Zheng *et al. Journal of Biological Engineering* (2026) 20:175
https://doi.org/10.1186/s13036-026-00712-2

Journal of Biological Engineering

RESEARCH Open Access

Dynamic sample augmentation discovers gene biomarker for IgA nephropathy

Fang Zheng[1,2], Juanjuan Zhao[2*], Baoping Jia[3], Yafeng Li[4*], Yaheng Li[5], Yan Geng[6], Yan Qiang[7], Yifang Zheng[8*] and Yuanchen Gao[2]

Abstract

Bulk RNA-seq data suffers from the issues of "high dimensionality and small sample size," which limits its application in disease research. This paper proposes a dynamic data augmentation method based on Layer-wise Relevance Propagation (LRP) aimed at improving classification performance and biological interpretability under small-sample conditions. The method utilizes the LRP algorithm to calculate the contribution weight of each gene to the classification result and uses this weight to guide sample generation. By systematically amplifying biologically meaningful signals, it constructs semantically reliable augmented samples, avoiding the semantic distortion caused by traditional random perturbations. Simultaneously, a dynamic augmentation mechanism is introduced that tightly couples sample generation with model training, providing difficult-to-classify samples with multiple iterative optimization opportunities and forming a virtuous cycle where classification performance and augmentation quality improve synergistically. On this basis, population-level gene biomarkers are identified from the trained model. Innovatively, an open-environment enrichment analysis method is proposed—that is, instead of being limited to a few feature genes of a single subtype, the union of feature genes from all disease subtypes is taken for enrichment analysis, revealing shared biological pathways from a systems-level perspective and providing a more comprehensive interpretation for subtype-specific mechanism research. Experimental results show that this method effectively improves classification accuracy, and through this open-environment enrichment approach, six hub genes were identified as gene markers for IgAN.

Keywords LRP attention mechanism, Dynamic augmentation, Whole-geneset enrichment analysis, Gene biomarkers

*Correspondence:
Juanjuan Zhao
zhaojuanjuan@tyut.edu.cn
Yafeng Li
dr.yafengli@gmail.com
Yifang Zheng
zhengyifang@sxbqeh.com.cn
[1]School of Information, Shanxi University of Finance and Economics, No. 140 Wucheng Road, Xiaodian District, Taiyuan City 030006, China
[2]College of Technology and Computer Science (College of Data Science), Taiyuan University of Technology, 209 University Street Yuci District, Jinzhong City 030600, China
[3]Shanxi Cardiovascular Hospital, No. 18, Yifen Street, Wanbailin District, Taiyuan City 030024, China
[4]The Eighth Affiliated Hospital of Sun Yat-Sen University, 3025 Shennan Middle Road, Shenzhen City 518033, China
[5]Department of Nephrology, The Fifth Hospital of Shanxi Medical University, No. 29 Shuangtasi Street, Taiyuan City 030001, China
[6]Clinical Laboratory, The Fifth Hospital of Shanxi Medical University, No. 29 Shuangtasi Street, Taiyuan City 030001, China
[7]School of Software, North University of China, No.3, Xueyuan Road, Jianchao District, Taiyuan City 030051, China
[8]Department of Obstetrics and Gynecology, Shanxi Bethune Hospital, Shanxi Academy of Medical Science, 99 Longcheng Street, Taiyuan City 030032, China

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Introduction

With the remarkable advancement of sequencing technologies, a variety of sequencing methods have emerged. Compared to proteomics sequencing [1], DNA sequencing [2], and microarray technology [3], transcriptome sequencing [4] has gained significant attention in molecular mechanism research, precision disease diagnosis, and targeted drug development due to its unique advantages of high sensitivity, full transcriptome coverage, and dynamic regulatory analysis. Currently, many molecular biomarker discovery methods based on transcriptome sequencing have been developed [5, 6]. Literature [7, 8] indicates that, in general, models with better classification performance tend to have better interpretability. Therefore, in classification-based biomarker discovery, improving the performance of the classification model can lead to the discovery of more accurate biomarkers. However, transcriptome sequencing data is characterized by high-dimensionality and small-sample size—although each disease may have tens to hundreds of samples, it contains tens of thousands of gene features, posing significant challenges for gene sequencing data processing.

Dimensionality reduction and data augmentation are the most commonly used approaches to address this problem [9–15] are all excellent feature selection methods, while [6, 7, 16–23] are typical data augmentation techniques. Both feature selection and data augmentation have played significant roles in improving final classification performance. However, important challenges still persist. This paper focuses on data augmentation as the primary research object.

Another more direct solution to the small-sample problem is data augmentation. However, current augmentation strategies for omics data suffer from critical flaws that prevent them from becoming reliable solutions: (1) Semantic Unreliability: Methods such as random Gaussian noise or salt-and-pepper noise [24, 25] lack biological rationale, and their indiscriminate perturbations can easily corrupt the phenotypic label semantics of the samples. (2) Marginal Benefits: Mixup-based methods [26, 27] generate interpolated samples with ambiguous labels, resulting in very limited performance improvements. (3) Susceptibility to "Mode Collapse": Data augmentation based on GANs or diffusion models [28–31] inherently suffers from "mode collapse," leading to unequal opportunities for sample augmentation. This ultimately results in augmented data exhibiting severe long-tail distribution characteristics. (4) Static and Passive Workflow: Most critically, existing methods treat augmentation as a one-time static preprocessing step. This paradigm fails to leverage the continuously evolving knowledge of the model; difficult-to-classify samples receive only a single, blind augmentation attempt, leaving their potential for being correctly learned far from fully exploited.

Therefore, the core research gap is not simply the "lack of augmentation," but rather the absence of an augmentation framework that is theoretically grounded, dynamically adaptive, and actively collaborates with the classifier to overcome sample scarcity.

To fill this gap, we propose a paradigm shift from static, blind augmentation to dynamic, attention-guided augmentation. Our core innovation lies in utilizing Layer-wise Relevance Propagation (LRP)—not merely as a post-hoc explainability tool—but as a real-time, differentiable attention mechanism to guide the sample generation process. This design ensures that the generated samples are not random perturbations but systematic amplifications of biologically meaningful signals, thereby preserving semantic information intact.

Furthermore, we introduce a dynamic augmentation strategy integrated within the training loop. In this setup, model training and the augmenter are tightly coupled, forming a virtuous cycle: improvements in classifier performance yield higher-quality relevance scores, which in turn produce more reliable synthetic samples. This allows difficult samples to receive multiple, progressively refined opportunities for correction.

Within this augmented data environment, we are finally able to restore deep learning to its essential purpose—active feature discovery. By applying LRP to models trained on the augmented data, we identify gene biomarkers at the population level [8]. This workflow naturally leads to our final contribution: a novel, open-environment enrichment analysis. By performing enrichment on the union of feature genes across all disease subtypes, rather than on narrow top-K lists from a single subtype, we provide a systems-level view of disease pathology. This enables researchers to interpret subtype-specific biomarkers within the broader context of shared biological pathways, yielding deeper mechanistic insights.

In summary, this study no longer treats data augmentation as a supplementary technique, but establishes it as a core methodological lever to unlock the potential of deep learning for small-sample omics data, thereby achieving simultaneous improvements in classification performance, model interpretability, and subsequent biological discovery.

Finally, it is summarized that the overall workflow of our framework is as follows: First, data preprocessing is performed, and random forest is used for binary classification feature selection between disease and healthy groups, identifying 5,701 feature genes. Subsequently, deep learning is employed for model training and dynamic sample augmentation. Then, using an established method for identifying group-level gene biomarkers, 20 feature genes are identified for each disease. Finally, enrichment analysis is conducted on the union of feature genes from all diseases, leading to the

identification of 6 hub genes for IgA nephropathy. The overall flow chart of this paper is shown in Fig. 1.

The main contributions of our work are as follows:

(1) We introduce a sample augmentation strategy based on the Layer-wise Relevance Propagation (LRP) algorithm, where each gene's relevance score to the label is computed and used as a weight to generate augmented samples. Unlike random noise injection, this approach functions similarly to an attention mechanism, adaptively reinforcing the importance of key genes. It offers strong interpretability and follows a structured, rule-based computation.
(2) We innovatively develop a dynamic augmentation technique that performs classification and augmentation simultaneously. This allows difficult-to-classify samples multiple opportunities for refinement through iterative augmentation, improving model robustness and accuracy.
(3) Novely propose a gene enrichment method for a disease subtype in a more open environment, i.e. using the union of feature genes of all disease subtypes to do gene enrichment analysis rather than only the feature genes of the subtype which constrains the biological analysis in a narrow range including a few genes. After that, query the enrichment items select the related molecular BPs and pathways for every disease group according to its own feature genes.

Methodology

Data preprocessing based on feature selection

Since binary classification consistently outperforms multi-class classification under the same model framework, we propose the following approach: For each disease subtype in the classification task, we perform binary classification against the control group (normal samples) to identify the top-K most discriminative genes as the feature genes for that specific disease group. Subsequently, we compute the union of all these disease-specific feature genes across all subtypes to form the final feature gene set for the entire sample cohort. The detailed algorithm is as follows:

Let $S = \{S_1, S_2, \cdots S_i \cdots S_{N-1}\}$, represent the set of all disease group samples, where Sdenotes the sample set for the ith disease class, and S_N represents the control group (normal) samples. Here, N is the total number of classes.

Given a binary classification algorithmϕ, the $topU$ feature genes for the ith disease class are selected as follows:

$$M_i = topU\left(\phi\left(S_i, S_N\right)\right) \tag{1}$$

Then the features of all samples are as follows:

$$M = \cup_{i=1}^{N-1} M_i \tag{2}$$

Sample augmentation based on LRP relevance score attention mechanism

Review of LRP theory

A pixel-wise decomposition methodology of Layer-wise Relevance Propagation (LRP) [32] is proposed that can judge the contribution of a pixel wise to the classification of an image. Figure 2a and b respectively show the forward and backward propogation process.

We will use this example to explain the calculation principles of LRP.

In Fig. 2a, y_0 and y_1 represent the classification results, with $y_0 + y_1 = 1$. In Fig. 1b, ${R_7}^{(3)}$ and ${R_8}^{(3)}$ denote the relevance scores of Node 7 and Node 8 respectively, and ${R_7}^{(3)} = y_0$,${R_8}^{(3)} = y_1$.

