

## Article

# ClearF++: Improved Supervised Feature Scoring Using Feature Clustering in Class-Wise Embedding and Reconstruction

Sehee Wang <sup>1</sup> , So Yeon Kim <sup>1,2</sup>  and Kyung-Ah Sohn <sup>1,2,\*</sup> 

<sup>1</sup> Department of Artificial Intelligence, Ajou University, Suwon 16499, Republic of Korea; wsh0509@ajou.ac.kr (S.W.); jebi1771@ajou.ac.kr (S.Y.K.)

<sup>2</sup> Department of Software and Computer Engineering, Ajou University, Suwon 16499, Republic of Korea

\* Correspondence: kasohn@ajou.ac.kr

**Abstract:** Feature selection methods are essential for accurate disease classification and identifying informative biomarkers. While information-theoretic methods have been widely used, they often exhibit limitations such as high computational costs. Our previously proposed method, ClearF, addresses these issues by using reconstruction error from low-dimensional embeddings as a proxy for the entropy term in the mutual information. However, ClearF still has limitations, including a nontransparent bottleneck layer selection process, which can result in unstable feature selection. To address these limitations, we propose ClearF++, which simplifies the bottleneck layer selection and incorporates feature-wise clustering to enhance biomarker detection. We compare its performance with other commonly used methods such as MultiSURF and IFS, as well as ClearF, across multiple benchmark datasets. Our results demonstrate that ClearF++ consistently outperforms these methods in terms of prediction accuracy and stability, even with limited samples. We also observe that employing the Deep Embedded Clustering (DEC) algorithm for feature-wise clustering improves performance, indicating its suitability for handling complex data structures with limited samples. ClearF++ offers an improved biomarker prioritization approach with enhanced prediction performance and faster execution. Its stability and effectiveness with limited samples make it particularly valuable for biomedical data analysis.

**Keywords:** feature selection; feature scoring; information theory; entropy; mutual information (MI); dimension reduction; low-dimensional embedding; reconstruction error; principal component analysis (PCA); clustering



**Citation:** Wang, S.; Kim, S.Y.; Sohn, K.-A. ClearF++: Improved Supervised Feature Scoring Using Feature Clustering in Class-Wise Embedding and Reconstruction.

*Bioengineering* **2023**, *10*, 824.

<https://doi.org/10.3390/bioengineering10070824>

<https://doi.org/10.3390/bioengineering10070824>

Academic Editor: Luca Mesin

Received: 19 May 2023

Revised: 28 June 2023

Accepted: 4 July 2023

Published: 10 July 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

In the field of bioinformatics, accurate disease classification is crucial for effective diagnosis and treatment. Furthermore, the precise identification and selection of relevant biomarkers is essential to predict disease risk or aid in drug development [1]. As a result, a significant amount of research is currently being conducted in biomarker detection. Feature selection methods [2–4] are widely used in this context to identify and prioritize biomarkers from large and complex datasets [5,6]. These methods are particularly valuable in bioinformatics, where datasets often have a high number of features relative to the number of samples. By reducing the dimensionality of such data, feature selection can help identify the most informative biomarkers, facilitating accurate disease classification. Many feature selection algorithms have been developed to precisely select the most relevant biomarkers. This is crucial to better understand the underlying mechanisms of disease development and prognosis and to develop more targeted therapies.

Feature selection methods can be broadly classified into supervised and unsupervised approaches, where supervised approaches utilize class labels to identify relevant features, while unsupervised approaches do not [7]. As supervised approaches are more suitable for targeting specific diseases and finding relevant biomarkers, this study utilizes supervised feature selection methods for effective biomarker prioritization. There are various

supervised feature selection methods, such as statistical methods [8,9], similarity-based approaches [10,11], and information-theoretic methods. Information-theoretic methods perform feature selection by quantifying the amount of mutual information, which is a measure of entropy and conditional dependencies between data variables and their labels. Information-theoretic methods have been widely studied for feature selection, and their effectiveness has been demonstrated by promising experimental results [12–16]. Recently, innovative approaches like MI-VIF [17] have emerged, which combine variance inflation factor and mutual information, offering a solution to the collinearity problem that leads to unstable parameter estimation. In addition, a methodology named Relevance based on Weight Feature Selection (RWFS) [18] has been proposed. This method is based on two types of changed ratios in relation to feature relevance evaluation: one for the undetermined amount of information and the other for the established amount of information. These strategies have demonstrated their effectiveness by improving performances

However, these methods often suffer from high computational costs, and they may require discretization of continuous variables, which may lead to information loss [19]. To address these issues, we have previously proposed ClearF [20], which uses the reconstruction error of a low-dimensional embedding method as a proxy for the mutual information. ClearF assigns supervised scores to features by applying unsupervised class-wise low-dimensional embedding, which has been demonstrated to be effective in several benchmark datasets. However, ClearF has a limitation in that the selection process of the bottleneck layer is not transparent, requiring the selection of feature size in advance, followed by a greedy search. Consequently, the process can be complicated, unstable, and time-consuming, depending on the experimental setup. Furthermore, due to the partitioning of the entire dataset based on class labels and the subsequent embedding of each partition, the sample size becomes significantly smaller. This may introduce the risk of generating unstable outcomes during the feature selection process.

In this paper, we propose ClearF++ to address the limitations of ClearF. ClearF++ simplifies the process of determining the number of uncertain bottleneck layers and further improves performance through feature clustering. First, we propose a method to increase convenience and stabilize the process by simply fixing the number of bottleneck layers to a single value. In addition, we apply a feature-wise clustering method to mitigate the problem of embedding too many features at once and only reflecting the importance of a few features. This method allows for the selection of important features by clustering similar features together, thus reducing the number of embedded features. In summary, ClearF++ addresses the limitations of ClearF by simplifying the selection process of bottleneck layers and improving performance through feature clustering. Figure 1 illustrates the proposed architecture, and the entire process is shown in the pseudocode presented in Algorithm 1.

---

**Algorithm 1** Algorithm ClearF++: Supervised feature scoring method using feature clustering in the class-wise embedding and reconstruction method.

