mathematics

MDPI

Article

A Novel Meta-Analysis-Based Regularized Orthogonal Matching Pursuit Algorithm to Predict Lung Cancer with Selected Biomarkers

Sai Wang [1,2], Bin-Yuan Wang [1] and Hai-Fang Li [1,2,*]

[1] College of Computer Science and Technology, Taiyuan University of Technology, Taiyuan 030024, China; wangsai@tyut.edu.cn (S.W.); wangbinyuan0647@link.tyut.edu.cn (B.-Y.W.)
[2] College of Data Science, Taiyuan University of Technology, Taiyuan 030024, China
* Correspondence: lihaifang@tyut.edu.cn

Abstract: Biomarker selection for predictive analytics encounters the problem of identifying a minimal-size subset of genes that is maximally predictive of an outcome of interest. For lung cancer gene expression datasets, it is a great challenge to handle the characteristics of small sample size, high dimensionality, high noise as well as the low reproducibility of important biomarkers in different studies. In this paper, our proposed meta-analysis-based regularized orthogonal matching pursuit (MA-ROMP) algorithm not only gains strength by using multiple datasets to identify important genomic biomarkers efficiently, but also keeps the selection flexible among datasets to take into account data heterogeneity through a hierarchical decomposition on regression coefficients. For a case study of lung cancer, we downloaded GSE10072, GSE19188 and GSE19804 from the GEO database with inconsistent experimental conditions, sample preparation methods, different study groups, etc. Compared with state-of-the-art methods, our method shows the highest accuracy, of up to 95.63%, with the best discriminative ability (AUC 0.9756) as well as a more than 15-fold decrease in its training time. The experimental results on both simulated data and several lung cancer gene expression datasets demonstrate that MA-ROMP is a more effective tool for biomarker selection and learning cancer prediction.

Keywords: biomarker selection; meta-analysis; regularized orthogonal matching pursuit; lung cancer

MSC: 62P10; 68Uxx

Citation: Wang, S.; Wang, B.-Y.; Li, H.-F. A Novel Meta-Analysis-Based Regularized Orthogonal Matching Pursuit Algorithm to Predict Lung Cancer with Selected Biomarkers. *Mathematics* **2023**, *11*, 4171. https://doi.org/10.3390/math11194171

Academic Editor: Ian Morilla

Received: 24 August 2023
Revised: 25 September 2023
Accepted: 29 September 2023
Published: 5 October 2023

1. Introduction

Biomarker selection [1] is the process of detecting relevant genes and removing irrelevant and redundant ones in order to obtain a subset of biomarkers that properly describes individuals' predisposition to particular types of cancers with a minimum degradation in performance. Cancer can be described as a group of diseases associated with uncontrollable cell growth that spreads into surrounding tissues. When cancer starts in the lungs, it is called lung cancer, and this is considered the cause of the greatest number of deaths from 2006 through 2020 [2]. With the development of new molecular and cell technologies, public gene expression databases have growth exponentially, providing an unprecedented number of datasets and studies. The main challenge with gene expression data is identifying new methods that can cope with information with a small sample size, high dimensionality, high noise as well as low reproducibility or lack of generalizability, due to small sample sizes relative to the large number of genes and low signal-to-noise ratios in most gene expression datasets.

Feature selection has been shown to provide an effective tool for analyzing omics data for the removal of irrelevant, noisy and redundant data while increasing the learning accuracy and improving the quality of the classification results [3,4]. To avoid this problem, the importance of feature selection has been stressed and several high-dimensional algorithms

have been proposed. Particularly, regularization methods use different penalty functions embedded in the learning procedure as a single process and have lower risk of over-fitting. The least absolute shrinkage and selection operator (lasso, ℓ_1-norm) [5], which performs continuous shrinkage and feature selection at the same time, is perhaps the most widely used signal processing algorithm in bioinformatics, machine learning and statistics. Other ℓ_1-norm-based learning methods typically include smoothly clipped absolute deviation (SCAD) [6], minimax concave penalty (MCP) [7] and group lasso [8]. Specifically, lasso solves a global optimization problem in a greedy way [9], similarly to another well-known algorithm in signal processing—the orthogonal matching pursuit (OMP) or orthogonal greedy algorithm.

The OMP algorithm [10] solves the problem via iterative methods using signal coefficients and tends to select only one from correlated features. Compared with other alternative methods, major advantages of the OMP are its simplicity and fast implementation. This method was originally employed to select features for binary-class classification and has been used for feature selection and signal recovery. For example, Tawfic and Kayhan [11] proposed least support denoising–orthogonal matching pursuit (LSD-OMP) to reconstruct an original signal with the presence of noise. Then, Ji and Zhang [12] improved the reconstruction accuracy of the LSD-OPM through setting the threshold, eliminating some wrong atoms and combining some support sets to locate the optimal support set. In this paper, we utilize the MS-ROMP strategy for the biomarker selection problem. With the development of bioinformatics, Shi et al. [13] proposed structured OMP by considering correlations among distinct features for multi-class classification. Recently, Tsagris et al. [14] proposed a general OMP feature selection algorithm that can handle binary outcomes, multi-class outcomes, continuous outcomes and censored time-to-event outcomes with applications on gene expression data. It is usually assumed that the same features in multiple studies should make the same contribution to their corresponding results. However, most existing methods identify a feature (gene) that is important in some studies but may not affect other studies. This may be due to the different experimental conditions, sample preparation tools, the sensitivity and accuracy of instruments, etc. Hence, it is important to make full use of different datasets and maintain the flexibility of feature selection at the same time.

Over the past few years, gene expression meta-analysis techniques have combined multiple and independent studies to improve reproducibility or obtain more reliable biomarkers [15]. There are three main types of meta-analysis methods [16]: meta-analysis based on combining results from different studies (e.g., effect sizes, p values or ranks), meta-analysis based on particular cross-platform normalization and a unified model on multiple datasets without data merging that can account for the joint effects of genes on clinical outcomes. Considering the joint modeling of multiple genes, Li et al. [17] proposed meta-lasso, and then meta-nonconvex methods emerged gradually for solving the heterogeneity problem [18,19]. Thus, we combine the MS-ROMP strategy with meta-analysis techniques to improve the strength across multiple datasets.

In this paper, a novel meta-analysis-based regularized orthogonal matching pursuit, namely MA-ROMP, is presented and the workflow of our proposed algorithm is shown in Figure 1. Several microarray gene expression data points for lung cancer are obtained from disparate platforms, standardized data and extracted common genes. After training a unified model on multiple datasets without data merging, MA-ROMP divides the results into tumor and nontumor disease with ROC evaluation, overlapping selected features (genes) and the highest-ranked gene alterations as a possible biological explanation. In this way, the MA-ROMP method is suitable for combining studies from different platforms or conditions, and obtains reliable results with a small number of studies.