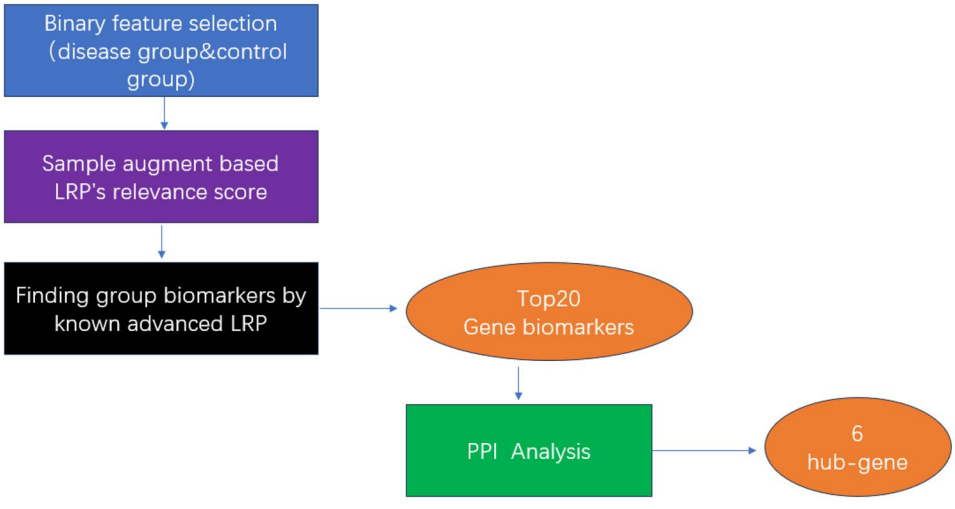


Fig. 1 The overall workflow of this paper

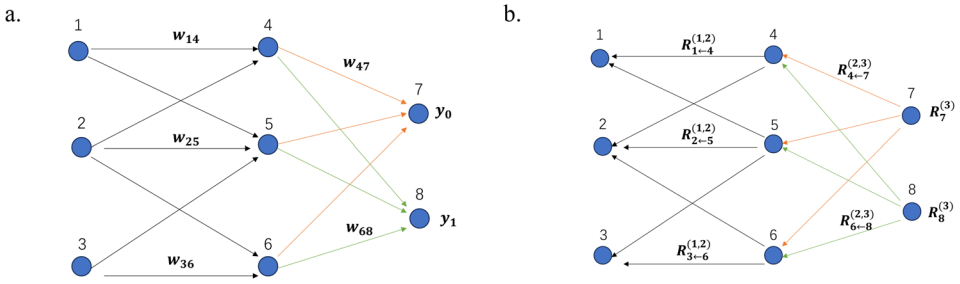


Fig. 2 The computation processes of relevance score.score. The green denotes the starting nodes of the computation and the yellow denotes the middle nodes. **a**. Forward propogation process. **b**. Backward propogation process

As can be seen from Fig. 2b,

$$R_7{}^{(3)} = R_{4\leftarrow 7}{}^{(2,3)} + R_{5\leftarrow 7}{}^{(2,3)} + R_{6\leftarrow 7}{}^{(2,3)} \tag{3}$$

$$R_8{}^{(3)} = R_{4\leftarrow 8}{}^{(2,3)} + R_{5\leftarrow 8}{}^{(2,3)} + R_{6\leftarrow 8}{}^{(2,3)} \tag{4}$$

Among them, $R_7{}^{(3)}$ denotes the relevance score of Node 7 in the 3rd layer. $R_{4\leftarrow 7}{}^{(2,3)}$ denotes the relevance score from Node 7 (in the 3rd layer) to Node 4 (in the 2nd layer).

According to the forward propagation, shown in Fig. 2a, the following is held.

$$a_4 w_{47} + a_5 w_{57} + a_6 w_{67} = a_7 = y_0 = R_7{}^{(3)} \tag{5}$$

So, according to the backward propagation(a_i is the input value of node i), get:

$$\begin{aligned} R_{4\leftarrow 7}{}^{(2,3)} &= a_7 \frac{a_4 w_{47}}{a_4 w_{47} + a_5 w_{57} + a_6 w_{67}} \\ &= R_7 \frac{a_4 w_{47}}{\sum_{i=4,5,6} a_i w_{i7}} \end{aligned} \tag{6}$$

$$\begin{aligned} R_{5\leftarrow 7}{}^{(2,3)} &= a_7 \frac{a_5 w_{57}}{a_4 w_{47} + a_5 w_{57} + a_6 w_{67}} \\ &= R_7 \frac{a_5 w_{57}}{\sum_{i=4,5,6} a_i w_{i7}} \end{aligned} \tag{7}$$

$$\begin{aligned} R_{6\leftarrow 7}{}^{(2,3)} &= a_7 \frac{a_6 w_{67}}{a_4 w_{47} + a_5 w_{57} + a_6 w_{67}} \\ &= R_7 \frac{a_6 w_{67}}{\sum_{i=4,5,6} a_i w_{i7}} \end{aligned} \tag{8}$$

Based on Eqs. (6), (7), and (8), we can calculate the contribution score of each node (Node 4, Node 5, and Node 6) in the 2nd layer to Class 7. Following this logic, we can also calculate $R_{1\leftarrow 4}{}^{(1,2)}$, $R_{1\leftarrow 5}{}^{(1,2)}$, and $R_{1\leftarrow 6}{}^{(1,2)}$. Then, the contribution score of Node 1 to Node 7 is as follows:

$$\begin{aligned} R_{1\leftarrow 7}{}^{(1,3)} = R_{1\leftarrow 4}{}^{(1,2)} \cdot R_{4\leftarrow 7}{}^{(2,3)} + R_{1\leftarrow 5}{}^{(1,2)} \\ \cdot R_{5\leftarrow 7}{}^{(2,3)} + R_{1\leftarrow 6}{}^{(1,2)} \cdot R_{6\leftarrow 7}{}^{(2,3)} \end{aligned} \tag{9}$$

Based on the above examples, the contribution score of any node in any layer to the classification result can be calculated. The formulas related to the above examples in Reference [32] are Eqs.(10), (11), (12), and (13):

$$R_{i\leftarrow k}{}^{(l,l+1)} = R_k{}^{(l+1)} \sum_j \frac{a_i w_{ik}}{\sum_i a_i w_{ik}} \tag{10}$$

Among them, $R_{i\leftarrow k}{}^{(l,l+1)}$ represents the relevance score from Node i to Node k between any two adjacent layers (e.g. $(ll+1)$).

$$\begin{aligned} \mathrm{y_i} = \mathrm{R}_i^{(o)} = \cdots \mathrm{R} &= \sum\nolimits_{d\in l+1,i} R_d{}^{(l+1)} \\ &= \sum\nolimits_{d\in l,i} R_d{}^{(l)} = \cdots = \sum\nolimits_{d1,i} R_d{}^{(1)} \end{aligned} \tag{11}$$

In Eq.(11), $R_d{}^{(1)}$ represents the relevance score of the d-th element x_d in the input vector of sample s for the classification result y_i.

$R_i^{(o)}$ denotes the ith output in the last layer and equals y_i. $\sum\limits_{d\in l,i} R_d{}^{(l)}$ denotes the relevance scores in the lth layer to the ith classification result.

$$\mathrm{y} = [y_1, y_2 \cdots y_i \cdots y_n] \tag{12}$$

$$\sum_{i=0}^{n-1} y_i = 1 \tag{13}$$

y_i is the ith item in prediction vector y, which denotes the probability of a certain sample being predicted as the ith category.

According to Eqs.(10), (11), (12), and (13), the relevance score of the d-th element in the input vector of sample s for the i-th classification result can be calculated.

When the model is a muti classification one and suppose there is n-class, the category corresponding to the maximum of y is the one that the sample is predicted by the model. i.e. to a sample s, if the following equations are held,

$$y_{max} = \max(y) \tag{14}$$

$$\mathrm{i} = \mathrm{loc}(y_{max}), \mathrm{i} \in [0, \mathrm{n}) \tag{15}$$

the category of the sample s predicted by the model is i.

y_{max} denotes the maximum of prediction vectory and is the location index of y_{max} in y.

Data augment based on relevance score

We know from the above section's discription, the relevance score $R_{d,i}^{(1)}$ denotes the contribution of the item x_d in the input vector X to the ith prediction value in the prediction vector y. Let

$$\alpha_{di} = R_{d,i}^{(1)} \tag{16}$$

α_{di} denotes the proportion of contribution of the item x_d to y_i, in which $\mathrm{x}_d \in \mathrm{X}$. The bigger $R_{d,i}^{(1)}$ is, α_{di} is the bigger and the contribution of d to y_i is the bigger.

Then the contribution score of all the item in input vector $X$to the category i is α_i:

$$\alpha_i = [\alpha_{1i}, \alpha_{2i} \cdots \alpha_{di} \cdots \alpha_{ni}] \tag{17}$$

$$X = [x_1, x_2 \cdots \mathrm{x}_d \cdots x_n] \tag{18}$$

Where n is the number of the input vector X.

α_i represents the attention scores, indicating the influence of each input element on the classification result.

Therefore, we define:

$$\mathrm{X}' = \mathrm{X} + \alpha_i \tag{19}$$

α_i fine-tunes the value of X according to the contribution of X to y_i. X_i' highlights the influence of the contribution to y_i.

In Eq.(19), we integrate the attention scores α_i with the original features X via element-wise addition. This approach is analogous to the residual connections used in networks like ResNet [15]. The role and benefits of residual connections have been rigorously demonstrated, both mathematically and empirically, in literature [15].

Dynamic data augmentation during training

Unlike traditional approaches where augmented data is pre-generated before model training, static data augmentation faces two critical challenges: (1) Not all augmented samples improve classification performance, when the original sample is misclassified (ground truth=i but predicted value=j), the correctness of its augmented counterpart remains uncertain; (2) Each sample only gets augmented once, depriving hard-to-classify samples of multiple enhancement opportunities. To address these limitations, we propose a dynamic augmentation technique.

