

Received 13 April 2023, accepted 6 May 2023, date of publication 15 May 2023, date of current version 18 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3275967

RESEARCH ARTICLE

A New Approach Based on Feature Selection of Light Gradient Boosting Machine and Transformer to Predict circRNA-Disease Associations

CHEN MA¹, YUHONG CHI¹, DONGLAI HAO¹, AND XIONGFEI JI²

¹School of Computer Science, Xijing University, Xi'an, Shaanxi 710123, China

²The 20th Research Institute of China Electronics Technology Group Corporation, Xi'an, Shaanxi 710119, China

Corresponding author: Chen Ma (997528107@qq.com)

ABSTRACT Circular RNA (circRNA) is a type of single-stranded RNA with a closed circular structure. Recent studies have shown that circRNA has a relatively more stable structure than its linear counterparts. The circRNA has become a biological marker in medicine and plays a crucial role in disease prediction. However, traditional biological experiments are often time-consuming and laborious. More researchers are taking computational approaches to predict the circRNA-disease associations more rapidly and reliably. In this paper, we propose a novel method for predicting the circRNA-disease associations based on the feature selection using Light Gradient Boosting Machine (LightGBM) and a self-attention neural network-Transformer (LGFRCD). Firstly, the histogram-based decision tree algorithm in LightGBM is used to discretize the continuous floating-point features in circRNA-disease into the histogram of integer numbers. While traversing samples, the difference between histograms is used to optimize the calculation, greatly improving the construction speed. Then a leaf-wise algorithm is employed to calculate the node with the maximum split gain, resulting in the final feature vector. Finally, these features are sorted in order of importance and introduced into the Transformer for information fusion and prediction. Our study demonstrates that after feature processing and dimension reduction, LGFRCD achieved a prediction accuracy of 95.44% for AUC (Area Under the receiver operating characteristic Curve), which is 3.11% higher than the latest algorithms for the same dataset. We also conducted a search in published literature to cross-validate the predicted result. Out of the top 15 circRNA-disease pairs predicted by the LGFRCD model, 13 were confirmed by existing literature. These results indicate that the proposed model is suitable for predicting circRNA-disease associations and can provide reliable candidates for biological experiments.

INDEX TERMS circRNA-disease, circRNA, light gradient boosting machine, self-attention neural network, transformer, feature selection.

I. INTRODUCTION

Circular RNA (circRNA) is a new type of RNA with a closed loop structure, distinguishing it from linear RNA. CircRNA is not susceptible to RNA exonuclease, and its expression is more stable and not easy to degradation. Additionally, it is widely present and diverse in eukaryotic cells, playing

a role in regulating endogenous RNA molecules through gene expression. The initial discovery of CircRNA in RNA viruses dates back to the 1970s. In 1976, it was first identified by electron microscopy in plant viruses [1]. In 1979, Hsu and Coca-Prados used electron microscopy to observe for the first time that RNA can exist in a circular form in the cytoplasm of eukaryotic cells [2]. Subsequently, various forms and categories of circRNA have been discovered. By 1993, some exon-based circRNAs were identified in human

The associate editor coordinating the review of this manuscript and approving it for publication was Kumaradevan Punithakumar¹.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see <https://creativecommons.org/licenses/by-nc-nd/4.0/>

cell transcripts. However, at the time, circRNA was predominantly viewed as a class of low-abundance RNA molecules resulting from exon transcript missplicing. Consequently, in-depth studies on circRNA were limited, and by 1999 only a few circRNAs had been identified.

Although circRNA was previously regarded as an error in the normal splicing process, research related to circRNA has experienced a significant surge since the early 2000s. In 2012, Salzman et al. detected a large number of highly expressed circular transcription RNA molecules [3] in pediatric acute lymphoblastic leukemia and bone marrow from human embryonic stem cells [4]. Their statistical estimates and biochemical analyses showed that a significant proportion of spliced transcripts from hundreds of genes in the human body were circRNAs. Furthermore, circRNAs were first confirmed to be widespread in human embryonic stem cells and malignant tissue cells, which drew the attention of the scientific community. In 2013, Salzman et al. expanded their data to include data from mice and fruit flies [5]. They found that circRNA was ubiquitous and accounted for about 1% of linear RNA in cells. In some cases, the amount of circRNAs even exceeded that of linear RNAs in the corresponding allelic sites. Since then, increasing numbers of circRNAs have been widely recognized and found in many eukaryotes, including humans, mice, and nematodes ([6]; [7]; [8]). Of course, with increasing research, various types of circRNAs have been shown to have specific biological functions. Rybak-Wolf et al. found that circRNAs are highly abundant in the central nervous system of mice and humans, and their levels of expression in the brain increase as the brain develops, particularly in synapses [9]. Liyun et al. discovered that the expression level of hsa_circ_0005986 was associated with chronic hepatitis B family history, tumor size, microvascular invasion, and BCLC stage [8]. More importantly, its down-regulation accelerated cell proliferation by promoting the G0/G1 to S phase transition. It has ultimately been proven that hsa_circ_0005986 plays a role in tumorigenesis and can be used as a new biomarker for hepatocellular carcinoma. Leire et al. analyzed the expression profiles of multiple sclerosis patients and screened out differentially expressed circRNAs associated with the disease [10]. They found that the expression of circRNA from anx2 (circ-anx2) was lower in patients with multiple sclerosis and could serve as a novel disease biomarker. The aforementioned research demonstrates that the study of circRNA-disease associations can provide the theory and basis for predicting and treating complex diseases. However, traditional biological experiments are expensive and time-consuming. Therefore, we require reliable and effective computational models to conduct large-scale studies of these associations more rapidly.

Lu et al. proposed a model based on edge-weighted graph attention and heterogeneous graph neural networks for predicting probable circRNA-disease correlations [11]. This model utilizes revised node features to learn meta-path contextual information and employs heterogeneous neural

networks to assign attention weights to different types of edges. Fan et al. utilized the KATZ measure to compute associations in heterogeneous networks formed based on various biological data [12]. This model uses similarities between nodes in heterogeneous networks [13] to predict associations, and it has been found that hsa_circ_0006054 may play a role as a miRNA sponge in the carcinogenesis of colorectal cancer. This indicates that a suitable computational model can not only improve prediction efficiency but also provide further scientific basis for studying the relationships between circRNAs and diseases. Lu et al. utilized deep matrix decomposition to predict circRNA-disease associations [14]. Their model takes into account both explicit and implicit feedback and employs a projection layer to automatically learn the latent representations of circRNAs and diseases. Finally, the model uses a multi-layer neural network to capture the nonlinear correlation between them. Although the prediction accuracy of Lu et al.'s model has improved, their comparison experiment used fewer indicators, and only the AUC values were presented. Wei et al. proposed a new predictor, iCricDA-LTR, for circRNA-disease association prediction [15]. This model uses a ranking framework to query global ranking associations between circRNAs and diseases and employs a Learning to Rank (LTR) algorithm to perform supervised sorting of associations based on various predictors and features. Although this model predicted a very high AUC value, other values used in comparative experiments were very low, even lower than those in some models proposed in earlier studies. Wang et al. proposed a deep learning model based on Stack Auto Encoder (SAE) and Rotating Forest (RF) to predict correlations [16]. This model combines three similarity measures: semantic similarity, GIP kernel similarity, and Jaccard similarity. SAE is used to extract implicit features from these measures and identify associations. The comparative data obtained from the model are relatively complete; however, the Matthews correlation coefficient (MCC) is relatively low. All other evaluation metrics have been optimized and improved. Therefore, we adopted the same complete-data comparison method in our study. Wang et al. proposed a model based on Extreme Learning Machine (ELM) and Convolutional Neural Network (CNN) for circRNA-disease correlation prediction [17]. Their model is based on numerical descriptors of the circRNA similarity network and disease semantic similarity network, including the Gaussian interaction spectral kernel similarity network. Finally, the convolutional neural network (CNN) was used to automatically and objectively extract depth features of circRNAs-disease descriptors, which were then put into the extreme learning machine for prediction. The comparative data predicted by this model is also relatively perfect, and the optimization steps of feature extraction are explained and demonstrated in detail. In our study, we also adopted methods of gradual optimization and comparison, which can more reasonably demonstrate the advantages of the current model.

By analyzing the literatures mentioned above, we found the following points to be optimized:

1) Firstly, the analysis of the evaluation criteria is not comprehensive. For some prediction models, the predicted value of some criteria may be high or low. When evaluating the model, we should ensure the quantity of evaluation criteria as well as the quality of the prediction effect.

2) Secondly, the analysis of the model's stability is not comprehensive. A good prediction model should not only have a good prediction effect, but also ensure the stability of calculation results, that is, the standard deviation of the forecast average should be as small as possible. And for the neural networks, the change in their accuracy and loss also needs to be analyzed. If the prediction rate is high, but the model has the problem of under-fitting or over-fitting, it is also unqualified. The model with a high degree of fitting has high stability.

3) Thirdly, the feature dimension reduction analysis is not comprehensive. Before feature fusion and prediction, it is possible to simplify the feature dimension. For the features ranked by importance, we can do multiple comparisons to find the optimal dimension. That is, the model not only maintains a good prediction effect, but also simplifies the feature dimension in the incoming classifier, so that the model can do prediction with high correlation.

This paper proposes a new model, LGFRCD, for predicting circRNA-disease associations by combining feature selection of light gradient boosting machine and self-attention neural network-Transformer. We used the dataset processed by Wang et al. [17], which merges disease semantic similarity, disease Gaussian interaction kernel similarity, and circRNA Gaussian interaction kernel similarity from the original dataset [18]. This dataset includes known associations [18] and additional unknown associations for testing. To extract important features from the circRNA-disease dataset, we first use LightGBM for feature traversal and importance extraction. The histogram-based decision tree algorithm used in LightGBM can discretize continuous floating point features into histograms of integer counts. During sample traversal, the algorithm uses these integers as indices to accumulate statistics in histograms for each corresponding feature. Additionally, the difference between histograms of the current node and its parent node is used to calculate the histogram of the sibling node, following the leaf-wise algorithm. Finally, the node with the maximum split gain is calculated, and the corresponding feature is selected as the final feature. These features are ordered by their importance based on the statistics of their associated histograms.

Then the Transformer analyzes the correlations between features in the original circRNA-disease dataset using a self-attention mechanism, generating a weighted correlation score for each feature. Using encoder and decoder components, it then objectively and automatically integrates and filters these feature correlations to effectively predict potential associations between circRNAs and diseases. As a popular and innovative deep learning technology, the Transformer has

been widely applied in various fields, including pedestrian trajectory prediction [19], prediction of drug-target interactions [20], and small object detection in satellite images [21]. Encouraging results have been achieved in these applications.