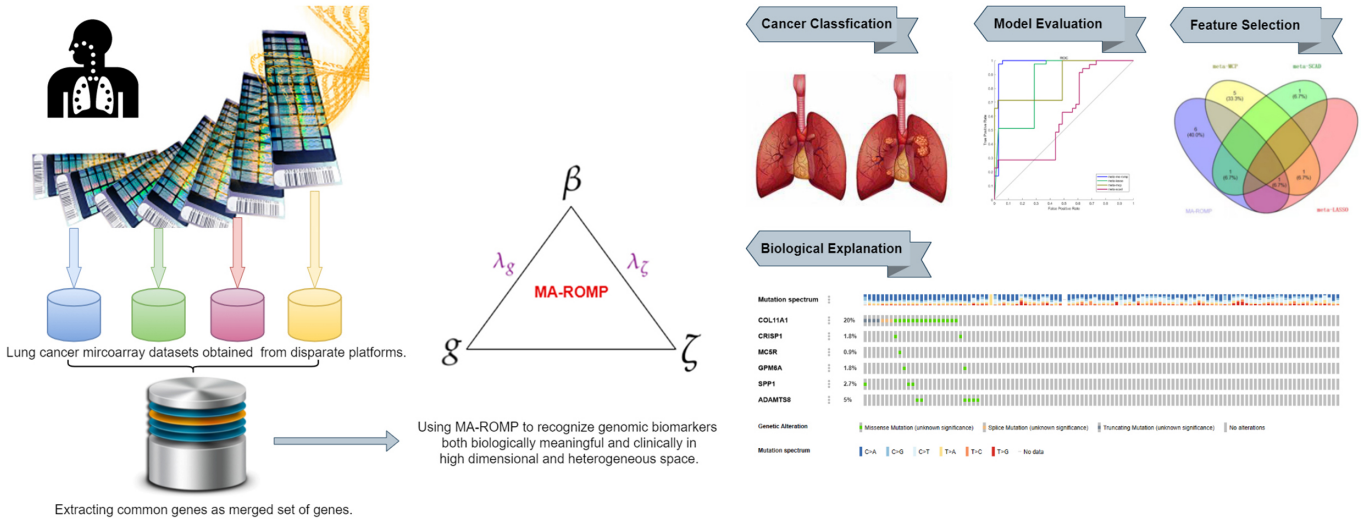

Figure 1. The workflow of the meta-analysis-based regularized orthogonal matching pursuit (MA-ROMP) to predict lung cancer with selected biomarkers. Several microarray gene expression data points for lung cancer were obtained from disparate platforms, data were standardized and common genes were extracted. After training a unified model on multiple datasets without data merging, the MA-ROMP divides the results into tumor and nontumor disease with ROC evaluation, overlapping selected features (genes) and the highest-ranked gene alterations as a possible biological explanation.

The results showing that our proposed MA-ROMP not only borrows strength from across multiple datasets to identify important genomic biomarkers efficiently, but it also mainains selection flexibility among datasets to take into account data heterogeneity through a hierarchical decomposition on regression coefficients. Furthermore, we apply the MA-ROMP in a real RNA expression data experiment. Compared with the state-of-the-art methods, our proposed method MA-ROMP has good performance.

The remainder of this paper is organized as follows: The regularized orthogonal matching pursuit algorithm for biomarker selection is presented in Section 2. In Section 3, we describe the implementation of novel meta-analysis-based regularized orthogonal matching pursuit (MA-ROMP). Then, we measure and discuss the performance of MA-ROMP on both simulated data and three publicly available lung cancer gene expression datasets in Section 4. A concluding remark is finally provided in Section 5.

2. The Regularized Orthogonal Matching Pursuit Algorithm for Biomarker Selection

As one of the main greedy pursuit algorithms, OMP is a forward search algorithm that was first proposed for continuous outcomes in the context of signal reconstruction [20]. Tawfic and Kayhan [11] proposed that least support denoising–orthogonal matching pursuit (LSD-OMP) enhanced OMP by choosing an optimum support set out of many in each iteration. Through setting the threshold, eliminating some wrong atoms and combining some support sets to locate the optimal support set, Ji and Zhang [12] proposed a regularized orthogonal matching pursuit-based multiple support (MS-ROMP) to improve the reconstruction accuracy of the LSD-OMP.

We utilize MS-ROMP in the biomarker selection problem. Suppose y denotes a continuous outcome and a data matrix $X = \{X_i\}_{i=1}^{n}$ of continuous features with n observations. The ith column of X is denoted as $X_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ corresponding to the p dimensional coefficient $\boldsymbol{\beta}$ and $X_{\mathbb{S}}$ denotes the matrix with all selected genes (variables), where $\mathbb{S}$ is the index set of selected genes. The algorithm initiates the current selection $\mathbb{S} \leftarrow \phi$, least support set $J \leftarrow \phi$ and residuals $r \leftarrow y$. At each iteration, the gene with the largest (in absolute value) Pearson linear correlation with r is selected for inclusion. If X is standardized, then it suffices to compute $\langle r, X_i \rangle$, i.e., the inner product of the two vectors. We can choose the index of the L (usually 10) biggest inner product as the optimum support set in

each iteration. By setting the threshold $\lambda \geq 0$, we eliminate some wrong atoms and find J^* to satisfy $\left|\langle r, X_{J^*}\rangle\right| \geq \lambda\left|\langle r, X_J\rangle\right|$. After selection, a new least squares regression model is fitted, resulting in new $\boldsymbol{\beta}$ coefficients, and is used to update the residuals. The procedure stops when the ℓ_2-norm of the residuals is below the smallest positive number ε, which is 0.001 in this paper.

The computational complexity of each step of MS-ROMP is provided in Algorithm 1. The computational complexity of lines 4–11 is $O(l \times (n-1+1+1+n^2p+1+1))$. Hence, the time taken by Algorithm 1 is $O(ln^2p+ln)$. For "high-dimensional-small-sample-size" data, $l, n \ll p$, the computational complexity of Algorithm 1 is further reduced to $O(p)$.

Algorithm 1 The MS-ROMP Algorithm for Biomarker Selection

Input: Outcome values y, dataset X, stop condition ε;
Hyper-Parameters: $\lambda,\ L$;
Output: A subset $X_{\mathbb{S}} \subseteq \mathbf{X}$ of selected genes corresponding with its coefficient $\boldsymbol{\beta}$.

