



EnRank: An Ensemble Method to Detect Pulmonary Hypertension Biomarkers Based on Feature Selection and Machine Learning Models

Xiangju Liu¹, Yu Zhang¹, Chunli Fu¹, Ruochi Zhang² and Fengfeng Zhou^{2*}

¹Department of Geriatric Medicine & Shandong Key Laboratory Cardiovascular Proteomics, Qilu Hospital of Shandong University, Jinan, China, ²College of Computer Science and Technology, and Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, China

OPEN ACCESS

Edited by:

Robert Friedman,
Retired, Columbia, SC, United States

Reviewed by:

Wanjun Gu,
Southeast University, China
Jianbo Pan,
Johns Hopkins Medicine,
United States
Weiqun Peng,
George Washington University,
United States

*Correspondence:

Fengfeng Zhou
fengfengzhou@gmail.com;
ffzhou@jlu.edu.cn

Specialty section:

This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: 05 January 2021

Accepted: 30 March 2021

Published: 27 April 2021

Citation:

Liu X, Zhang Y, Fu C, Zhang R and
Zhou F (2021) EnRank: An Ensemble
Method to Detect Pulmonary
Hypertension Biomarkers Based on
Feature Selection and Machine
Learning Models.
Front. Genet. 12:636429.
doi: 10.3389/fgene.2021.636429

Pulmonary hypertension (PH) is a common disease that affects the normal functioning of the human pulmonary arteries. The peripheral blood mononuclear cells (PMBCs) served as an ideal source for a minimally invasive disease diagnosis. This study hypothesized that the transcriptional fluctuations in the PMBCs exposed to the PH arteries may stably reflect the disease. However, the dimension of a human transcriptome is much higher than the number of samples in all the existing datasets. So, an ensemble feature selection algorithm, EnRank, was proposed to integrate the ranking information of four popular feature selection algorithms, i.e., T-test (Ttest), Chi-squared test (Chi2), ridge regression (Ridge), and Least Absolute Shrinkage and Selection Operator (Lasso). Our results suggested that the EnRank-detected biomarkers provided useful information from these four feature selection algorithms and achieved very good prediction accuracy in predicting the PH patients. Many of the EnRank-detected biomarkers were also supported by the literature.

Keywords: EnRank, ensemble feature selection, filter, pulmonary hypertension, biomarker detection

INTRODUCTION

Pulmonary hypertension (PH) shows the symptom of high blood pressure in the lung arteries which impedes the delivery of blood from the heart to the lungs (Mandras et al., 2020). PH is diagnosed by at least 20 mmHg (millimeter of mercury) of the rest-state mean pulmonary arterial pressure (mPAP) and the right-sided heart catheterization (Simonneau et al., 2019). Although PH may be caused by various factors, the PH patients painfully suffer from shortness of breath and increased mortality (Mandras et al., 2020). As many as, 10% of people over age 65 are affected by PH, and more than half of them develop heart failure (Hoepfer et al., 2016). Detection of novel transcriptomic biomarkers may facilitate the understanding of the PH molecular mechanisms and serve as candidates for investigation and prognosis of the disease (Jardim and Souza, 2015; Swaminathan et al., 2015).

The high throughput DNA sequencing technology generates the expression levels of all human protein-coding and non-coding genes (Jandl et al., 2019; Tzimas et al., 2019), and machine learning methods rely on this data (Stephens et al., 2015; Mirza et al., 2019). The sequenced samples may be lesion tissues, e.g., the endothelial cells or the small remodeled arteries (Jandl et al., 2019; Tzimas et al., 2019). The peripheral blood mononuclear cells (PBMCs) serve ideally as the targets for the transcriptome sequencing because it is less invasive than use of lesion tissue (Tzouveleakis et al., 2018).

Transcriptome based disease prediction is often limited by sufficient sampling toward detection of disease features. This is mostly caused by the high cost of sequencing, a transcriptome and the difficulty in recruiting a cohort of individuals with and without the disease (Diao and Vidyashankar, 2013). Building a prediction model using all features may lead to overfitting and loss of applicability to a non-training data set (Schinkel et al., 2019). This problem is addressed by an algorithm for selecting a subset of the features in building the disease prediction model (Schinkel et al., 2019; Shi et al., 2019; McCabe et al., 2020).

There are two main categories of feature selection algorithms, filter and wrapper (Ye et al., 2017). A filter feature selection algorithm evaluates the association of each feature with a class label and then ranks features based on the significance of this association (Hall and Smith, 1999). These filter algorithms are commonly used for detection of biomarkers since their time complexity is linear. However, the filters ignore the inter-feature correlations and cannot detect the subset of low ranked features with good prediction performance (Ye et al., 2017). A wrapper feature selection algorithm heuristically generates the feature subsets and evaluates the prediction performance of a given feature subset using a user defined classifier (Das, 2001). The wrappers usually have higher time complexities than the filters and tend to deliver the feature subsets with better prediction performance than the filters (Ge et al., 2016).

This study proposes an ensemble feature selection algorithm, EnRank, to take advantage of both filters and wrappers. The main idea of EnRank is to integrate the ranks of multiple feature selection algorithms and verify that the final feature subset is efficient for use in a prediction model. A comprehensive evaluation was carried out to test which classifier achieved the best prediction performance. Our experimental data suggested that different feature selection algorithms may contribute complementary information to each other and the orchestration of the features selected by these algorithms are efficient for use in a predictive model.

MATERIALS AND METHODS

Collection of Data

This study used the transcriptome dataset GSE33463 of pulmonary hypertension patients and controls (Cheadle et al., 2012). The gene expression levels were profiled from the PBMCs of the recruited participants on the microarray platform GPL6947 (Illumina HumanHT-12 V3.0 expression beadchip). Each sample

had 49,576 transcriptomic features, and the feature annotations were retrieved from the platform definition file.

This dataset consisted of 140 samples in total. There were 30 idiopathic pulmonary arterial hypertension (PAH) patients, 19 patients with systemic sclerosis (SSc) without pulmonary hypertension, 42 scleroderma-associated PAH patients, and eight patients with SSc complicated by interstitial lung diseases and pulmonary hypertension. The remaining 41 samples were non-disease controls. This study investigated the binary classification problem between the 99 patients (positive samples) and 41 non-disease controls (negative samples).

Feature Selection Algorithms

Feature selection algorithms were used to find the biomarkers with the best disease detection performance. Each sample had 49,576 transcriptomic features, and the overall dataset had 140 samples in total. A classification model may have a large chance of overfitting for this “large p small n ” situation (Keel et al., 2019; Ren et al., 2020). A feature selection algorithm may be used to find a subset of features for building an accurate and stable classification model. This would also make the model easier to be interpreted along with better performance during the training step. The following four feature selection algorithms were utilized to find a good subset of features.

The Chi-squared test (Chi2) helps to test the relationships or dependence between two variables. Chi2 may be used to remove the features without dependency on the class labels (Xiao et al., 2020). In other words, these removed features will have a small contribution to any classification model.

The T-test (Ttest) is widely used to evaluate the statistical significance (p value) of the null hypothesis that a feature of the positive samples has the same normal distribution as that of the negative samples (Govindan et al., 2019; Soh et al., 2020). A feature with the value of $p < 0.05$ is typically considered a candidate for differential expression between the positive and negative samples. In addition, a feature with a lower p value is considered to have increased power for binary classification.

The ridge regression (Ridge) evaluates a subset of features for their connections with the class labels (Gao et al., 2020; Xu et al., 2020). Ridge provides a model-based trade-off between the fitting and complexity of the features by adding the L2 regularization to the regression model.

The Least Absolute Shrinkage and Selection Operator (Lasso) algorithm adds the L1 regularization to the regression model along with a penalty value for number of features (Deshpande et al., 2019). So, Lasso tends to select a small subset of features and weights them for building a robust regression model.

Binary Classification Methods

The models for predicting disease were trained using five binary classifiers.

Logistic regression (LR) is a statistics model using a logistic function to model a binary classification problem (Cuadrado-Godia et al., 2019; Khandezamin et al., 2020).

The logistic model calculates the log-odds for the class label by a linear combination of one or multiple predictors.

Support vector machine (SVM) is a supervised machine learning algorithm originally designed for binary classification (Jin et al., 2019; Wang et al., 2019). SVM searches for a hyperplane to separate two classes of samples with the maximal margins. It enriches the feature space through a kernel function to quantify the inter-sample similarities.

A simple algorithm K Nearest Neighbor (KNN) is a popular supervised machine learning framework for both classification and regression tasks (Wang et al., 2020; Yuan et al., 2020). KNN determines the class label of a query sample through the majority voting strategy of the KNNs of the query sample.

Decision tree (DT) uses a tree structure to solve the classification problem (Prieto-Gonzalez et al., 2020). Each node except for the leaves in a DT classifier exerts a feature evaluation, and the evaluation result determines which sub-branch of this current node to follow. DT is a simple and easy-to-interpret classifier.