Let M denote the model and S represent the original sample set. During initial training, when model predictions are unreliable (high error rate), premature augmentation would waste computational resources by generating many incorrect samples. Therefore, we first pre-trained a model using the non-augmented data, and then performed augmentation based on this model. Since k-fold cross-validation was employed, the data was divided into k folds. During the training of each fold, k-1 folds were selected as the training subset (train_subset), and the remaining one fold was used as the validation subset (val_subset). Augmentation was applied only to the training data each time, meaning only the k-1 folds were augmented, while the validation fold was not augmented. Since training was conducted k times in total, each sample had k-1 opportunities to be augmented. Before training each fold, the algorithm described in Section Data augment based on relevance score was called to augment the train_subset of that fold. A buffer NS was constructed to store the new samples that were successfully augmented in this fold. The criterion for successful augmentation was Equ.(15). Let the set of successfully augmented samples in each fold be NS,

and the set of unsuccessfully augmented samples be FS. Then $|FS| = |S| - |NS|$, where S represents the set of all original unaugmented samples. Let AS be the set of all augmented samples. Then:

$$AS = \sum_{i=1}^{k} NS_i \tag{20}$$

As the model's predictive capability improves through training, the quality of augmentation progressively enhances, creating a virtuous cycle that boosts both augmentation effectiveness and classification performance. The detailed algorithms are as. Algorithm I demenstrates the augment method for every sample s and Alogrithm II shows the dynamic augment process.

Algorithm I: Augment the samples

function name: augment_samples

S: sample set

X: input vector of a sample s

y: prediction vector

T: ground truth vector encoded with one-hot

M: the model has been trained

NS: the set of successfully augmented samples

$train_dataloader$: DataLoader of training dataset

Begin:

$NS \leftarrow [\,]$

for every X, label in train_dataloader:

$\mathrm{y} = \mathrm{M}(X)$

$R_i^{(o)} = y_i$

According to Equs.(10)(11)(12) and Equ(13) to compute relative score as weight of every item:

for every item in X:

$\alpha_i = [\alpha_{1i}, \alpha_{2i} \cdots \alpha_{di} \cdots \alpha_{ni}]$

$X' \leftarrow \mathrm{X} + \alpha_i$

$y' \leftarrow M(X')$

if $y_i' == label$, it proves that the label semantics are correct, then:

$NS.append(X')$

return NS

The algorithm of dynamicly augment is shown in Algorithm II:

Algorithm II: training process
S: the original sample set
FS: the sample set in which the samples are not augmented
NS: the sample set in which the samples are augmented
AS: all augmented samples
X: input vector of a sample s
y: prediction vector
M_P: the model has been trained
folds: the cross validation number
train_dataloader: training dataset, which is obtained by splitting the total dataset in an 8:2 ratio
train_subset: the k-1 folds used for training in each round of k-fold cross-validation
val_subset: the 1 fold used for validation in each round of k-fold cross-validation
Begin:
train_dataset, test_dataset ← split(dataset)
$model \leftarrow M_P$
for fold in kfolds:
    train_subset, val_subset ← split(train_dataset)
    train_loader ← DataLoader(train_subset)
    val_loader=DataLoader(val_loader)
    $NS \leftarrow$ **augment_samples(model, s)**
    train_subset.add_sample(NS)
    train_loader ← DataLoader(train_subset)
    for epoch in range(epoches):
        train_epoch(model,train_loader)
    test(model, test_dataloader)
    $AS.append(NS)$
    $NS \leftarrow [\,]$
End

TopK gene biomarkers discovery

Currently, most commonly used interpretability algorithms are fundamentally based on Layer-wise Relevance Propagation (LRP). However, LRP can only identify features for individual samples and fails to discover population-level characteristics. To address this limitation [7], proposed a median prototype-based algorithm for identifying group gene markers. In our study, we adopt the method introduced in [7] to calculate gene biomarkers at the population level. The related equations about gene biomarkers in [7] are as follows:

$$X^{(i)} = median_prototype \left(X_1^{(i)}, X_2^{(i)}, \cdots X_N^{(i)} | model\left(X_j\right) == y_i\right) \tag{21}$$

$$y = M\left(X\right) \tag{22}$$

$$C^{(i)} = LRP\left(y_i\right) \tag{23}$$

$$G^{(i)} = topK\left(C^{(i)}\right) \tag{24}$$

Where $median_prototype$ denotes the median prototype, M denotes the model, $X^{(i)}$ as input vector of gene sequence that is the average of all samples that are predicted correctly, y as the multi-classification result, y_i as the ith category's classification value, $C^{(i)}$ as the weights of every gene inferenced by LRP and $G^{(i)}G_i$ as the topK genes of the ith category.

Novel gene enrichment analysis

Define the features of all categories as:

$$G = \cup_{i=0}^{N} G^{(i)} \tag{25}$$

So, we think $G^{(i)}$ is the feature genes of the ith category and also at most the biomarkers of the ith group. But how to find the molecular mechanism of the ith disease group from the biomarkers. The traditional method is doing gene enrichment analysis or protein interaction analysis directly to the biomarkers [5]. However, we know the biomarkers is relatively small and they may be just the partial genes of some BPs or pathways. If only considerating the features of $G^{(i)}$ and just enriching them, the BPs and pathways only include the genes completely from $G^{(i)}$. But in fact, we know as long as a gene joins in some BP or pathway, we can think the change of the gene can influence the BP or pathway without requiring all the genes in the BP or pathway that belongs to it. Therefore, we hope place the biomarkers into an open environment to explore the molecular mechenism related with the disease. The total sample features G is just the suitable object. It includes the features of all categories and is justable for every sample because every sample has the same feature input. So, we propose to enrich the genes by G and then in the enriched items such as BPs, CCs, MFs, KEGG_pathways, find the genes which are involved into the ith features $G^{(i)}$. If the enriched items include the gene from $G^{(i)}$, we think the item may be related with the ith category. To make further ensure the probability by seaching in the PUBMED [13] through entering the key words: the item and the ith disease name, for example, *CP and IgA nephropathy*. The detailed steps are shown in Algorithm III:

Algorithm III

find the molecular mechenism for the ith disease

Begin:

$G^{(i)}$: the feature of the ith category and also the biomarkers of the ith category

G: the features of all the categories and also the input features of every sample

E: the set of the enriched items to $G^{(i)}$, in our experiment, items only denotes the biological BPs and KEGG_PATHWAYs

E_i: the subset of E and the item in which at least one gene belongs to $G^{(i)}$

F_i: the subset of E_i and selected by PUBMED from E_i

ST_i: the most potiential target set of the ith disease

num: the number of gene

Initialization:

$E = \emptyset$, $E_i = \emptyset$, $F = \emptyset$, num=0

Steps:

1) Do gene encirched analysis for G including GO annotations and KEGG_PATHWAY and store them into E;
2) for $\forall$gene $\in$ $G^{(i)}$:
   if $\exists$item $\in$ E, make item include gene:
   add gene into E_i
3) Take every item $\in$ E_i and the ith disease name as the key words into PUBMED
   if the result is not none:
   add item into F_i

Novely gene enrichment analysis

All statistical analyses were performed as follows. Classification accuracy confidence intervals (95% CI) were estimated using the bootstrap method with 1000 resamples. Differential expression analysis between IgAN and control samples was performed using the limma package (version 3.48.0) with Benjamini-Hochberg false discovery rate (FDR) correction for multiple testing; adjusted p-values < 0.05 were considered statistically significant. For GO/KEGG enrichment analysis, FDR-adjusted p-values were calculated using the DAVID online tool. PPI network edge confidence scores were obtained from the STRING database (version 11.5); only edges with confidence score > 0.4 were considered for analysis.

Data and experiment setup

The kidney subtypes comprise 9 categories with a total of 974 samples (see Table 1). The bulk RNA-seq data were processed and converted into a CSV file with samples as columns and genes as rows. Subsequently, following the

Table 1 Data types and sample number of kidney disease

Disease name	Data sources		samples	label
Lupus Nephritis, LN	**GSE175759:3**	**GSE157293:6**	**43**	**0**
	GSE98422:20	**GSE296560:14**		
Membranous Glomerulonephritis, MN	GSE197307:50	GSE175759:4	129	1
	GSE175759:4	GSE182380:50		
	GSE216841:13	GSE216841:12		
nephrectomy control,Normal	GSE98422:7	GSE175759:22	95	2
	GSE216841: 8	GSE296560:5		
	GSE289843:4	GSE216841:8		
	GSE242476:4	GSE197307:8		
	GSE228971:3	GSE274774:21		
	GSE133530:4	GSE139061:11		
	GSE285469:11	GSE261018:5		
Minimal Change Disease, MCD	GSE246204:11	GSE216841:14	135	3
	GSE175759:9	GSE197307:80		
	GSE182380:30			
Focal Segmental Glomerulosclerosis, FSGS)	GSE176465:23	GSE197307:70	198	4
	GSE175759:3	GSE140989:68		
	GSE182380:60			
IgA Nephropathy, IgAN	GSE243078:9	GSE210098:17	93	5
	GSE141295: 24	GSE175759:43		
Diabetic Nephropathy, DN	GSE142025:36	GSE162830:32	68	6
Acute Kidney Injury, AKI	GSE261018:34	GSE142077:15	86	7
	GSE139061:37			
Chronic Kidney Disease, CKD	6		127	8
total			974	9

data processing method provided by scBERT, the data were transformed into an h5ad file, ultimately obtaining a dataset containing 974 samples and 16,909 genes. It was then converted back to a CSV file. The purpose of this conversion was to obtain the final set of genes from the data (as the original data contained over 50,000 genes). Then, using the data preprocessing method described in Section Data preprocessing based on feature selection, binary classification was performed for each disease group and control group. The binary classification method used was Random Forest. Genes with binary classification accuracy higher than 85% were selected. Statistics showed that the maximum number of genes (U) meeting the condition of classification accuracy > 85% for each subtype was 900; that is, $U = 900$ in Eq. (1). The main reason for choosing 85% classification accuracy was that, among these controls, it was ultimately found that the classification accuracy was mostly around 85%. The highest reached 90%, and the lowest was 78%. Therefore, 85% was finally chosen. The number of features selected for each disease was 900 because, through multiple experiments, it was found that for most diseases, when the feature count reached 900, the binary classification accuracy for that disease group essentially reached its maximum. The maximum of U was 938, the minimum was 793, and most fluctuated around 900. Consequently, 900 was ultimately selected. A sensitivity analysis confirmed that the six hub genes remained stable when varying the threshold between 80 and 90% (Supplementary Figure 1). Next, the union set of genes from all disease groups was calculated as the final feature set for all categories, as described in Eq. (2). Subsequently, the data after feature selection were split into training, validation and test sets in an 8:1:1 ratio. Using 5-fold cross-validation, the augmented dataset was divided into 5 folds. Each fold was trained for 100 epochs. For each fold, 4 folds were used as the training subset (train_subset), and the remaining fold was held out as the validation subset (val_subset). Data augmentation was applied exclusively to the train_subset in each fold using Algorithm I. Algorithm II operates by first calling Algorithm 1 to augment the data and then proceeding with model training.

