Cho, J. H., Choi, J. W., et al. (2020). Single-cell rna sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat. Commun.* 11 (1), 2285–2315. doi:10.1038/s41467-020-16164-1

Kruglyak, K. M., Lin, E., and Ong, F. S. (2016). "Next-generation sequencing and applications to the diagnosis and treatment of lung cancer," in *Lung cancer and personalized medicine: Novel therapies and clinical management*, 123–136. doi:10.1007/978-3-319-24932-2_7

Li, S., Paweł, P. Ł., Paul, Z., Peter, S., Shi, W., Shi, L., et al. (2014). Detecting and correcting systematic variation in large-scale rna sequencing data. *Nat. Biotechnol.* 32 (9), 888–895. doi:10.1038/nbt.3000

Lu, Y., Wang, L., Liu, P., Yang, P., and You, M. (2012). Gene-expression signature predicts postoperative recurrence in stage i non-small cell lung cancer patients. *PloS one* 7 (1), e30880. doi:10.1371/journal.pone.0030880

Ma, L., Bajic, V. B., and Zhang, Z. (2013). On the classification of long non-coding rnas. *RNA Biol.* 10 (6), 924–933. doi:10.4161/rna.24604

Nikolas Kather, J., Pearson, A. T., Halama, N., Jäger, D., Krause, J., Loosen, S. H., et al. (2019). Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* 25 (7), 1054–1056. doi:10.1038/s41591-019-0462-y

Passiglia, F., Galvano, A., Castiglia, M., Incorvaia, L., Calò, V., Listì, A., et al. (2019). Monitoring blood biomarkers to predict nivolumab effectiveness in nsclc patients. *Ther. Adv. Med. Oncol.* 11, 175883591983992. doi:10.1177/1758835919839928

Peng, L., Wu Bian, X., Li, D. K., Xu, C., Wang, G. M., Xia, Q. Y., et al. (2015). Large-scale rna-seq transcriptome analysis of 4043 cancers and 548 normal tissue controls across 12 tcga cancer types. *Sci. Rep.* 5 (1), 13413–13418. doi:10.1038/srep13413

Qi, L., Li, T., She, G., Wang, J., Li, X., Zhang, S., et al. (2017). An individualized gene expression signature for prediction of lung adenocarcinoma metastases. *Mol. Oncol.* 11 (11), 1630–1645. doi:10.1002/1878-0261.12137

Reynders, K., Wauters, E., Moisse, M., Herbert, D., De Leyn, P., Peeters, S., et al. (2018). Rna-sequencing in non-small cell lung cancer shows gene downregulation of therapeutic targets in tumor tissue compared to non-malignant lung tissue. *Radiat. Oncol.* 13 (1), 131–138. doi:10.1186/s13014-018-1075-1

Sharma, A., Vans, E., Shigemizu, D., Boroevich, K. A., and Tsunoda, T. (2019). Deepinsight: A methodology to transform a non-image data to an image for convolution neural network architecture. *Sci. Rep.* 9 (1), 11399–11407. doi:10.1038/s41598-019-47765-6

Soda, M., Choi, Y. L., Enomoto, M., Takada, S., Yamashita, Y., Ishikawa, S., et al. (2007). Identification of the transforming eml4–alk fusion gene in non-small-cell lung cancer. *Nature* 448 (7153), 561–566. doi:10.1038/nature05945

Su, R., Zhang, J., Liu, X., and Wei, L. (2020). Identification of expression signatures for non-small-cell lung carcinoma subtype classification. *Bioinformatics* 36 (2), 339–346. doi:10.1093/bioinformatics/btz557

Sun, N., Chu, J., Hu, W., Chen, X., Yi, N., and Shen, Y. (2022). A novel 14-gene signature for overall survival in lung adenocarcinoma based on the bayesian hierarchical cox proportional hazards model. *Sci. Rep.* 12 (1), 27–11. doi:10.1038/s41598-021-03645-6

Sun, W., Jiang, M., Dang, J., Chang, P., and Yin, F. F. (2018). Effect of machine learning methods on predicting nsclc overall survival time based on radiomics analysis. *Radiat. Oncol.* 13 (1), 197–198. doi:10.1186/s13014-018-1140-9

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA a cancer J. Clin.* 71 (3), 209–249. doi:10.3322/caac.21660

Tao, X., Wu, X., Huang, T., and Mu, D. (2020). Identification and analysis of dysfunctional genes and pathways in cd8+ t cells of non-small cell lung cancer based on rna sequencing. *Front. Genet.* 11, 352. doi:10.3389/fgene.2020.00352

Tian, S. (2017). Classification and survival prediction for early-stage lung adenocarcinoma and squamous cell carcinoma patients. *Oncol. Lett.* 14 (5), 5464–5470. doi:10.3892/ol.2017.6835

Travers, C., Zhu, X., and Garmire, L. X. (2018). Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS Comput. Biol.* 14 (4), e1006076. doi:10.1371/journal.pcbi.1006076

Volckmar, A. L., Jonas, L., Kirchner, M., Christopoulos, P., Neumann, O., Jan, B., et al. (2019). Combined targeted dna and rna sequencing of advanced nsclc in routine molecular diagnostics: Analysis of the first 3,000 heidelberg cases. *Int. J. Cancer* 145 (3), 649–661. doi:10.1002/ijc.32133

Wang, X., Wang, H., Liu, D., Wang, N., He, D., Wu, Z., et al. (2022). Deep learning using bulk rna-seq data expands cell landscape identification in tumor microenvironment. *OncoImmunology* 11 (1), 2043662. doi:10.1080/2162402x.2022.2043662

Wiesweg, M., Mairinger, F., Reis, H., Goetz, M., Walter, R. F. H., Hager, T., et al. (2019). Machine learning-based predictors for immune checkpoint inhibitor therapy of non-small-cell lung cancer. *Ann. Oncol.* 30 (4), 655–657. doi:10.1093/annonc/mdz049

Xiong, Y., Feng, Y., Qiao, T., and Han, Y. (2020). Identifying prognostic biomarkers of non-small cell lung cancer by transcriptome analysis. *Cancer Biomarkers* 27 (2), 243–250. doi:10.3233/cbm-190222

Yuan, L., Sun, N., Lu, Z., Sun, S., Huang, J., Chen, Z., et al. (2017). Prognostic alternative mrna splicing signature in non-small cell lung cancer. *Cancer Lett.* 393, 40–51. doi:10.1016/j.canlet.2017.02.016

Zhao, X., Wang, A., Walter, V., Patel, N. M., Eberhard, D. A., Hayward, M. C., et al. (2015). Combined targeted dna sequencing in non-small cell lung cancer (nsclc) using uncseq and ngscopy, and rna sequencing using uncqer for the detection of genetic aberrations in nsclc. *PloS one* 10 (6), e0129280. doi:10.1371/journal.pone.0129280

Zhou, Z., Folkert, M., Cannon, N., Iyengar, P., Westover, K., Zhang, Y., et al. (2016). Predicting distant failure in early stage nsclc treated with sbrt using clinical parameters. *Radiotherapy Oncol.* 119 (3), 501–504. doi:10.1016/j.radonc.2016.04.029