The adaptive boosting tree (AdaBoost) is an integrated machine learning technique (Qiao and Xie, 2019; Dou et al., 2020). The weight of a sample will be increased if this sample leads to a misclassified base classifier. Each iteration will add new base classifiers. The final goal is to find a strong classifier with sufficiently small error rate.

Performance Evaluation Metrics

The supervised machine learning algorithms were evaluated by the following performance metrics. These metrics are essential to measure a prediction model from different aspects. This study used specificity (Sp), sensitivity (Sn), accuracy (Acc), and the area under the receiver operating characteristics curve (AUC). The number of correctly predicted positive samples was defined as the true positive (TP) and that of the incorrectly predicted positives was the false negative (FN). The true negative (TN) and the false positive (FP) defined the numbers of correctly and incorrectly predicted negative samples, respectively.

The overall accuracy is calculated as the number of all the correct predictions divided by the total number of samples in the dataset. That is to say, $Acc = (TP + TN) / (TP + FN + TN + FP)$. The value of Acc is between 0.0 and 1.0. The two metrics Sp and Sn describe the ratios of correctly predicted negative and positive samples, respectively. So $Sp = TN / (TN + FP)$ and $Sn = TP / (TP + FN)$. Both metrics are between 0.0 and 1.0. A larger value of the three metrics Acc/Sp/Sn suggests a better prediction performance. The Matthews' Correlation Coefficient (MCC; Matthews, 1975) was introduced by the biochemist Brian W. Matthews in 1975 and MCC is generally regarded as a balanced measurement which can be used even if the classes are of very different sizes. The metric AUC is a parameter independent metric for the prediction model and shows a trade-off between Sp and Sn (Shao et al., 2020).

The Proposed Feature Ranking Algorithm, EnRank

This study proposed the ensemble feature selection algorithm, EnRank, by calculating the weighted ranks of the four feature

selection algorithms, i.e., Ttest, Chi2, Ridge, and Lasso. The two filter algorithms Ttest and Chi2 rank the features by their individual association values of p with the class labels. The two linear fitting algorithms Ridge and Lasso rank the features based on the absolute values of the fitted model's coefficients. The values of the feature ranks start from 1, i.e., the best ranked feature has the rank 1. Each feature selection algorithm selects top-ranked $pTopK = 100$ features for further screening.

The proposed algorithm EnRank defines a weight Aim_i for each feature selection algorithm, where $i \in \{Ttest, Chi2, Ridge, Lasso\}$. The $pTopK$ features selected by each algorithm were loaded into the five classification algorithms, i.e., LR, SVM, KNN, DT, and AdaBoost. The stratified 5-fold cross validation (S5FCV) strategy was used to calculate the metric AUC, and each feature selection algorithm received five AUC values. This study aimed to find a feature subset with stably high AUC values for five classification algorithms, and defined $Aim_i = Avg_i / Var_i$, where Avg_i and Var_i were the averaged value and variance of the five AUC values of the feature selection algorithm i , respectively.

Finally, EnRank generated an integrated rank for each feature f . To avoid the case of very low ranking features, the rank of feature $Rank_i(f)$ was redefined as the penalization rank $pPenaltyRank = 1,000$, if $Rank_i(f) > pTopK$. The integrated rank $EnRank(f) = Average(Rank_i(f) \times Aim_i)$ was defined as the EnRank metric, where the function $Average()$ is the averaged value, and $i \in \{Ttest, Chi2, Ridge, Lasso\}$.

Then, any filter-based feature selection frameworks, e.g., the incremental feature selection (IFS), may be used to find the best subset of top-ranked features generated by EnRank.

Workflow of This Study

This study proposed an ensemble feature selection algorithm, EnRank, by integrating the feature ranks from different algorithms (Figure 1). The experimental data in the following section suggested that different feature selection algorithms performed differently, and it is necessary to integrate the ranking information calculated by different feature selection algorithms.

RESULTS AND DISCUSSION

Comparison of the Feature Ranks by Ttest and Chi2

Table 1 illustrated the top-ranked 10 features delivered by the two filter algorithms Ttest and Chi2. Firstly, the statistical significance p values of the two algorithms Ttest and Chi2 were different to each other. The minimum p value of Ttest was $2.83e-19$ while Chi2 only calculated the minimum p value $4.08e-4$ for its null hypothesis. Actually, even the rank-100 feature ILMN_1698668 by Ttest had value of $p = 2.28e-12$, which was much smaller than the minimum value of $p = 4.08e-4$ of the algorithm Chi2.

And there was only one feature ILMN_1789074 shared among the top-ranked 10 features by Ttest and Chi2. The p value for

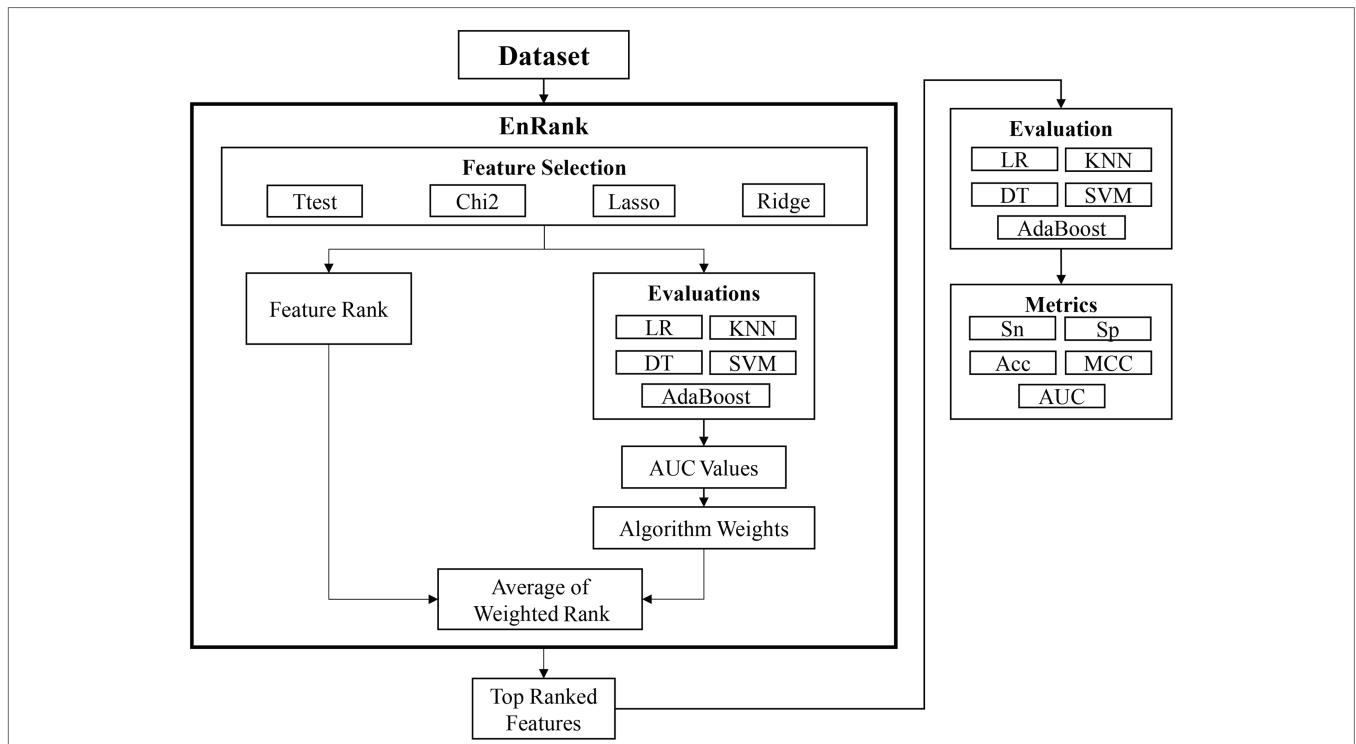


FIGURE 1 | Workflow of this study. The proposed algorithm EnRank integrates the ranks of both feature selection algorithms and classification algorithms. The finally generated feature subset is further evaluated by five different classification algorithms.

TABLE 1 | The top-10 features ranked by Ttest and Chi2. The two columns “Ttest” and “Chi2” gave the names of the ranked features.

Ttest	Ttest-p value	Rank	Chi2	Chi2-p value
ILMN_1812970	2.83E-19	1	ILMN_1806023	4.08E-04
ILMN_1875248	9.71E-19	2	ILMN_1656011	8.59E-04
ILMN_1704335	2.28E-18	3	ILMN_1702691	1.48E-03
ILMN_1794233	6.85E-18	4	ILMN_2367126	2.42E-03
ILMN_1765725	1.06E-17	5	ILMN_2339955	3.21E-03
ILMN_1758687	1.83E-17	6	ILMN_1815527	3.44E-03
ILMN_1687526	5.60E-17	7	ILMN_1789074	5.21E-03
ILMN_1767168	7.40E-17	8	ILMN_1782305	5.49E-03
ILMN_2159384	1.38E-16	9	ILMN_2088437	8.62E-03
ILMN_1789074	5.47E-16	10	ILMN_1751607	9.14E-03

Column “Rank” provided the rank values. “Ttest-p value” and “Chi2-p value” provided the statistical p values as calculated by the two algorithms Ttest and Chi2. The feature was highlighted in bold if it was among the top-ranked 10 features of both algorithms.

the Ttest null hypothesis was 5.47e-16 for feature ILMN_1789074 (Ttest rank 10), while Chi2 recommended ILMN_1789074 as the rank 7 feature with p value 5.21e-3.