---

- 1: **function** CLEARF++( $X, Y, k, l$ )
  - 2:   **Input:**  
 $X = \{X_1, X_2, \dots, X_s\} \in \mathbb{R}^{n \times s}$ : Data matrix (n features and s samples)  
 $Y = \{y_1, y_2, \dots, y_s\}$ : Label vector  
 $k$ : Number of clusters  
 $l$ : Number of classes
  - 3:   **Output:**  
 $F = \{f_1, f_2, \dots, f_n\}$ : Feature scores
  - 4:   Perform feature-wise clustering on data  $X$ :
  - 5:   Apply DEC clustering method that divides n features into k clusters to obtain  $C = \{C_1, C_2, \dots, C_k\}$ , where each cluster  $C_i \in \mathbb{R}^{C_i \times s}$ .
-

**Algorithm 1** *Cont.*

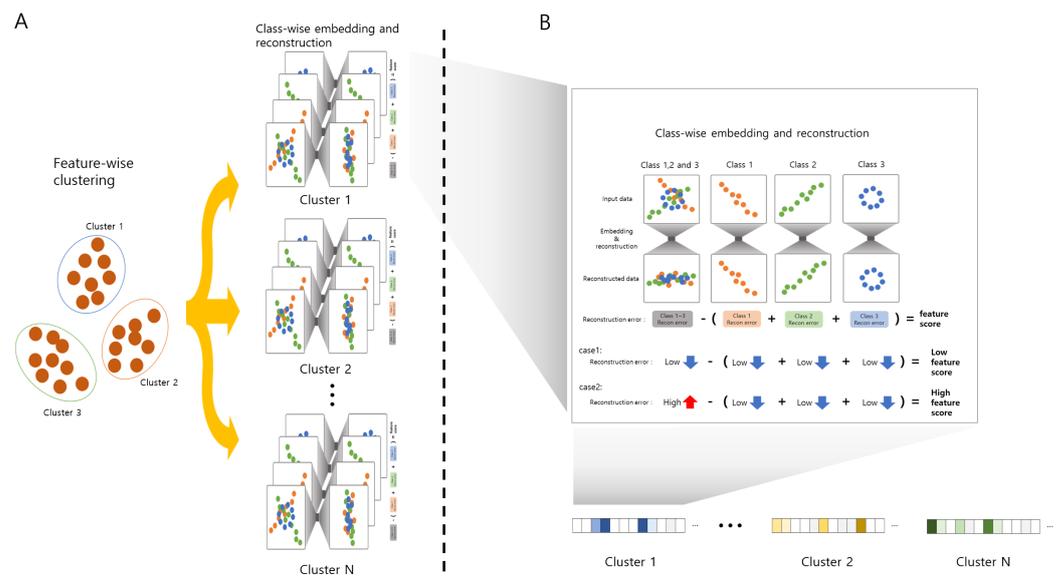
```

6:   for  $i = 1, \dots, k$  do
7:      $F_i \leftarrow \text{CLEARF}(C_i, Y, l, 1)$ 
8:   end for
9:   Aggregate feature scores for each cluster and rank features to obtain  $F$ 
10:  return  $F$ 
11: end function

12: function CLEARF( $X, Y, l, d$ )
13:  Input:
       $X = \{X_1, X_2, \dots, X_s\} \in \mathbb{R}^{n \times s}$ : Data matrix
       $Y = \{y_1, y_2, \dots, y_s\}$ : Label vector
       $l$ : Number of classes
       $d$ : Number of components
14:  Output:
       $F = \{f_1, f_2, \dots, f_n\}$ : Feature scores

15:  Using label vector  $Y$ , divide  $X$  into  $L = \{L_1, L_2, \dots, L_l\}$ , where each divided data
       $L_j \in \mathbb{R}^{n \times l_j}$ .
16:  Perform low-dimensional embedding on  $X$  with  $d$  components and reconstruct to
      calculate the feature-wise reconstruction error:
17:   $R_X = \{r_{(X,1)}, r_{(X,2)}, \dots, r_{(X,n)}\}$ 
18:  for  $j = 1, \dots, l$  do
19:    Perform low-dimensional embedding on  $L_j$  and reconstruct to calculate the
      feature-wise reconstruction error:
20:     $R_j = \{r_{(j,1)}, r_{(j,2)}, \dots, r_{(j,n)}\}$ 
21:  end for
22:   $R_{\text{sum}} = \text{sum}(R_1, R_2, \dots, R_l)$ 
23:   $F = R_X - R_{\text{sum}}$ 
24:  return  $F$ 
25: end function

```



**Figure 1.** Overview of ClearF++, a supervised feature scoring method that utilizes feature clustering in a class-wise embedding and reconstruction method: **(A)** Description of how the entirety of the data are divided into multiple partitions using feature-wise clustering. **(B)** Description of the process of calculating feature importance using ClearF++.

## 2. Materials and Methods

### 2.1. ClearF-One: Simplifying Bottleneck Layer Selection

To tackle the instability issue in the bottleneck layer selection process within ClearF, we propose a refined approach called ClearF-one. In this modified method, the bottleneck layer is set to a single layer rather than employing a greedy search to determine the optimal number of bottleneck layers. As the size of the bottleneck layer increases, a broader range of information is selected, resulting in the dilution of focusing important parts. By constraining the number of layers to one, only the most informative features from each class's embedding are selected, aligning with the theoretical foundation of ClearF. In summary, ClearF-one serves as an enhanced version of ClearF that addresses instability in the feature selection process by simplifying the bottleneck layer to a single layer.

### 2.2. ClearF++: Advanced Feature Selection via Feature-Wise Clustering

As described above, ClearF-one fixes the bottleneck layer to a single layer, resulting in features with strong signals for each class that are likely to have high scores. However, ClearF-one can be disadvantageous in selecting multiple features due to the limited amount of expressed information. To overcome this limitation, we propose a novel method that divides the data into several partitions through feature-wise clustering and applies ClearF-one to each cluster. As shown in Figure 1A, feature-wise clustering is performed to divide the data into units of each cluster with similar features. When we perform feature-wise clustering, the Deep Embedding Clustering (DEC) method [21] is applied, which is a method of unsupervised learning that combines deep neural networks with clustering algorithms. Next, ClearF-one is applied to each of the clustered data to calculate the feature score. Finally, ClearF++ produces a high feature score when it exhibits a significant difference between classes, such as in case 2 of Figure 1B. Features with no significant difference between classes, such as case 1, are not scored high. This approach allows the most informative features to produce high scores by calculating a class-wise reconstruction error. The above process is performed for each cluster, as depicted in Figure 1B, extracting features that encapsulate important characteristics unique to each cluster.

Our proposed method ClearF++ has several advantages over ClearF and ClearF-one. It is particularly useful when an appropriate number of features must be selected from data with a large number of features, such as in biomarker identification. Additionally, it can be applied when the number of samples is too small compared with the number of features, making it difficult to learn ClearF stably. In summary, ClearF++ divides the data into feature-wise clusters and applies ClearF-one to each cluster, enabling us to select multiple informative features from a large number of features.

## 3. Results

### 3.1. Datasets

We conducted an experiment on the gene expression data of lung cancer patients using the ARCHS4 dataset [22], which has been used in several studies [23,24]. We removed genes that had more than 25% zero expression across all samples. The experiment was tested with 8710 genes and 3079 samples. Out of the 3079 samples, 1158 samples belong to the A549 cells (non-small-cell lung cancer) and 1921 samples belong to the IMR90 cells (normal lung fibroblast). Additionally, we performed experiments on several benchmark datasets. To externally validate our results, we further conducted the experiments over two additional benchmark datasets, colon and ALL/AML leukemia datasets [7]. The ALL/AML dataset consists of 72 samples, 47 samples belong to acute lymphoblastic leukemia (ALL) and 25 samples belong to acute myeloid leukemia (AML) [25]. Gene expression levels were measured using Affymetrix high-density oligonucleotide arrays containing 7129 genes. The colon dataset consists of 62 samples, of which 22 are normal and 40 are colon tumor tissue samples [26]. Gene expression levels were measured using Affymetrix oligonucleotide arrays containing expression levels for the 2000 genes with the highest minimal intensity across the samples, as it is prepared in the paper [26].