The contribution of LGFRCD proposed in this paper are as follows:

(1) We used disease semantic similarity, disease Gaussian interaction kernel similarity, and circRNA Gaussian interaction kernel similarity as attributes in a heterogeneous biological network, aiming to extract as much hidden information from the network as possible to fully train the prediction model.

(2) To extract important features from the heterogeneous biological network, we performed feature traversal and importance extraction using LightGBM. The algorithm ranked the features according to their importance and reduced the feature dimensionality of the dataset.

(3) The Transformer utilizes a self-attention mechanism to generate a weighted correlation score for each feature. By integrating the correlations of each feature with encoders and decoders, it can effectively predict potential associations between circRNAs and diseases.

(4) To comprehensively and fairly evaluate the stability of LGFRCD, we used ten-fold cross-validation in this study. The results showed that the standard deviation of predictions was very small and the accuracy and loss curves of LGFRCD changed normally, indicating that the model was highly stable.

(5) To evaluate LightGBM and Transformer, we compared their performance with other traditional classifiers and feature extraction algorithms. Our results indicated that, with the same feature dimensions, LightGBM's feature extraction method outperformed other algorithms in terms of accuracy. Additionally, Transformer's feature fusion and prediction capabilities were superior to those of other classifiers.

(6) We also performed feature dimension reduction to determine the optimal prediction performance. As a result, we achieved an AUC (Area Under the receiver operating characteristic Curve) of 95.44% for LGFRCD. Furthermore, when compared with known prediction models using the same dataset, LGFRCD demonstrated superior prediction accuracy. Notably, 13 out of 15 pairs with the highest predicted scores were confirmed by published literature, demonstrating the reliability and feasibility of our method.

II. MATERIALS AND METHODS

A. CALCULATION OF DISEASE SIMILARITY

For the model of calculating disease similarity, we decided to choose the model of disease semantic similarity, disease Gaussian interaction kernel similarity and circRNA Gaussian interaction kernel similarity proposed by Wang et al. [17]. After feature extraction and prediction, Wang et al. obtained a high prediction accuracy. In addition, values of other comparison indicators are maintained at a relatively reasonable and

high level, and the overall model has a comprehensive and good expression effect.

1) CONSTRUCTION OF DISEASE SEMANTIC SIMILARITY

First, in Mesh we selected, the relationship between diseases is described by directed acyclic graph (DAG). Nodes in the diagram represent diseases, and edges represent relationships between the diseases. If we take a current disease $curr_dis$, and its corresponding DAG can be represented as $\{curr_dis, Set_{par}, Set_{edg}\}$. Set_{par} is the set of ancestral nodes of $curr_dis$ that includes itself, and Set_{edg} is the set of edges that connect these nodes. If there is a disease d in DAG_{curr} , its contribution to $curr_dis$ is represented as follows:

$$Con_{curr}(d) = \begin{cases} 1, & d = curr_dis; \\ \max\{k * Con_{curr}(d') | d' \in children\ of\ d\}, & d \neq curr_dis \end{cases} \quad (1)$$

where k is the semantic contribution factor linking disease d and its child disease d' . In $curr_dis$'s DAG, its own contribution is valued at 1. The contribution of disease d to $curr_dis$ is the maximum product of semantic factor k and $Con_{curr}(d')$. Thus, the total semantic contribution SC_{curr} of $curr_dis$ is as follows:

$$SC_{curr} = \sum_{d \in Set_{par}} Con_{curr}(d) \quad (2)$$

So for disease d_i and d_j , the formula of first semantic similarity value $SV_1(d_i, d_j)$ between them in DAG is as follows:

$$SV_1(d_i, d_j) = \frac{\sum_{d \in (Set_{par_i} \cap Set_{par_j})} (Con_{d_i}(d) + Con_{d_j}(d))}{SC_{d_i} + SC_{d_j}} \quad (3)$$

where Set_{par_i} is the ancestral node set of d_i , including itself, and Set_{par_j} is the ancestral node set of d_j , including itself. It can be seen from the above formula that the larger the intersection between the DAGs of two diseases is, the larger the first semantic similarity value will be. However, for some frequently occurring disease associations, their first semantic similarity values in DAGs become very large, which makes the values of some infrequently occurring associations very small. This means that the first semantic similarity does not consider the number of diseases. To balance the data, we need to add a model similar to the penalty function to reflect the importance of contribution by the number of diseases:

$$Con'_{curr}(d) = -\log \frac{\text{num}(DAGs(d))}{\text{num}(\text{diseases})} \quad (4)$$

where $\text{num}(DAGs(d))$ represents the number of DAGs including disease d , $\text{num}(\text{diseases})$ represents the number of total diseases. So, we can obtain the second semantic similarity model according to formula 3 and 4:

$$SV_2(d_i, d_j) = \frac{\sum_{d \in (Set_{par_i} \cap Set_{par_j})} (Con'_{d_i}(d) + Con'_{d_j}(d))}{SC_{d_i} + SC_{d_j}} \quad (5)$$

2) DISEASE GAUSSIAN INTERACTION KERNEL SIMILARITY

The above disease semantic similarity models, SV_1 and SV_2 , are created based on Mesh. However, for CircR2Disease, there are always some diseases that cannot be found in Mesh. But similar diseases can be found, and we can take that into account. Therefore, in order to ensure that similar disease information is not ignored, Gaussian interaction kernel similarity [22] is applied. It is a scalar function that is symmetric along the radial direction, and it can better measure the similarity between samples, so that these samples can be better clustered based on similarity. Finally, the goal of linear divisibility is achieved.

In CircR2Disease, we assumed that the association matrix is denoted by M_{ij} , which consists of 676 rows and 100 columns. Here i represents the index of 676 circRNAs, and j represents the index of 100 diseases. We then used a binary vector, $R(d_j)$, whose value is determined by the association of disease d_j with each of the 676 circRNAs. When d_j is associated with a certain circRNA, the corresponding position of this circRNA in $R(d_j)$ is assigned to 1, otherwise it is assigned to 0. In other words, R is the row vector in CircR2Disease. Thus, the Gaussian interaction profile kernel similarity for diseases $GS(d_i, d_j)$ of disease d_i and disease d_j can be calculated using the following formula:

$$GS(d_i, d_j) = \exp(-\theta_d ||R(d_j) - R(d_i)||^2) \quad (6)$$

where θ_d is an adjustable parameter, which controls the kernel bandwidth and can be calculated by normalizing the original parameters. Its formula is as follows:

$$\theta_d = \frac{1}{r} \sum_{i=1}^r ||R(d_i)||^2 \quad (7)$$

where r is the number of rows of the association matrix M_{ij} .

3) CircRNA GAUSSIAN INTERACTION KERNEL SIMILARITY

Similarly, we used a binary vector $C(c_i)$, whose value was determined by the association of circRNA c_i with each of the 100 diseases. When a circRNA is associated with a certain disease, the corresponding position of the disease in $C(c_i)$ is assigned to a value of 1, and 0 otherwise. In other words, C is the column vector in CircR2Disease. Thus, the Gaussian interaction profile kernel similarity for circRNAs $GS(c_i, c_j)$ can be calculated using the following formula, which measures the similarity between circRNA c_i and circRNA c_j :

$$GS(c_i, c_j) = \exp(-\theta_c ||C(c_j) - C(c_i)||^2) \quad (8)$$

where θ_c is an adjustable parameter, which controls the kernel bandwidth and can be calculated by normalizing the original parameters. Its formula is as follows:

$$\theta_c = \frac{1}{\text{num}} \sum_{j=1}^{\text{num}} ||C(c_j)||^2 \quad (9)$$

where num is the number of columns of the association matrix M_{ij} .

The above similarity calculations have been concluded for diseases and circRNAs. It can be observed that the most complex similarity calculation is to determine SV_1 and SV_2 . We further assume that the final disease similarity model, denoted by $DSM(d_i, d_j)$, represents the ultimate similarity between disease d_i and d_j . If there exists semantic similarity between two diseases, we select the combination of SV_1 and SV_2 as the value of $DSM(d_i, d_j)$; otherwise, we choose $GS(d_i, d_j)$. The formula is as follows:

$$DSM(d_i, d_j) = \begin{cases} GS(d_i, d_j), & d_i \text{ \& } d_j \text{ has no semantic similarity;} \\ \frac{(SV_1(d_i, d_j) + SV_2(d_i, d_j))}{2}, & \text{or not} \end{cases} \quad (10)$$

For circRNA, its similarity model is simple. Assuming that the circRNA similarity matrix is denoted as $CSM(c_i, c_j)$, we can express it directly using $GS(c_i, c_j)$. Therefore, the final similarity model $FSM(c_i, d_j)$ between circRNA c_i and disease d_j is expressed as follows:

$$FSM(c_i, d_j) = [CSM(c_i), DSM(d_j)] \quad (11)$$

where $CSM(c_i)$ represents Gaussian interaction profile kernel similarity of circRNA c_i , and $DSM(d_j)$ represents Gaussian interaction profile kernel similarity of disease d_j . After normalization, the dimension of $FSM(c_i, d_j)$ becomes dim .

At this point, the extraction and calculation of the similarity are completed, and its processed matrix $FSM(c_i, d_j)$ is obtained. We chose this similarity calculation method [17] because it considers multiple aspects: the model SV_1 accounts for the semantic similarity between diseases, while the penalty function model SV_2 is based on the number of diseases. Both models optimize multiple facets for a single attribute. Additionally, to handle similar diseases in CircR2Disease and Mesh, Gaussian interaction kernel similarity is applied to deal with circRNAs and diseases, respectively. In this way, the final similarity matrix will be more reasonable and reliable, laying the foundation for our subsequent feature processing.

B. FEATURE PROCESSING AND IMPORTANCE RANKING

Gradient Boosting Decision Tree (GBDT) [23] is a well-established model in machine learning, whose main idea is to get the optimal model by using weak classifiers (decision tree). This model has the advantages of good training effect, and it is not easy to overfit. It is not only widely used in industry, but also popular in various data mining competitions. Light Gradient Boosting Machine (LightGBM) [24] is a framework that implements GBDT and supports the efficient parallel training. Additionally, it offers advantages such as faster training speed, lower memory consumption, better accuracy, support for distribution, and fast processing of massive data. Therefore, in this paper, we adopt LightGBM for the following processing steps. As per equation 11, the final value dimension of $FSM(c_i, d_j)$ is dim , which represents the number of features of each sample. Hence, we can also

express $FSM(c_i, d_j)$ using the following formula:

$$FSM(c_i, d_j) = (f_1, f_2, \dots, f_{dim}) \quad (12)$$

where f represents each feature. There is also a histogram-based decision tree algorithm in LightGBM. Its main idea is to convert feature values into *bin* values before training. In other words, it creates a piecewise function for each feature value and divides all sample values for that feature into a certain segment (*bin*). Finally, these features are transformed from continuous values to discrete integer ones. That is, for each feature f_k , there is a bin_k that acts like an index. But there are still a *bins* on f_k , as shown in the following formula:

$$f_k.bins[k] = bin_k \quad (13)$$

where *bins* is the set of each *bin* corresponding to its own feature.