So, the statistical tests Ttest and Chi2 generated significantly different p values for the features, and we had to integrate the features by their rank values.

Comparison of the Feature Ranks by Ridge and Lasso

Only six out of the top-10 ranked features by the absolute values of their model coefficients were shared by the two algorithms Ridge and Lasso (Table 2). This study assumed that both positive and negative correlations of the features

with the class labels were important, and the absolute values of the model correlation coefficients of these features were used to rank the features in descending order. The feature ILMN_1697499 was the best ranked feature by Ridge, but it was not even within the top-10 ranked features by Lasso. Actually, the feature ILMN_1697499 was ranked 26 by Lasso. And the best ranked feature ILMN_1678859 by Lasso was only the ninth ranked feature by Ridge.

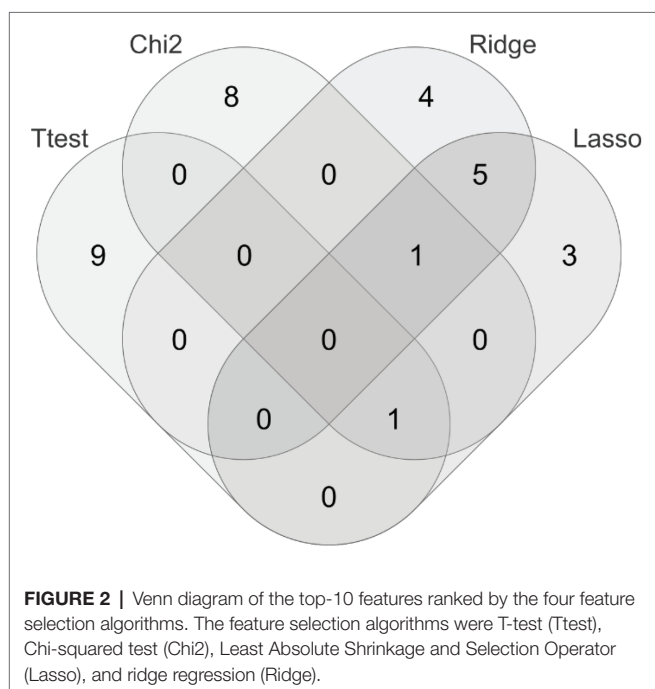
Venn diagram (Figure 2) shows that very few features were shared by these four feature selection algorithms, i.e., Ttest, Chi2, Lasso, and Ridge, except between Lasso and Ridge.

The data in Tables 1, 2 suggested that the top-ranked features by the four algorithms Ttest, Chi2, Ridge, and Lasso described

TABLE 2 | The top-10 features ranked by the model coefficients of the regression models Ridge and Lasso.

Ridge	RidgeCoef	Rank	Lasso	LassoCoef
ILMN_1697499	0.0671	1	ILMN_1678859	0.1663
ILMN_2165753	0.0563	2	ILMN_1781236	0.1589
ILMN_1807491	0.0484	3	ILMN_2058782	0.1362
ILMN_1781236	0.0435	4	ILMN_2083066	0.1307
ILMN_1806023	0.0435	5	ILMN_1807491	0.1269
ILMN_2083066	0.0428	6	ILMN_1806023	0.1142
ILMN_1721113	0.0425	7	ILMN_1801216	0.1120
ILMN_2229649	0.0408	8	ILMN_1822671	0.1078
ILMN_1678859	0.0398	9	ILMN_1789074	0.1077
ILMN_2323933	0.0395	10	ILMN_2165753	0.1072

The two columns "Ridge" and "Lasso" identified the top-10 ranked features. Column "Rank" provided the rank values. The two columns "RidgeCoef" and "LassoCoef" gave the absolute values of the model coefficients calculated by the two algorithms Ridge and Lasso. The features were highlighted in bold if they were among the top-ranked 10 features of both algorithms.



the class correlations of the features from different aspects. **Figure 3** evaluated different value choices of the parameter pTopK. Both pTopK = 75 and 100 achieved the best averaged AUC = 0.9446. In order to introduce more feature diversity, this study focused on the four lists of top-ranked pTopK = 100 features by the above four algorithms, and their union consisted of 269 features.

Evaluation of the Four Feature Selection Algorithms

Figure 4A demonstrated that the classification algorithm DT had low performance on all four feature lists. And the other four classification algorithms achieved at least 0.9000 in the metric AUC for all four feature lists. Although the Lasso-selected 100 features achieved the best mean AUC value 0.9571 by the five classification algorithms, its SD 0.0701 was larger than that (0.0594) of another algorithm Ridge. So the Lasso's Aim 13.6620

was slightly larger than that (15.9508) of Ridge, as shown in **Figure 4B**. The filter Ttest was assigned the Aim 10.6090 due to its largest SD 0.0877.

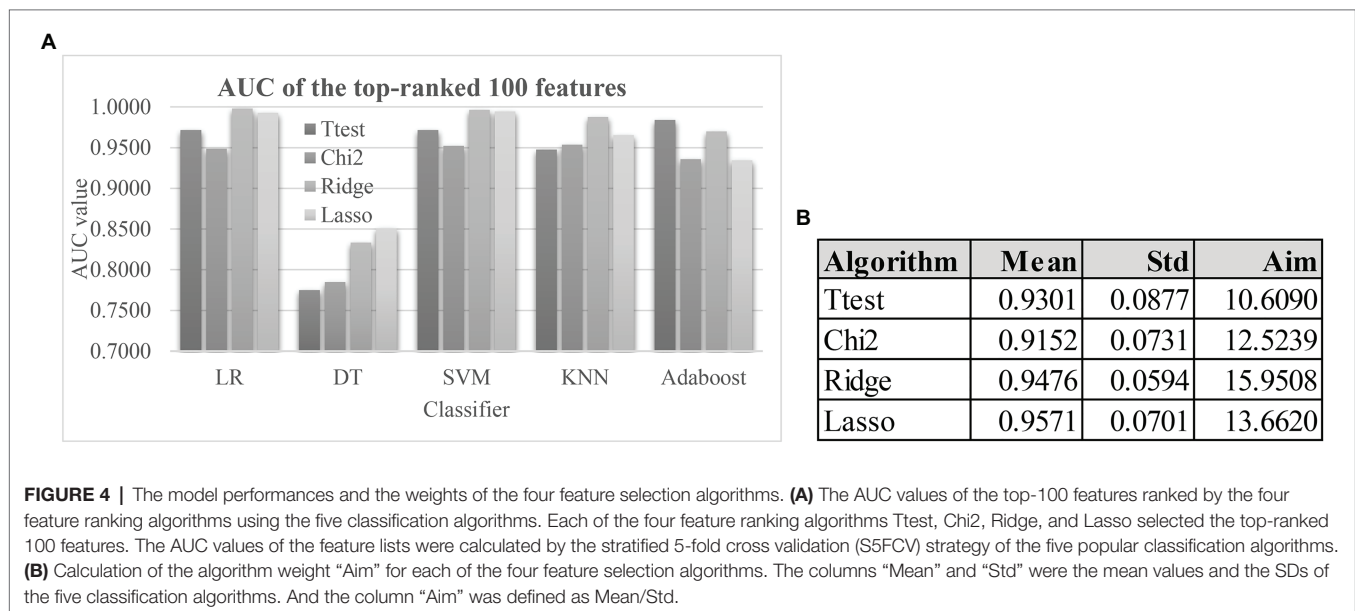
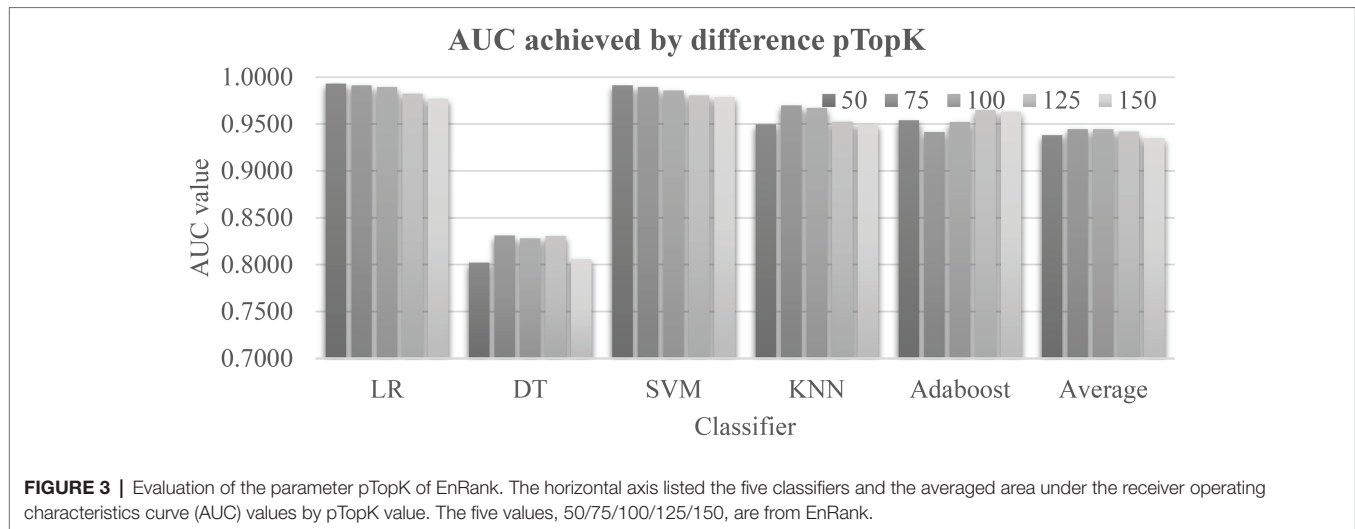
Distribution of the Calculated EnRank Metrics