### 3.2. Performance Evaluation on Multiple Benchmark Datasets

We compared the performance of several feature selection algorithms, including MultiSURF, IFS, and ClearF, to demonstrate the effectiveness of our proposed method in extracting the most relevant features for lung cancer classification using the ARCHS4 dataset. The selected features from each method were used for classification and their AUCs were compared. We performed 10-fold cross-validation by dividing the entire dataset into 10 folds, with one fold for test data and the remaining folds for training data. Each feature selection algorithm was applied solely to the training data to select important features. A classification algorithm was then applied using only the selected features, and the average AUC of the 10-fold cross-validation was measured. The classification model is a basic four-layer DNN, consisting of an input layer, two hidden layers, and an output layer. The sizes of the hidden layers were determined as the number of selected features \* 2 and the number of selected features, respectively. The hyperbolic tangent served as the activation function. The Adam optimizer was employed for learning with a learning rate of  $1 \times 10^{-3}$  and 500 epochs using a full batch.

To validate the stability and effectiveness of our proposed method, ClearF++, we conducted a performance comparison using the ARCHS4 lung cancer dataset and several benchmark datasets, such as colon and ALL/AML. The results are displayed in Table 1. These experiments demonstrate that ClearF++ mostly outperforms ( $p$ -value < 0.05) other feature selection methods, such as MultiSURF [11], IFS [27], and ClearF [20], across the ARCHS4, colon, and ALL/AML datasets. ClearF++ achieved the highest performance across most of the feature subsets, reaching its best performance at 60 features both in the colon dataset (AUC = 0.826) and the ARCHS4 lung dataset (AUC = 0.983). Likewise, in the ALL/AML dataset, excluding the comparison with ClearF when the number of features was 60, ClearF++ outperformed other methods ( $p$ -value < 0.05), achieving the best performance at both 45 and 60 features (AUC = 0.949). Overall, these results highlight the consistent and enhanced performance of ClearF++ across varying numbers of features, showing the robustness and effectiveness of ClearF++. To show the statistical significance of the improvement, we included the results of a paired  $t$ -test between ClearF++ and other methods in Table S2 of the Supplementary Material, aligning with the results in Table 1. The results predominantly affirmed the notable superiority of our proposed method compared with other methods, with the exceptions of the case where ClearF++ vs. ClearF selected 60 features in the ALL/AML dataset, and the case where 45 features were selected in the colon dataset.

**Table 1.** Performance comparison on several benchmark datasets across a varying number of features. The performance was measured with the average AUC of 10-fold cross-validation.

n <sup>1</sup>	Colon				ALL/AML				ARCHS4			
	MultiSURF	IFS	ClearF	ClearF++	MultiSURF	IFS	ClearF	ClearF++	MultiSURF	IFS	ClearF	ClearF++
15	0.648	0.749	0.707	<b>0.805</b>	0.906	0.913	0.912	<b>0.947</b>	0.925	0.927	0.949	<b>0.959</b>
30	0.711	0.672	0.765	<b>0.773</b>	0.927	0.926	0.921	<b>0.940</b>	0.944	0.953	0.943	<b>0.973</b>
45	0.703	0.658	<b>0.801</b>	0.751	0.938	0.915	0.927	<b>0.949</b>	0.955	0.955	0.952	<b>0.980</b>
60	0.723	0.761	0.815	<b>0.826</b>	0.927	0.915	<b>0.949</b>	<b>0.949</b>	0.969	0.949	0.951	<b>0.983</b>

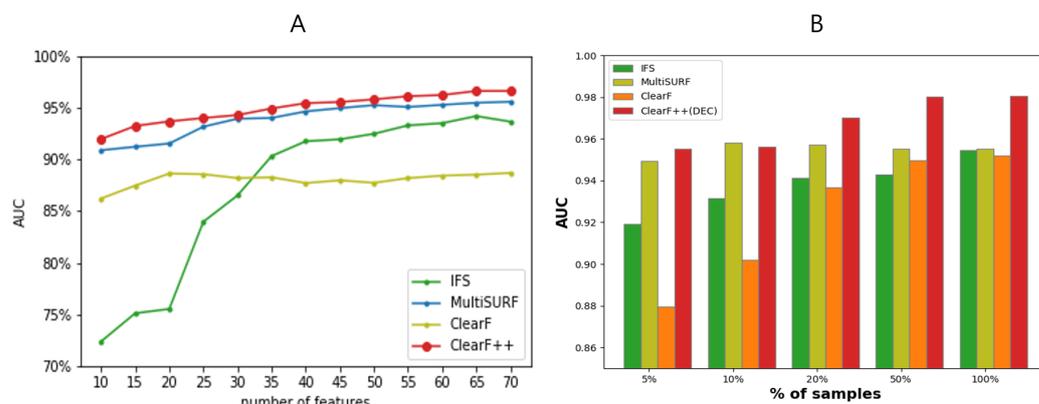
<sup>1</sup> The number of features.

### 3.3. Performance Evaluation Across Varying Feature and Sample Sizes

To assess the stability of our proposed method with a limited number of data samples, we evaluated lung cancer classification performance using only 5% of the training data samples from the ARCHS4 dataset. The number of features to be selected increased by 5, starting from 10, in accordance with the experimental procedure employed in the previous study [20].

Figure 2 presents the experimental results on the ARCHS4 lung cancer dataset, using only 5% of the training samples. Our proposed method, ClearF++, exhibited superior performance compared with ClearF and other comparable feature selection algorithms. As displayed in Figure 2A, we observed that the stability of ClearF++ was preserved, while

other methods yielded relatively unstable and poor performances with a limited number of samples.