For the entire decision tree model, we assume that the current leaf is represented by *curr_leaf*. The algorithm will first traverse all features. We assume that the current feature is represented by f_{curr} , and then create a corresponding histogram *His* for f_{curr} . This histogram stores two types of information: the sum of the gradient g and the number n of samples in each *bin*. Then, the algorithm iterates through all samples and accumulates the above statistical values for each sample. The formula for this process is as follows:

$$\begin{cases} His[f_{curr}.bins[x]].g += g_x \\ His[f_{curr}.bins[x]].n += 1 \end{cases} \quad (14)$$

where the value of x the index of each sample. $His[f_{curr}.bins[x]].g$ represents the sum of gradients of f_{curr} under the current leaf, and $His[f_{curr}.bins[x]].n$ represents the sum of the number of all samples. Finally, the algorithm starts to traverse all *bins*, taking the current *bin* as the segmentation point, and adding the gradient g and number n from its left point respectively to the sum of gradient G_L and number N_L . Its formula is as follows:

$$\begin{cases} G_L += His[z].g \\ N_L += His[z].n \end{cases} \quad (15)$$

where z represents the index of the *bin*. Then there is an optimization technique in LightGBM, namely differential acceleration. The histogram of a leaf can be obtained by taking the difference between the histogram of its parent node and that of its sibling, which can effectively double the build speed. Normally, when constructing a histogram, it is necessary to traverse all the data points on the leaf. By using the histogram difference, we only need to traverse some *bins* of the current histogram. LightGBM can also calculate the leaf with small histograms, and then use the difference between the histograms to obtain the large histogram. Thus, the histogram of its sibling leaf can be obtained at a very low cost. Therefore, G_L and N_L are subtracted from the total gradient G_P and the number of samples N_P on the parent

node, respectively, to obtain the gradient G_R and the sample number N_R of the right *bin*. The formula is as follows:

$$\begin{cases} G_R = G_P - G_L \\ N_R = N_P - N_L \end{cases} \quad (16)$$

Finally, we can start to calculate and find the maximum gain point. We assume that the gain is $\Delta loss$, the current leaf is L , feature is f , and corresponding value of bin, v , contains the sum of gradients and the sum of numbers calculated by the previous equation 14. Then the formula is as follows:

$$\Delta loss = \frac{G_L^2}{N_L} + \frac{G_R^2}{N_R} - \frac{G_P^2}{N_P} \quad (17)$$

$$(L_m, f_m, v_m) = (L, f, v), \text{ if } \Delta loss > \Delta loss(L_m, f_m, v_m) \quad (18)$$

where the leaf with maximum gain is L_m , feature is f_m , and value of *bin* is v_m . That is to say, this algorithm computes the loss of each *bin* in turn while traversing all *bins* until it finds its maximum gain. The above method is also called Leaf-wise, and it's another optimization for LightGBM. Most GBDT tools use a decision-tree growth strategy called Level-wise, which means traversing the data once can split the same layer of leaves simultaneously. But in fact, many leaves have low split gains, so there is no need to search and split, which consumes too much computing power. At each iteration, Leaf-wise finds the one with the biggest split gain from all the current leaves, and then this leaf continues to be divided, and so on. Compared with Level-wise, Leaf-wise can reduce more errors and obtain better precision with the same number of splits, and LightGBM also adds a maximum depth limit to prevent overfitting while ensuring high efficiency.

The above concludes the processing part of the feature.

For the feature f_m at the current maximum gain point, the sum n of numbers of samples in v_m is the basis for feature ranking. In other words, the importance of f_m calculated as the segmentation point is decided on the times it occurs in all trees. This also means that the more a feature is used to build the decision tree in the model, the more important it is. And finally we can figure out the ranking of the importance of the features, which called *FR*.

C. INFORMATION FUSION AND PREDICTION

In the previous section, we completed the feature processing and importance ranking to get *FR*. The next step is to use the Transformer [25] to extract the information from *FR* for the final information fusion and prediction.

The algorithm architecture diagram of Transformer is shown in Figure 1. Firstly, source data is input in accordance with the sequence of lightGBM feature values *FR*, and each of data can be represented as a word vector. The expression of each word vector is as follows:

$$data = (fr_1, fr_2, \dots, fr_{t-1}, fr_t), \quad t \text{ in range}(FR) \quad (19)$$

where fr represents the sorted feature vector of each sample in *FR*, and t is its index. After embedding the inputs, positional encoding needs to be added to each word. This is because

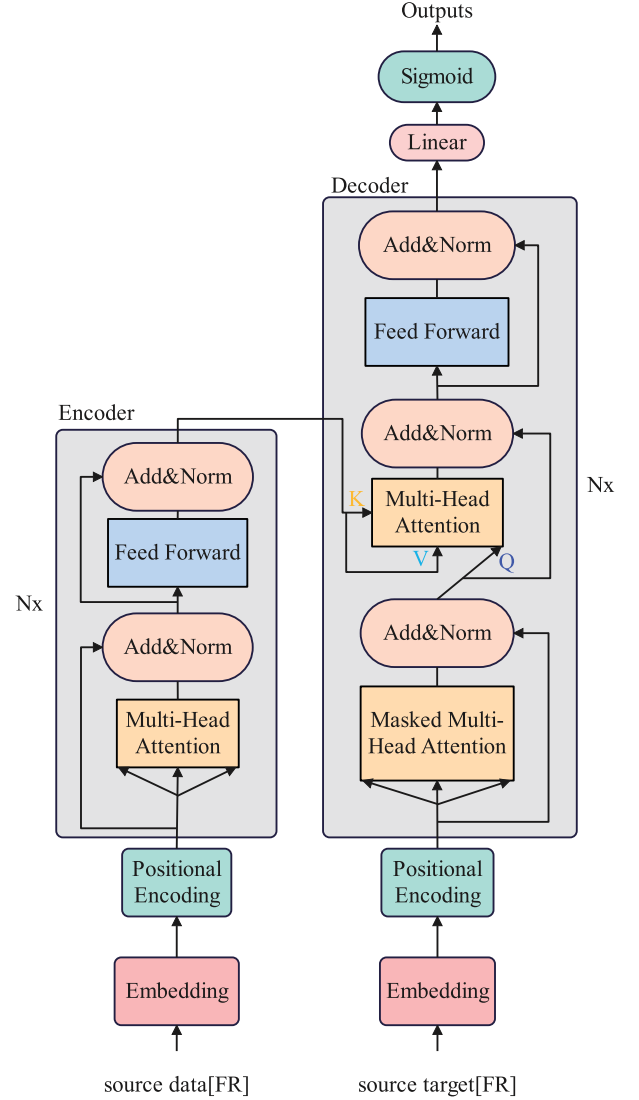


FIGURE 1. The architecture diagram of Transformer.

the Transformer model is based on self-attention and does not capture word placement information. If the positions of the features in a word vector change, it may result in different prediction results. Therefore, positional coding must be added so that attention values between words can be properly calculated even if their order is disrupted. The positional encoding is generated using positive and cosine functions of different frequencies [25], which are then added to the word vector before being input into the encoder. This can be expressed as follows:

$$\begin{cases} PE_1 \{pos, 2y\} = \sin\left(\frac{pos}{10000^{2y/d_{model}}}\right) \\ PE_2 \{pos, 2y+1\} = \cos\left(\frac{pos}{10000^{2y/d_{model}}}\right) \end{cases} \quad (20)$$

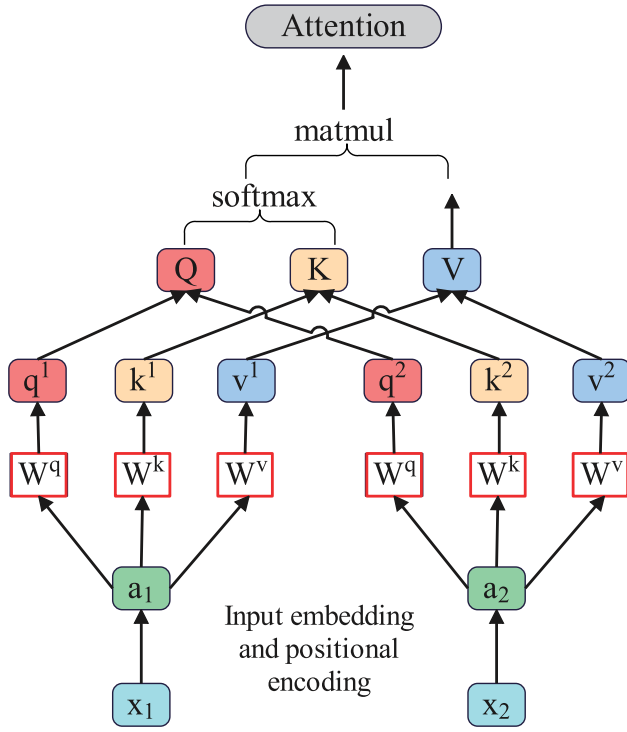


FIGURE 2. The working flow of self-attention.

where pos represents the absolute position of a word in the word vector, d_{model} represents the dimension of the word vector, and y represents the y th of dimensions in the word vector. $2y$ represents even dimensions and $2y + 1$ represents odd dimensions, and both of them are less than or equal to d_{model} . The final inputs entered into the encoder is the sum of input embedding and positional encoding, expressed as follows:

$$inputs = data + PE_1 + PE_2 \quad (21)$$

The above inputs then will be fed into the Multi-Head Attention in Figure 1, but we need to understand how self-attention works firstly. Its working flow is shown in Figure 2. (x_1, x_2) represents the original data, which will be subsequently transformed into a vector (a_1, a_2) after inputs embedding and positional encoding. Then self-attention will multiply (a_1, a_2) by the three weight matrices w_q, w_k and w_v to get the corresponding q^n, k^n and v^n . And q represents the target vector to be queried, k represents the feature vector to be queried, and v represents the queried value vector. Then we can combine these three vectors individually and get the overall matrix Q, K , and V . Finally, we can calculate the value of attention between input word vectors, whose formula is as follows:

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_{wv}}}\right)V \quad (22)$$

where d_{wv} represents the dimension of a word vector. QK^T represents the multiplication of Q and transpose K , which