The ranking metric EnRank was defined in the above subsection "The proposed feature ranking algorithm, EnRank." EnRank used the EnRank metrics to rank the features in ascending order, and the features were roughly separated into four groups, as shown in **Figure 5**. The EnRank metrics of the ordered features were within these four ranges, i.e., [1, 1,000], [2,500, 3,300], [5,000, 5,700], and [7,500, 7,900]. The experimental data suggested that these four groups of features consisted of features recommended by four, three, two, and one feature selection algorithms, respectively. That is to say, a feature recommended by four feature selection algorithms was not penalized by the penalization rank pPenaltyRank, and algorithm aims were between 10 and 16. Such a feature had an EnRank smaller than 1,000. So the metric EnRank reasonably described how each feature was ranked by multiple feature selection algorithms.

Literature Supportive of the EnRank-Detected 50 Biomarkers

The metric Literature Support (LR) of a feature was defined by the number of PubMed (Fiorini et al., 2017) publications matching the gene symbol of this feature and the key word "pulmonary disease" in both title and abstract. The query term was "term={}[tiab] AND pulmonary disease [tiab]," and the queried date was November 16, 2020. The cumulative LR (CLR) of the top-k ranked features was defined as the sum of the LR values of these k features.

In order to compare with the biomarkers selected by EnRank, we randomly selected the same number of genes among the remaining genes as a control group, and then compared the metrics CLR and LR in the two groups. **Figure 6** illustrated that the EnRank-detected top-ranked features were investigated for their roles in pulmonary diseases many more times than the randomly-chosen features. The randomly-chosen features were supported



by at most two PubMed publications, and only four out of the 50 randomly-selected features had literature support. And the EnRank-detected top-ranked 50 features were more significantly supported by the scientific literature. Some features were supported by as many as nine PubMed publications, and 14 out of the 50 features had literature support. So the EnRank-detected features were consistently supported by the literature.

Model Evaluation Based on the EnRank-Detected Biomarkers

A comparative study was carried out to evaluate whether the proposed algorithm EnRank recommended features with good prediction performance of pulmonary hypertension (Figure 7). The baseline models in Figure 7A showed that the classifier DT achieved the worst PH prediction accuracy ($Acc = 0.7545$), while the classifier LR achieved the best $Acc = 0.9000$. SVM achieved the same $Sn = 0.9560$ as LR, but much worse $Sp = 0.5361$

than that ($Sp = 0.8056$) of LR. So, it is necessary to find a subset of biomarker features with a better PH prediction accuracy.

Figure 7B showed that the 50 EnRank-detected biomarkers improved the prediction accuracies of all five classification algorithms. The largest improvement in Acc (0.1364) was achieved for both SVM and KNN. The classification algorithm LR achieved the best $Acc = 0.9545$ again using the 50 EnRank-detected biomarkers. The parameter-independent metric $AUC = 0.9894$ of LR was also the best among the five classification algorithms.

So this study delivered a PH prediction model using the 50 EnRank-detected biomarkers and the LR classification algorithm.

Further Validation of the Proposed PH Biomarkers

Firstly, the proposed PH biomarkers were validated using an independent dataset GSE22356 (Risbano et al., 2010)

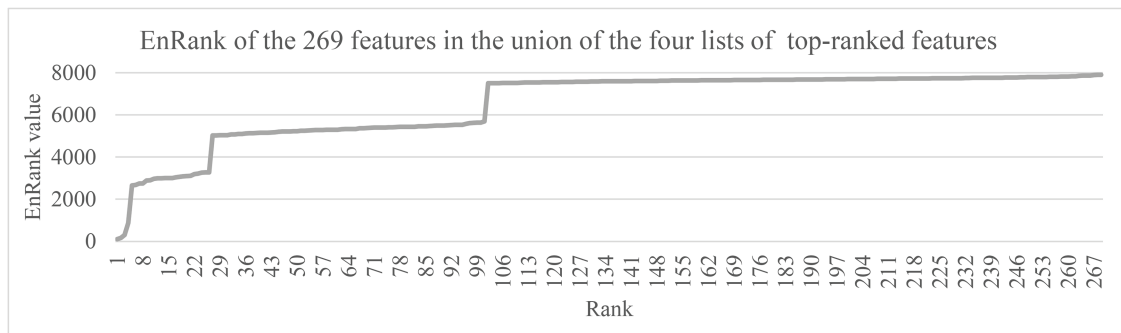


FIGURE 5 | The EnRank metrics of the 269 features in the union of the four lists of top-100 ranked features. The horizontal axis gave the feature ranks ordered by the EnRank metric, and the vertical axis gave the EnRank metrics of the top-ranked 269 features. These features were among the union of the top-100 ranked features recommended by the four algorithms, Ttest, Chi2, Ridge, and Lasso.

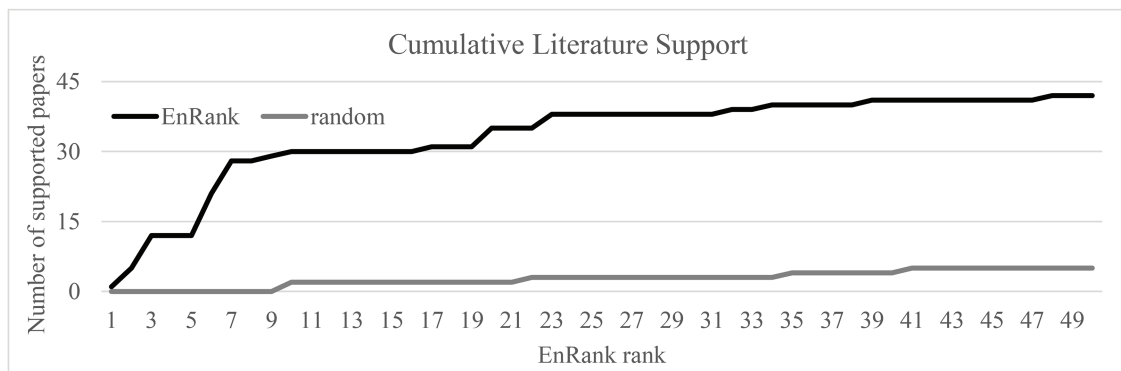


FIGURE 6 | Evaluation of the cumulative literature support LR (CLR) of the top-50 EnRank-ranked features. The horizontal axis gave the EnRank-recommended ranks and the vertical axis shows the metric CLR.