To sum up, feature selection was performed first, followed by cross-validation training. Each fold was trained on the data after feature selection. No feature selection was conducted during the training process. A single, unified feature selection was performed only once at the beginning of the entire work, in the data preprocessing section.

Regarding the computation of LRP-relevant scores, our implementation was based on the code provided by Wang et al. [33], specifically by calling the `inner_model.investigator(x)` function. The model architecture is $S_M \rightarrow 2000 \rightarrow 100 \rightarrow 9$, where each layer incorporates Batch Normalization and Dropout. Before performing data augmentation, we pre-trained a model on the original data using 5-fold cross-validation.

The specific procedural workflow for all the steps described above is as follows: (1) For each disease group and control group, perform binary classification using Random Forest, and select the top U genes with the highest feature importance weights as the characteristic genes for that disease group. This process follows Equ. (1), where the function in Eq. (1) is the Random Forest algorithm. Genes with binary classification accuracy higher than 85% were selected. Statistics indicated that the maximum number of genes (U) satisfying the condition of classification accuracy > 85% for each subtype was 900; hence, $U = 900$ in Eqs. (1). (2) Calculate the union of genes from all disease groups as the final feature set for all categories, according to Eqs. (2). (3) Train a pre-trained model using the dataset after feature selection. (4) Based on Algorithm I and Algorithm II, perform training with integrated data augmentation using the 5-fold cross-validation method on the Multilayer Perceptron (MLP) model (architecture: $S_M \rightarrow 2000 \rightarrow 100 \rightarrow 9$, with Batch Normalization and Dropout in each layer). This experiment is conducted on the basis of the pre-trained model. (5) Identify the top 20 key genes for each subtype using the

method described in Section Dynamic data augmentation during training, Eqs.(21-24) of reference [7]. (6) Perform gene enrichment analysis according to Algorithm III. Comparative and ablation experiments are conducted subsequently.

Contrast experiment

Figure 3 presents a comparative analysis of classification performance under different data augmentation approaches. The results demonstrate that GAN-based augmentation yields the lowest classification accuracy, likely attributable to the "mode collapse" phenomenon in GANs. The baseline model using non-augmented data achieves 89.0% accuracy (95% CI: 83.5–93.2%). The introduction of random Gaussian noise improves performance to 90.0% (95% CI: 84.7–94.1%). Salt-and-pepper noise further enhances accuracy to 91.3% (95% CI: 86.2–95.0%). The LRP-augment(S) method achieves 93.8% accuracy (95% CI: 89.5–96.8%). The LRP-augment(D) method reaches the peak performance of 95.3% (95% CI: 91.2–97.8%). All confidence intervals were estimated using the bootstrap method with 1000 resamples.

The baseline model using non-augmented data achieves 89.0% accuracy. The introduction of random Gaussian noise improves performance to 90.0%, presumably due to increased sample quantity. Salt-and-pepper noise further enhances accuracy to 91.3%. While Mixup shows additional improvement, it doesn't reach maximum performance, possibly because it expands both sample size and category complexity. Notably, all these augmentation methods risk introducing label errors where augmented samples might be misclassified. The LRP-augment(S) method achieves 93.8% accuracy, where 'S' denotes static augmentation - a one-time enhancement combined with original samples for training. This approach utilizes LRP scores as feature-specific weights for augmentation, systematically strengthening features based on their contribution to classification. The LRP-augment(D) method reaches the peak performance of 95.3%, where 'D' represents dynamic augmentation that provides multiple enhancement opportunities for challenging samples, increasing the probability of correct classification through iterative attempts.

Ablation experiment

Feature selection

Table 2 compares the classification results with and without feature selection, showing that the classification accuracy increases by 6.1% after feature selection.

LRP-augment

Table 3 compares classification results with and without sample augmentation, showing that the classification accuracy increases by 6.2% when using augmented data.

Figure 4 shows the ROC Curves before augment (shown in Fig. 4a) and after augment (shown in Fig. 4b). As observed in Fig. 4, the macro-average accuracy improves from 93% in panel (a) to 97% in panel (b). Furthermore, the classification accuracy for every subtype in Fig. 4b is equal to or higher than the corresponding accuracy in Fig. 4a. A case in point is IgAN, which demonstrates an accuracy of 93% in Fig. 4b compared to only 90% in Fig. 4a.

Classification comparison of every Fold

Table 4 shows the classification effects of different folds. $|S|$ denotes the training sample set. In k-fold cross-validation, the subset used for training in each fold is

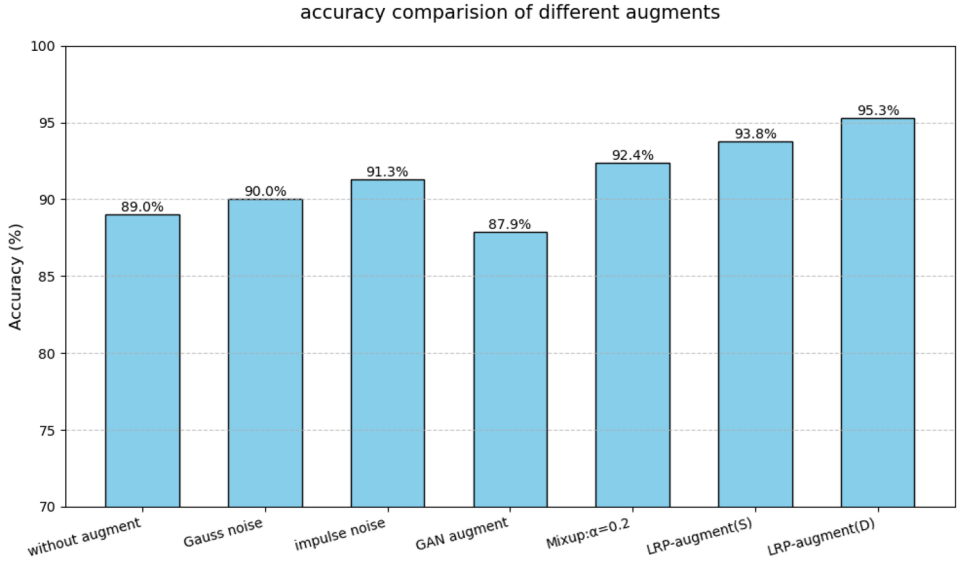


Fig. 3 Comparison of different augments

Table 2 Comparison of different gene number

Gene number	model	F1_score	Accuracy	Recall	Prediction
16909	MLP	82.3%	83.6%	84.1%	84.8%
5701	MLP	**88.3%**	**89.0%**	**89.4%**	**89.8%**

of size $(k-1)|S|$; this is the portion of the data that is augmented. We can discover that the fifth fold reaches the best result. It may be because the number of non-augmented samples ($|FS|$) reaches the minimum, while the number of augmented samples ($|NS|$) reaches the maximum. The number after being augmented for training in every fold is |AS|. From Table 4, we can discover that |AS| increases with the fold and reaches 2,816 in the five fold. The accuracy also reaches the maximum 94.6% in the five fold, which is because of the increase of number of samples.

4.2.4 Comparison of $|S|, |NS|, |FS|$ in different folds

Fig. 5 shows the value of $|AS|$, $|NS|$ and $|FS|$ which change from fold one to fold five. From the figure, we can discover that with the fold number increase, the $|NS|$ becomes bigger and bigger and $|FS|$ becomes smaller and smaller. It means that there are more samples augment correctly with the fold increase.

Proof that the performance gain is due to the proposed augmentation strategy or simply increased sample size

The improvement in performance could be attributed either to the proposed augmentation strategy or merely to the increase in sample size. We conducted the following experiment to investigate this. Random sampling was performed on the original samples until the number of samples reached 2,816. This value corresponds to the maximum value of |AS| in Fig. 5, i.e., the maximum number of samples after augmentation using LRP-relevant scores. The sampled data was then trained on the pre-trained model (training only, without augmentation), and the classification accuracy stabilized at 91.8%. This is lower than the 95.4% achieved under the augmentation strategy. Furthermore, 91.8% is the value observed for the 2nd fold under the augmentation strategy, as shown in Table 5. From the table, it is evident that increasing the number of samples can indeed improve classification accuracy, but the improvement is limited. The augmentation strategy demonstrably enhances classification performance.

Time complexity Analysis

The time cost of this study primarily consists of the augmentation time and the training time. Let the number of original data samples be $|S|$, which is split into a training set (train_dataset) and a test set (test_dataset) in an 8:2 ratio. During the training phase of k-fold cross-validation, the train_dataset is partitioned into k-1 folds as the training subset (train_subset) and 1 fold as the validation subset (val_subset). Let the number of samples in train_dataset be $|TS|$; then the number of samples in train_subset is $\frac{k-1}{k} * |TS|$.

Assume that the time complexity of running one epoch for an MLP model is $O(n*d)$, and the time complexity of running one epoch of LRP is also $O(n*d)$, where n represents the number of samples and d represents the number of features. Let each fold be trained for c epochs. Then, the time complexity of k-fold cross-validation is $k*c*O(n*d)$.

In this study, prior to augmentation, a pre-training of the model using k-fold cross-validation is conducted, with a time complexity of $k*c*O(|TS|*d)$.

Subsequently, a new k-fold cross-validation process begins, which includes data augmentation and training. The augmentation step uses 1 epoch, and the training step uses c epochs, with the training set size being $\frac{k-1}{k} * |TS|$. The time complexity per fold for augmentation is $O\left(\frac{k-1}{k} * |TS| * d\right)$, and for training, it is $c * O\left(\frac{k-1}{k} * |TS| * d\right)$. Thus, the time complexity for k folds is $k * \left(O\left(\frac{k-1}{k} * |TS| * d\right) + cO\left(\frac{k-1}{k} * |TS| * d\right)\right)$.

Therefore, the overall time complexity is: $k*c*O(|TS|*d) + k * \begin{pmatrix} O\left(\frac{k-1}{k} * |TS| * d\right) \\ +c * O\left(\frac{k-1}{k} * |TS| * d\right) \end{pmatrix}$, namely:

$$O \sim k*c*O(n*d) + k*(O(n*d) + c*O(n*d))$$

$$\sim k*c*O(n*d) + k*(O(n*d) + c*O(n*d))$$

$$\sim (2k+1)*c*O(n*d)$$

$$\sim k*c*O(n*d)$$

The term $k*c*O(n*d)$ represents the time complexity of training a k-fold cross-validation model without augmentation, where each fold is trained for c*c* epochs. This implies that after introducing augmentation, the time complexity only increases linearly, rather than changing in a fundamental or exponential way.