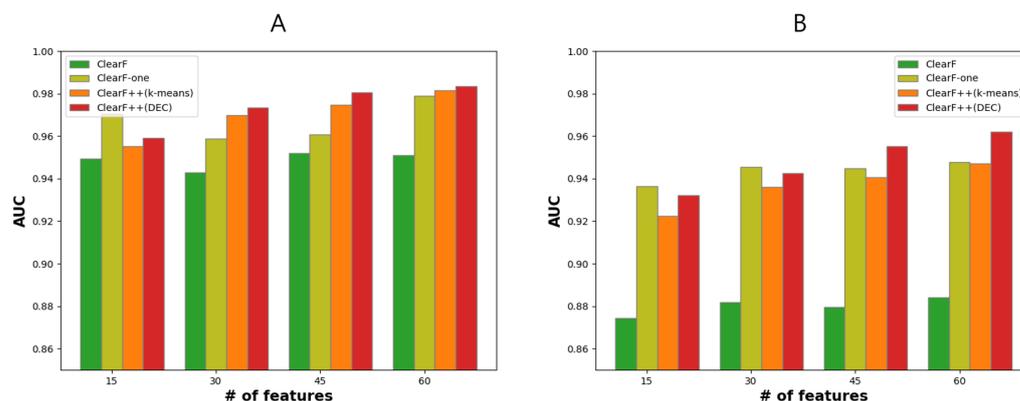
**Figure 2.** (A) Performance comparison across varying numbers of features between ClearF++ and other algorithms. The experiment was obtained using only 5% of the sample for training. (B) A comparative experiment measuring performances as the number of samples changes. The experiment was conducted by fixing the number of features at 45.

In Figure 2B, we conducted experiments with varying numbers of samples to investigate the stability of ClearF++ across different sample sizes, fixing the number of features at 45. The experimental results reveal that ClearF++ and MultiSURF showed stable and improved performance across varying numbers of features, even with a small sample size. However, ClearF++ outperformed other methods when more than 10% of the samples were used, whereas MultiSURF exhibited no improvement when larger sample sizes were used and even suffered from slight performance degradation.

These results indicate that our proposed method, ClearF++, demonstrates impressive stability and performance even when dealing with a limited number of data samples. In comparison with other feature selection algorithms, ClearF++ consistently outperforms them, particularly when utilizing more than 10% of the training samples. This highlights the robustness and effectiveness of ClearF++ in various feature or sample size scenarios. These findings emphasize the potential of ClearF++ as a robust and effective feature selection technique, capable of maintaining its performance across a range of feature or sample sizes.

### 3.4. Effect of Feature-Wise Clustering Algorithms

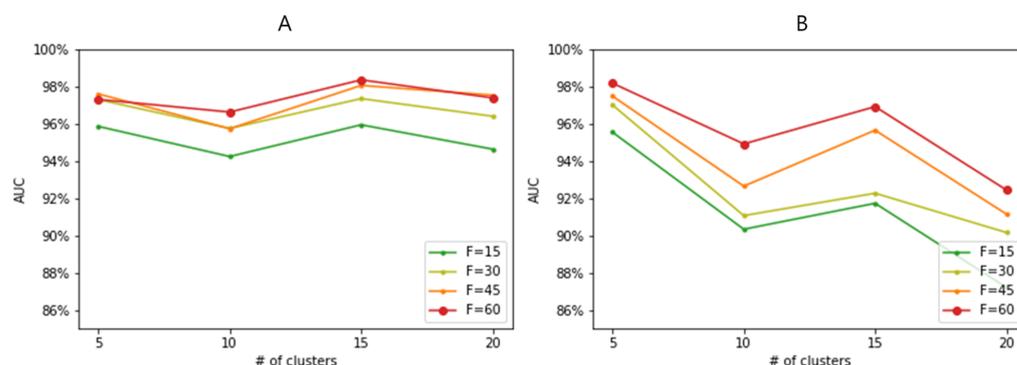
We conducted an ablation study to confirm that our proposed method, ClearF++, indeed contributes to performance improvement compared with the previously proposed method, ClearF. For low-dimensional embedding, KernelPCA with an RBF kernel that showed the best performance in ClearF was utilized in ClearF-one and ClearF++. For the clustering method in ClearF++, k-means and DEC [21] were used. The ablation study was conducted on the ARCHS4 lung cancer dataset, with the results shown in Figure 3. Figure 3A displays the results for the entire samples, while Figure 3B presents the result using only 5% of the samples for training.



**Figure 3.** Performance evaluation of ClearF-based methods with various experimental settings in a lung cancer dataset. ClearF-one limits the number of bottleneck layers to 1, and ClearF++ applies feature-wise clustering in the proposed algorithm; thus, two clustering algorithms were compared. (A) Results using entire samples. (B) Results using 5% of samples for training.

The experimental results show that ClearF-one, which substantially restricts the number of bottleneck layers, contributes to performance improvement, particularly when only a small portion of samples (5%) is used for training. This suggests that ClearF-one is effective in handling limited data samples and can still yield improved performance by simplifying the architecture, reducing the complexity, and focusing on the most relevant features. It is noteworthy that in the context of k-means clustering, the results yielded from training with only 5% of the samples shortly underperformed in comparison with those obtained from ClearF-one without clustering. Under the constraints of a small data size, k-means clustering appeared to struggle in achieving effective clustering. However, when we employed the more sophisticated clustering algorithm, DEC, we observed stable and enhanced performances when feature-wise clustering was applied. Particularly, DEC outperformed k-means clustering in terms of performance when a larger number of features were being selected, both in the scenarios with all samples in Figure 3A and with a small number of samples in Figure 3B. This suggests that DEC may be more suitable for handling complex data structures and capturing underlying patterns in the data when compared with the k-means clustering algorithm, particularly in situations with limited data samples.

We evaluated ClearF++ performance by employing two different clustering methods with varying numbers of clusters. Figure 4A presents the results using the DEC clustering algorithm, while Figure 4B shows the results using k-means clustering. The results reveal that DEC with 15 clusters yielded the best performance, while k-means clustering achieved optimal results with 5 clusters. Although the best performances of these two clustering methods were comparable (close to AUC = 0.98), k-means clustering showed considerable variance depending on the number of clusters. In contrast, DEC demonstrated smaller variance and consistently higher performance across different cluster numbers. Consequently, while both clustering methods (DEC and k-means) can achieve comparable performances, DEC not only improved performances but showed more consistent and reliable results across different numbers of clusters. It suggests that the choice of clustering algorithm is important and that the proposed feature-wise clustering idea of ClearF++ contributes to more robust and improved performances.



**Figure 4.** Performance evaluation based on varying numbers of clusters in the lung cancer dataset. F represents the number of selected features. (A) Performances of ClearF++ when DEC clustering is applied. (B) Performances of ClearF++ when k-means clustering is applied.

### 3.5. Computational Cost Validation

To verify the computational efficiency of ClearF++ over other methods, we measured and compared the CPU time by running each method 10 times. The experiment utilized data containing 5000 randomly generated features and 5000 samples. As shown in Table 2, ClearF++ showed a substantially faster execution time than MultiSURF and IFS. While it showed a slightly slower execution time than ClearF due to the inclusion of clustering time, the improved performance ensures its competitiveness and effectiveness in comparison with other methods.