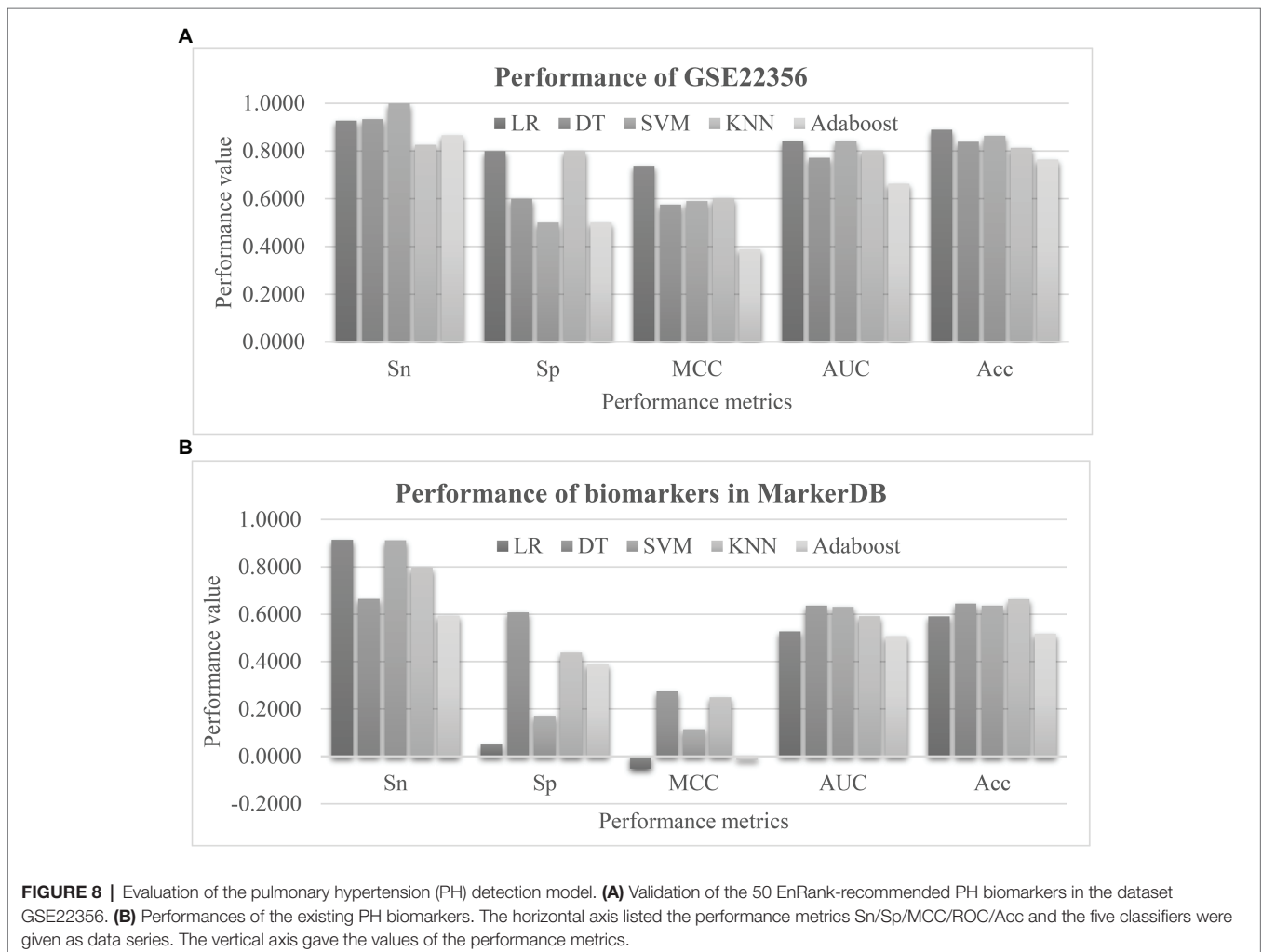
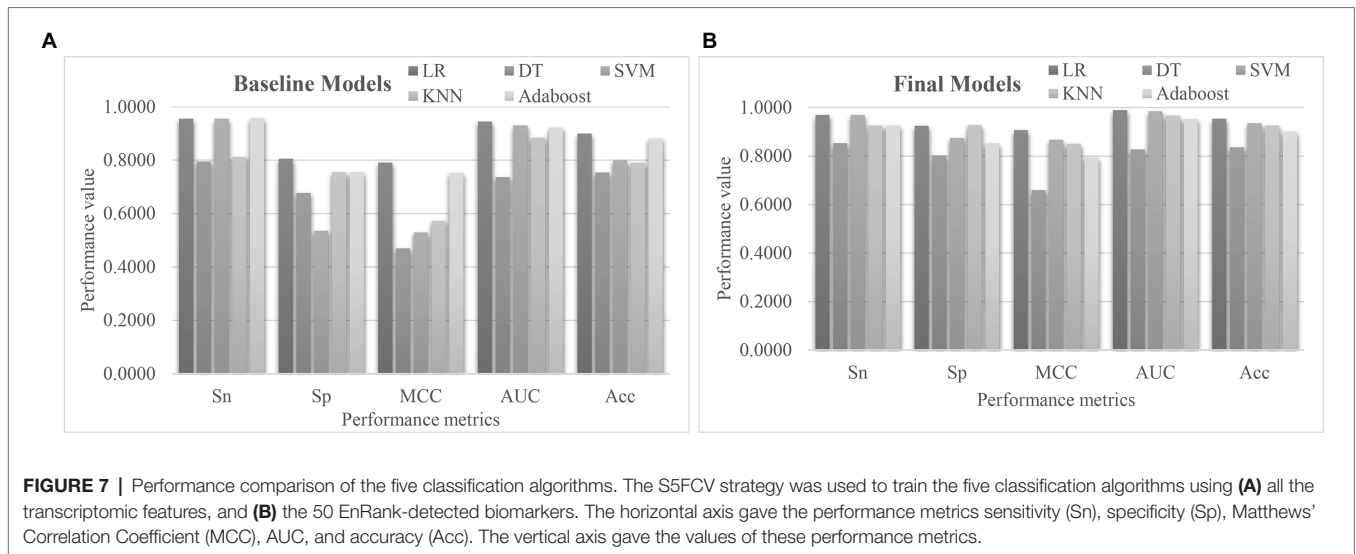
from the Gene Expression Omnibus (GEO) database (Clough and Barrett, 2016). This dataset consists of 38 PBMC samples profiled by the Affymetrix Human Genome U133 Plus 2.0 Array (GPL570), and investigates the altered immune phenotypes of the scleroderma-associated pulmonary hypertension (Risbano et al., 2010). The normal controls were assumed as the negative samples, and the other samples were regarded as the positive ones. The 50 EnRank-recommended features matched 236 features through 36 unique genes in the independent dataset. The same settings of training and evaluation as EnRank were used. **Figure 8A** showed that four of the five classifiers achieved AUC values at least 0.8000. The classifier LR achieved the largest AUC = 0.8433, and the largest Acc = 0.8893. Considering that this independent dataset was profiled using a different transcriptome platform than our original dataset GSE33463, the independent validation results supported the robustness of the EnRank-recommended PH biomarkers.

We searched the literature database PubMed using the keywords “pulmonary hypertension” and “biomarker” in the titles, and only 41 publications were detected. Most of them focused on the protein (Wu et al., 2020), vocal (Sara et al., 2020), and

imaging (Jivraj et al., 2017; Jose et al., 2020) data. So we collected the PH marker genes from the recently updated database MarkerDB (Wishart et al., 2021). Three unique genes were annotated as the PH biomarkers, including Bone Morphogenetic Protein Receptor Type 2 (BMP2), Activin A Receptor Like Type 1 (ACVRL1), and Endoglin (ENG). Four features were associated with these three genes. The prediction performances of these four biomarker features were shown in **Figure 8B**. Unfortunately, no classifiers showed larger than 0.7000 in either AUC or Acc using these biomarkers. This should be due to that the existing biomarkers were screened for their individual associations with the phenotype PH, and their combined PH prediction performances were not investigated in the existing studies.

Further Evaluation of Other Feature Selection Combinations

The proposed algorithm EnRank is a feature selection framework that may integrate the ranking data of multiple feature selection algorithms. The above sections integrated four feature selection algorithms, i.e., Ttest, Chi2, Ridge, and Lasso. **Figure 9A** evaluated the proposed ensemble algorithm EnRank and its



four individual feature selection algorithms using the same training and testing settings. The parameter-independent metric AUC was used to compare the performances of the feature

selection algorithms. EnRank achieved the best AUC values using three out of the five classifiers. The Lasso-recommended features achieved the best AUC = 0.9946 while the