1: Standardize the data: Center each gene expression data and y, then scale them to unit norm. ▷ $O(p)$
2: Initialize index set $\mathbb{S} \leftarrow \phi$, least support set $J \leftarrow \phi$, residuals $r \leftarrow y$. ▷ $O(1)$
3: $\mathbb{S} \leftarrow \phi,\ l \leftarrow 1$. ▷ $O(1)$
4: **while** $\| r \|^2 > \varepsilon$ **do**
5: $\quad J \leftarrow sliceMax(|\langle r, X_i\rangle|,\ L),\ i \in X \setminus \mathbb{S}$; ▷ $O(n-1)$
6: $\quad \left|\langle r, X_{J^*}\rangle\right| \geq \lambda\left|\langle r, X_J\rangle\right|$; ▷ $O(1)$
7: $\quad \mathbb{S} \leftarrow \mathbb{S} \cup \{J^*\}$; ▷ $O(1)$
8: $\quad \boldsymbol{\beta} \leftarrow$ least squares regression of y on $X_{\mathbb{S}}$; ▷ $O(n^2p)$
9: $\quad$ update residuals $r \leftarrow y - X_{\mathbb{S}} \cdot \boldsymbol{\beta}$; ▷ $O(1)$
10: $\quad l \leftarrow l+1$. ▷ $O(1)$
11: **end while**
12: **return** $X_{\mathbb{S}}, \boldsymbol{\beta}$.

3. The MA-ROMP Algorithm for Biomarker Selection

Machine learning classifiers are very popular for predicting cancer disease using microarray gene expression data. Logistic regression, one of the most well-known algorithms for binary classification, has no turning parameters and its prediction equation is simple and easy to implement. In this paper, to demonstrate the effectiveness of the proposed MA-ROMP algorithm, we use logistic regression to predict cancer risk—whether a person will develop lung cancer or not.

Given M independent datasets, $D = \{\tilde{X}, \tilde{Y}\}$, where $\tilde{X} = diag(X_1, X_2, \cdots, X_M)$, $\tilde{Y} = (Y_1^T, Y_2^T, \cdots, Y_M^T)^T$ and the superscript T denotes the standard vector transpose. Denote $X_{mi} = (x_{mi,1}, x_{mi,2}, \cdots, x_{mi,p})$ as a vector of expression profiles of the ith sample with p genes in the mth dataset and y_{mi} the corresponding dependent variable with the value of 0 or 1 for binary classification. In logistic regression, a logit transformation $\theta(s) = e^s/(1+e^s)$ is applied on the odds—that is, the probability of success divided by the probability of failure. We assume the conditional probability that y_{mi} takes value 1 given the gene expression vector X_{mi}, and the logistic function is represented by the following formulas:

$$\begin{aligned} \mathbb{P}(y_{mi} = 1 | X_{mi}) &= \theta(\beta_{m0} + X_{mi}^T \boldsymbol{\beta_m} + \varepsilon_{mi}) \\ &= \frac{e^{\beta_{m0} + X_{mi}^T \boldsymbol{\beta_m} + \varepsilon_{mi}}}{1 + e^{\beta_{m0} + X_{mi}^T \boldsymbol{\beta_m} + \varepsilon_{mi}}}, \end{aligned} \tag{1}$$

where $i = 1, \ldots, n_m,\ m = 1, \ldots, M, \rightarrow \mathbb{R}$, β_{m0} is the interception, $\boldsymbol{\beta_m} \subseteq \mathbb{R}^p$ is a vector of regression coefficient for the mth data and ε_{mi} are n_m independent random errors with a normal distribution function. We hope to find the true nonzero components of $\boldsymbol{\beta_m}$ for each dataset.

Because of the data heterogeneity, we utilize a joint fitting approach [17] that borrows strength from across different datasets. Consider the following hierarchical reparameterization:

$$\beta_{mj} = g_j \zeta_{mj}, \tag{2}$$

where $j = 1, 2, \ldots, p$, $m = 1, \ldots, M$, the parameter g_j is an effect of the jth gene at the first level of the hierarchy and $\zeta_{mj} \geq 0$ with different m's reflects effect differences for the jth gene among M datasets at the second level of the hierarchy. If there is no heterogeneity, i.e., $\zeta_{mj} = 1$, then $\beta_{mj} = g_j$.

This proposed MA-ROMP selects genomic biomarkers by solving

$$\max_{\beta_{m0}, \mathbf{g}, \boldsymbol{\zeta_m}} \left\{ \sum_{m=1}^{M} \mathcal{L}_m(\beta_{m0}, \mathbf{g}, \boldsymbol{\zeta_m}) - \sum_{j=1}^{p} \mathcal{F}(g_j; \lambda_g) - \sum_{j=1}^{p} \sum_{m=1}^{M} \mathcal{F}(\zeta_{mj}; \lambda_\zeta) \right\}, \tag{3}$$

where $\mathcal{F}$ is the MS-ROMP function,

$$\mathcal{L}_m(\beta_{m0}, \mathbf{g}, \boldsymbol{\zeta_m}) = \sum_{i=1}^{n_m} y_{mi} \left\{ \beta_{m0} + X_{mi}^T (\mathbf{g} \cdot \boldsymbol{\zeta_m}) \right\} - \log \left[1 + e^{\beta_{m0} + X_{mi}^T (\mathbf{g} \cdot \boldsymbol{\zeta_m})} \right],$$

$\mathbf{g} = (g_1, g_2, \cdots, g_p)^T$, $\boldsymbol{\zeta_m} = (\zeta_{m1}, \zeta_{m2}, \cdots, \zeta_{mp})^T$, $\boldsymbol{\beta_0} = (\beta_{10}, \beta_{20}, \cdots, \beta_{M0})^T$, $\boldsymbol{\zeta} = (\boldsymbol{\zeta}_1^T, \boldsymbol{\zeta}_2^T, \cdots, \boldsymbol{\zeta}_M^T)^T$ and $\mathbf{g} \cdot \boldsymbol{\zeta_m}$ means the element-wise product. Furthermore, there are two hyper-parameters: λ_g controls the biomarker selection algorithm at the entire-gene level and can effectively remove unimportant genes in all M datasets, while λ_ζ controls each individual dataset. Hence, if g_j is not equal to zero, the corresponding β_{mj} can still be shrunken to zero by some ζ_{mj}.

As shown in Algorithm 2, the MA-ROMP algorithm solves β_0, g and ζ iteratively. First we fix β_0 and ζ to update g via MS-ROMP function $\mathcal{F}$. Then, β_0 and g are fixed to update ζ. We next maximize the log-likelihood $\mathcal{L}_m$ for the mth dataset over β_0 by fixing g and ζ. These above training steps are repeated until the algorithm converges. At last, we predict the cancer risk with above selected biomarkers.

Algorithm 2 The MA-ROMP Algorithm for Biomarker Selection

Input: Datasets $\mathcal{D} = \{X_m, Y_m\}_{m=1}^M$, stop condition;
Hyper-Parameters: λ_g, λ_ζ;
Output: A subset $X_{\mathbb{S}} \subseteq X$ of selected genes.