**Table 2.** Computational costs comparison of ClearF++ and other feature selection methods.

Methods	IFS	MultiSURF	ClearF	ClearF++
CPU times (s)	398.91 ± 17.22	78,515.59 ± 346.53	132.68 ± 6.51	166.85 ± 10.77

### 3.6. Functional Enrichment Analysis

To identify high-scoring features, we analyzed the ARCHS4 dataset using ClearF++, which demonstrated improved performance in our experiments. Since scores were calculated for each of the 10 folds, we defined an integrated scoring method. The top 100 features in each fold were assigned scores in descending order, ranging from 100 to 1 point. We then calculated the rank scores by averaging the scores obtained across all folds. The 50 highest-scoring genes are shown in Table S1. Furthermore, to investigate the biological relationships among the selected genes, we performed pathway and gene ontology enrichment analysis using ToppGene [28] on the top 50 genes. The results are shown in Table 3, respectively.

In biomarker detection, high classification accuracy of selected features does not guarantee that features associated with the disease are selected. Considering that the purpose of our algorithm is to select features for identifying important biomarkers, it is crucial to determine whether the top-scoring features are associated with the target disease. The 50 genes with high scores in our method are shown in Table S1. We performed enrichment analysis using these 50 genes.

The enrichment analysis results conducted with ToppGene [28] reveal that several genes related to the glycosaminoglycan (GAG) metabolism pathway received high scores. The extracellular matrix (ECM) regulates cell fate, and glycosaminoglycans (GAGs) are major macromolecules that compose the ECM, which play well-known roles in cancer angiogenesis, proliferation, invasion, and metastasis [29]. GAGs have been widely studied as treatments for cancer, inflammation, infection, and lung diseases, and one study [30] clarified the role of GAGs, contributing to future research.

**Table 3.** Pathway and gene ontology enrichment analysis results using ToppGene on the top 50 ranked genes. The 10 most significant gene ontology (GO) terms that have the lowest p-values are shown, as well as pathway and disease terms with significant p-values ( $p < 0.05$ ) from the enrichment analysis.

Category	ID	Name	p-Value	q-Value <sup>1</sup>	q-Value <sup>2</sup>	HC <sup>3</sup>	HCG <sup>4</sup>
BP	GO:0032963	collagen metabolic process	$2.89 \times 10^{-8}$	$5.13 \times 10^{-5}$	$5.13 \times 10^{-5}$	7	144
BP	GO:0030042	actin filament depolymerization	$5.03 \times 10^{-7}$	$8.95 \times 10^{-4}$	$3.68 \times 10^{-4}$	5	71
BP	GO:0032964	collagen biosynthetic process	$6.63 \times 10^{-7}$	$1.18 \times 10^{-3}$	$3.68 \times 10^{-4}$	5	75
MF	GO:0044877	protein-containing complex binding	$7.35 \times 10^{-7}$	$2.09 \times 10^{-4}$	$2.09 \times 10^{-4}$	16	1726
BP	GO:0001568	blood vessel development	$8.27 \times 10^{-7}$	$1.47 \times 10^{-3}$	$3.68 \times 10^{-4}$	13	1152
BP	GO:0035904	aorta development	$1.47 \times 10^{-6}$	$2.62 \times 10^{-3}$	$4.71 \times 10^{-4}$	5	88
BP	GO:0001944	vasculature development	$1.87 \times 10^{-6}$	$3.33 \times 10^{-3}$	$4.71 \times 10^{-4}$	13	1239
BP	GO:0030198	extracellular matrix organization	$2.21 \times 10^{-6}$	$3.93 \times 10^{-3}$	$4.71 \times 10^{-4}$	8	394
BP	GO:0043062	extracellular structure organization	$2.25 \times 10^{-6}$	$4.00 \times 10^{-3}$	$4.71 \times 10^{-4}$	8	395
BP	GO:0045229	external encapsulating structure organization	$2.38 \times 10^{-6}$	$4.24 \times 10^{-3}$	$4.71 \times 10^{-4}$	8	398
Disease	C0268362	Osteogenesis imperfecta type III (disorder)	$1.90 \times 10^{-6}$	$3.42 \times 10^{-3}$	$3.42 \times 10^{-3}$	3	11
Pathway	1269980	Heparan sulfate/heparin (HS-GAG) metabolism	$1.82 \times 10^{-5}$	$5.99 \times 10^{-3}$	$1.82 \times 10^{-3}$	4	54
Pathway	1309217	Defective B3GALT6 causes EDSP2 and SEMDJL1	$2.21 \times 10^{-5}$	$7.26 \times 10^{-3}$	$1.82 \times 10^{-3}$	3	19
Pathway	1269015	Defective B3GAT3 causes JDSSDHD	$2.21 \times 10^{-5}$	$7.26 \times 10^{-3}$	$1.82 \times 10^{-3}$	3	19
Pathway	1269014	Defective B4GALT7 causes EDS, progeroid type	$2.21 \times 10^{-5}$	$7.26 \times 10^{-3}$	$1.82 \times 10^{-3}$	3	19
Pathway	1269981	A tetrasaccharide linker sequence is required for GAG synthesis	$5.84 \times 10^{-5}$	$1.92 \times 10^{-2}$	$3.20 \times 10^{-3}$	3	26
Pathway	1269011	Diseases associated with glycosaminoglycan metabolism	$5.84 \times 10^{-5}$	$1.92 \times 10^{-2}$	$3.20 \times 10^{-3}$	3	26
Pathway	1269982	HS-GAG biosynthesis	$9.99 \times 10^{-5}$	$3.29 \times 10^{-2}$	$4.69 \times 10^{-3}$	3	31
Pathway	M39870	Type I collagen synthesis in the context of osteogenesis imperfecta	$1.21 \times 10^{-4}$	$3.97 \times 10^{-2}$	$4.97 \times 10^{-3}$	3	33
Pathway	1268756	Unfolded Protein Response (UPR)	$1.48 \times 10^{-4}$	$4.88 \times 10^{-2}$	$5.42 \times 10^{-3}$	4	92

<sup>1</sup> Bonferroni q-value, <sup>2</sup> FDR B&H q-value, <sup>3</sup> Hit Count in the query list, <sup>4</sup> Hit count in the genome.

Among the GAG-associated genes that received high scores in our method are GPC1, NDST1, CSPG4, and SDC3. An experiment involving CSPG4-specific mAb 225.28 demonstrated the regression induction of tumor metastasis in a lung metastasis model [31]. Endothelial cell (ECs) junction disassembly, a key step in inflammation, allows for vascular leakage during disease, and thrombin-cleaved fragments of the SDC3 ectodomain promote this process in human lung microvessels in certain cases [32]. NDST1 participates in the synthesis of the heparan sulfate (HS) chain of HSPG, and a study [33] found that it may provide an explanation for the clinical observation that heparin can improve outcomes in small-cell lung cancer (SCLC). Another study [34] suggested that NDST1 is associated with angiogenesis and tumor growth in lung tumors. There is also a study that recommended the use of glypican-1 (GPC1) as an additional positive marker for lung squamous cell carcinoma [35]. These findings suggest that analyzing the effects of NDST1 and SDC3 expression on pulmonary blood vessels in relation to GAGs may be helpful in diagnosing and treating lung cancer.

