TYPE Original Research
PUBLISHED 13 July 2023
DOI 10.3389/fgene.2023.1232038

EMDL-ac4C: identifying N4-acetylcytidine based on ensemble two-branch residual connection DenseNet and attention

Jianhua Jia*, Zhangying Wei* and Xiaojing Cao

School of Information Engineering, Jingdezhen Ceramic University, Jingdezhen, China

Check for updates

OPEN ACCESS

EDITED BY
Margherita Mutarelli,
National Research Council (CNR), Italy

REVIEWED BY
Hilal Tayara,
Jeonbuk National University, Republic of Korea
Kunqi Chen,
Fujian Medical University, China

*CORRESPONDENCE
Jianhua Jia,
✉ jjh163yx@163.com
Zhangying Wei,
✉ weizy5003@163.com

RECEIVED 31 May 2023
ACCEPTED 29 June 2023
PUBLISHED 13 July 2023

CITATION
Jia J, Wei Z and Cao X (2023), EMDL-ac4C: identifying N4-acetylcytidine based on ensemble two-branch residual connection DenseNet and attention. *Front. Genet.* 14:1232038. doi: 10.3389/fgene.2023.1232038

Introduction: N4-acetylcytidine (ac4C) is a critical acetylation modification that has an essential function in protein translation and is associated with a number of human diseases.

Methods: The process of identifying ac4C sites by biological experiments is too cumbersome and costly. And the performance of several existing computational models needs to be improved. Therefore, we propose a new deep learning tool EMDL-ac4C to predict ac4C sites, which uses a simple one-hot encoding for a unbalanced dataset using a downsampled ensemble deep learning network to extract important features to identify ac4C sites. The base learner of this ensemble model consists of a modified DenseNet and Squeeze-and-Excitation Networks. In addition, we innovatively add a convolutional residual structure in parallel with the dense block to achieve the effect of two-layer feature extraction.

Results: The average accuracy (Acc), mathews correlation coefficient (MCC), and area under the curve Area under curve of EMDL-ac4C on ten independent testing sets are 80.84%, 61.77%, and 87.94%, respectively.

Discussion: Multiple experimental comparisons indicate that EMDL-ac4C outperforms existing predictors and it greatly improved the predictive performance of the ac4C sites. At the same time, EMDL-ac4C could provide a valuable reference for the next part of the study. The source code and experimental data are available at: https://github.com/13133989982/EMDLac4C.

KEYWORDS

ac4C site identification, ensemble deep learning, DenseNet, attention mechanism, residual structure

1 Introduction

RNAs from both eukaryotic and prokaryotic cells may include a broad range of nucleoside modifications (Tardu et al., 2019), and it is statistically known that there are more than 170 types (Boccaletto et al., 2018). Song et al. (2023) have developed the RMDisease database and identified a large number of disease-associated variants to elucidate the important regulatory role of RNA modifications. Among them, N4-acetylcytidine (ac4C) is a highly conserved RNA modification, and at the same time, he is the sole acetylation modification of eukaryotic RNA that has been identified (Zhao et al., 2019a; Jin et al., 2020). Ac4C plays an important role in biology, and it has different functions on different RNAs. On tRNA, ac4C helps to improve the accuracy of protein translation and maintain the heat resistance of the organism (Kumbhar et al., 2013); the role of ac4C on rRNA likewise includes maintaining high fidelity of protein translation (Sharma et al., 2015), while it is also a marker of thermophilic organisms, which is significant; on mRNA, ac4C is required to safeguard the stability of mRNA while increasing the efficiency of protein translation (Arango et al., 2018; Dominissini and Rechavi, 2018). Meanwhile, Chen et al. (2022) demonstrated that the only known ac4C writer, N-acetyltransferase 10 (NAT10), has an important effect in male reproduction. In addition, ac4C has regulatory effects on viruses (Tsai et al., 2020; Hao et al., 2022) and has been associated with several human diseases, including: osteoporosis (Yang et al., 2021), pancreatic cancer (Feng et al., 2022), *etc.*

The recognition of ac4C sites has gradually become a popular topic in biology and computer research. In the context of biological experiments, multiple testing methods exist for ac4C studies. Previously, partially enzymatic hydrolysis and two-dimensional paper chromatography were commonly used to identify ac4C modifications in RNA (Thomas et al., 2019). In the past few years, researchers had found that the combination of LC-MS and HPLC-MS analyses is more efficient in isolating partially modified nucleic acids, including ac4C (Ito et al., 2014; Sharma et al., 2015; Sharma et al., 2017). In addition, RPHPLC (Mezzar et al., 2014) is widely used by the biological community as it requires only a small number of samples and does not rely on expensive mass spectrometry assays or the use of radioactive substrates. There are also specific ac4C sequencing approaches, including ac4C-seq (Gamage et al., 2021) and RedaC:T-seq (Sturgill et al., 2022), both of which sequence ac4C through a series of experiments under certain conditions. Nevertheless, these experimental methods always have several problems, such as time-consuming and high expensive.