1: Standardize the data: Center each gene expression data X_m, then scale them to unit norm. ▷ $O(p)$
2: Initialize the hierarchical parameter $\zeta_{mj}^{(0)} \leftarrow 1$, the interception $\boldsymbol{\beta}_{m0}^{(0)} \leftarrow 0$ and the iteration index $k \leftarrow 1$. ▷ $O(1)$
3: **while** $\max |\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}^{(k-1)}| > \varepsilon$ **do**
4: $\quad X_m \leftarrow X_m \zeta_m^{(k-1)}$; ▷ $O(n_m p)$
5: $\quad$ update $g^{(k)} \leftarrow \mathcal{F}(g; \lambda_g)$; ▷ $O(p)$
6: $\quad X_m \leftarrow X_m g^{(k)}$; ▷ $O(n_m p)$
7: $\quad$ update $\zeta_m^{(k)} \leftarrow \mathcal{F}(\zeta_m; \lambda_\zeta)$; ▷ $O(p)$
8: $\quad \boldsymbol{\beta_m}^{(k)} \leftarrow g^{(k)} \zeta_m^{(k)}$; ▷ $O(1)$
9: $\quad$ update $\beta_{m0}^{(k)} \leftarrow \underset{\beta_{m0}}{arg\max} \mathcal{L}(\beta_{m0}^{(k-1)}, \boldsymbol{\beta_m}^{(k)})$; ▷ $O(n_m^2 p)$
10: $\quad \hat{\boldsymbol{\beta}}^{(k)} \leftarrow (\beta_{m0}^{(k)}; \boldsymbol{\beta_m}^{(k)})$. ▷ $O(1)$
11: **end while**
12: $\mathbb{S} \leftarrow$ index of nonzero elements in $\hat{\boldsymbol{\beta}}$. ▷ $O(1)$
13: **return** $X_{\mathbb{S}}$.

Regarding the overall computational complexity of MA-ROMP, we have analyzed the function $\mathcal{F}$ in Algorithm 1. The details of computational complexity are given in Algorithm 2. Lines 3–11 consume the time $O(k \times (n_m p + p + n_m p + p + 1 + n_m^2 p + 1))$, i.e., $O(k n_m^2 p + 2k(n_m + 1)p + 2k)$. For "high-dimensional-small-sample-size" data, $k, n_m \ll p$, the computational complexity of Algorithm 2 is further reduced to $O(p)$, which is the same as Algorithm 1. Combining gene expression meta-analysis techniques greatly enhances MS-ROMP's ability to define common biomarkers across multiple datasets.

4. Experiments

To demonstrate the performance of our proposed MA-ROMP algorithm, we compare it with seven state-of-the-art methods, that is, lasso, SCAD, MCP, meta-lasso, meta-SCAD, meta-MCP and MS-ROMP.

4.1. Simulations

In order to simulate multiple datasets, we suppose that the number of datasets M is 10, and n_m is used to represent the sample size where $m = 1, \ldots, M$. Moreover, $n_m = 50$ and the number of genes (features) p is 1000. We generate the gene expression data X_m in normal distribution, associated with the response y_m, based on the following logistic model:

$$\mathbb{P}(y_{mi} = 1 | X_{mi}) = \frac{e^{\beta_{m0} + X_{mi}^T \boldsymbol{\beta_m^*} + \varepsilon_{mi}}}{1 + e^{\beta_{m0} + X_{mi}^T \boldsymbol{\beta_m^*} + \varepsilon_{mi}}}, \tag{4}$$

where the interception $\beta_{m0} = 0$, the independent random noise $\varepsilon_{mi} \sim N(0, 1)$ and the nonzero regression coefficient $\boldsymbol{\beta_m^*}$ of the first 10 genes are generated via a joint fitting approach in Equation (2). To allow possible different covariance structures, the effect of the gene g for all M datasets is generated from $N(3, 0.52)$ and the discrepancy among different datasets ζ is generated from $Bernoulli(\pi_0)$, where $\pi_0 :$ 0.2, 0.5 and 0.9 from low to high heterogeneity. This means that for the first 10 genes, 1 gene is important with a probability of π_0 in each dataset. For each case, we run 50 replicates.

The optimal hyperparameters of all methods are selected by minimizing the Bayesian information criterion (BIC) defined as:

$$BIC(\lambda) = \sum_{m=1}^{M} \{-2\mathcal{L}_m(\hat{\beta}_{m,\lambda}) + \log(n_m) df_\lambda\}, \tag{5}$$

where the mth dataset's estimated coefficients with hyperparameter λ are expressed as $\hat{\beta}_{m,\lambda}$, df_λ is the number of nonzero parameters, n_m is the sample size and $\hat{\beta}_{m,\lambda}$ is the log-likelihood with $\boldsymbol{\beta_m^*}$ being replaced by its estimate $\hat{\beta}_{m,\lambda}$.

The biomarker selection performances of the eight methods are summarized using selection sensitivity, specificity and accuracy of coefficient $\boldsymbol{\beta}$ in Table 1. The sensitivity is the proportion of nonzero β_{mj}^*s that are correctly estimated as nonzero, the specificity is the proportion of zero β_{mj}^*s that are correctly estimated as zero and the accuracy is the proportion of β_{mj}^*s that are correctly estimated. The specificity and accuracy of the coefficient in Table 1 for all eight methods are similar. The sensitivity of the methods without a meta-analysis learning policy, which are lasso, SCAD, MCP and MS-ROMP, dramatically decreases as π_0 increases, while meta-lasso, meta-SCAD, meta-MCP and MA-ROMP still have steady performances. Especially when data heterogeneity is strong ($\pi_0 = 0.2$), our proposed MA-ROMP has the superior performance.

Table 1. Results of the synthetic data, sensitivity, specificity and accuracy of coefficient β of the four methods are based on 50 simulations. Standard errors are given in parentheses.

Methods		π_0 = 0.9	π_0 = 0.5	π_0 = 0.2
SCAD	Sensitivity	0.470 (0.021)	0.400 (0.001)	0.310 (0.011)
	Specificity	0.946 (0.011)	0.942 (0.001)	0.940 (0.002)
	Accuracy	0.941 (0.001)	0.937 (0.001)	0.934 (0.001)
lasso	Sensitivity	0.540 (0.049)	0.520 (0.156)	0.360 (0.197)
	Specificity	0.947 (0.013)	0.948 (0.011)	0.945 (0.011)
	Accuracy	0.943 (0.013)	0.944 (0.013)	0.940 (0.013)
MCP	Sensitivity	0.560 (0.002)	0.520 (0.001)	0.320 (0.101)
	Specificity	0.899 (0.001)	0.894 (0.001)	0.890 (0.001)
	Accuracy	0.896 (0.001)	0.890 (0.001)	0.885 (0.001)
MS-ROMP	Sensitivity	0.610 (0.002)	0.570 (0.011)	0.410 (0.002)
	Specificity	0.944 (0.001)	0.922 (0.001)	0.943 (0.001)
	Accuracy	0.939 (0.001)	0.917 (0.001)	0.937 (0.001)
meta-SCAD	Sensitivity	0.890 (0.001)	0.870 (0.001)	0.690 (0.056)
	Specificity	1 (0)	0.999 (0.001)	0.981 (0.004)
	Accuracy	0.999 (0.001)	0.998 (0.001)	0.978 (0.003)
meta-lasso	Sensitivity	0.960 (0.001)	0.860 (0.001)	0.670 (0.048)
	Specificity	0.986 (0.001)	0.984 (0.001)	0.984 (0.001)
	Accuracy	0.986 (0.001)	0.983 (0.001)	0.981 (0.001)
meta-MCP	Sensitivity	0.990 (0.032)	0.930 (0.001)	0.700 (0.067)
	Specificity	0.964 (0.002)	0.999 (0.001)	0.976 (0.004)
	Accuracy	0.965 (0.002)	0.998 (0.001)	0.975 (0.004)
MA-ROMP	Sensitivity	0.900 (0.001)	1 (0)	0.790 (0.001)
	Specificity	0.994 (0.001)	0.995 (0.001)	0.993 (0.001)
	Accuracy	0.993 (0.001)	0.995 (0.001)	0.991 (0.001)