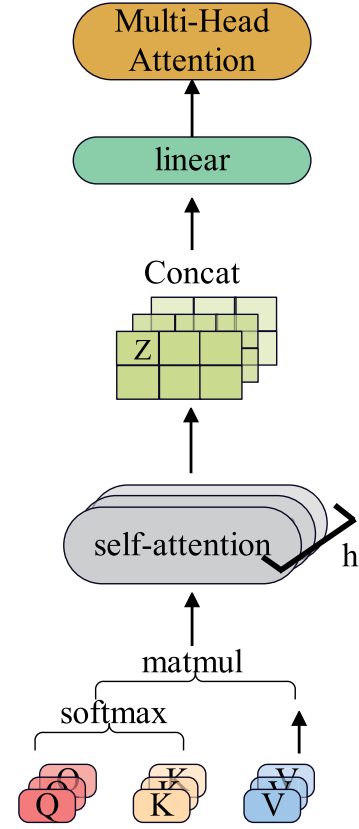


FIGURE 3. The working flow of Multi-Head Attention.

means the correlation score between words in the word vector. Next, QK^T needs to be divided by $\sqrt{d_{wv}}$ in order to ensure gradient stability during training. Softmax [26] is then applied to convert the correlation score into a probability distribution between 0 and 1. Finally, the result is multiplied by the matrix V to obtain the attention value between word vectors. This is how self-attention works, and understanding this process makes it easier to comprehend Multi-Head Attention. The working process of Multi-Head Attention is shown in Figure 3. As can be seen, Multi-Head Attention involves parallel calculations of h combinations of Q, K and V . In other words, Multi-Head Attention is composed of multiple self-attentions. After calculating their respective attention values, represented by Z , we combine them into a multidimensional matrix which is then linearly transformed to map features back to their original dimension, resulting in the final output. Therefore, Multi-Head Attention can be expressed as follows:

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{concat}(Z_1, Z_2, \dots, Z_h)W_0 \\ Z_i &= \text{Attention}(Q, K, V) \end{aligned} \quad (23)$$

where W_0 means the linear change.

The output of the Multi-Head Attention is then passed into the Add&Norm layer. The first layer in this architecture is the residual block, which is the core component of ResNet [27]. Its function is to prevent poor learning effects caused by certain layers, which can lead to a lack of improvement in overall

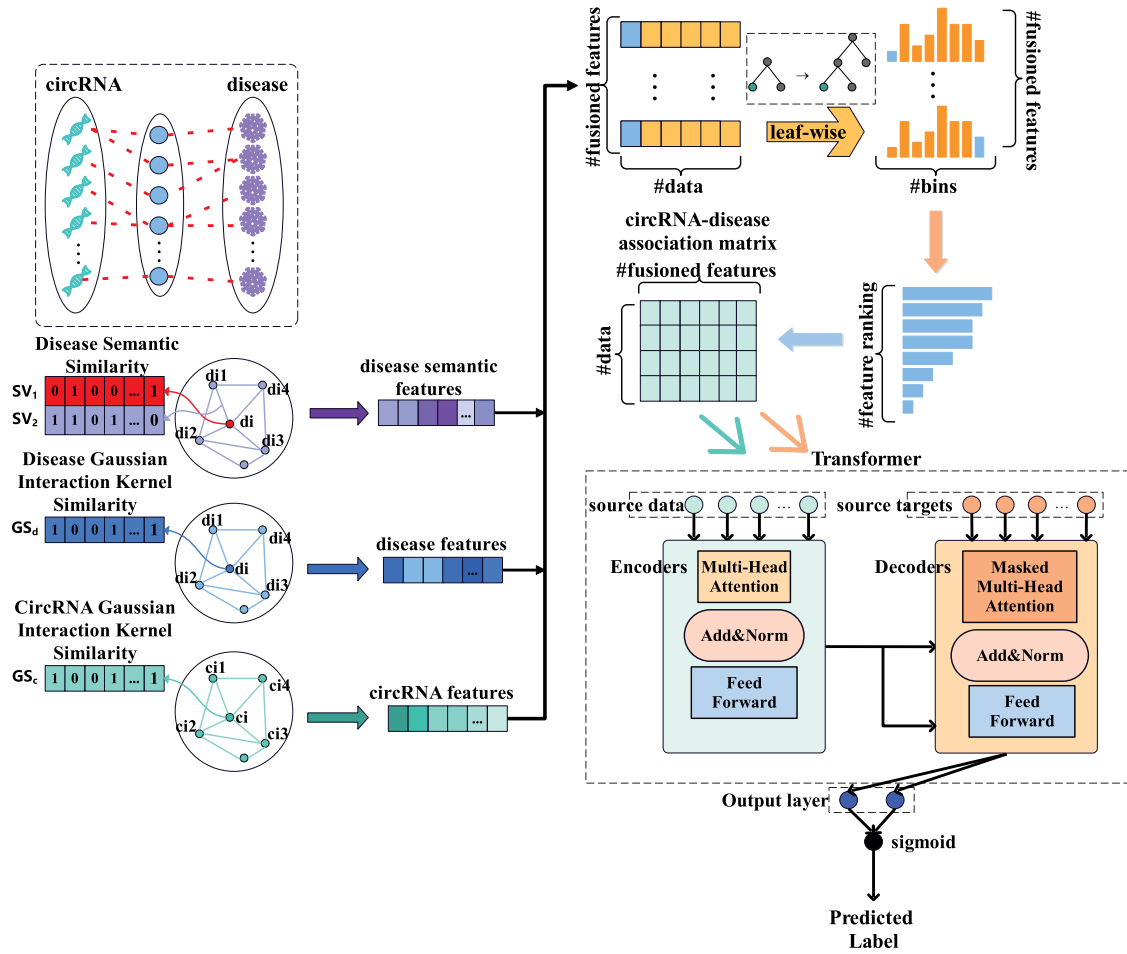


FIGURE 4. The flow chart of LGFRCDA.

training accuracy. The residual block can skip these layers through a residual edge and set their weights to 0. In this way, it ensures steady improvement in learning effect and also avoids the situation where the training falls into a local minimum value and gradients disappear. Norm refers to the layer normalization [28], which normalizes input data from the previous layer. Its function is to maintain the stability of feature distribution during model training and can effectively accelerate the convergence rate during gradient descent. It is also a common optimization technique for neural network training. Therefore, as shown in Figure 1, the data expression processed by Add&Norm is as follows:

$$\text{LayerNorm}(X) = \begin{cases} X + \text{MultiHead}(X) \\ X + \text{FeedForward}(X) \end{cases} \quad (24)$$

where X represents the input data or output of the previous layer, FeedForward represents the forward feedback network. FeedForward is essentially a two-tier, fully connected network, with the first layer being Relu [29] and the second layer having no functions. And its formula is as follows:

$$\text{FeedForward}(Y) = \text{Linear}(\max(0, \text{Linear}(Y))) \quad (25)$$

where Linear represents the linear transformation, and Y represents the input of the previous layer.

For the decoder, most processes are the same as those of an encoder, but there are three main differences. The first is the input data. In a supervised training, data accepted by the decoder are real targets of the dataset. And after inputs embedding and positional encoding, it is passed into the decoder.

The second difference is Masked Multi-Head Attention. Its workflow is no different from that of Multi-Head Attention, except for the addition of a mask to prevent the decoder from seeing future information. For an input sequence, the output affected by the mask should only depend on the data prior to the current time. In training, we provide real targets to the decoder and only need to get the output for the previous moment when predicting. The mask is designed to ensure that the current predicted value is not influenced by other targets.

The third difference concerns the input of the second Multi-Head Attention layer in the decoder. As shown in Figure 3, Multi-Head Attention requires Q, K, and V matrices. In this layer, Q represents the actual target matrix generated by the decoder, while K and V are the feature matrix to be queried

and the queried value matrix generated by encoders during training. These three matrices are then trained using Multi-Head Attention, Add&Norm, and FeedForward layers, before finally being output.

Finally, outputs by decoder are first mapped to low-dimensional values through linear changes. After sigmoid [30] processing, the final probability value of binary classification is output. Its calculation formula is as follows:

$$\text{sigmoid}(\hat{a}) = \frac{1}{1 + e^{-\text{Linear}(\hat{a})}} \quad (26)$$

where \hat{a} is the output by decoder. At this point, the whole algorithm process is over. The entire flow chart of LGFRCD is shown in Figure 4.

In the next section, we will discuss the impact of predictions and the degree of fit of the model using a loss function. Mean Absolute Error (MAE) [22] has been chosen as the loss function in this paper. Assuming that p is the number of samples in datasets, the calculation formula of MAE is as follows:

$$\text{MAE} = \frac{1}{p} \sum_{q=1}^p |y_{\text{pre}_q} - y_{\text{tru}_q}| \quad (27)$$

where y_{pre_q} is the predicted value of the q th sample, and y_{tru_q} is its true value. In this dataset, y_{tru_q} of samples known to be associated is 1, otherwise it is 0. MAE calculates absolute values of errors, and these values' penalty is fixed for the difference. No matter what kinds of input values, their stable gradients can be obtained after MAE, which will not lead to gradient explosion problem. And it can get a relatively robust solution.

D. EVALUATION CRITERIA

In this study, the ten-fold cross-validation method [31] was used to verify the model in order to obtain a more reliable one for comparison with other advanced models. This method involves dividing circRNA-disease associations into ten parts, nine of which are used as training data and one as test data. In our study, we carried out 10 rounds of ten-fold cross-validation to reduce the changes brought by sample division, and the final result is the average of these experiments. In this study, we use accuracy (Acc), precision (Pre), recall (Rec), F1-Score (F1) and Matthews Correlation Coefficient (MCC) as the evaluation criteria to evaluate the proposed model. These evaluation metrics have been validated in previous studies such as heart disease detection [32], early detection of cervical cancer [33], ASDs in children [34] and the stroke prediction [35]. They are defined as follows:

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (28)$$

$$\text{Pre} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (29)$$

$$\text{Rec} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (30)$$

$$\text{F1-Score} = \frac{2 * \text{Pre} * \text{Rec}}{\text{Pre} + \text{Rec}} \quad (31)$$

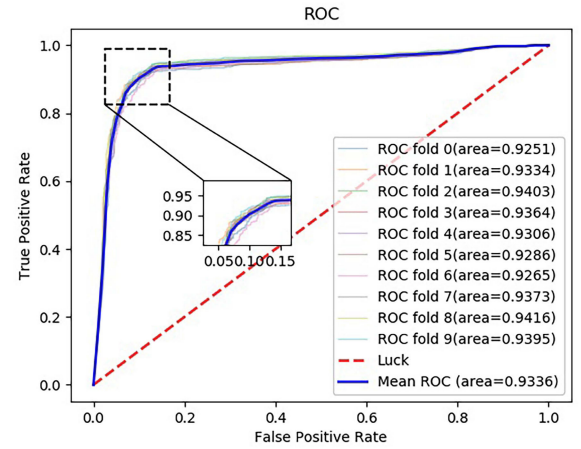


FIGURE 5. Receiver operating characteristic curves generated by LGFRCD.

$$\text{MCC} = \frac{2 * \text{Pre} * \text{Rec}}{\text{Pre} + \text{Rec}} \quad (32)$$

where TP indicates the number of true positives, TN indicates the number of true negatives, FP indicates the number of false positives, FN indicates the number of false negatives. In addition, we also plotted the Receiver Operating Characteristic [36] (ROC) and Precision-Recall [37] (PR) curves generated by the proposed model and calculated the area under the ROC curve (AUC) and PR curve (AUPR) respectively.