In addition to the theoretical analysis, we also provide a time curve (Fig. 6) showing the duration of each fold throughout the augmentation and training process. As shown in the figure, the average time per fold remains around 6.5 minutes, resulting in a total of approximately 40 minutes for all 5 folds. Compared to current architectures such as Transformer or GNN—which often require

Table 3 Comparison of whether using LRP augments

augment	model	F1_score	Accuracy	Recall	Prediction
no	MLP	88.3%	89.0%	89.4%	89.8%
yes	MLP	**94.6%**	**95.3%**	**96.1%**	**96.8%**

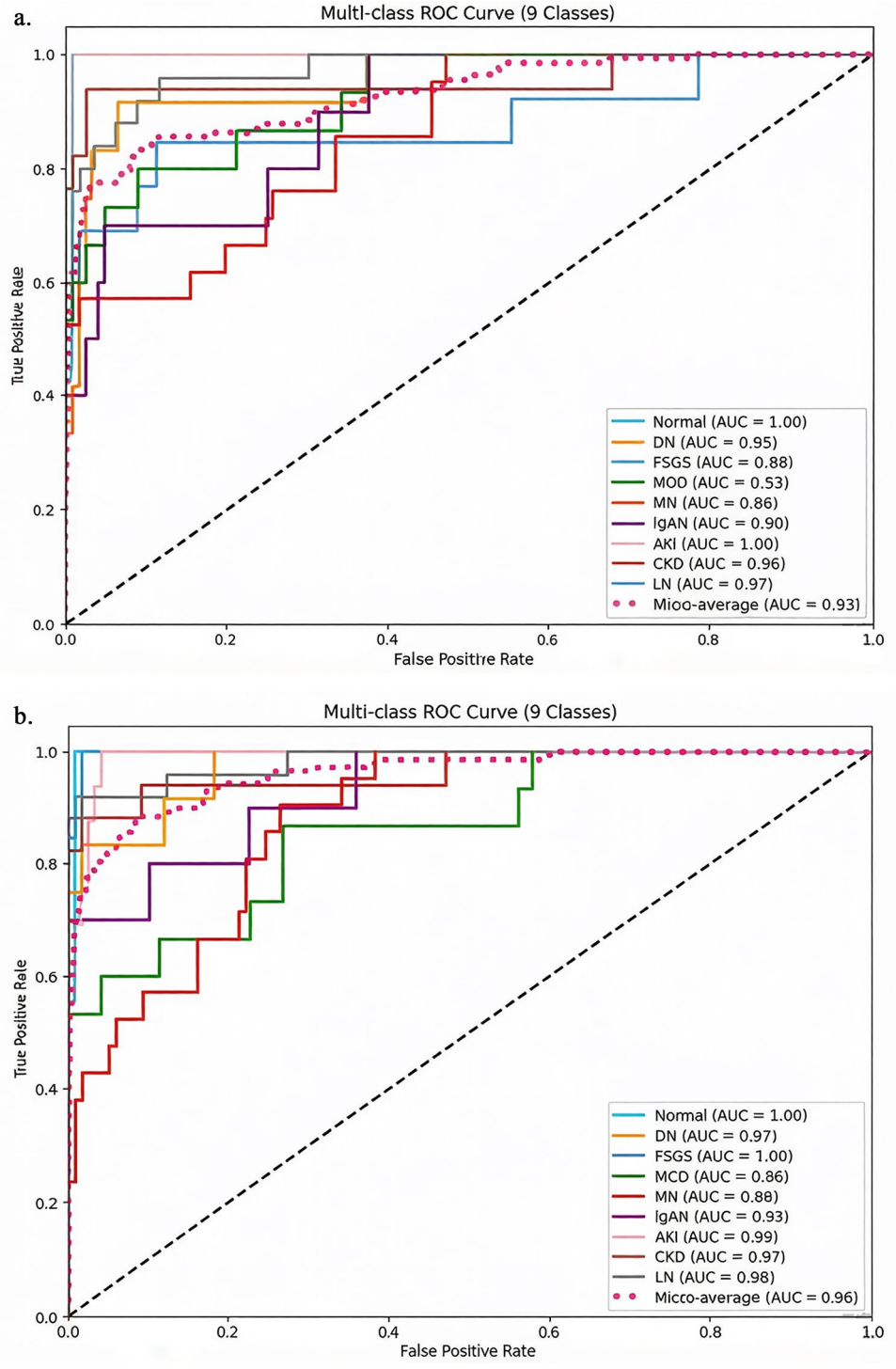


Fig. 4 Comparison of ROC curves before and after augment

Table 4 Comparison of different folds

Fold	$\|S\|$	|AS|	(k-1)/k*|S|	$\|NS\|$	$\|FS\|$	model	F1_score	Accuracy	Recall	Prediction
1	779	724	623	101	522	MLP	89.6%	90.1%	91.0%	91.2%
2	779	1,101	623	377	246	MLP	91.1	91.8	92.0	92.2
3	779	1,593	623	492	131	MLP	92.3	93.4	94.0	94.5
4	779	2,178	623	585	62	MLP	93.6	94.2	95.0	95.4
5	779	2,816	623	607	16	MLP	**94.6**	**95.3**	**96.1**	**96.8**

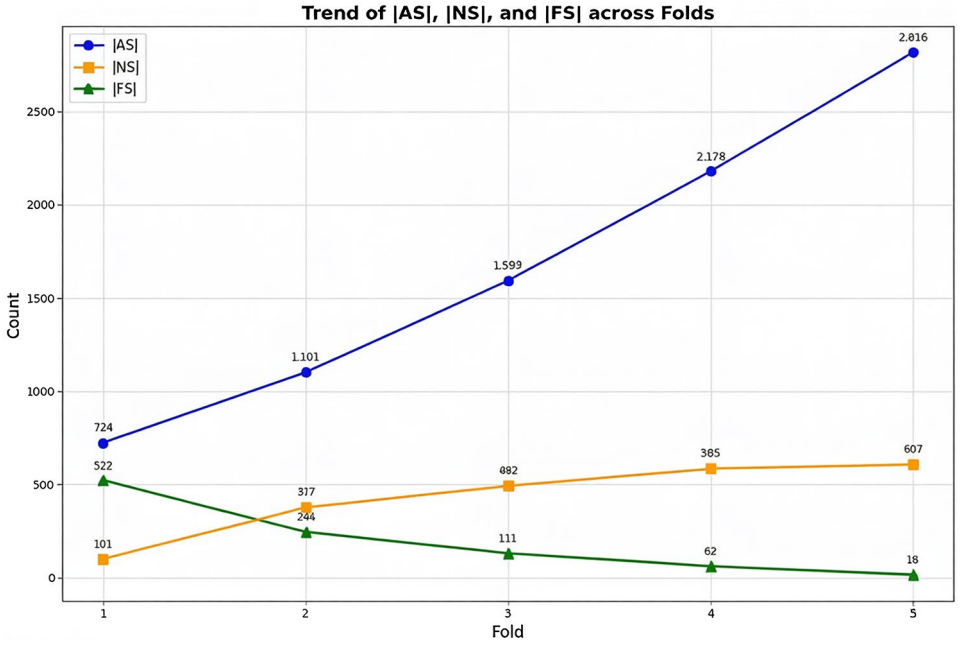


Fig. 5 Comparison of $|AS|$, $|NS|$ and $|FS|$

Table 5 Comparison of different folds

Method	Sample number	model	F1_score	Accuracy	Recall	Prediction
random sampling	2,816	MLP	91.1	91.8	92.0	92.2
LRP augmentation	2,816	MLP	**94.6**	**95.3**	**96.1**	**96.8**

days or even weeks to train on petabyte-scale data—our approach remains relatively simple and efficient. This is primarily thanks to our small dataset and simple model.

Gene set enrichment analysis

Take top20 gene as features of every disease. Figure 7 shows the heatmap of the top 20 genes in IgAN subtype. The total genes number selected for all diseases according to top20 is 90. Do gene enrichment analysis to the 90 genes using GO ontology and KEGG pathway databases. Take the 90 genes as an input into the webset: https://david.ncifcrf.gov/summary.jsp and get the sets of BPs (Biological Process), CCs(Cellular Component), MFs(Molecular Function) and KEGG_PATHWAYs. Figure 7a shows the enrichment results of total 90 genes. Figure 7b shows the BPs, CCs, MFs and KEGG_PATHWAYs for IgA Nephropathy(IgAN), which are gained by querying the enriched results of 90 genes according to the method in section TopK gene biomarkers discovery.

And then take every item in the enrichement sets and IgAN as the key words input into PUBMED: https://pubmed.ncbi.nlm.nih.gov/. If the query results is not NULL, we think the item is related with the disease IgAN. Figure 7b shows the enrichment items related with IgAN.

Table 6 shows top20 genes of IgAN according to the method in section Dynamic data augmentation during training and the related information about the 20 genes respectively from relevance score order, CC number, BP and MF number, BP number and KEGG_PATHWAY number they joined. Relevance score order reflects the contribution of the gene to the classification result. In principle, the closer to 0 the location is, the contribution is bigger and the probability of being biomarker of the category trends bigger; CC number reflects the cell types number of the gene locates. The bigger the number is, all related cell component will be influenced easier once the gene mutations; BP,MF and KEGG_PATHWAY number reflect the number that the gene join in the

reactions and are reflections of the activity of a gene. Figure 7a shows the GO annotation and KEGG_PATHWAY of total 90 sample genes selected according to top20 and Fig. 7b shows thoes for IgAN selected according to Algorithm $= 3\backslash*$ROMANIII. In Fig. 6b, there are eight BPs and two KEGG pathways. The first KEGG_PATHWAY is hsa01100:Metabolic pathways (FDR = 0.015) and the enriched genes related with IgAN are GPX2, AGL, FMO2, CYP3A5 and PDE3B. The related reference is [17]. The second KEGG_PATHWAY is hsa00140:Steroid hormone biosynthesis (FDR = 0.028) and the enriched genes with IgAN are GPX2 and CASP8. The related reference is [18] etc.