4.2. Real-Data Analysis

Cancer is a major public health problem worldwide. Moreover, lung cancer was the most common cause of cancer death according to [21]. We demonstrate our proposed method by analyzing lung cancer microarray expression data from NCBI's gene expression omnibus (GEO, https://www.ncbi.nlm.nih.gov/gds/, accessed on 24 September 2023) for the different experimental datasets with the accession numbers in Table 2. GSE10072 [22] contains 107 samples from 58 tumor and 49 nontumor tissues with 22,283 genes obtained using GEO Platform GPL96. GSE19188 [23] contains 91 patients from 91 tumor and 65 adjacent normal lung tissue samples with 54,675 genes obtained using GEO Platform GPL570. GSE19804 [24] contains 60 pairs of tumor and adjacent normal lung tissue specimens with 54,675 genes obtained using GEO Platform GPL570.

Table 2. Summaries of datasets in the lung cancer.

Datasets	GSE10072	GSE19188	GSE19804
Platform	GPL96	GPL570	GPL570
Total sample size	107	156	120
No. of genes	22,283	54,675	54,675

These above three balanced datasets come from inconsistent experimental conditions, sample preparation methods, measurement sensitivities or precision, and also from different study groups and biological sample selections. We extracted 22,277 common genes from these three lung cancer gene expression datasets as the merged set of genes. In order to keep class balance, we randomly selected from two types of samples respectively, and divided the data into 70% training and 30% test sets.

We measure the computational efficiency of each meta-analysis-based algorithm during the the $\mathcal{K}$-fold cross-validation. Typically, $\mathcal{K}$ is chosen to be 5 or 10 in medium-sized datasets. The special case of $\mathcal{K}$-fold cross-validation is called leave-one-out cross-validation, which is similar to the jack-knife method only for small data, and costly for data of other scales. Our training data were medium-sized, so we used 10-fold cross validation.

Figure 2 shows the boxplot for training times of four meta-analysis-based algorithms during the 10-fold cross-validation. The median training times of MA-ROMP, meta-lasso, meta-MCP and meta-SCAD (the red lines within each blue box) are 402.65 s, 4964.9 s, 6327.05 s and 4641 s respectively. The blue box length is the interquartile range and the crosses in red denote outliers. Our proposed MA-ROMP requires the shortest training time, and the computational complexity of MA-ROMP has been illustrated in Algorithm 2. Unlike the orthogonal matching pursuit-based method, meta-lasso, meta-MCP and meta-SCAD utilized an approximate message passing algorithm [25] containing a pseudo-inverse operation, which consumes considerable time in high dimensions and sparse data.

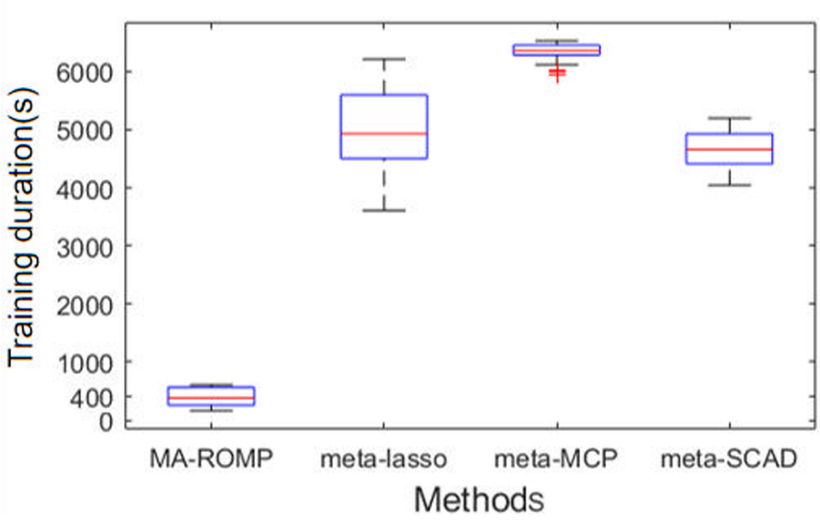

Figure 2. Boxplot for the training times of four meta-analysis-based algorithms during the 10-fold cross-validation. The median training times of MA-ROMP, meta-lasso, meta-MCP and meta-SCAD (the red lines within each blue box) are 402.65 s, 4964.9 s, 6327.05 s and 4641 s respectively. The blue box length is the interquartile range and the crosses in red denote outliers.

As shown in Table 3, when the hyper-parameters $\lambda_g = 0.01$, $\lambda_\zeta = 0.01$, the MA-ROMP performs better than other three meta-analysis-based methods with the highest accuracy of up to 95.63%, and is on average 11 times faster than meta-SCAD. Moreover, we used the area under the curve (AUC) as the performance metric in binary classification. Ideally, ROC curves that are closer to the upper-left corner (0, 1) indicate superior performance, as this signifies higher sensitivity and lower false-positive rates. Our proposed MA-ROMP exhibited a shape closest to the upper-left corner, along with the best discriminative ability with AUC 0.9756, in Figure 3, that moved up 11.64%, compared with the meta-MCP, method and cut its training time more than 15 times.

Table 4 gives the names of selected genes in each dataset. To further validate the genes selected by our proposed method, we performed an alterations analysis using cBioPortal (https://www.cbioportal.org/, accessed on 24 September 2023), illustrated in Figure 4, and the overlap of commonly selected genes across the different methods is shown in Figure 5. As seen in Table 4 and Figures 4 and 5, COL11A1 is associated with angiogenesis, invasion and drug resistance in cancer, and it is the most altered gene, representing 20% of all alteration in all patients, and was selected by meta-SCAD and MA-ROMP in Table 4. Perhaps that is why, so far, meta-SCAD has shown a better performance than meta-lasso and meta-MCP, as shown in Table 3, despite the limited amount of selected genomic biomarkers. Additionally, the growth factor FGF-14 negatively regulates COL11A1 expression in lung

cancer cells [26]. Yi et al. [27] identified that COL11A1 acted downstream of secreted phosphoprotein 1 (SPP1), which facilitates cell migration and invasion. This SPP1, also called osteopontin, is expressed in tumor cells, and alterations were detected in 2.7% of patients. Many studies demonstrated that its methylation variability and mRNA expression level are both correlated with prognosis in multiple cancer types [28,29]. SPP1-related signals are considered to represent a potential target for anti-cancer immunotherapies [30,31].