III. RESULTS

A. DATASET

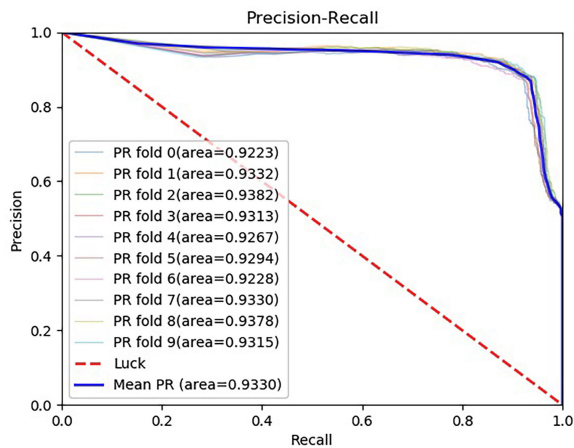
In this study, we used two datasets. The first dataset, CircR2Disease [18], was provided by Fan et al., and consisted of 739 associations between 676 circRNAs and 100 diseases. The second dataset, Mesh [38], was provided by Fodor et al., and comprised two columns. The first column contains the disease name and the second column contains the unique ID. Although the Disease name can be repeated in Mesh, each ID is unique. The whole dataset served like a catalog, indicating the affiliations among different diseases. For this study, we considered all known associations in the CircR2Disease dataset as positive samples. Additionally, we randomly combined correlations between circRNAs and diseases to create negative samples. The final dataset comprised 1478 samples and 776 features.

B. ASSESSMENT OF PREDICTION ABILITY

In this part, we evaluated the proposed model by using the processed circRNA-disease association to verify its practicability and performance. Experimental results are shown in Table 1. The average accuracy and standard deviation of ten verification methods are 89.69% and 0.54%, respectively. Regarding the accuracy rate Pre of true positive samples, our model achieved 86.74% with a standard deviation of 0.60%. In terms of recall rate Rec of true positive samples, our model achieved 93.72% with a standard deviation of 0.77%.

TABLE 1. Ten-fold cross-validation results performed by LGFRCDA.

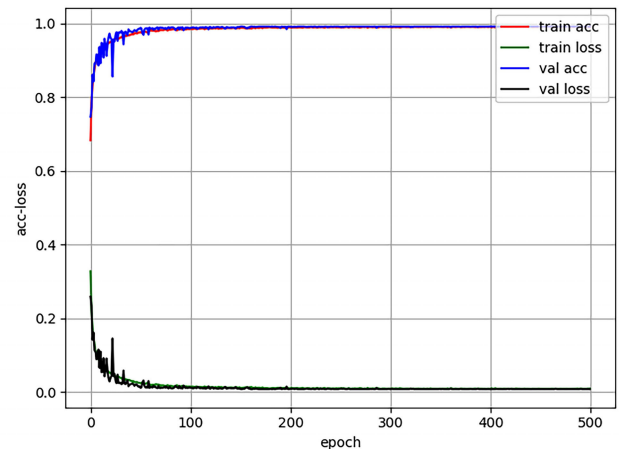
Acc(%)	Pre(%)	Rec(%)	F1-Score(%)	MCC(%)	AUC(%)	AUPR(%)
89.32	87.07	92.37	89.64	78.79	92.51	92.23
89.76	87.17	93.24	90.10	79.71	93.34	93.32
90.38	87.24	94.58	90.77	90.38	94.03	93.82
90.14	87.32	93.91	90.49	80.51	93.64	93.13
90.09	87.31	93.81	90.44	80.41	93.06	92.67
88.84	85.82	93.04	89.28	77.95	92.86	92.94
88.88	85.97	92.95	89.32	78.02	92.65	92.28
89.80	86.59	94.20	90.24	79.91	93.73	93.31
89.71	86.18	94.59	90.19	79.79	94.16	93.78
90.38	87.12	94.78	90.79	81.08	93.95	93.15
89.69±0.54	86.74±0.60	93.72±0.77	90.09±0.52	79.65±1.09	93.36±0.56	93.30±0.51

**FIGURE 6.** Precision-recall curves generated by LGFRCDA.

The evaluation criteria, F1-score and MCC, which reflect the comprehensive performance of the models, were 90.09% and 79.65% respectively, for our model. Their standard deviations are 0.52% and 1.09%. Additionally, we plotted the receiver operating characteristic curve (ROC) and precision-recall (PR) for visualization. We calculated the area under ROC (AUC) and area under PR (AUPR) separately to facilitate comparison with other methods. The AUC and PR curves are shown in Table 1, Figure 5 and Figure 6. As can be seen from figures, the average AUC is 93.36% with a standard deviation of 0.56%, while the average of AUPRs is 93.30% with a standard deviation of 0.51%. The above experimental results demonstrate that the proposed model performs well on the CircR2Disease dataset and can effectively predict potential circRNA-disease association.

C. ASSESSMENT OF PREDICTION STABILITY

For LGFRCDA, we evaluated its stability using two methods. The first method involved calculating the standard deviation of each metric, as shown in Table 1. The small standard deviation observed for each metric indicate that the mean values predicted by LGFRCDA are highly representative. Additionally, the low standard deviation suggests that the predictions have low dispersion, minimal errors, and the model exhibits strong stability.

**FIGURE 7.** LGFRCDA's accuracy and loss curves.

The second method involved checking the fitting degree of LGFRCDA. For a stable neural network model, not only should the accuracy be high, but the accuracy and loss curves on the train set and test set should also be reasonable. As the number of epochs increased, we plotted the accuracy and loss curves of LGFRCDA's [22] on the train set and test set in Figure 7. As shown in the figure, at the beginning of the training, the train accuracy is lower than the validation accuracy. However, with the increase in epochs, both gradually approach 1. Similarly, the train loss is initially higher than the validation loss, but both gradually approach 0 as the epochs increase. Despite some abnormal changes of samples in these curves, they don't significantly affect the overall fitting process of LGFRCDA. Overall, the accuracy of the test set is higher than that of the training set, while the loss of the test set is lower than that of the training set. In summary, LGFRCDA demonstrates a good ability to fit data with high accuracy and low loss.

D. COMPARISON WITH TRADITIONAL CLASSIFIER

In order to fully highlight the classification advantage of LGFRCDA, we compared the Transformer used in this study with some traditional classifiers, including the logistic

TABLE 2. The performance comparison with different feature dimensions.

Dimension	Acc(%)	Pre(%)	Rec(%)	F1-Score(%)	MCC(%)	AUC(%)	AUPR(%)
776	89.76	86.87	93.71	90.16	79.78	93.40	93.36
726	91.93	91.26	92.74	91.99	83.87	95.42	95.04
676	87.73	90.89	83.87	87.23	75.69	92.84	91.94
626	88.40	88.31	88.53	88.41	76.82	91.58	91.06
576	79.53	92.94	63.91	75.74	62.71	89.82	90.20
526	82.88	78.94	89.69	83.97	66.39	87.86	87.08
476	77.36	74.60	82.30	78.57	55.08	81.28	77.84
426	72.48	72.28	72.94	72.60	44.97	78.71	78.78
376	68.40	63.97	84.25	72.72	38.81	77.35	75.84

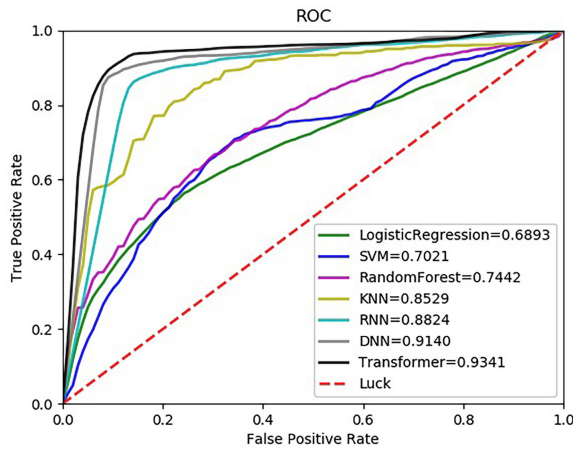


FIGURE 8. The performance comparison on different classifiers in terms of AUC values.

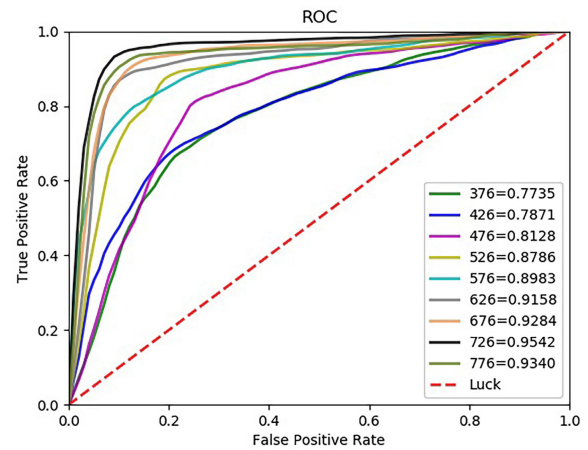


FIGURE 9. The performance comparison with different feature dimensions in terms of AUC values.

regression classifier [39], the support vector machine classifier [40], the random forest classifier [41], the k-nearest neighbor classifier [42], the recurrent neural network [43] and deep neural network [44]. Under the condition that the feature processing of lightGBM remains unchanged, we used the average receiver operating characteristic curve of each independent experiment as the comparison index through the ten-fold cross-validation. The comparison is shown in the Figure 8:

As observed in the figure, the prediction effect of Transformer surpasses that of the other six traditional classifiers. This experiment provides evidence of the Transformer's advantages in data characteristics fusion and data result prediction. Moreover, in the feature processing method proposed in this paper, Transformer effectively utilizes the strengths of attribute features. Consequently, Transformer demonstrates superior performance compared to the other traditional classifiers.