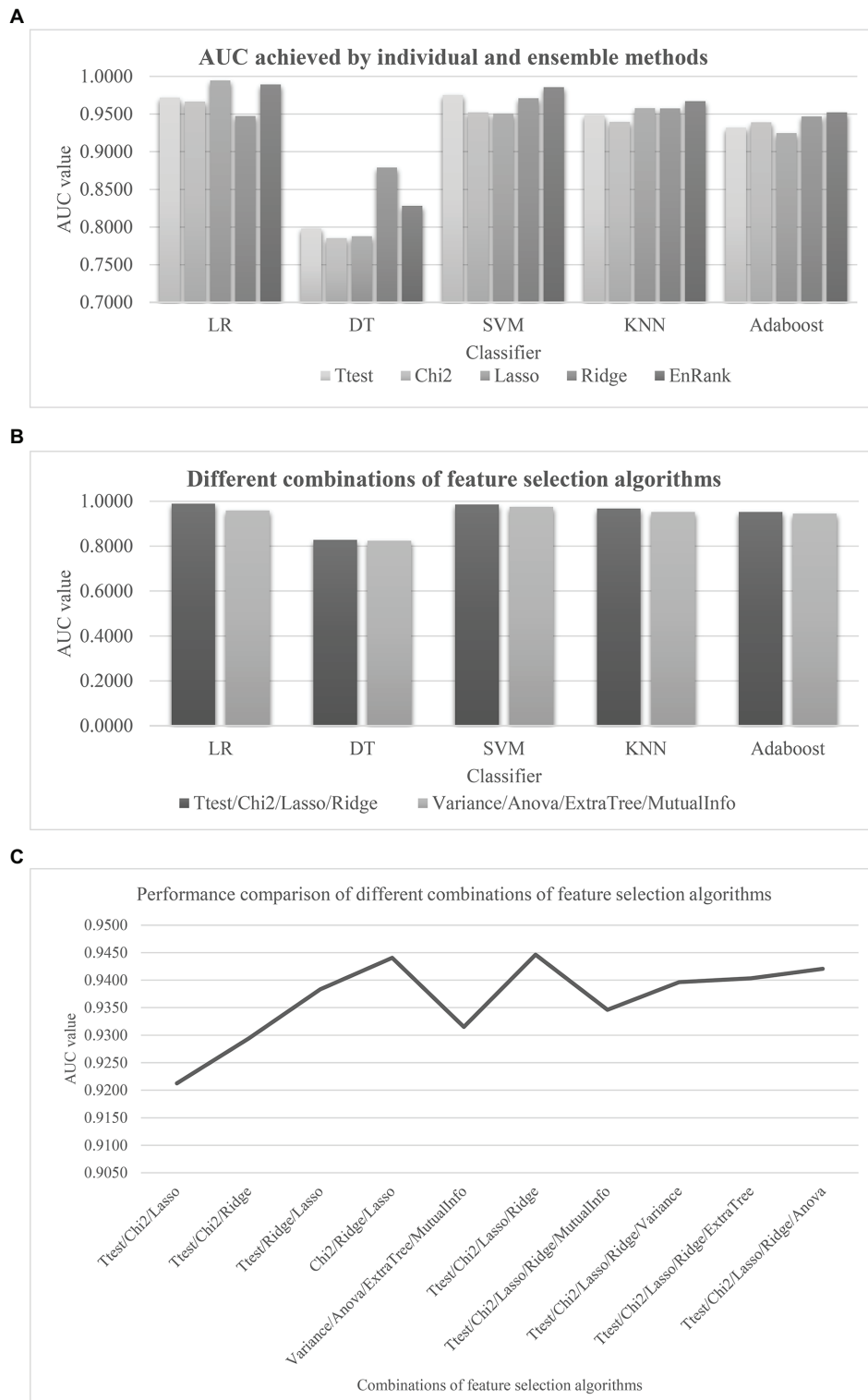


FIGURE 9 | Comparison of EnRank with the other feature selection algorithms and their combinations. **(A)** Evaluation of EnRank and its individual feature selection algorithms. **(B)** Two groups of feature selection algorithms were integrated by EnRank. The original version of EnRank was “Ttest/Chi2/Lasso/Ridge,” and the new version was “Variance/Anova/ExtraTree/MutualInfo.” The horizontal axis listed the classifiers and the vertical axis gave the AUC values of the evaluated models. The ensemble algorithm EnRank and its four individual feature selection algorithms. **(C)** The AUC values of different combinations of feature selection algorithms averaged over the five classifiers logistic regression (LR)/decision tree (DT)/support vector machine (SVM)/k nearest neighbor (KNN)/adaptive boosting tree (AdaBoost). The horizontal axis listed the algorithm combinations, and the vertical axis gave the AUC values.

EnRank-recommended features achieved the second best AUC = 0.9894. EnRank achieved the second best AUC = 0.8283 using the classifier DT, while Ridge-recommended features achieved the slightly better AUC = 0.8790.

The original version of EnRank integrated four feature selection algorithms Ttest/Chi2/Lasso/Ridge, which was compared with the new version integrating four new feature selection algorithms, as shown in **Figure 9B**. The four new feature selection algorithms were Variance Threshold (Variance), Mutual Information (MutualInfo), Extra Trees (ExtraTree), and ANOVA (Anova). The same model training and testing setting were carried out. The original version of EnRank outperformed the new version for all five classifiers. The best classifier LR was even improved by 0.0302 in the parameter-independent performance metric AUC.

The EnRank's performance relied on the including efficient feature selection algorithms (**Figure 9C**). So a comparison was carried out for the performances of different combinations of feature selection algorithms. Here, we investigated the combinations of three or five algorithms. **Figure 9C** showed that the original version of EnRank achieved the best AUC value = 0.9446, although a slightly worse AUC = 0.9441 was achieved by removing Ttest.

Biological Involvement of the EnRank-Detected Biomarkers

Table 3 listed the 50 EnRank-detected biomarkers and their corresponding gene information. Many transcriptomic biomarkers are from chromosomes 19 and 2. And two biomarkers

ILMN_1807491 and ILMN_2323933 are from the same gene Leukocyte Associated Immunoglobulin Like Receptor 2 (LAIR2). Limited knowledge was known about the roles of LAIR2 in the PH patients, based on the information from PubMed (Fiorini et al., 2017) and MalaCards (Rappaport et al., 2017). There were five transcriptomic biomarkers with unknown chromosomal locations.

The feature ILMN_2088437 was from the gene C-X3-C Motif Chemokine Receptor 1 (CX3CR1), which was known to be involved in HIV proliferation (Mhandire et al., 2014; Guo et al., 2020). The absence of CX3CR1 was observed to provide protection from tissue destruction from chronic obstructive pulmonary disease (COPD; Lee, 2012). And the gene CX3CR1 also demonstrated differential expressions in the COPD patients (Huang et al., 2019). Another feature ILMN_1740875 was within the gene Formyl Peptide Receptor 2 (FPR2) encoded on chromosome 19, which was actively involved in the mononuclear phagocyte responses in Alzheimer disease (Iribarren et al., 2005). FPR2 also demonstrated its capability of promoting the chemotaxis and survival of neutrophils in the COPD patients (Iribarren et al., 2005).

The EnRank-recommended genes were analyzed using the online tool DAVID version 6.8 (Jiao et al., 2012). The list of genes was annotated to cover the top 50 EnRank-recommended features and was screened against the human genome. The statistical significance p values were adjusted by the multi-test Benjamini corrections, and only the functional terms with the Benjamini-corrected values of $p < 0.05$ were kept for further analysis. It is interesting to observe that no GO terms were significantly enriched in PH biomarkers; while seven KEGG pathways were enriched

TABLE 3 | Detailed information of the 50 EnRank-detected biomarkers.

Rank	Feature	Gene	Chr	Strand	Rank	Feature	Gene	Chr	Strand
41	ILMN_1804350	LOC644852	1	+	39	ILMN_1711786	NFE2	12	-
1	ILMN_1806023	JUN	1	-	20	ILMN_2207291	IFNG	12	-
15	ILMN_1723912	IFI44L	1	+	25	ILMN_2388547	EPST11	13	-
43	ILMN_2339955	NR4A2	2	-	40	ILMN_2229649	KCTD12	13	-
11	ILMN_1782305	NR4A2	2	-	5	ILMN_2058782	IFI27	14	+
23	ILMN_1800602	GCA	2	+	38	ILMN_1763364	WHDC1	15	+
8	ILMN_1733998	DHRS9	2	+	10	ILMN_2057836	RNU2	17 NT_113932.1	-
44	ILMN_1755643	MGAT4A	2	-	12	ILMN_1772796	DYNLL2	17	+
22	ILMN_1801307	TNFSF10	3	-	47	ILMN_1742618	XAF1	17	+
3	ILMN_2088437	CX3CR1	3	-	26	ILMN_1749722	RNF213	17	+
7	ILMN_1745788	CX3CR1	3	-	24	ILMN_2413331	TMEM107	17	-
33	ILMN_1801216	S100P	4	+	18	ILMN_1775304	DNAJB1	19	-
48	ILMN_1745522	PF4V1	4	+	46	ILMN_2302757	FCGBP	19	-
49	ILMN_1710734	GZMK	5	+	16	ILMN_1751607	FOSB	19	+
42	ILMN_1779147	ENC1	5	-	6	ILMN_1740875	FPR2	19	+
9	ILMN_1702691	TNFAIP3	6	+	28	ILMN_1807491	LAIR2	19	+
32	ILMN_1721113	HLA-C	6	-	45	ILMN_2323933	LAIR2	19	+
2	ILMN_1789074	HSPA1A	6	+	17	ILMN_1664861	ID1	20	+
34	ILMN_1697499	HLA-DRB5	6	-	29	ILMN_2083066	IGLL3	22	+
4	ILMN_1748473	GIMAP4	7	+	35	ILMN_1796830	UBE2L3	22	+
14	ILMN_1799467	SAMD9L	7	-	13	ILMN_1852793	UniGene BC067908		
21	ILMN_1684982	PDK4	7	-	27	ILMN_1781236	RefSeq XR_001116.1		
37	ILMN_1716733	MYOM2	8	+	30	ILMN_1678859	RefSeq XM_938277.1		
50	ILMN_1773313	USMG5	10	-	31	ILMN_2165753	RefSeq NM_001080840.1		
19	ILMN_1674063	OAS2	12	+	36	ILMN_1822671	UniGene BC020840		

Column "Gene" gave the gene symbol for each biomarker. Some biomarkers may not reside in a protein-coding gene, and they may have no annotated gene information.

TABLE 4 | Enriched functional terms of the 50 EnRank-detected PH biomarkers.