Table 3. Results of three lung cancer datasets; sensitivity, specificity and accuracy of different methods are based on 50 simulations. Standard errors are given in parentheses.

Methods	**Training Data**		
	Sensitivity	**Specificity**	**Accuracy**
MA-ROMP	0.9788 (0.001)	0.9465 (0.001)	0.9644 (0.001)
meta-lasso	0.4722 (0.001)	0.6667 (0.002)	0.6683 (0.001)
meta-MCP	0.4722 (0.001)	0.6667 (0.001)	0.6683 (0.002)
meta-SCAD	0.5470 (0.004)	0.9167 (0.001)	0.7178 (0.001)
Methods	**Training Data**		
	Sensitivity	**Specificity**	**Accuracy**
MA-ROMP	1 (0)	0.9188 (0.001)	0.9563 (0.001)
meta-lasso	0.4119 (0.003)	0.6528 (0.002)	0.5249 (0.011)
meta-MCP	0.4050 (0.013)	0.6528 (0.002)	0.5215 (0.001)
meta-SCAD	0.5278 (0.001)	0.9231 (0.001)	0.6571 (0.004)

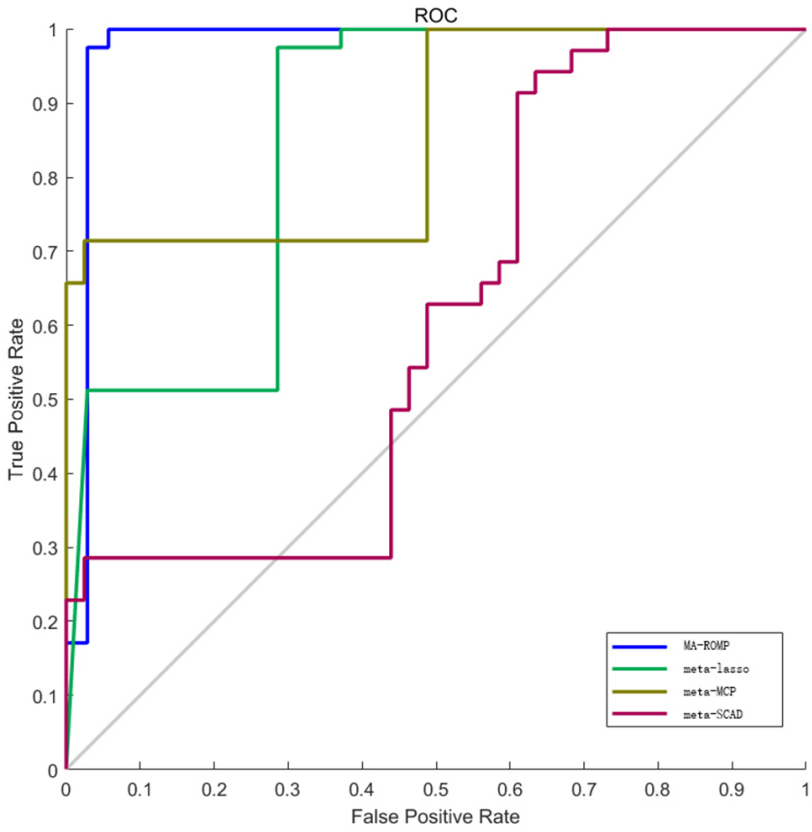

Figure 3. The ROC curves obtained by different methods with meta-analysis learning policy in three lung cancer datasets. The area under the curve (AUC) of MA-ROMP, meta-lasso, meta-MCP and meta-SCAD are 0.9756, 0.8551, 0.8592 and 0.6118 respectively.

Table 4. The genomic biomarkers selected by different meta-analysis-based methods in three lung cancer datasets.

Methods	GSE10072	GSE19188	GSE19804
MA-ROMP	COL11A1 COL10A1 CRISP1 MC5R GPM6A SPP1 ADAMTS8 LINC00216	COL11A1 COL10A1 CRISP1 MC5R GPM6A SPP1 ADAMTS8 LINC00216	COL11A1 COL10A1 CRISP1 GPM6A SPP1 ADAMTS8 LINC00216
meta-MCP	EDNRB CA4 GPM6A ADH1B TNNC1 AGER TMEM100	EDNRB CA4 GPM6A ADH1B TNNC1 AGER TMEM100	EDNRB CA4 GPM6A ADH1B TNNC1 AGER TMEM100
meta-SCAD	GREM1 COL11A1	GREM1 COL11A1	GREM1 COL11A1
meta-lasso	GPM6A AGER	GPM6A AGER	GPM6A AGER

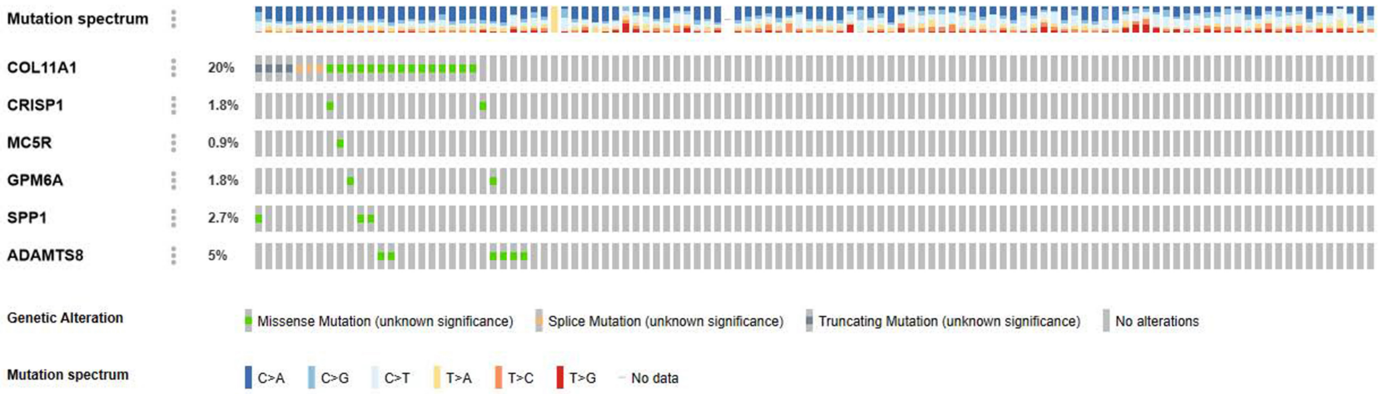

Figure 4. The highest ranked gene alterations in the lung caner dataset selected by MA-ROMP.