E. COMPARISON WITH DIFFERENT FEATURE DIMENSION

For lightGBM in this paper, it can extract and rank the importance of features in the dataset. So if we reduce some features in order of importance, will the prediction effect of LGFRCDA change? The feature dimension of the input was set at 776 initially. We then reduced the dimension by

50 features each time and compared the results. We continued using the ten-fold cross-validation, drew AUC curves, and recorded all indicators. This comparison of AUC values is shown in the Figure 9, and all indicator data is shown in Table 2:

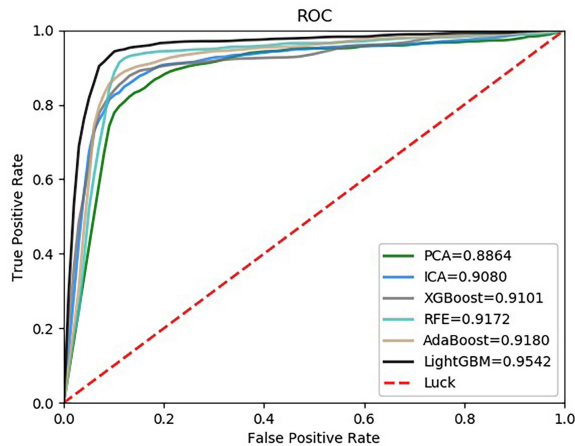
From the figure, it's evident that LGFRCDA achieves the maximum AUC value when the feature dimension is set to 726. Additionally, Table 2 also shows that when the feature dimension is 726, LGFRCDA outperforms other models in most metrics, with the exception of slightly lower recall compared to the dimension of 776. In summary, lightGBM not only ranks the importance of features, but also reduce the feature dimension of the incoming classifier. This comparison shows that lightGBM can not only optimize the prediction effect, but also reduce the feature dimension and extract strongly correlated features.

F. COMPARISON WITH DIFFERENT FEATURE EXTRACTION

In order to fully highlight the feature extraction advantage of lightGBM, we compared it with some traditional feature extraction algorithms, including independent component analysis(ICA) [45], principal component analysis(PCA) [45], recursive feature elimination(RFE) [46], adaboost [47] and xgboost [48]. Under the condition of a feature dimension of 726 and an unchanged Transformer, we employ the average

TABLE 3. The performance comparison with related models.

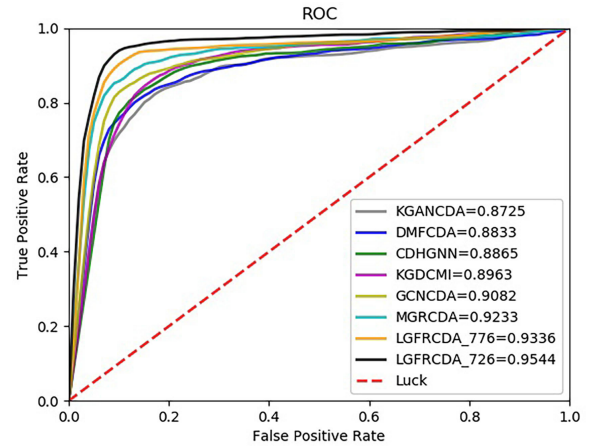
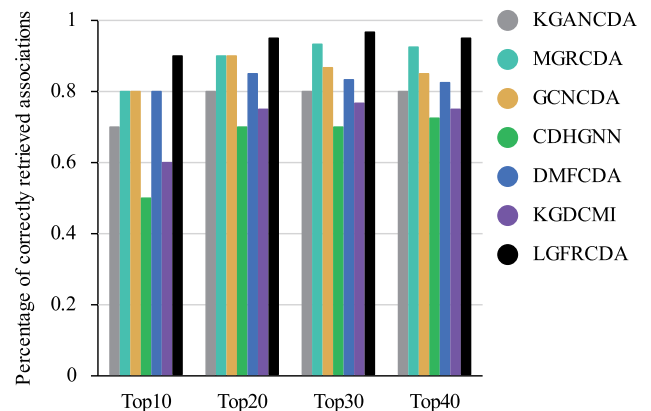
Method	KGANCD	DMFCDA	CDHGNN	KGDCMI	GCNCDA	MGRCD	LGFRCD ₇₇₆	LGFRCD ₇₂₆
AUC(%)	87.25	88.33	88.65	89.63	90.82	92.33	93.36	95.44

**FIGURE 10.** The performance comparison with different feature extraction algorithms in terms of AUC values.

receiver operating characteristic curve of each independent experiment as the comparison index through the ten-fold cross-validation. Its comparison is presented in the Figure 10. From the figure, it is evident that the prediction effect of lightGBM surpasses that of the other five feature extraction algorithms. This experiment provides evidence of lightGBM's advantages in extracting feature importances and achieving accurate predictions. Furthermore, by performing dimension reduction of selected features, Transformer effectively integrate highly correlated features for prediction. Therefore, lightGBM shows better performance than other traditional feature extraction algorithms.

G. COMPARISON WITH RELATED MODELS

To demonstrate the superiority of LGFRCD in predicting the circRNA-disease association, we compared this model with existing models in related fields. We still use the ten-fold cross-validation to verify each model, including CDHGNN [11], DMFCDA [14], KGANCD [62], MGRCD [63], GCNCDA [64] and KGDCMI [65]. To ensure fairness, we use the AUC value, which comprehensively reflects the model performance, as a comparison indicator. The comparison of AUC curves is presented in Figure 11 and their data are shown in Table 3. The superior performance of LGFRCD is evident in two dimensions, as it achieves the highest scores (776_Auc:93.36%, 726_Auc:95.44%). Detailed AUC results are shown in Figure 11. The AUC value of LGFRCD is much higher than that of other models, indicating its superior accuracy. Further more, we conducted a comparison of the percentage of correct retrievals for the top 10 and top 40 prediction associations, as depicted in Figure 12. The

**FIGURE 11.** The performance comparison with different related models in terms of AUC values.**FIGURE 12.** Percentage of correctly retrieved associations.

results show that LGFRCD significantly improved the prediction accuracy. These findings further validate the effectiveness of lightGBM's feature selection and Transformer's information fusion, which lead to enhanced information inference and prediction outcomes in this study.

IV. CASE STUDIES

To evaluate the ability of LGFRCD to predict potential circRNA-disease associations, we conducted case studies using the CircR2Disease dataset. Specifically, we initially trained the proposed model using all known circRNA-disease associations within CircR2Disease. Subsequently we utilized LGFRCD to predict unknown circRNA-disease associations collected from CircR2Disease. We ranked circRNA-disease pairs based on their predicted scores and

TABLE 4. Top 15 circRNA-disease pairs were predicted by LGFRCD A based on known associations in CircR2Disease.

circRNA	disease	evidence (PMID)	literature
hsa_circ_0000745	Oral squamous cell carcinoma	unconfirmed	Nan
hsa_circ_0020397	Colorectal cance	28707774	[49]
hsa_circ_0001821	Oral squamous cell carcinomas	31485648	[50]
hsa_circ_0005986	Hepatocellular carcinoma	28410211	[51]
circPVT1	Breast cancer	33223849	[52]
hsa_circ_0001821	Breast cancer	27928058	[53]
circRNA9777	Esophageal squamous cell carcinoma	29218114	[54]
circ-Foxo3	Osteosarcoma	29805712	[55]
hsa_circ_0001566	Glioblastoma	28236760	[56]
chr18:107887-108499	Glioblastoma	unconfirmed	Nan
ciRS-7	Neuropathic pain	31978512	[57]
circRNA-CDR1as	Osteoarthritis	32099534	[58]
hsa_circ_0001666	Pancreatic ductal adenocarcinoma	29454093	[59]
hsa_circ_0005273	Acute myeloid leukemia	33126007	[60]
hsa_circ_0043278	Non-small cell lung cancer	30873868	[61]

cross-referenced them with recently published literature. The results showed that among the 15 unknown circRNA-disease pairs with the highest scores, 13 pairs were confirmed by relevant literature, as illustrated in Table 4. It’s important to note that while the remaining unknown associations lacked support from the literature, the possibility of their existence cannot be ruled out.

V. CONCLUSION

In this study, we proposed an efficient method based on lightGBM feature extraction and ranking using dataset of disease semantic similarity and disease and circRNA Gaussian interaction kernel similarities, and utilized self-attention Transformer for information fusion and predictive classification. The model demonstrates excellent performance on the CircR2Disease dataset. Furthermore, to fully demonstrate the robustness and reliability of this model, we compared it with other relevant advanced models. Experimental results show that this model exhibits strong predictive performance, with various predictive indices consistently maintained at high levels. Additionally, we conducted a study on the feature processing method of lightGBM. Through a comparison of feature exaction algorithms and feature dimensions in this paper, we not only discovered that lightGBM feature extraction and dimension reduction had a positive impact on prediction ability, but also reaffirmed the significance of the Transformer in enhancing prediction performance. Further more, case studies were designed to demonstrate predictive performance of the model. Remarkably, out of the 15 circRNA-disease pairs with the highest predictive scores, 13 of them are supported by published literature. These findings suggest that the proposed model is effective in predicting circRNA-disease correlations and can offer valuable theoretical support for biological experiments.

Of course, there are several aspects that can be improved in future researches and studies. For example, in terms of feature extraction, there’s room to explore faster and more efficient methods to extract the existing or hidden features from the dataset. Regarding dataset selection, it would be valuable to compare different datasets in order to find a better suited

one that effectively matches diseases in CircR2Disease rather than just similar diseases. Additionally, for the prediction model, we can find a model or a combination of models that is more efficient than Transformer, capable of achieving not only high prediction accuracy, but also high robustness and reliability. The aforementioned aspects will be in urgent need of improvement in the future.

REFERENCES

[1] H. L. Sanger, G. Klotz, D. Riesner, H. J. Gross, and A. K. Kleinschmidt, “Viroids are single-stranded covalently closed circular RNA molecules existing as highly base-paired rod-like structures,” *Proc. Nat. Acad. Sci. USA*, vol. 73, no. 11, pp. 3852–3856, Nov. 1976. [Online]. Available: <https://www.pnas.org/doi/epdf/10.1073/pnas.73.11.3852>

[2] M.-T. Hsu and M. COCA-PRADOS, “Electron microscopic evidence for the circular form of RNA in the cytoplasm of eukaryotic cells,” *Nature*, vol. 280, no. 5720, pp. 339–340, Jul. 1979. [Online]. Available: <https://www.nature.com/articles/280339a0>

[3] J. Salzman, C. Gawad, P. L. Wang, N. Lacayo, and P. O. Brown, “Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types,” *PLoS ONE*, vol. 7, no. 2, Feb. 2012, Art. no. e30733. [Online]. Available: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0030733>

[4] L. Yang, M. Duff, B. Graveley, G. Carmichael, and L. Chen, “Genomewide characterization of non-polyadenylated RNAs,” *Genome Biol.*, vol. 12, no. 2, pp. 1–14, 2011, doi: 10.1186/gb-2011-12-2-r16.

[5] J. Salzman, R. E. Chen, M. N. Olsen, P. L. Wang, and P. O. Brown, “Cell-type specific features of circular RNA expression,” *PLoS Genet.*, vol. 9, no. 9, Sep. 2013, Art. no. e1003777. [Online]. Available: <https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003777>

[6] L. Chen, S. Zhang, J. Wu, J. Cui, L. Zhong, L. Zeng, and S. Ge, “circRNA-100290 plays a role in oral cancer by functioning as a sponge of the miR-29 family,” *Oncogene*, vol. 36, no. 32, pp. 61–4551, 2017, doi: 10.1038/onc.2017.89.