The first BP is GO:0008202~steroid metabolic process (FDR = 0.012) and the enriched genes related with IgAN are AKRIC2 and CYP3A5. The related reference is [19] etc. The second BP is GO:0001525~angiogenesis (FDR = 0.035) and the enriched genes are CASP8 and PDE3B. The related reference is [20]. The third BP is GO:0010628~positive regulation of gene expression (FDR = 0.028) and the enirched genes related with IgAN is TFAP2A. The related reference is [21] etc. The fourth BP is GO:0060395~SMAD protein signal transduction (FDR = 0.018) and the enriched gene related with IgAN is GDF15. The related reference is [22] etc. The fifth BP is GO:0097190~apoptotic signaling pathway (FDR = 0.022) and the enriched genes related with IgAN are GDF15. The related reference is [23] etc. The sixth BP is GO:0007179~transforming growth factor beta receptor signaling pathway (FDR = 0.015) and the enriched genes related with IgAN is GDF15. The related reference are [22, 23, 30, 31] etc. The seventh BP is GO:0001822~kidney development (FDR = 0.032) and the enriched gene is TFAP2A. The related reference is [31] etc. The eighth BP is GO:0006508~proteolysis (FDR = 0.025) and the enriched genes are CASP8, F12 and LCN1. The related reference is [12].

Finally, from Table 5, we can see that AGL, FMO2, CYP3A5, PDE3B, GPX2,AKRIC2, PDE3B, GDF15, TFAP2A, CASP8, F12 and LCN1 appear on the many KEGG and BPs. So, we guess that they are the most likely candidate targets of IgAN.

PPI analysis

To further validate our enrichment method, we input the identified top 20 gene markers into the STRING [14] network. Figure 9 shows the results of the PPI network analysis. Figure 9a displays the enrichment graph with only 20 nodes, revealing that GPX2 is a hub gene connecting AKR1C2 and CASP8. Meanwhile, LYPLA1 and ABHD2 are also connected. These five genes play important roles. When the number of nodes increases to 25, as shown in Fig. 9b, the role of AKR1C2 becomes significantly more prominent, connecting with many genes. In Fig. 9b, CYP3A5 also participates in connections and maintains relationships with several other genes. When the node count increases to 31, as shown in Fig. 9c, we observe that AKR1C2 and CYP3A5 become surrounded by more genes, emerging as true hub genes. The PPI network analysis further validates our enrichment method, demonstrating that using genes from a single disease subtype cannot reveal complete gene relationships. The true picture of gene relationships only emerges when we increase the number of genes.

The specific analysis of the different results of the 20, 25, and 31 genes is as follows:

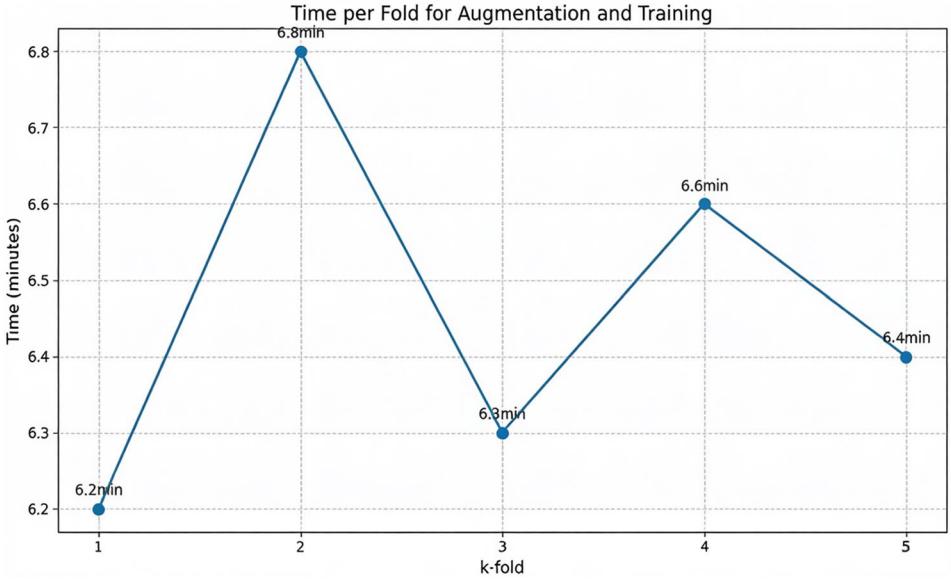


Fig. 6 Comparison of different noise augments

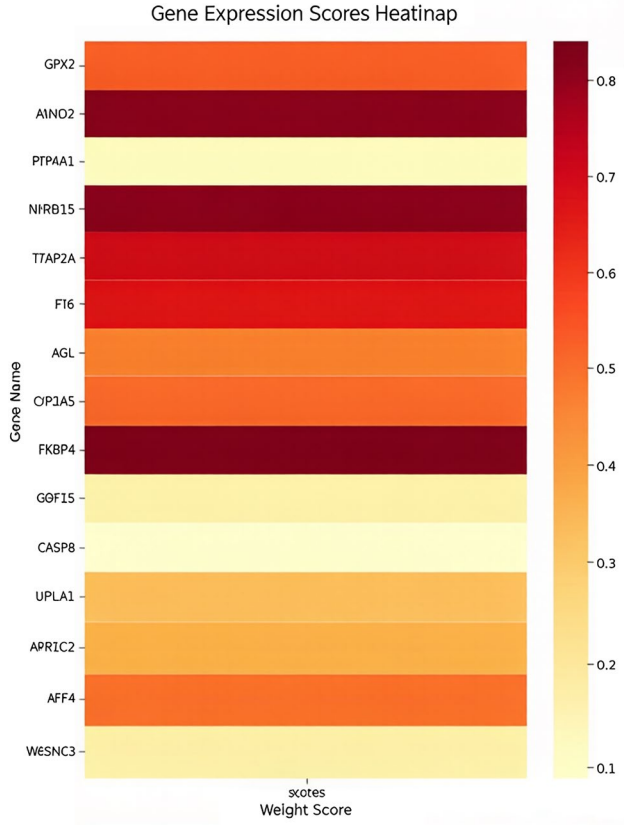


Fig. 7 The heatmap of the top 20 genes in IgAN subtypes

20-node network (Figure 9a): At this minimal scale, only two connections were identified:

- AKR1C2—GPX2—CASP8 chain: AKR1C2 (steroid metabolism) connects to GPX2 (glutathione peroxidase, an antioxidant enzyme), which in turn connects to CASP8 (a key executor of apoptosis). This chain forms a coherent functional cascade linking steroid metabolism → oxidative stress → apoptosis—three processes centrally implicated in IgAN pathogenesis.
- LYPLA1—ABHD2 direct connection: Both genes belong to the serine hydrolase superfamily and are involved in lipid metabolism. Their direct connection at the minimal network scale suggests they may function as a coordinated enzymatic module.

The presence of this biologically coherent cascade at the minimal network scale strongly suggests that our PPI analysis captures genuine biological relationships rather than statistical artifacts.

(2) 25-node network (Fig. 9b): When the network expands to 25 nodes, the connection between GDF15 and its functional receptor GFRAL emerges. GDF15-GFRAL is a well-characterized ligand-receptor pair involved in metabolic stress and inflammatory responses. The capture of this known biological relationship further validates the biological relevance of our PPI analysis.

(3) 31-node network (Fig. 9c): At this scale, the core structure from the 20-node network remains stable. Notably, AKR1C2 and CYP3A5 become surrounded by an increasing number of interacting partners, emerging as true hub genes. Additionally, TFAP2A localizes at the periphery of the AKR1C2-centered network, consistent with its role as a transcription factor that may act downstream of metabolic signals to regulate kidney development (GO:0001822).

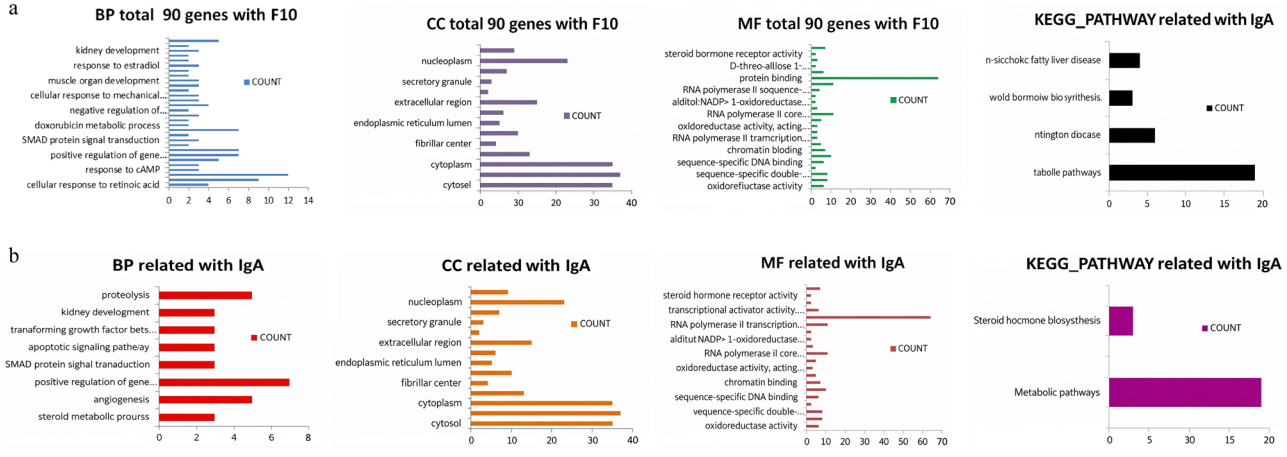


Fig. 8 Visualization of GO annotation and KEGG_PATHWAY. **a**. Visualization of GO annotation and KEGG_PATHWAY with F10 for 90 genes from all categories. **b**. Visualization of GO annotation and KEGG_PATHWAY for genes IgAN with F10, which is a subset of thoese of genes 90 and makes further selection through PUBMED. All reported GO terms and KEGG pathways have FDR < 0.05 (Benjamini-Hochberg correction)

Table 6 20 genes of IgAN selected according to F10 embody on molecular level which are enriched with 90 sample features

gene	KPNA3	GPX2	ABHD2	PTP4A1	NPIPB15	LCN1	TFAP2A	IFI6	AGL	CYP3A5	FKBP4	GDF15	CASP8	LYPLA1	F12	PDE3B	AKR1C2	AFF4	WASHC3	FMO2
order according to reference score	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
CC number	4	2		3	2	1	3	2	4		5	3	5	3	1	2	1	2		
BP，MF,KEGG number	1	2	3	1		3	16	3	3	7	2	4	9	2	2	4	8	1	1	2
riched on BPs			1			1	7	2	1	1	1	2	6		1	1	4			
riched on pathway		2							1	2			2			1	1			1

So, from PPI analysis we can gain 8 hub genes for IgAN ans they are respectively AKR1C2,CYP3A5,GPX2,CASP 8,LYPLA1,ABHD2,GDF15,TFAP2A.