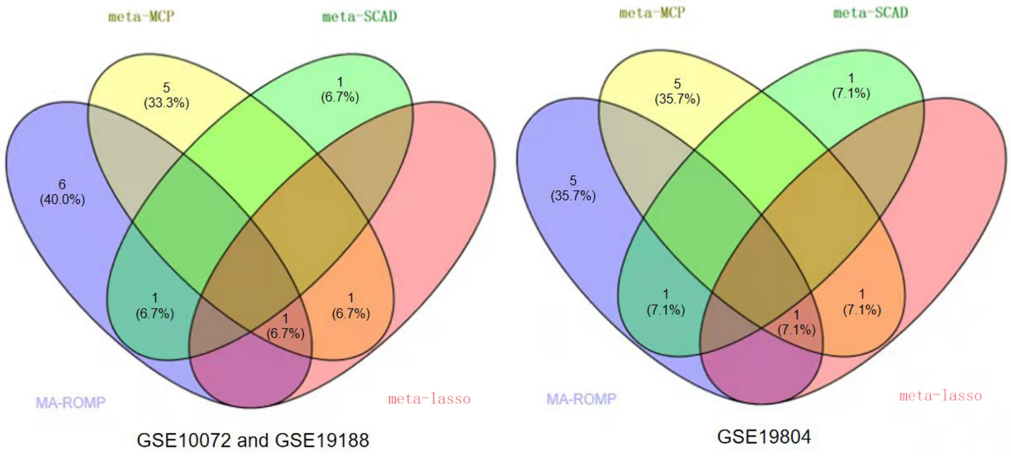

Figure 5. Overlap of commonly selected genes across the different meta-analysis-based methods in lung cancer datasets.

Another unique genomic biomarker selected by MA-ROMP is ADAMTS8, which can also inhibit lung cancer by targeting vascular endothelial growth factor (VEGFA) [32,33] with a maximum gene alteration of 5% in all patients. ADAMTS8 is under the regulation of GATA1, which has been linked to survival time in lung cancer patients [34]. The ADAMTS (A disintegrin and metalloproteinase with thrombospondin motifs) family members play important roles in tumor progression through regulating angiogenesis.

Although meta-lasso and meta-SCAD selected just two genomic biomarkers, GPM6A was selected by the meta-lasso, MA-ROMP and meta-MCP methods and alterations were detected in 1.8% of patients. GPM6A is a transmembrane protein widely distributed on the surface of neuronal cells in the central nervous system and is closely related to tumor growth [35]. Zhang et al. [36] identified that GPM6A downregulation promotes the progression of lung adenocarcinoma cells.

5. Conclusions

In this paper, we have proposed a meta-analysis-based regularized orthogonal matching pursuit (MA-ROMP) algorithm to recognize genomic biomarkers with both biological and clinical significance. This MA-ROMP method is suitable for combining studies from different platforms or conditions, and obtains reliable results with a small number of studies. It is pointed out that the successful running this method needs data standardization.

Compared with state-of-the-art methods, our method shows the highest accuracy of up to 95.63% with the best discriminative ability (AUC 0.9756) as well as a more than 15-fold decrease in its training time. Therefore, our method not only borrows strengths from across multiple datasets to identify important genomic biomarkers efficiently, but also maintains selection flexibility among datasets to take into account data heterogeneity through a hierarchical decomposition on regression coefficients. The experimental results demonstrate that our proposed MA-ROMP is a more effective tool for biomarker selection and learning prediction.

Author Contributions: S.W.: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, validation, visualization and writing—original draft; B.-Y.W.: conceptualization, formal analysis, investigation, validation and writing—review and editing; H.-F.L.: conceptualization, formal analysis, funding acquisition, methodology, project administration, resources, supervision and writing—review and editing. All authors gave final approval for publication and agreed to be held accountable for the work performed herein. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Natural Science Foundation of China (61976150), the Central Government's Guide to Local Science and Technology Development Fund (YDZJSX2021C005, YDZJSX2022A016), the Natural Science Foundation of Shanxi Province (20210302124272) and the Foundation of Taiyuan University of Technology (2022QN029).

Data Availability Statement: The datasets analyzed during the current study are available in NCBI's gene expression omnibus with the accession numbers GSE10072, GSE19188 and GSE19804.

Acknowledgments: The authors would like to thank members of the Brain Science and Intelligent Computing team at the College of Computer Science and Technology at the Taiyuan University of Technology (TYUT) for the useful discussions and support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bolón-Canedo, V.; Sánchez-Maroño, N.; Alonso-Betanzos, A. Feature selection for high-dimensional data. *Prog. Artif. Intell.* **2016**, *5*, 65–75. [CrossRef]
2. Siegel, R.L.; Miller, K.D.; Wagle, N.S.; Jemal, A. Cancer Statistics, 2023. *CA Cancer J. Clin.* **2023**, *73*, 17–48. [CrossRef] [PubMed]
3. Dokeroglu, T.; Deniz, A.; Kiziloz, H.E. A comprehensive survey on recent metaheuristics for feature selection. *Neurocomputing* **2022**, *494*, 269–296. [CrossRef]
4. Hu, L.; Yang, Y.; Tang, Z.; He, Y.; Luo, X. FCAN-MOPSO: An Improved Fuzzy-based Graph Clustering Algorithm for Complex Networks with Multi-objective Particle Swarm Optimization. *IEEE Trans. Fuzzy Syst.* **2023**, *14*, 1–16. [CrossRef]

5. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* **1996**, *58*, 267–288. [CrossRef]
6. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
7. Zhang, C.H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **2010**, *38*, 894–942. [CrossRef] [PubMed]
8. Yuan, M.; Lin, Y. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2006**, *68*, 49–67. [CrossRef]
9. Efron, B.; Hastie, T.; Johnstone, I.; Tibshirani, R. Least angle regression. *Ann. Stat.* **2004**, *32*, 407–499. [CrossRef]
10. Mallat, S.; Zhang, Z. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.* **1993**, *41*, 3397–3415. [CrossRef]
11. Tawfic, I.; Kayhan, S. Compressed sensing of ECG signal for wireless system with new fast iterative method. *Comput. Methods Programs Biomed.* **2015**, *122*, 437–449. [CrossRef] [PubMed]
12. Ji, C.; Zhang, X. Reguladzation orthogonal matclling pursuit based on multiple support. *Syst. Eng. Electron.* **2020**, *42*, 8.
13. Shi, X.; Xing, F.; Guo, Z.; Su, H.; Liu, F.; Yang, L. Structured orthogonal matching pursuit for feature selection. *Neurocomputing* **2019**, *349*, 164–172. [CrossRef]
14. Tsagris, M.; Papadovasilakis, Z.; Lakiotaki, K.; Tsamardinos, I. The γ-OMP algorithm for feature selection with application to gene expression data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2022**, *19*, 1214–1224. [CrossRef] [PubMed]
15. Toro-Domínguez, D.; Villatoro-García, J.A.; Martorell-Marugán, J.; Román-Montoya, Y.; Alarcón-Riquelme, M.E.; Carmona-Sáez, P. A survey of gene expression meta-analysis: Methods and applications. *Brief. Bioinform.* **2021**, *22*, 1694–1705. [CrossRef]
16. Huang, H.H.; Rao, H.; Miao, R.; Liang, Y. A novel meta-analysis based on data augmentation and elastic data shared lasso regularization for gene expression. *BMC Bioinform.* **2022**, *23*, 353. [CrossRef]
17. Li, Q.; Wang, S.; Huang, C.C.; Yu, M.; Shao, J. Meta-analysis based variable selection for gene expression data. *Biometrics* **2014**, *70*, 872–880. [CrossRef]
18. Zhang, H.; Li, S.J.; Zhang, H.; Yang, Z.Y.; Ren, Y.Q.; Xia, L.Y.; Liang, Y. Meta-Analysis Based on Nonconvex Regularization. *Sci. Rep.* **2020**, *10*, 5755. [CrossRef]
19. Hu, Z.; Zhou, Y.; Tong, T. Meta-Analyzing Multiple Omics Data With Robust Variable Selection. *Front. Genet.* **2021**, *12*, 1–16. [CrossRef]
20. Khosravy, M.; Gupta, N.; Patel, N.; Duque, C.A. Recovery in compressive sensing: A review. *Compressive Sens. Healthc.* **2020**, *2020*, 25–42.
21. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef] [PubMed]
22. Landi, M.T.; Dracheva, T.; Rotunno, M.; Figueroa, J.D.; Liu, H.; Dasgupta, A.; Mann, F.E.; Fukuoka, J.; Hames, M.; Bergen, A.W.; et al. Gene Expression Signature of Cigarette Smoking and Its Role in Lung Adenocarcinoma Development and Survival. *PLoS ONE* **2008**, *3*, e1651. [CrossRef] [PubMed]
23. Hou, J.; Aerts, J.; den Hamer, B.; van IJcken, W.; den Bakker, M.; Riegman, P.; van der Leest, C.; van der Spek, P.; Foekens, J.A.; Hoogsteden, H.C.; et al. Gene Expression-Based Classification of Non-Small Cell Lung Carcinomas and Survival Prediction. *PLoS ONE* **2010**, *5*, e10312. [CrossRef]
24. Lu, T.P.; Tsai, M.H.; Lee, J.M.; Hsu, C.P.; Chen, P.C.; Lin, C.W.; Shih, J.Y.; Yang, P.C.; Hsiao, C.K.; Lai, L.C.; et al. Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women. *Cancer Epidemiol. Biomarkers Prev.* **2010**, *19*, 2590–2597. [CrossRef] [PubMed]
25. Donoho, D.L.; Maleki, A.; Montanari, A. Message passing algorithms for compressed sensing: I. motivation and construction. In Proceedings of the 2010 IEEE Information Theory Workshop on Information Theory (ITW 2010), Cairo, Egypt, 6–8 January 2010; pp. 1–5.
26. Nallanthighal, S.; Heiserman, J.P.; Cheon, D.J. Collagen Type XI Alpha 1 (COL11A1): A Novel Biomarker and a Key Player in Cancer. *Cancers* **2021**, *13*, 935. [CrossRef]
27. Yi, X.; Luo, L.; Zhu, Y.; Deng, H.; Liao, H.; Shen, Y.; Zheng, Y. SPP1 facilitates cell migration and invasion by targeting COL11A1 in lung adenocarcinoma. *Cancer Cell Int.* **2022**, *22*, 324. [CrossRef]
28. Liu, Y.; Ye, G.; Dong, B.; Huang, L.; Zhang, C.; Sheng, Y.; Wu, B.; Han, L.; Wu, C.; Qi, Y. A pan-cancer analysis of the oncogenic role of secreted phosphoprotein 1 (SPP1) in human cancers. *Ann. Transl. Med.* **2022**, *10*, 279. [CrossRef]
29. Tang, H.; Chen, J.; Han, X.; Feng, Y.; Wang, F. Upregulation of SPP1 Is a Marker for Poor Lung Cancer Prognosis and Contributes to Cancer Progression and Cisplatin Resistance. *Front. Cell Dev. Biol.* **2021**, *9*, 646390. [CrossRef]
30. Zhang, Y.; Du, W.; Chen, Z.; Xiang, C. Upregulation of PD-L1 by SPP1 mediates macrophage polarization and facilitates immune escape in lung adenocarcinoma. *Exp. Cell Res.* **2017**, *359*, 449–457. [CrossRef]
31. Matsubara, E.; Yano, H.; Pan, C.; Komohara, Y.; Fujiwara, Y.; Zhao, S.; Shinchi, Y.; Kurotaki, D.; Suzuki, M. The Significance of SPP1 in Lung Cancers and Its Impact as a Marker for Protumor Tumor-Associated Macrophages. *Cancers* **2023**, *15*, 2250. [CrossRef]
32. Zhang, Y.; Hu, K.; Qu, Z.; Xie, Z.; Tian, F. ADAMTS8 inhibited lung cancer progression through suppressing VEGFA. *Biochem. Biophys. Res. Commun.* **2022**, *598*, 1–8. [CrossRef] [PubMed]

33. Wang, F.; Su, Q.; Li, C. Identidication of novel biomarkers in non-small cell lung cancer using machine learning. *Sci. Rep.* **2022**, *12*, 16693. [CrossRef] [PubMed]
34. Wang, L.L.; Chen, Z.S.; Zhou, W.D.; Shu, J.; Wang, X.H.; Jin, R.; Zhuang, L.L.; Hoda, M.A.; Zhang, H.; Zhou, G.P. Down-regulated GATA-1 up-regulates interferon regulatory factor 3 in lung adenocarcinoma. *Sci. Rep.* **2017**, *7*, 2551. [CrossRef] [PubMed]
35. Falch, C.M.; Sundaram, A.Y.M.; Øystese, K.A.; Normann, K.R.; Lekva, T.; Silamikelis, I.; Eieland, A.K.; Andersen, M.; Bollerslev, J.; Olarescu, N.C. Gene expression profiling of fast- and slow-growing non-functioning gonadotroph pituitary adenomas. *Eur. J. Endocrinol.* **2018**, *178*, 295–307. [CrossRef]
36. Zhang, Q.; Deng, S.; Li, Q.; Wang, G.; Guo, Z.; Zhu, D. Glycoprotein M6A Suppresses Lung Adenocarcinoma Progression via Inhibition of the PI3K/AKT Pathway. *J. Oncol.* **2022**, *2022*, 4601501. [CrossRef]