[7] M. Qin, G. Liu, X. Huo, X. Tao, X. Sun, Z. Ge, J. Yang, J. Fan, L. Liu, and W. Qin, “Hsa_circ_0001649: A circular RNA and potential novel biomarker for hepatocellular carcinoma,” *Cancer Biomarkers*, vol. 16, no. 1, pp. 161–169, 2017, doi: 10.1038/onc.2017.89.

[8] L. Fu, Q. Chen, T. Yao, T. Li, S. Ying, Y. Hu, and J. Guo, “Hsa_circ_0005986 inhibits carcinogenesis by acting as a miR-129-5p sponge and is used as a novel biomarker for hepatocellular carcinoma,” *Oncotarget*, vol. 8, no. 27, pp. 43878–43888, Jul. 2017, doi: 10.18632/oncotarget.16709.

[9] A. Rybak-Wolf, C. Stottmeister, M. Jens, N. Pino, S. Giusti, M. Hanan, M. Behm, O. Bartok, R. Ashwal-Fluss, and M. Herzog, “Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed,” *Mol. Cell*, vol. 58, no. 5, pp. 85–870, 2015, doi: 10.1016/j.molcel.2015.03.027.

- [10] L. Iparraguirre, M. Muñoz-Culla, I. Prada-Luengo, T. Castillo-Triviño, J. Olascoaga, and D. Otaegui, "Circular RNA profiling reveals that circular RNAs from ANXA2 can be used as new biomarkers for multiple sclerosis," *Human Mol. Genet.*, vol. 26, no. 18, pp. 72–3564, 2017, doi: [10.1093/hmg/ddx243](https://doi.org/10.1093/hmg/ddx243).
- [11] C. Lu, L. Zhang, M. Zeng, W. Lan, and J. Wang, "Identifying disease-associated circRNAs based on edge-weighted graph attention and heterogeneous graph neural network," *bioRxiv*, to be published, doi: [10.1101/2022.05.04.490565](https://doi.org/10.1101/2022.05.04.490565).
- [12] C. Fan, X. Lei, and F.-X. Wu, "Prediction of CircRNA-disease associations using Katz model based on heterogeneous networks," *Int. J. Biol. Sci.*, vol. 14, no. 14, pp. 1950–1959, 2018. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6299360/>
- [13] X. Chen, Y.-A. Huang, Z.-H. You, G.-Y. Yan, and X.-S. Wang, "A novel approach based on Katz measure to predict associations of human microbiota with non-infectious diseases," *Bioinformatics*, vol. 33, no. 5, pp. 733–739, Mar. 2017, doi: [10.1093/bioinformatics/btw715](https://doi.org/10.1093/bioinformatics/btw715).
- [14] C. Lu, M. Zeng, F. Zhang, F. Wu, M. Li, and J. Wang, "Deep matrix factorization improves prediction of human CircRNA-disease associations," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 3, pp. 891–899, Mar. 2021, doi: [10.1109/JBHI.2020.2999638](https://doi.org/10.1109/JBHI.2020.2999638).
- [15] H. Wei, Y. Xu, and B. Liu, "iCircDA-LTR: identification of circRNA-disease associations based on learning to rank," *Bioinformatics*, vol. 37, no. 19, pp. 10–3302, 2021, doi: [10.1093/bioinformatics/btab334](https://doi.org/10.1093/bioinformatics/btab334).
- [16] L. Wang, Z. You, J. Li, and Y. Huang, "IMS-CDA: Prediction of CircRNA-disease associations from the integration of multisource similarity information with deep stacked autoencoder model," *IEEE Trans. Cybern.*, vol. 51, no. 11, pp. 5522–5531, Nov. 2021. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/9216542>
- [17] L. Wang, Z.-H. You, Y.-A. Huang, D.-S. Huang, and K. C. C. Chan, "An efficient approach based on multi-sources information to predict circRNA-disease associations using deep convolutional neural network," *Bioinformatics*, vol. 36, no. 13, pp. 4038–4046, Jul. 2020, doi: [10.1093/bioinformatics/btz825](https://doi.org/10.1093/bioinformatics/btz825).
- [18] C. Fan, X. Lei, Z. Fang, Q. Jiang, and F.-X. Wu, "CircR2Disease: A manually curated database for experimentally supported circular RNAs associated with various diseases," *Database*, vol. 2018, p. 44, Jan. 2018, doi: [10.1093/database/bay044](https://doi.org/10.1093/database/bay044).
- [19] L. Li, M. Pagnucco, and Y. Song, "Graph-based spatial transformer with memory replay for multi-future pedestrian trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2231–2241.
- [20] G. Wang, X. Zhang, Z. Pan, A. Rodríguez Patón, S. Wang, T. Song, and Y. Gu, "Multi-TransDTI: Transformer for drug-target interaction prediction based on simple universal dictionaries with multi-view strategy," *Biomolecules*, vol. 12, no. 5, p. 644, Apr. 2022, doi: [10.3390/biom12050644](https://doi.org/10.3390/biom12050644).
- [21] H. Gong, T. Mu, Q. Li, H. Dai, C. Li, Z. He, W. Wang, F. Han, A. Tuniyazi, H. Li, X. Lang, Z. Li, and B. Wang, "Swin-transformer-enabled YOLOv5 with attention mechanism for small object detection on satellite images," *Remote Sens.*, vol. 14, no. 12, p. 2861, Jun. 2022, doi: [10.3390/rs14122861](https://doi.org/10.3390/rs14122861).
- [22] C. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Res.*, vol. 30, no. 1, pp. 79–82, 2005. [Online]. Available: <https://www.int-res.com/abstracts/cr/v30/n1/p79-82>
- [23] H. Rao, X. Shi, A. K. Rodrigue, J. Feng, Y. Xia, M. Elhoseny, X. Yuan, and L. Gu, "Feature selection based on artificial bee colony and gradient boosting decision tree," *Appl. Soft Comput.*, vol. 74, pp. 634–642, Jan. 2019, doi: [10.1016/j.asoc.2018.10.036](https://doi.org/10.1016/j.asoc.2018.10.036).
- [24] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
- [25] A. Vaswani, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–21.
- [26] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," 2016, *arXiv:1612.02295*.
- [27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [28] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.
- [29] A. Fred Agarap, "Deep learning using rectified linear units (ReLU)," 2018, *arXiv:1803.08375*.
- [30] J. Kilian and H. T. Siegelmann, "On the power of sigmoid neural networks," in *Proc. 6th Annu. Conf. Comput. Learn. Theory*, 1993, pp. 137–143.
- [31] T. Fushiki, "Estimation of prediction error by using K-fold cross-validation," *Statist. Comput.*, vol. 21, no. 2, pp. 137–146, Apr. 2011, doi: [10.1007/s11222-009-9153-8](https://doi.org/10.1007/s11222-009-9153-8).
- [32] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," *Comput. Biol. Med.*, vol. 136, Sep. 2021, Art. no. 104672, doi: [10.1016/j.compbiomed.2021.104672](https://doi.org/10.1016/j.compbiomed.2021.104672).
- [33] M. M. Ali, K. Ahmed, F. M. Bui, B. K. Paul, S. M. Ibrahim, J. M. W. Quinn, and M. A. Moni, "Machine learning-based statistical analysis for early stage detection of cervical cancer," *Comput. Biol. Med.*, vol. 139, Dec. 2021, Art. no. 104985, doi: [10.1016/j.compbiomed.2021.104985](https://doi.org/10.1016/j.compbiomed.2021.104985).
- [34] A. A. Hemu, R. B. Mim, M. M. Ali, M. Nayer, K. Ahmed, and F. M. Bui, "Identification of significant risk factors and impact for ASD prediction among children using machine learning approach," in *Proc. 2nd Int. Conf. Adv. Electr., Comput., Commun. Sustain. Technol. (ICAECT)*, Apr. 2022, pp. 1–6, doi: [10.1109/ICAECT54875.2022.9808043](https://doi.org/10.1109/ICAECT54875.2022.9808043).
- [35] M. S. Hossain, "ML_SPS: Stroke prediction system employing machine learning approach," in *Proc. Artif. Intell. Data Sci., First Int. Conf.*, vol. 1673, 2022, pp. 215–226.
- [36] M. H. Zweig and G. Campbell, "Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine," *Clin. Chem.*, vol. 39, no. 4, pp. 561–577, Apr. 1993, doi: [10.1093/clinchem/39.4.561](https://doi.org/10.1093/clinchem/39.4.561).
- [37] P. Flach and M. Kull, "Precision-recall-gain curves: PR analysis done right," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–10, doi: [10.1093/clinchem/39.4.561](https://doi.org/10.1093/clinchem/39.4.561).
- [38] E. L. Folador, S. S. Hassan, N. Lemke, D. Barh, A. Silva, R. S. Ferreira, and V. Azevedo, "An improved interolog mapping-based computational prediction of protein-protein interactions with increased network coverage," *Integr. Biol.*, vol. 6, no. 11, pp. 1080–1087, 2014, doi: [10.1039/c4ib00136b](https://doi.org/10.1039/c4ib00136b).
- [39] H. K. Andi, "An accurate Bitcoin price prediction using logistic regression with LSTM machine learning model," *Journal of Soft Computing Paradigm*, vol. 3, no. 3, pp. 205–217, Sep. 2021, doi: [10.36548/jscp.2021.3.006](https://doi.org/10.36548/jscp.2021.3.006).
- [40] T. M. Ghazal, M. Anam, M. K. Hasan, M. Hussain, M. S. Farooq, H. M. A. Ali, M. Ahmad, and T. R. Soomro, "Hep-Pred: Hepatitis C staging prediction using fine Gaussian SVM," *Comput., Mater. Continua*, vol. 69, no. 1, pp. 191–203, 2021. [Online]. Available: <http://www.techscience.com/cmc/v69n1/42725>
- [41] M. A. Khan, S. A. Memon, F. Farooq, M. F. Javed, F. Aslam, and R. Alyousef, "Compressive strength of fly-ash-based geopolymer concrete by gene expression programming and random forest," *Adv. Civil Eng.*, vol. 2021, pp. 1–17, Jan. 2021, doi: [10.1155/2021/6618407](https://doi.org/10.1155/2021/6618407).
- [42] J. Lu, W. Qian, S. Li, and R. Cui, "Enhanced K-nearest neighbor for intelligent fault diagnosis of rotating machinery," *Appl. Sci.*, vol. 11, no. 3, p. 919, Jan. 2021, doi: [10.3390/app11030919](https://doi.org/10.3390/app11030919).
- [43] K. Shang, Z. Chen, Z. Liu, L. Song, W. Zheng, B. Yang, S. Liu, and L. Yin, "Haze prediction model using deep recurrent neural network," *Atmosphere*, vol. 12, no. 12, p. 1625, Dec. 2021, doi: [10.3390/atmos12121625](https://doi.org/10.3390/atmos12121625).
- [44] A. Aldahdooh, W. Hamidouche, S. A. Fezza, and O. Déforges, "Adversarial example detection for DNN models: A review and experimental comparison," *Artif. Intell. Rev.*, vol. 55, no. 6, pp. 4403–4462, Aug. 2022.
- [45] B. A. Draper, K. Baek, M. S. Bartlett, and J. R. Beveridge, "Recognizing faces with PCA and ICA," *Comput. Vis. Image Understand.*, vol. 91, nos. 1–2, pp. 115–137, Jul./Aug. 2003, doi: [10.1016/S1077-3142\(03\)00077-8](https://doi.org/10.1016/S1077-3142(03)00077-8).
- [46] H. Sanz, C. Valim, E. Vegas, J. M. Oller, and F. Reverter, "SVM-RFE: Selection and visualization of the most relevant features through non-linear kernels," *BMC Bioinf.*, vol. 19, no. 1, pp. 1–18, Dec. 2018.
- [47] X. Huang, Z. Li, Y. Jin, and W. Zhang, "Fair-AdaBoost: Extending AdaBoost method to achieve fair classification," *Exp. Syst. Appl.*, vol. 202, Sep. 2022, Art. no. 117240, doi: [10.1016/j.eswa.2022.117240](https://doi.org/10.1016/j.eswa.2022.117240).
- [48] S. Pan, Z. Zheng, Z. Guo, and H. Luo, "An optimized XGBoost method for predicting reservoir porosity using petrophysical logs," *J. Petroleum Sci. Eng.*, vol. 208, Jan. 2022, Art. no. 109520, doi: [10.1016/j.petrol.2021.109520](https://doi.org/10.1016/j.petrol.2021.109520).
- [49] X. Zhang, L. Xu, and F. Wang, "Hsa_circ_0020397 regulates colorectal cancer cell viability, apoptosis and invasion by promoting the expression of the miR-138 targets TERT and PD-L1," *Cell Biol. Int.*, vol. 41, no. 9, pp. 64–1056, 2017, doi: [10.1002/cbin.10826](https://doi.org/10.1002/cbin.10826).