Results

To select the most robust candidate biomarkers for IgAN, we integrated the results from LRP ranking, GO/KEGG enrichment, and PPI network analysis (Table 7). The intersection of these three lines of evidence yielded 6 common genes:

These six genes satisfy all three criteria: (i) high LRP contribution to IgAN classification, (ii) central or functionally relevant positions in the PPI network.

Table 8 shows differential expression analysis of the six hub genes between IgAN and control samples. Differential expression analysis was performed using the limma package with Benjamini-Hochberg FDR correction. All six hub genes showed significant differential expression between IgAN and control samples (FDR < 0.05).

Discussion

Summary of Principal Findings

In this study, we proposed a dynamic data augmentation framework guided by Layer-wise Relevance Propagation (LRP) to identify gene biomarkers for IgA nephropathy (IgAN) in small-sample, high-dimensional data. By integrating LRP-based ranking, GO/KEGG functional enrichment, and PPI network analysis, we identified six hub genes—AKR1C2, CYP3A5, GPX2, CASP8, GDF15, and TFAP2A—as the most likely candidate biomarkers for IgAN. These genes form an integrated network connecting steroid metabolism, oxidative stress, apoptosis, TGF-β/SMAD signaling, and kidney development.

Biological interpretation

The AKR1C2–GPX2–CASP8 cascade. Our PPI analysis revealed a direct chain linking these three genes at the 20-node scale (Fig. 9a). AKR1C2 is involved in steroid metabolism; GPX2 is an antioxidant enzyme bridging metabolism to oxidative stress; and CASP8 is a key executor of apoptosis. This cascade supports a mechanistic model in which steroid metabolism dysregulation → oxidative stress → apoptosis contributes to IgAN pathogenesis [17–19, 23].

The TGF-β/SMAD axis: GDF15. GDF15 was enriched in TGF-β receptor and SMAD signaling pathways (GO:0060395, GO:0007179) and formed a connection with its receptor GFRAL at the 25-node scale (Fig. 9b). Given the central role of TGF-β signaling in renal fibrosis, GDF15 represents a promising therapeutic target or diagnostic marker for IgAN [22, 23, 30, 31].

Transcriptional regulation: TFAP2A. TFAP2A localized at the periphery of the AKR1C2 network (31-node, Fig. 9c) and was enriched in kidney development

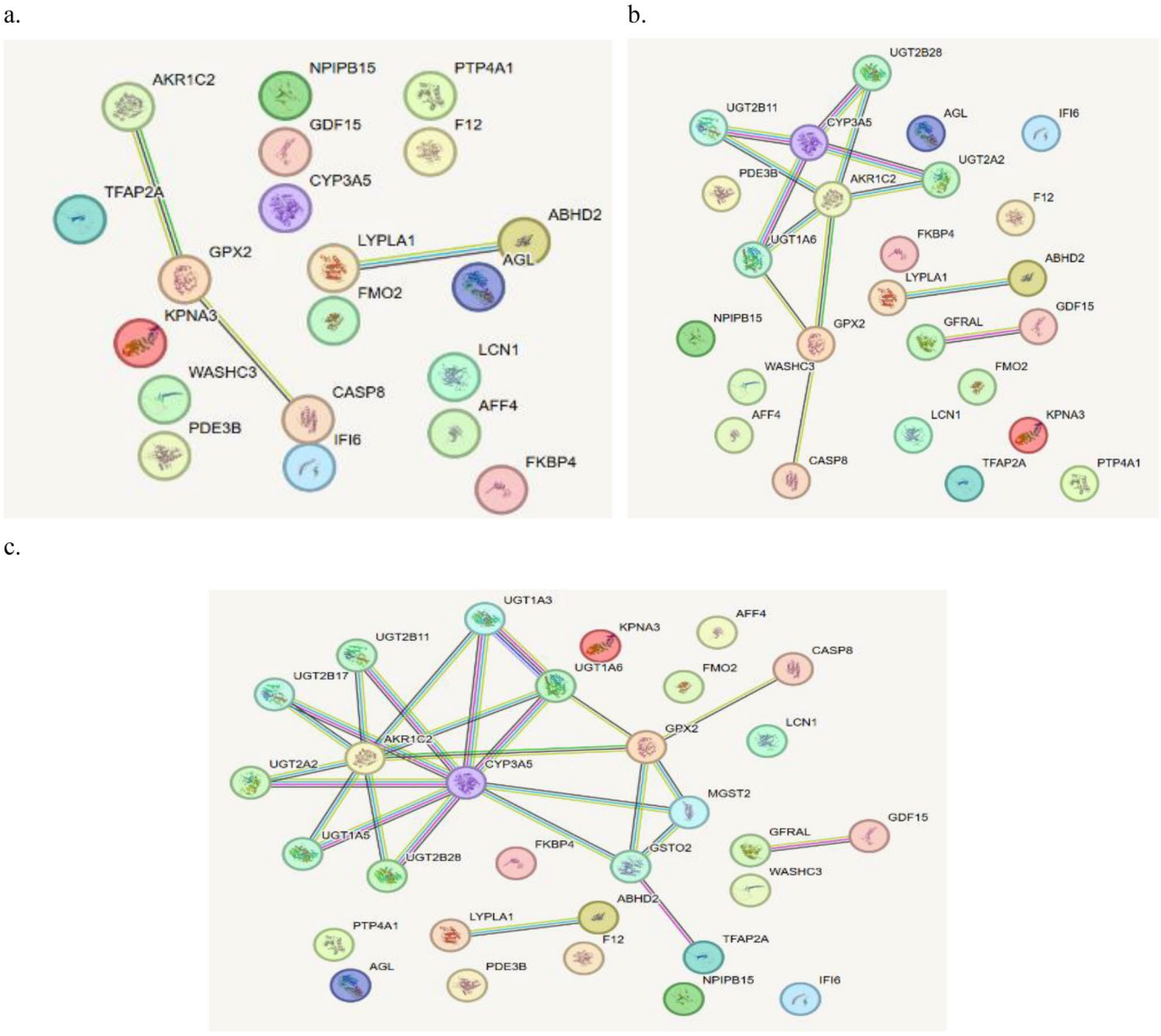


Fig. 9 PPI network analysis diagram. **a**. 20-node analysis diagram. **b**. 25-node analysis diagram. **c**. 31-node analysis diagram. Edge confidence scores are from the STRING database (version 11.5); only edges with confidence score > 0.4 are shown. Key edge scores: AKR1C2–GPX2: 0.72, GPX2–CASP8: 0.68, LYPLA1–ABHD2: 0.65, GDF15-GFRAL:0.71

(GO:0001822) and gene expression regulation (GO:0010628). Its peripheral position is consistent with its role as a downstream transcription factor [12, 31].

Metabolic hub: CYP3A5. Together with AKR1C2, CYP3A5 formed a steroid metabolism module enriched in KEGG pathways hsa01100 and hsa00140. This finding suggests that steroid metabolism abnormalities may influence IgAN susceptibility or treatment response [18, 19].

The LYPLA1–ABHD2 Module

The early PPI connection between LYPLA1 and ABHD2 (20-node, Fig. 9a) represents a serine hydrolase module involved in lipid metabolism. These genes were not included in the final six because they lacked enrichment in IgAN-relevant BP/KEGG pathways (e.g., TGF-β signaling, apoptosis). We interpret this as a lipid metabolism module that may reflect general metabolic stress rather than IgAN-specific pathogenesis. Nevertheless, its early emergence validates the sensitivity of our PPI approach.

Methodological implications

A key strength of this study is the integration of three independent criteria for biomarker selection shown in Table 9:

Genes satisfying all three criteria were selected as final candidates. This triangulation approach reduces false positives. Additionally, the threshold-dependent

Table 7 The six hub genes from the intersection of three analysises

Gene	LRP Rank	PPI Role	GO/KEGG Enrichment
AKR1C2	#17	Core hub (31-node)	Steroid metabolism
CYP3A5	#10	Core hub (31-node)	Steroid metabolism; Metabolic pathways
GPX2	#2	Bridge (AKR1C2–CASP8)	Oxidative stress; Metabolic pathways
CASP8	#13	Cascade terminal	Apoptosis; Proteolysis; Steroid hormone biosynthesis
GDF15	#12	Ligand-receptor (with GFRAL)	TGF-β/SMAD signaling; Apoptosis
TFAP2A	#7	Peripheral gene in AKR1C2 network	Kidney development; Gene regulation

Table 8 Differential expression analysis of the six hub genes between IgAN and control samples

Gene	log2FC	*p*-value	FDR
AKR1C2	1.82	1.2e-04	0.012
CYP3A5	1.45	3.5e-04	0.018
GPX2	0.98	2.1e-03	0.035
CASP8	1.21	8.7e-04	0.022
GDF15	1.67	2.5e-04	0.015
TFAP2A	0.87	4.3e-03	0.042

Table 9 The integration of three independent criteria for biomarker selection

Criterion	Purpose
LRP	contribution scoreIdentifies genes critical for IgAN classification
GO/KEGG	enrichmentEstablishes functional relevance to IgAN
PPI network position	Reveals gene-gene interactions and hub status

emergence of biological relationships (20-node: core cascade; 25-node: ligand-receptor; 31-node: hierarchical regulation) supports the authenticity of our network [12, 31].

Limitations

Several limitations should be acknowledged. First, the lack of an external validation dataset limits generalizability; future work should validate our findings in independent IgAN cohorts (e.g., GEO datasets). Second, our sample size is small, though dynamic augmentation was designed to mitigate this. Third, no functional experiments (e.g., gene knockdown) were performed to establish causality; our conclusions are correlational and hypothesis-generating. Fourth, the 85% threshold may not be universally optimal. Finally, PPI analysis relies on STRING database biases.

Clinical implications and conclusions

If validated, the six hub genes could serve as a diagnostic panel, prognostic markers, or therapeutic targets (e.g., GDF15-neutralizing antibodies, caspase inhibitors). The enrichment of AKR1C2 and CYP3A5 in steroid metabolism also suggests potential pharmacogenetic applications for steroid therapy response.