KEGG	Term	<i>p</i> values	Benjamini
hsa05164	Influenza A	3.50E-06	2.80E-04
hsa05162	Measles	3.30E-04	1.30E-02
hsa04612	Antigen processing and presentation	9.40E-04	2.20E-02
hsa05168	Herpes simplex infection	1.10E-03	2.20E-02
hsa05332	Graft-versus-host disease	3.30E-03	4.70E-02
hsa05169	Epstein-Barr virus infection	3.70E-03	4.70E-02
hsa05330	Allograft rejection	4.10E-03	4.70E-02

The screening was carried out using the online tool DAVID in the Gene Ontology (GO) terms and KEGG terms. The first two columns gave the KEGG IDs and the corresponding terms. The last two columns gave the statistical significance *p* values and the Benjamini-corrected *p* values. Only the terms with the Benjamini-corrected *p* values < 0.05 were kept for further analysis.

with PH biomarkers. Many of these KEGG pathways were associated with antiviral immunity. The most significant KEGG pathway was hsa05164 (Influenza A) with the Benjamini-corrected *p* value = 2.80e-4. The infection of influenza A caused a patient's death after 3 months of treatment with the popular drug bosentan for pulmonary hypertension in a clinical trial (Hoepfer et al., 2005). As of now, no direct link was presented in the literature. But virus infection is known to be closely connected with pulmonary hypertension (Kimura et al., 2019; Miyasaka et al., 2020; **Table 4**).

CONCLUSION

This study proposed a novel ensemble filter feature selection algorithm EnRank by the weighted integration of four popular filter algorithms. Five classification algorithms were used to evaluate the filter algorithms. The EnRank-detected biomarkers demonstrated very good performances on the PH prediction problem. And most of these biomarkers also demonstrated close connections with the disease PH from the literature.

The proposed algorithm EnRank is a feature selection framework, and may integrate feature selection algorithms with feature weights. The main limitation of EnRank is the choices

REFERENCES

- Cheadle, C., Berger, A. E., Mathai, S. C., Grigoryev, D. N., Watkins, T. N., Sugawara, Y., et al. (2012). Erythroid-specific transcriptional changes in PBMCs from pulmonary hypertension patients. *PLoS One* 7:e34951. doi: 10.1371/journal.pone.0034951
- Clough, E., and Barrett, T. (2016). The gene expression omnibus database. *Methods Mol. Biol.* 1418, 93–110. doi: 10.1007/978-1-4939-3578-9_5
- Cuadrado-Godia, E., Jamthikar, A. D., Gupta, D., Khanna, N. N., Araki, T., Maniruzzaman, M., et al. (2019). Ranking of stroke and cardiovascular risk factors for an optimal risk calculator design: logistic regression approach. *Comput. Biol. Med.* 108, 182–195. doi: 10.1016/j.combiomed.2019.03.020
- Das, S. (2001). "Filters, wrappers and a boosting-based hybrid for feature selection." 74–81.
- Deshpande, S., Shuttleworth, J., Yang, J., Taramonli, S., and England, M. (2019). PLIT: An alignment-free computational tool for identification of long non-coding RNAs in plant transcriptomic datasets. *Comput. Biol. Med.* 105, 169–181. doi: 10.1016/j.combiomed.2018.12.014
- Diao, G., and Vidyashankar, A. N. (2013). Assessing genome-wide statistical significance for large *p* small *n* problems. *Genetics* 194, 781–783. doi: 10.1534/genetics.113.150896

of feature selection algorithms to be integrated. The parameter *p*TopK may also impact the final model performances. Others may want to carry out a series of comparable experiments to find the best parameters for their own datasets.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE33463> [The dataset has the accession GSE33463 in the Gene Expression Omnibus (GEO) database] and <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE22356> [The dataset has the accession GSE22356 in the Gene Expression Omnibus (GEO) database].

AUTHOR CONTRIBUTIONS

FZ, XL, and RZ designed the project, carried out the experiments, and drafted the manuscript. XL, YZ, and CF were involved in the clinical annotations and results discussion. RZ carried out the coding of the computational analysis. RZ and FZ revised and polished the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Natural Science Foundation of the Shandong Province (ID-ZR2017MH122).

ACKNOWLEDGMENTS

We appreciate the constructive comments from the handling editor and the three reviewers that have substantially improved this manuscript.

- Dou, L., Li, X., Zhang, L., Xiang, H., and Xu, L. (2020). iGlu_AdaBoost: identification of lysine glutarylation using the AdaBoost classifier. *J. Proteome Res.* 20, 191–201. doi: 10.1021/acs.jproteome.0c00314
- Fiorini, N., Lipman, D. J., and Lu, Z. (2017). Towards PubMed 2.0. *elife* 6:e28801. doi: 10.7554/eLife.28801
- Gao, X., Liu, S., Song, H., Feng, X., Duan, M., Huang, L., et al. (2020). AgeGuess, a methylopic prediction model for human ages. *Front. Bioeng. Biotechnol.* 8:80. doi: 10.3389/fbioe.2020.00080
- Ge, R., Zhou, M., Luo, Y., Meng, Q., Mai, G., Ma, D., et al. (2016). McTwo: a two-step feature selection algorithm based on maximal information coefficient. *BMC bioinformatics* 17:142. doi: 10.1186/s12859-016-0990-0
- Govindan, R. B., Massaro, A., Vezina, G., Chang, T., and du Plessis, A. (2019). Identifying an optimal epoch length for spectral analysis of heart rate of critically-ill infants. *Comput. Biol. Med.* 113:103391. doi: 10.1016/j.combiomed.2019.103391
- Guo, N., Chen, Y., Su, B., Yang, X., Zhang, Q., Song, T., et al. (2020). Alterations of CCR2 and CX3CR1 on three monocyte subsets during HIV-1/treponema pallidum coinfection. *Front. Med.* 7:272. doi: 10.3389/fmed.2020.00272
- Hall, M.A., and Smith, L.A. (1999). "Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper." 235–239.