Additionally, enrichment analysis results using the top 50 genes with high scores (Table 3) reveal that 7 genes related to the collagen metabolic process are included in the biological process. These genes are P3H3, MRC2, MMP14, ENG, EMILIN1, CREB3L1, and COL1A1, with CREB3L1, EMILIN1, and COL1A1 ranking 1st, 3rd, and 4th, respectively. Impaired collagen metabolism is accompanied by increased prolidase activity in lung cancer squamous epithelium [36]. Furthermore, idiopathic pulmonary fibrosis (IPF) is associated with an increased risk of lung cancer with elevated collagen and prolidase activity [36–38]. On the other hand, Prolidase Deficiency (PD) and osteogenesis imperfecta (OI) share similar phenotypes [37]. Notably, our enrichment analysis includes CREB3L1 and COL1A1, which, out of the top 10 genes, are associated with osteogenesis imperfecta type III (disorder). The expression of  $\beta 1$  integrin, which has been shown to regulate prolidase activity, is decreased in OI [37,39,40]. However, there is no difference in the levels of  $\beta 1$  integrin between healthy lung cells and cancer cells, suggesting that prolidase regulation in lung cancer may involve

a different mechanism [36,37]. Therefore, studying the role of prolidase, CREB3L1, and COL1A1 gene expression in lung cancer appears to be significant.

In addition to the aforementioned genes, we found several high-scoring genes (shown in Table S1) that have been linked to lung cancer in multiple studies. The CREB3L1 gene has been associated with lung cancer growth due to its involvement in the activation of alpha-smooth muscle actin ( $\alpha$ -SMA)-positive cancer-associated fibroblasts (CAFs) [41]. SYDE1 is associated with epithelial–mesenchymal transition (EMT) reversal, which is associated with the progression of various tumors, including lung cancer [42]. Reduced EMILIN-1 production in some tumor types is associated with higher proliferation of tumor cells in breast and lung cancer [43]. Another study [44] suggested that COL1A1 can be a potential biomarker for poor progression-free survival and chemoresistance in metastatic lung cancer. Serum CKAP4 levels can distinguish lung cancer patients from healthy controls, making it a potential serum diagnostic marker for lung cancer [45]. Carbohydrates associated with LAMP1 play a crucial role in determining lung metastasis [46]. A potential target of TAF15 concerning resistance to radiotherapy, essential for non-small-cell lung cancer treatment, has been proposed [47]. TBX2 subfamily methylation may serve as a potential biomarker for early detection and intervention in non-small-cell lung cancer [48]. Consequently, the genes selected by our method are shown to be biomarker candidates for lung cancer.

#### 4. Discussion

We evaluated the suitability of our methodology for biomarker detection from a machine learning perspective. The results in Figure 2 demonstrate that our proposed method is effective when selecting a small number of features. Particularly, when combined with Figure 3, ClearF-one generally yields favorable results when selecting from 10 to 25 features, and ClearF++ shows improvement when selecting from 30 to 50 features. Given the importance of selecting a small number of features in biomarker discovery, our method can be considered suitable. Moreover, our approach demonstrates stable performance even with a small sample size, as shown in Figure 2B. In the biomedical field, insufficient learning samples are often encountered, and our method proves effective in such cases. Additionally, as shown in Table 2, our method can be effectively employed in environments with limited computational power due to its advantageous execution time.

The results in Figure 3 indicate that ClearF++ can show degraded performances compared with the model without clustering (ClearF-one) when the number of features is small. This is likely because ClearF-one effectively selects a small number of features when only certain information remains after embedding the entirety of the data into a single bottleneck layer. However, when the number of features increases, it suffers from performance degradation due to information loss. In contrast, ClearF++ extracts information for each cluster, which provides more robust and improved performances when selecting multiple features.

Our method addresses the sensitivity issue related to the number of bottleneck layers in the previously proposed method, ClearF, but still requires many parameter adjustments. In particular, determining the number of clusters remains a challenge in clustering. Although Figure 4 shows that our method is not highly sensitive to the number of clusters within a range of 5 to 20 clusters, an exceptionally higher number of clusters, such as 50 or 100, led to instability in clustering results and a substantial performance decrease. Through our experiments, we discerned that the optimal number of clusters likely resides within the 5 to 20 range. However, this range may vary with different datasets according to their sample sizes. Accordingly, future research could focus on the automatic selection of the number of clusters. Further, there is still an issue in setting the model structure or learning method in the part that utilizes DEC. This issue can be addressed in future studies. Furthermore, our method exhibits flexibility towards a range of clustering algorithms. Our experimental findings, as illustrated in Figure 3, indicated that employing a more sophisticated clustering technique could result in more stable and improved performances. Thus, future research that utilizes advanced clustering methods could potentially enhance performances.

## 5. Conclusions

In this study, we developed an improved feature selection algorithm for identifying biomarkers that can be used for disease prediction and biomedical data analysis. Our experimental results demonstrate several advantages of our method, including improved prediction performance and faster execution. Furthermore, it shows substantially stable performance even with a limited number of samples, making it particularly effective for biomedical data analysis, where the available sample size is often insufficient.

One limitation of our method is that it requires determining the optimal number of clusters, which can vary across different datasets. In this study, we experimented with several scenarios to select the most appropriate parameters. However, automatic parameter selection methods can be exploited in future work to address this issue.

**Supplementary Materials:** The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/bioengineering10070824/s1>, Table S1: List of the highest-scoring 50 genes in the ARSCH4 lung cancer dataset; Table S2: Results of statistical significance tests between ClearF++ and three other methods: MultiSURF, IFS, and ClearF, corresponding to the results presented in Table 1. The  $p$ -value was measured through a paired  $t$ -test between ClearF++ and other methods across three different datasets.

**Author Contributions:** Conceptualization, S.W.; methodology, S.W. and S.Y.K.; software, S.W.; validation, S.W. and S.Y.K.; formal analysis, S.Y.K. and K.-A.S.; investigation, S.W., S.Y.K. and K.-A.S.; resources, S.W.; writing—original draft preparation, S.W. and S.Y.K.; writing—review and editing, S.Y.K. and K.-A.S.; visualization, S.W. and S.Y.K.; supervision, K.-A.S.; project administration, K.-A.S.; funding acquisition, K.-A.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(NRF-2022R1A2C1007434), and also by the Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00255968) grant funded by the Korea government(MSIT).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** All data that are not presented in the main paper are available from the corresponding author on request.