- [50] T. He, X. Li, D. Xie, and L. Tian, "Overexpressed circPVT1 in oral squamous cell carcinoma promotes proliferation by serving as a miRNA sponge," *Mol. Med. Rep.*, vol. 20, pp. 3509–3518, Aug. 2019, doi: [10.3892/mmr.2019.10615](https://doi.org/10.3892/mmr.2019.10615).
- [51] L. Fu, Q. Chen, T. Yao, T. Li, S. Ying, Y. Hu, and J. Guo, "Hsa_circ_0005986 inhibits carcinogenesis by acting as a miR-129-5p sponge and is used as a novel biomarker for hepatocellular carcinoma," *Mol. Med. Rep.*, vol. 8, no. 27, pp. 43878–43888, Jul. 2017, doi: [10.18632/oncotarget.16709](https://doi.org/10.18632/oncotarget.16709).
- [52] J. Wang, K. Huang, L. Shi, Q. Zhang, and Z. S., "Circpvt1 promoted the progression of breast cancer by regulating MiR-29a-3p-mediated AGR2-HIF-1 α pathway," *Cancer Manage. Res.*, vol. 12, pp. 11477–11490, Nov. 2020, doi: [10.2147/CMAR.S265579](https://doi.org/10.2147/CMAR.S265579).
- [53] A. C. Panda, I. Grammatikakis, K. M. Kim, S. De, J. L. Martindale, R. Munk, X. Yang, K. Abdelmohsen, and M. Gorospe, "Identification of senescence-associated circular RNAs (SAC-RNAs) reveals senescence suppressor CircPVT1," *Nucleic Acids Res.*, vol. 45, no. 7, pp. 4021–4035, Apr. 2017, doi: [10.1093/nar/gkw1201](https://doi.org/10.1093/nar/gkw1201).
- [54] J. Sun, X. Yuan, X. Li, D. Wang, T. Shan, W. Wang, Q. Wan, X. Wang, J. Yan, and S. Gao, "Comparative transcriptome analysis of the global circular RNAs expression profiles between SHEE and SHEEC cell lines," *Amer. J. Transl. Res.*, vol. 9, no. 11, p. 5169, 2017. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5714800/>
- [55] M. Xiao-Long, Z. Kun-Peng, and Z. Chun-Lin, "Circular RNA circ_HIPK3 is down-regulated and suppresses cell proliferation, migration and invasion in osteosarcoma," *J. Cancer*, vol. 9, no. 10, pp. 1856–1862, 2018, doi: [10.7150/jca.24619](https://doi.org/10.7150/jca.24619).
- [56] J. Zhu, J. Ye, L. Zhang, L. Xia, H. Hu, H. Jiang, Z. Wan, F. Sheng, Y. Ma, W. Li, and J. Qian, "Differential expression of circular RNAs in glioblastoma multiforme and its correlation with prognosis," *Transl. Oncol.*, vol. 10, no. 2, pp. 9–271, 2017, doi: [10.1016/j.tranon.2016.12.006](https://doi.org/10.1016/j.tranon.2016.12.006).
- [57] W. Cai, Y. Zhang, and Z. Su, "CiRS-7 targeting miR-135a-5p promotes neuropathic pain in CCI rats via inflammation and autophagy," *Gene*, vol. 736, Apr. 2020, Art. no. 144386, doi: [10.1016/j.gene.2020.144386](https://doi.org/10.1016/j.gene.2020.144386).
- [58] W. Zhang, C. Zhang, C. Hu, C. Luo, B. Zhong, and X. Yu, "Circular RNA-CDRIas acts as the sponge of microRNA-641 to promote osteoarthritis progression," *J. Inflammation*, vol. 17, no. 1, pp. 1–10, Dec. 2020, doi: [10.1186/s12950-020-0234-y](https://doi.org/10.1186/s12950-020-0234-y).
- [59] W.-J. Huang, Y. Wang, S. Liu, J. Yang, S.-X. Guo, L. Wang, H. Wang, and Y.-F. Fan, "RETRACTED: Silencing circular RNA hsa_circ_0000977 suppresses pancreatic ductal adenocarcinoma progression by stimulating miR-874-3p and inhibiting PLK1 expression," *Cancer Lett.*, vol. 422, pp. 70–80, May 2018, doi: [10.1016/j.canlet.2018.02.014](https://doi.org/10.1016/j.canlet.2018.02.014).
- [60] L. Yi, L. Zhou, J. Luo, and Q. Yang, "Circ-PTK2 promotes the proliferation and suppressed the apoptosis of acute myeloid leukemia cells through targeting miR-330-5p/FOXO1 axis," *Blood Cells, Molecules, Diseases*, vol. 86, Feb. 2021, Art. no. 102506, doi: [10.1016/j.bcmd.2020.102506](https://doi.org/10.1016/j.bcmd.2020.102506).
- [61] J. Cui, W. Li, G. Liu, X. Chen, X. Gao, H. Lu, and D. Lin, "A novel circular RNA, hsa_circ_0043278, acts as a potential biomarker and promotes non-small cell lung cancer cell proliferation and migration by regulating miR-520f," *Artif. Cells, Nanomedicine, Biotechnol.*, vol. 47, no. 1, pp. 810–821, Dec. 2019, doi: [10.1080/21691401.2019.1575847](https://doi.org/10.1080/21691401.2019.1575847).
- [62] W. Lan, Y. Dong, Q. Chen, R. Zheng, J. Liu, Y. Pan, and Y.-P.-P. Chen, "KGANCA: Predicting circRNA-disease associations based on knowledge graph attention network," *Briefings Bioinf.*, vol. 23, no. 1, Jan. 2022, doi: [10.1093/bib/bbab494](https://doi.org/10.1093/bib/bbab494).
- [63] L. Wang, Z. You, D. Huang, and J. Li, "MGRCA: Metagraph recommendation method for predicting circRNA-disease association," *IEEE Trans. Cybern.*, vol. 53, no. 1, pp. 67–75, Jan. 2021. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/9477295>
- [64] L. Wang, Z.-H. You, Y.-M. Li, K. Zheng, and Y.-A. Huang, "GCNCDA: A new method for predicting circRNA-disease associations based on graph convolutional network algorithm," *PLOS Comput. Biol.*, vol. 16, no. 5, May 2020, Art. no. e1007568, doi: [10.1371/journal.pcbi.1007568](https://doi.org/10.1371/journal.pcbi.1007568).
- [65] X.-F. Wang, C.-Q. Yu, L.-P. Li, Z.-H. You, W.-Z. Huang, Y.-C. Li, Z.-H. Ren, and Y.-J. Guan, "KGDCMI: A new approach for predicting circRNA-miRNA interactions from multi-source information extraction and deep learning," *Frontiers Genet.*, vol. 13, Aug. 2022, doi: [10.3389/fgene.2022.958096](https://doi.org/10.3389/fgene.2022.958096).



CHEN MA received the B.E. degree from Shanxi University, Taiyuan, China, in 2021. He is currently pursuing the master's degree in computer technology with Xijing University, Xi'an, Shaanxi, China. He has published two journal articles in IEEE on the related subject. His research interests include big data analysis and software engineering.



YUHONG CHI received the master's degree in computer applied engineering from Northeastern University, Liaoning, China, in 2005, and the Ph.D. degree in computer science and technology from Tsinghua University, Beijing, China, in 2013. In recent years, she has participated in research projects. She has participated in and presided over several provincial projects and has presided over the Key Laboratory Opening Fund of Shaanxi Province. She has published more than 20 academic articles. Her research interests include computational intelligence and its applications and data analysis. She received one second prize and one third prize for scientific and technological progress.



DONGLAI HAO received the master's and Ph.D. degrees in communication and information systems from Xidian University, Xi'an, Shaanxi, China, in 2002 and 2010, respectively. In recent years, he has presided over three provincial projects, participated in two NSFC projects, and published more than 20 academic articles. His research interest includes artificial intelligence and its application in wireless communications. He received one third prize in provincial and ministerial science and technology progress.



XIONGFEI JI received the master's degree in electromagnetic field and microwave technology from Northwestern Polytechnical University, Xi'an, Shaanxi, China, in 2007. In recent years, he participated in and presided over several provincial and ministerial scientific research projects, and developed several anti-interference communication terminals. His research interest includes wireless communications. He received the Second Prize and Third Prize of the Provincial Science and Technology Progress Award.

...