- Hoepfer, M. M., Humbert, M., Souza, R., Idrees, M., Kawut, S. M., Sliwa-Hahnle, K., et al. (2016). A global view of pulmonary hypertension. *Lancet Respir. Med.* 4, 306–322. doi: 10.1016/S2213-2600(15)00543-3
- Hoepfer, M. M., Kramm, T., Wilkens, H., Schulze, C., Schafers, H. J., Welte, T., et al. (2005). Bosentan therapy for inoperable chronic thromboembolic pulmonary hypertension. *Chest* 128, 2363–2367. doi: 10.1378/chest.128.4.2363
- Huang, X., Li, Y., Guo, X., Zhu, Z., Kong, X., Yu, F., et al. (2019). Identification of differentially expressed genes and signaling pathways in chronic obstructive pulmonary disease via bioinformatic analysis. *FEBS Open Bio.* 9, 1880–1899. doi: 10.1002/2211-5463.12719
- Iribarren, P., Zhou, Y., Hu, J., Le, Y., and Wang, J. M. (2005). Role of formyl peptide receptor-like 1 (FPR1/FPR2) in mononuclear phagocyte responses in Alzheimer disease. *Immunol. Res.* 31, 165–176. doi: 10.1385/IR:31:3:165
- Jandl, K., Thekkekara Puthenparampil, H., Marsh, L. M., Hoffmann, J., Wilhelm, J., Veith, C., et al. (2019). Long non-coding RNAs influence the transcriptome in pulmonary arterial hypertension: the role of PAXIP1-AS1. *J. Pathol.* 247, 357–370. doi: 10.1002/path.5195
- Jardim, C., and Souza, R. (2015). Biomarkers and prognostic indicators in pulmonary arterial hypertension. *Curr. Hypertens. Rep.* 17:556. doi: 10.1007/s11906-015-0556-y
- Jiao, X., Sherman, B. T., Huang, D. W., Stephens, R., Baseler, M. W., Lane, H. C., et al. (2012). DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics* 28, 1805–1806. doi: 10.1093/bioinformatics/bts251
- Jin, H., Titus, A., Liu, Y., Wang, Y., and Han, A. Z. (2019). Fault diagnosis of rotary parts of a heavy-duty horizontal lathe based on wavelet packet transform and support vector machine. *Sensors* 19:4069. doi: 10.3390/s19194069
- Jivraj, K., Bedayat, A., Sung, Y. K., Zamanian, R. T., Haddad, F., Leung, A. N., et al. (2017). Left atrium maximal axial cross-sectional area is a specific computed tomographic imaging biomarker of World Health Organization Group 2 pulmonary hypertension. *J. Thorac. Imaging* 32, 121–126. doi: 10.1097/RTI.0000000000000252
- Jose, A., Kher, A., O'Donnell, R. E., and Elwing, J. M. (2020). Cardiac magnetic resonance imaging as a prognostic biomarker in treatment-naïve pulmonary hypertension. *Eur. J. Radiol.* 123:108784. doi: 10.1016/j.ejrad.2019.108784
- Keel, B. N., Snelling, W. M., Lindholm-Perry, A. K., Oliver, W. T., Kuehn, L. A., and Rohrer, G. A. (2019). Using SNP weights derived from gene expression modules to improve gwas power for feed efficiency in pigs. *Front. Genet.* 10:1339. doi: 10.3389/fgene.2019.01339
- Khandezamin, Z., Naderan, M., and Rashti, M. J. (2020). Detection and classification of breast cancer using logistic regression feature selection and GMDH classifier. *J. Biomed. Inform.* 111:103591. doi: 10.1016/j.jbi.2020.103591
- Kimura, D., McNamara, I. F., Wang, J., Fowke, J. H., West, A. N., and Philip, R. (2019). Pulmonary hypertension during respiratory syncytial virus bronchiolitis: a risk factor for severity of illness. *Cardiol. Young* 29, 615–619. doi: 10.1017/S1047951119000313
- Lee, J. S. (2012). Heterogeneity of lung mononuclear phagocytes in chronic obstructive pulmonary disease. *J. Innate Immun.* 4, 489–497. doi: 10.1159/000337434
- Mandras, S. A., Mehta, H. S., and Vaidya, A. (2020). Pulmonary hypertension: a brief guide for clinicians. *Mayo Clin. Proc.* 95, 1978–1988. doi: 10.1016/j.mayocp.2020.04.039
- Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* 405, 442–451. doi: 10.1016/0005-2795(75)90109-9
- McCabe, S. D., Lin, D. Y., and Love, M. I. (2020). Consistency and overfitting of multi-omics methods on experimental data. *Brief. Bioinform.* 21, 1277–1284. doi: 10.1093/bib/bbz070
- Mhandire, K., Duri, K., Kandawasvika, G., Chandiwana, P., Chin'ombe, N., Kanyera, R. B., et al. (2014). CCR2, CX3CR1, RANTES and SDF1 genetic polymorphisms influence HIV infection in a Zimbabwean pediatric population. *J. Infect. Dev. Ctries.* 8, 1313–1321. doi: 10.3855/jidc.4599
- Mirza, B., Wang, W., Wang, J., Choi, H., Chung, N. C., and Ping, P. (2019). Machine learning and integrative analysis of biomedical big data. *Gene* 10:87. doi: 10.3390/genes10020087
- Miyasaka, A., Yoshida, Y., Suzuki, A., Ueda, H., Morino, Y., and Takikawa, Y. (2020). A case of suspected portal-pulmonary hypertension due to hepatitis C virus infection. *Clin. J. Gastroenterol.* 13, 90–96. doi: 10.1007/s12328-019-01016-3
- Prieto-Gonzalez, D., Castilla-Rodriguez, I., Gonzalez, E., and Couce, M. L. (2020). Automated generation of decision-tree models for the economic assessment of interventions for rare diseases using the RaDiOS ontology. *J. Biomed. Inform.* 110:103563. doi: 10.1016/j.jbi.2020.103563
- Qiao, L., and Xie, D. (2019). MIonSite: ligand-specific prediction of metal ion-binding sites via enhanced AdaBoost algorithm with protein sequence information. *Anal. Biochem.* 566, 75–88. doi: 10.1016/j.ab.2018.11.009
- Rappaport, N., Twik, M., Plaschkes, I., Nudel, R., Iny Stein, T., Levitt, J., et al. (2017). MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Res.* 45, D877–D887. doi: 10.1093/nar/gkx1012
- Ren, J., Zhou, F., Li, X., Chen, Q., Zhang, H., Ma, S., et al. (2020). Semiparametric Bayesian variable selection for gene-environment interactions. *Stat. Med.* 39, 617–638. doi: 10.1002/sim.8434
- Risbano, M. G., Meadows, C. A., Coldren, C. D., Jenkins, T. J., Edwards, M. G., Collier, D., et al. (2010). Altered immune phenotype in peripheral blood cells of patients with scleroderma-associated pulmonary hypertension. *Clin. Transl. Sci.* 3, 210–218. doi: 10.1111/j.1752-8062.2010.00218.x
- Sara, J. D. S., Maor, E., Borlaug, B., Lewis, B. R., Orbelo, D., Lerman, L. O., et al. (2020). Non-invasive vocal biomarker is associated with pulmonary hypertension. *PLoS One* 15:e0231441. doi: 10.1371/journal.pone.0231441
- Schinkel, M., Paranjape, K., Nannan Panday, R. S., Skyttberg, N., and Nanayakkara, P. W. B. (2019). Clinical applications of artificial intelligence in sepsis: a narrative review. *Comput. Biol. Med.* 115:103488. doi: 10.1016/j.combiomed.2019.103488
- Shao, Y., Nir, G., Fazli, L., Goldenberg, L., Gleave, M., Black, P., et al. (2020). Improving prostate cancer classification in H&E tissue micro arrays using Ki67 and P63 histopathology. *Comput. Biol. Med.* 127:104053. doi: 10.1016/j.combiomed.2020.104053
- Shi, L., Westerhuis, J. A., Rosen, J., Landberg, R., and Brunius, C. (2019). Variable selection and validation in multivariate modelling. *Bioinformatics* 35, 972–980. doi: 10.1093/bioinformatics/bty710
- Simonneau, G., Montani, D., Celermajer, D. S., Denton, C. P., Gatzoulis, M. A., Krowka, M., et al. (2019). Haemodynamic definitions and updated clinical classification of pulmonary hypertension. *Eur. Respir. J.* 53:1801913. doi: 10.1183/13993003.01913-2018
- Soh, D. C. K., Ng, E. Y. K., Jahmunah, V., Oh, S. L., San, T. R., and Acharya, U. R. (2020). A computational intelligence tool for the detection of hypertension using empirical mode decomposition. *Comput. Biol. Med.* 118:103630. doi: 10.1016/j.combiomed.2020.103630
- Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., et al. (2015). Big data: astronomical or genetical? *PLoS Biol.* 13:e1002195. doi: 10.1371/journal.pbio.1002195
- Swaminathan, A. C., Dusek, A. C., and McMahon, T. J. (2015). Treatment-related biomarkers in pulmonary hypertension. *Am. J. Respir. Cell Mol. Biol.* 52, 663–673. doi: 10.1165/rcmb.2014-0438TR
- Tzimas, C., Rau, C. D., Buerger, P. E., Jean-Louis, G. Jr., Lee, K., Chukwunke, J., et al. (2019). WIP1 is a conserved mediator of right ventricular failure. *JCI Insight* 5:e122929. doi: 10.1172/jci.insight.122929
- Tzouvelekis, A., Herazo-Maya, J. D., Ryu, C., Chu, J. H., Zhang, Y., Gibson, K. F., et al. (2018). S100A12 as a marker of worse cardiac output and mortality in pulmonary hypertension. *Respirology* 23, 771–779. doi: 10.1111/resp.13302
- Wang, W., Ding, M., Duan, X., Feng, X., Wang, P., Jiang, Q., et al. (2019). Diagnostic value of plasma microRNAs for lung cancer using support vector machine model. *J. Cancer* 10, 5090–5098. doi: 10.7150/jca.30528
- Wang, C., Long, Y., Li, W., Dai, W., Xie, S., Liu, Y., et al. (2020). Exploratory study on classification of lung cancer subtypes through a combined K-nearest neighbor classifier in breathomics. *Sci. Rep.* 10:5880. doi: 10.1038/s41598-020-62803-4
- Wishart, D. S., Bartok, B., Oler, E., Liang, K. Y. H., Budinski, Z., Berjanskii, M., et al. (2021). MarkerDB: an online database of molecular biomarkers. *Nucleic Acids Res.* 49, D1259–D1267. doi: 10.1093/nar/gkaa1067
- Wu, X., You, W., Wu, Z., Ye, F., and Chen, S. (2020). Serum biomarker analysis at the protein level on pulmonary hypertension secondary to old anterior myocardial infarction. *Pulm. Circ.* 10:2045894020969079. doi: 10.1177/2045894020969079
- Xiao, M., Ma, F., Li, Y., Li, Y., Li, M., Zhang, G., et al. (2020). Multiparametric MRI-based radiomics nomogram for predicting lymph node metastasis in

- early-stage cervical cancer. *J. Magn. Reson. Imaging* 52, 885–896. doi: 10.1002/jmri.27101
- Xu, W., Liu, X., Leng, F., and Li, W. (2020). Blood-based multi-tissue gene expression inference with Bayesian ridge regression. *Bioinformatics* 36, 3788–3794. doi: 10.1093/bioinformatics/btaa239
- Ye, Y., Zhang, R., Zheng, W., Liu, S., and Zhou, F. (2017). RIFS: a randomly restarted incremental feature selection algorithm. *Sci. Rep.* 7:13013. doi: 10.1038/s41598-017-13259-6
- Yuan, Z., Zha, X., and Zhang, X. (2020). Adaptive multi-type fingerprint indoor positioning and localization method based on multi-task learning and weight coefficients k-nearest neighbor. *Sensors* 20:5416. doi: 10.3390/s20185416

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

Copyright © 2021 Liu, Zhang, Fu, Zhang and Zhou. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.