**Conflicts of Interest:** The authors declare that they have no conflict of interest.

## References

1. Group, B.D.W.; Atkinson, A.J., Jr.; Colburn, W.A.; DeGruttola, V.G.; DeMets, D.L.; Downing, G.J.; Hoth, D.F.; Oates, J.A.; Peck, C.C.; Schooley, R.T.; et al. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin. Pharmacol. Ther.* **2001**, *69*, 89–95. [[CrossRef](#)]
2. Lee, I.H.; Lushington, G.H.; Visvanathan, M. A filter-based feature selection approach for identifying potential biomarkers for lung cancer. *J. Clin. Bioinform.* **2011**, *1*, 11. [[CrossRef](#)]
3. Abeel, T.; Helleputte, T.; Van de Peer, Y.; Dupont, P.; Saeys, Y. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics* **2010**, *26*, 392–398. [[CrossRef](#)] [[PubMed](#)]
4. Le, T.T.; Blackwood, N.O.; Taroni, J.N.; Fu, W.; Breitenstein, M.K. Integrated machine learning pipeline for aberrant biomarker enrichment (i-mAB): characterizing clusters of differentiation within a compendium of systemic lupus erythematosus patients. *AMIA Annu. Symp. Proc.* **2018**, *2018*, 1358–1367. [[PubMed](#)]
5. He, Z.; Yu, W. Stable feature selection for biomarker discovery. *Comput. Biol. Chem.* **2010**, *34*, 215–225. . [[CrossRef](#)] [[PubMed](#)]
6. Hemphill, E.; Lindsay, J.; Lee, C.; Mändoiu, I.I.; Nelson, C.E. Feature selection and classifier performance on diverse biological datasets. *BMC Bioinform.* **2014**, *15* (Suppl. S13), S4. [[CrossRef](#)]
7. Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R.P.; Tang, J.; Liu, H. Feature Selection: A Data Perspective. *ACM Comput. Surv.* **2017**, *50*. [[CrossRef](#)]
8. Davis, J.C. *Statistics and Data Analysis in Geology*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1973.
9. Hall, M.A. Correlation-Based Feature Selection for Machine Learning. Ph.D. Thesis, The University of Waikato, Hamilton, New Zealand, 1999.

10. Robnik-Šikonja, M.; Kononenko, I. Theoretical and Empirical Analysis of Relief and RRelief. *Mach. Learn.* **2003**, *53*, 23–69. [[CrossRef](#)]
11. Urbanowicz, R.J.; Olson, R.S.; Schmitt, P.; Meeker, M.; Moore, J.H. Benchmarking relief-based feature selection methods for bioinformatics data mining. *J. Biomed. Inform.* **2018**, *85*, 168–188. [[CrossRef](#)]
12. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [[CrossRef](#)]
13. Fleuret, F. Fast Binary Feature Selection with Conditional Mutual Information. *J. Mach. Learn. Res.* **2004**, *5*, 1531–1555.
14. Leem, S.; hwan Jeong, H.; Lee, J.; Wee, K.; Sohn, K.A. Fast detection of high-order epistatic interactions in genome-wide association studies using information theoretic measure. *Comput. Biol. Chem.* **2014**, *50*, 19–28. Advances in Bioinformatics: Twelfth Asia Pacific Bioinformatics Conference (APBC2014). [[CrossRef](#)]
15. Zhou, H.; Wang, X.; Zhu, R. Feature selection based on mutual information with correlation coefficient. *Appl. Intell.* **2022**, *52*, 5457–5474. [[CrossRef](#)]
16. Al-Sarem, M.; Saeed, F.; Alkhamash, E.H.; Alghamdi, N.S. An aggregated mutual information based feature selection with machine learning methods for enhancing iot botnet attack detection. *Sensors* **2022**, *22*, 185. [[CrossRef](#)] [[PubMed](#)]
17. Cheng, J.; Sun, J.; Yao, K.; Xu, M.; Cao, Y. A variable selection method based on mutual information and variance inflation factor. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2022**, *268*, 120652. [[CrossRef](#)] [[PubMed](#)]
18. Hu, L.; Gao, L.; Li, Y.; Zhang, P.; Gao, W. Feature-specific mutual information variation for multi-label feature selection. *Inf. Sci.* **2022**, *593*, 449–471. [[CrossRef](#)]
19. Ohyr-Nielsen, M. *Loss of Information by Discretizing Hydrologic Series*; Colorado State University Hydrology Papers; Colorado State University: Fort Collins, CO, USA, 1972.
20. Wang, S.; Jeong, H.H.; Sohn, K.A. ClearF: a supervised feature scoring method to find biomarkers using class-wise embedding and reconstruction. *BMC Med. Genom.* **2019**, *12*, 95. [[CrossRef](#)]
21. Xie, J.; Girshick, R.; Farhadi, A. Unsupervised Deep Embedding for Clustering Analysis. In Proceedings of the 33rd International Conference on International Conference on Machine Learning–ICML’16, New York, NY, USA 19–24 June 2016; Volume 48, pp. 478–487.
22. Lachmann, A.; Torre, D.; Keenan, A.B.; Jagodnik, K.M.; Lee, H.J.; Wang, L.; Silverstein, M.C.; Ma’ayan, A. Massive mining of publicly available RNA-seq data from human and mouse. *Nat. Commun.* **2018**, *9*, 1366. [[CrossRef](#)]
23. Strobel, B.; Klein, H.; Leparc, G.; Stierstorfer, B.E.; Gantner, F.; Kreuz, S. Time and phenotype-dependent transcriptome analysis in AAV-TGFβ1 and Bleomycin-induced lung fibrosis models. *Sci. Rep.* **2022**, *12*, 12190. [[CrossRef](#)]
24. Kaur, N.; Oskotsky, B.; Butte, A.J.; Hu, Z. Systematic identification of ACE2 expression modulators reveals cardiomyopathy as a risk factor for mortality in COVID-19 patients. *Genome Biol.* **2022**, *23*, 15. [[CrossRef](#)]
25. Golub, T.R.; Slonim, D.K.; Tamayo, P.; Huard, C.; Gaasenbeek, M.; Mesirov, J.P.; Coller, H.; Loh, M.L.; Downing, J.R.; Caligiuri, M.A.; et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **1999**, *286*, 531–537. [[CrossRef](#)] [[PubMed](#)]
26. Alon, U.; Barkai, N.; Notterman, D.A.; Gish, K.; Ybarra, S.; Mack, D.; Levine, A.J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 6745–6750. [[CrossRef](#)] [[PubMed](#)]
27. Roffo, G.; Melzi, S.; Castellani, U.; Vinciarelli, A.; Cristani, M. Infinite Feature Selection: A Graph-based Feature Filtering Approach. *IEEE Trans. Pattern. Anal. Mach. Intell.* **2021**, *43*, 4396–4410. [[CrossRef](#)]
28. Chen, J.; Bardes, E.E.; Aronow, B.J.; Jegga, A.G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* **2009**, *37*, W305–W311. [[CrossRef](#)]
29. Wei, J.; Hu, M.; Huang, K.; Lin, S.; Du, H. Roles of Proteoglycans and Glycosaminoglycans in Cancer Development and Progression. *Int. J. Mol. Sci.* **2020**, *21*, 5983. [[CrossRef](#)]
30. Morla, S. Glycosaminoglycans and Glycosaminoglycan Mimetics in Cancer and Inflammation. *Int. J. Mol. Sci.* **2019**, *20*, 1963. [[CrossRef](#)] [[PubMed](#)]
31. Wang, X.; Osada, T.; Wang, Y.; Yu, L.; Sakakura, K.; Katayama, A.; McCarthy, J.B.; Brufsky, A.; Chivukula, M.; Khoury, T.; et al. CSPG4 protein as a new target for the antibody-based immunotherapy of triple-negative breast cancer. *J. Natl. Cancer Inst.* **2010**, *102*, 1496–1512. [[CrossRef](#)] [[PubMed](#)]
32. Arokiasamy, S.; Balderstone, M.J.M.; De Rossi, G.; Whiteford, J.R. Syndecan-3 in Inflammation and Angiogenesis. *Front. Immunol.* **2019**, *10*, 3031. [[CrossRef](#)]
33. Hu, Z.; Wang, C.; Xiao, Y.; Sheng, N.; Chen, Y.; Xu, Y.; Zhang, L.; Mo, W.; Jing, N.; Hu, G. NDST1-dependent heparan sulfate regulates BMP signaling and internalization in lung development. *J. Cell. Sci.* **2009**, *122*, 1145–1154. [[CrossRef](#)]
34. Marques, C.; Reis, C.A.; Vivès, R.R.; Magalhães, A. Heparan Sulfate Biosynthesis and Sulfation Profiles as Modulators of Cancer Signalling and Progression. *Front. Oncol.* **2021**, *11*, 778752. [[CrossRef](#)]
35. Kai, Y.; Amatya, V.J.; Kushitani, K.; Kambara, T.; Suzuki, R.; Fujii, Y.; Tsutani, Y.; Miyata, Y.; Okada, M.; Takeshima, Y. Glypican-1 is a novel immunohistochemical marker to differentiate poorly differentiated squamous cell carcinoma from solid predominant adenocarcinoma of the lung. *Transl. Lung Cancer Res.* **2021**, *10*, 766–775. [[CrossRef](#)] [[PubMed](#)]
36. Karna, E.; Surazynski, A.; Palka, J. Collagen metabolism disturbances are accompanied by an increase in prolydase activity in lung carcinoma planoepitheliale. *Int. J. Exp. Pathol.* **2000**, *81*, 341–347. [[CrossRef](#)]