In conclusion, using an LRP-guided dynamic augmentation framework integrated with GO/KEGG and PPI analyses, we identified AKR1C2, CYP3A5, GPX2, CASP8, GDF15, and TFAP2A as candidate biomarkers for IgAN. Our triangulation approach offers a generalizable strategy for biomarker discovery in small-sample, high-dimensional datasets.

Conclusion

This study addresses the core challenges of high-dimensional small-sample bulk RNA-seq data by proposing an integrated system combining innovative feature selection and sample augmentation strategies. The methodology employs binary case-control feature selection to construct disease-relevant gene sets, incorporates an interpretable attention-based augmentation mechanism using LRP, and implements dynamic iterative optimization for challenging samples, significantly improving the accuracy and robustness of biomarker discovery. The innovative whole-geneset enrichment analysis framework overcomes the biological interpretation limitations of traditional top-K gene approaches, providing multi-scale analytical perspectives (population-individual, pathway-function) for disease subtype mechanism research. The proposed method not only enhances classification performance but also establishes a crucial bridge between computational models and molecular biological interpretation, offering novel methodological support for precision medicine research.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13036-026-00712-2.

Supplementary material 1

Author contributions

Conceptualization, Juanjuan Zhao, Yafeng Li, Yan Qiang and Yifang Zheng. Methodology, Fang Zheng. Data processing, Wensong Yue. Writing—original draft: Fang Zheng. Codes: Fang Zheng. Medical biological knowledge guidance, Yan Geng and Yifang Zheng. Supervision, Yaheng Li. Writing—review and editing, Juanjuan Zhao, Yan Qiang, and Yifang Zheng. Funding acquisition, Juanjuan Zhao, Yan Qiang, and Yafeng Li. Code management-Yuanchen Gao.

Funding

This research was in part funded by the National Natural Science Foundation of China (82170716), the National Natural Science Foundation of China (62376183), the National Natural Science Foundation (U21A20469), the special fund for Science and Technology Innovation Teams of Shanxi Province (202304051001009), National Health Commission Key Laboratory of

Pneumoconiosis (2020-PT320-005), the National Natural Science Foundation of China (62476190).

Data availability
The data can be downloaded via the GEO (https://www.ncbi.nlm.nih.gov/gds/).

Declarations

Ethical approval
Not applicable.

Competing interest
The authors declare no competing interests.

Received: 10 December 2025 / Accepted: 6 June 2026
Published online: 19 August 2026

References

1. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature. 2003, Mar;422(6928):198–207. https://doi.org/10.1038/nature01511.
2. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA. 1977, Dec;74(12):5463–67. https://doi.org/10.1073/pnas.74.12.5463.
3. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995, Oct;270(5235):467–70. https://doi.org/10.1126/science.270.5235.467.
4. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010, May;28(5):511–15. https://doi.org/10.1038/nbt.1621.
5. Jin T, Talos FM, Wang D. Ecmarker: interpretable ML model identifies gene expression biomarkers predicting clinical outcomes and reveals molecular mechanisms of human disease in early stages. bioRxiv; 2019. https://doi.org/10.1101/825414.
6. Zhang T, Zhang D. Integrating omics data and protein interaction networks to prioritize driver genes in cancer. Oncotarget. 2017;8(35):58050–60. https://doi.org/10.18632/oncotarget.19481.
7. Zheng F, Zhao J, Yuan Z, et al. An interpretable two-branch network based on an prototype to promote group gene biomarkers discovery. J Supercomput. (2025);81:1080. https://doi.org/10.1007/s11227-025-07554-7.
8. Shen S, Qi Z, Sun J, Huang Q, Tian Q, Wang S. Enhancing pre-trained representation classifiability can boost its interpretability. arXiv; 2025. https://arxiv.org/abs/2510.2410.
9. Zhang K, Fan C, Cai D, Zhang Y, Zuo R, Zhu L, et al. Contribution of TGF-Beta-mediated NLRP3-HMGB1 activation to tubulointerstitial fibrosis in rat with angiotensin II-Induced chronic kidney disease. Front Cell Dev Biol. 2020;8:1–16. https://doi.org/10.3389/fcell.2020.00001.
10. Nam Y, Jhee JH, Cho J, Lee J-H, Shin H. Disease gene identification based on generic and disease-specific genome networks. Bioinformatics. 2019;35(11):1923–30. https://doi.org/10.1093/bioinformatics/bty882.
11. Zhang Y, Chen Y, Hu T. PANDA: prioritization of autismgenes using network-based deep-learning approach. Genet Epidemiol. 2020;44(4):382–94. https://doi.org/10.1002/gepi.22282.
12. Mucha K, Bakun M, Jaźwiec R, Dadlez M, Florczak M, Bajor M, et al. Complement components, proteolysisrelated, and cell communicationrelated proteins detected in urine proteomics are associated with IgA nephropathy. Pol Arch Med Wewn. 2014;124(7–8):380–86. https://doi.org/10.20452/pamw.2345. Epub 2014 Jun 18. PMID: 24938435.
13. https://pubmed.ncbi.nlm.nih.gov/.
14. https://cn.string-db.org/cgi/.
15. He K, Zhang X, Ren S, Sun J. Deep residual learning for image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016 (pp. 770–78). https://arxiv.org/abs/1512.03385.
16. Qing J, Zheng F, Zhi H, Yaigoub H, Tirichen H, Li Y, et al. Identification of unique genetic biomarkers of various subtypes of glomerulonephritis using machine learning and deep learning. Biomolecules. 2022;12, 1276. https://doi.org/10.3390/biom12091276.
17. Kalantari S, Nafar M, Samavat S, Parvin M. 1 H NMR-based metabolomics study for identifying urinary biomarkers and perturbed metabolic pathways associated with severity of IgA nephropathy: a pilot study. Magn Reson Chem. 2017, Aug;55(8):693–99. https://doi.org/10.1002/mrc.4573. Epub 2017 Mar 2. PMID: 28042675.
18. Coppo R. Treatment of IgA nephropathy: recent advances and prospects. Néphrologie & Thérapeutique. 2018, Apr;14 Suppl, 1: S13–21. https://doi.org/10.1016/j.nephro.2018.02.010. PMID: 29606258.
19. Abe K, Miyazaki M, Shioshita K, Harada T, Koji T, Kohno S. Clinical and immunohistochemical study of immunoglobulin a nephropathy (IgAN) before and after tonsillectomy. Acta Otolaryngol Suppl. 2004 Dec;(555:20–24. https://doi.org/10.1080/03655230410003378. PMID: 15768792.
20. Zhai Y, Long X, Gao J, Yao X, Wang X, Zhao Z. Elevated Endostatin expression is regulated by the pIgA immune complex and associated with disease severity of IgA nephropathy. Kidney Blood Press Res. 2021;46(1):31–40. https://doi.org/10.1159/000508079. Epub 2020 Dec 18. PMID: 33341801.
21. Nakamura T, Ebihara I, Nagaoka I, Takahashi T, Tomino Y, Koide H. Abnormal regulation of insulin-like growth factor gene expression in peripheral blood mononuclear cells from patients with IgA nephropathy. Am J Nephrol. 1992;12(5):292–302. https://doi.org/10.1159/000168463. PMID: 1283288.
22. Lai KN, Tang SCW, Guh JY, Chuang TD, Lam MF, Chan LYY, et al. Polymeric IgA1 from patients with IgA nephropathy Upregulates transforming growth factor-β synthesis and signal transduction in human mesangial cells via the renin-angiotensin System. J Am Soc Nephrol. 2003, Dec;14(12):3127–37. https://doi.org/10.1097/01.asn.0000095639.56212.bf. PMID: 14638911.
23. Deng F, Zhang J, Li Y, Wang W, Hong D, Li G, et al. Hirudin ameliorates immunoglobulin a nephropathy by inhibition of fibrosis and inflammatory response. Ren Fail. 2019, Nov;41(1):104–12. https://doi.org/10.1080/0886022X.2019.1583113. PMID: 30880546; PMCID: PMC6427573.
24. Ko T, Peddinti V, Povey D, Khudanpur S. Audio augmentation for speech recognition. Proc. Interspeech. 2015, pp. 3586–89.
25. Wang Y, Yao Q, Kwok JT, Ni LM. Generalizing from a few examples: a survey on few-shot learning. ACM Comput Surv. 2020, Jun;53(3):1–34. https://doi.org/10.1145/3386252.
26. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. Mixup: beyond empirical risk minimization. Proc Int Conf Learn Represent. (ICLR); 2018. [Online]. Available: https://arxiv.org/abs/1710.09412.
27. Verma V, Lamb A, Beckham C, Courville A, Mitliagkis I, Bengio Y. Manifold mixup: better representations by interpolating hidden states," in Proc Int Conf Mach Learn (ICML); 2019; p. 6438–47. [Online]. Available: https://arxiv.org/abs/1806.05236.
28. Antoniou A, Storkey A, Edwards H. Data augmentation generative adversarial networks," arXiv preprint arXiv: 1711.04340, 2017. [Online]. Available: https://arxiv.org/abs/1711.04340.
29. Mariani G, Scheidegger F, Istrate R, Bekas C, Malossi C. BAGAN: data augmentation with balancing GAN," arXiv preprint arXiv: 1803.09655, 2018. [Online]. Available: https://arxiv.org/abs/1803.09655.
30. Onetti Muda A, Feriozzi S, Rahimi S, Faraggiana T. Expression of TGF-beta receptors type I and II in human glomerulonephritis. Nephrol Dial Transpl. 1998, Feb;13(2):279–84. https://doi.org/10.1093/oxfordjournals.ndt.a027819. PMID: 9509435.
31. Haniuda K, Gommerman JL, Reich HN. The microbiome and IgA nephropathy. Semin Immunopathol. 2021, Oct;43(5):649–56. https://doi.org/10.1007/s00281-021-00893-6. Epub 2021 Oct 18. PMID: 34664087.
32. Bach S, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One. 2015, Jul;10(7):e0130140.
33. Böhle M, Eitel F, Weygandt M, Ritter K. Layer-wise relevance Propagation for Explaining deep Neural network decisions in MRI-Based Alzheimer's disease classification. Front Aging Neurosci. 2019;11:194–. https://doi.org/10.3389/fnagi.2019.00194. https://gitee.com/hqlvhq/Pytorch-LRP/.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.