37. Eni-Aganga, I.; Lanaghan, Z.M.; Balasubramaniam, M.; Dash, C.; Pandhare, J. PROLIDASE: A Review from Discovery to its Role in Health and Disease. *Front. Mol. Biosci.* **2021**, *8*, 723003. [[CrossRef](#)] [[PubMed](#)]
38. Ballester, B.; Milara, J.; Cortijo, J. Idiopathic Pulmonary Fibrosis and Lung Cancer: Mechanisms and Molecular Targets. *Int. J. Mol. Sci.* **2019**, *20*, 593. [[CrossRef](#)] [[PubMed](#)]
39. Galicka, A.; Wolczyński, S.; Anchim, T.; Surazyński, A.; Lesniewicz, R.; Palka, J. Defects of type I procollagen metabolism correlated with decrease of prolidase activity in a case of lethal osteogenesis imperfecta. *Eur. J. Biochem.* **2001**, *268*, 2172–2178. [[CrossRef](#)]
40. Galicka, A.; Wolczyński, S.; Gindzieński, A.; Surazyński, A.; Palka, J. Gly511 to Ser substitution in the COL1A1 gene in osteogenesis imperfecta type III patient with increased turnover of collagen. *Mol. Cell Biochem.* **2003**, *248*, 49–56. [[CrossRef](#)]
41. Pan, Z.; Xu, T.; Bao, L.; Hu, X.; Jin, T.; Chen, J.; Chen, J.; Qian, Y.; Lu, X.; Li, L.; et al. CREB3L1 promotes tumor growth and metastasis of anaplastic thyroid carcinoma by remodeling the tumor microenvironment. *Mol. Cancer* **2022**, *21*, 190. [[CrossRef](#)]
42. Huang, R.Y.J.; Kuay, K.T.; Tan, T.Z.; Asad, M.; Tang, H.M.; Ng, A.H.C.; Ye, J.; Chung, V.Y.; Thiery, J.P. Functional relevance of a six mesenchymal gene signature in epithelial-mesenchymal transition (EMT) reversal by the triple angiokinase inhibitor, nintedanib (BIBF1120). *Oncotarget* **2015**, *6*, 22098–22113. [[CrossRef](#)]
43. Amor López, A.; Mazariegos, M.S.; Capuano, A.; Ximénez-Embún, P.; Hergueta-Redondo, M.; Recio, J.Á.; Muñoz, E.; Al-Shahrour, F.; Muñoz, J.; Megías, D.; et al. Inactivation of EMILIN-1 by Proteolysis and Secretion in Small Extracellular Vesicles Favors Melanoma Progression and Metastasis. *Int. J. Mol. Sci.* **2021**, *22*, 7406. [[CrossRef](#)]
44. Hou, L.; Lin, T.; Wang, Y.; Liu, B.; Wang, M. Collagen type 1 alpha 1 chain is a novel predictive biomarker of poor progression-free survival and chemoresistance in metastatic lung cancer. *J. Cancer* **2021**, *12*, 5723–5731. [[CrossRef](#)]
45. Yanagita, K.; Nagashio, R.; Jiang, S.X.; Kuchitsu, Y.; Hachimura, K.; Ichinoe, M.; Igawa, S.; Fukuda, E.; Goshima, N.; Satoh, Y.; et al. Cytoskeleton-Associated Protein 4 Is a Novel Serodiagnostic Marker for Lung Cancer. *Am. J. Pathol.* **2018**, *188*, 1328–1333. [[CrossRef](#)] [[PubMed](#)]
46. Agarwal, A.K.; Srinivasan, N.; Godbole, R.; More, S.K.; Budnar, S.; Gude, R.P.; Kalraiya, R.D. Role of tumor cell surface lysosome-associated membrane protein-1 (LAMP1) and its associated carbohydrates in lung metastasis. *J. Cancer Res. Clin. Oncol.* **2015**, *141*, 1563–1574. [[CrossRef](#)] [[PubMed](#)]
47. Singh, A.K.; Kapoor, V.; Thotala, D.; Hallahan, D.E. TAF15 contributes to the radiation-inducible stress response in cancer. *Oncotarget* **2020**, *11*, 2647–2659. [[CrossRef](#)] [[PubMed](#)]
48. Nehme, E.; Rahal, Z.; Sinjab, A.; Khalil, A.; Chami, H.; Nemer, G.; Kadara, H. Epigenetic Suppression of the T-box Subfamily 2 (TBX2) in Human Non-Small Cell Lung Cancer. *Int. J. Mol. Sci.* **2019**, *20*, 1159. [[CrossRef](#)] [[PubMed](#)]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.