Several machine learning-based predictive models (Zhou et al., 2016; Wei et al., 2019; Basith et al., 2019; Lv et al., 2020; Hasan et al., 2021) for the predictive identification of RNA post-translational modification sites have been developed by researchers in the past several years. Among them, there are two predictors used to identify ac4C sites, firstly, Zhao et al. (2019a) developed PACES based on position-specific dinucleotide sequences as well as K-nucleotide frequency coding, which was trained using two random forest classifiers. Secondly, Alam et al. (2020b) proposed XG-ac4C predictor based on this, which used multiple encoding methods, including one-hot, nucleotide chemistry and density, Kmer, *etc.*, and used extreme gradient boosting (XGboost) to train the dataset to identify ac4C loci. Nonetheless, neither of these two predictors' ability in making predictions is sufficient.

With the widely use of deep learning, researchers have introduced different deep learning models to the field of RNA post-transcriptional modification site prediction (Yu and Dai, 2019; Liu et al., 2020; Hasan et al., 2021; Rehman et al., 2021; Hasan et al., 2022; Tsukiyama et al., 2022; Yang et al., 2023), and a lot of experimental results show that deep learning models perform better than machine learning models for datasets with a large number of samples. Muhammad et al. (Iqbal et al., 2022) used a deep learning network-based CNN model using an encoding approach similar to the XG-ac4C predictor DL-ac4C predictor was proposed to identify ac4C sites, and compared with machine learning methods, the results showed that DL-ac4C has better prediction performance. Recently, a new deep learning model DeepAc4C (Wang et al., 2021a) had also been developed to increase the efficiency of ac4C locus recognition in mRNA. It also used CNN to extract information and classify the feature maps, and encoded a combination of physicochemical features and nucleotide semantic information. Yet, the classification performance of these two predictors still needs to be improved. As a result, a more analytically precise model to anticipate ac4C sites is urgently required.

TABLE 1 Distribution of data set D1.

Dataset	Positive	Negative
Training	1,148	1,148
Testing	467	467

In order to more accurately predict ac4C sites in mRNAs, an effective prediction model EMDL-ac4C was developed in this paper, and the contributions of this paper were various: 1) EMDL-ac4C used only the simplest encoding method one-hot to represent nucleotides. This encoding showed the distribution probability of each nucleotide, making it easier to calculate the distance between nucleotides. 2) It proposed a powerful deep learning that used DenseNet in combination with convolutional residuals to form two-branch residual connection DenseNet, and the effectiveness of feature extraction was increased by this way. 3) For cases like this paper, where unbalanced datasets were processed into multiple balanced datasets, downsampling integration was used to achieve significantly superior generalization performance than a single learner. 4) The attention mechanism was carried out throughout the network structure to give greater attention to the important information at each stage. 5) We compared the performance of various encoding techniques, different numbers of dense blocks, multiple model architectures, and several predictors, respectively, to confirm the efficacy of the EMDL-ac4C model.

Therefore, the final predictor is called "EMDL-ac4C," in which "EM" stands for "ensemble," "DL" represents "deep learning," and "ac4C" means "N4-acetylcytidine."

2 Materials and methods

2.1 Benchmark dataset

The base dataset for this study was extracted from 2,134 genes offered by Zhao et al. (2019a) from a highly throughput dataset previously presented. In the training set, there were a total of 1,160 positive and 10,855 negative samples, and in the test set, the counts of positive and negative samples were: 469, 4,343, respectively. Alam et al. (2020b) also built the XG-ac4C predictor from this dataset. Wang et al. (2021b) performed a de-sampling redundancy with a threshold of 0.4 using CD-HIT (Weizhong et al., 2006) software in order to remove redundant sequences from these datasets, resulting in 1,615 positive and 7,590 negative samples. These samples were separated into a set for training and one for testing, with 1,148 positive samples and 5,439 negative samples in the former. On the other hand, there were 467 positive samples and 2,151 negative samples in the independent test set. Finally,

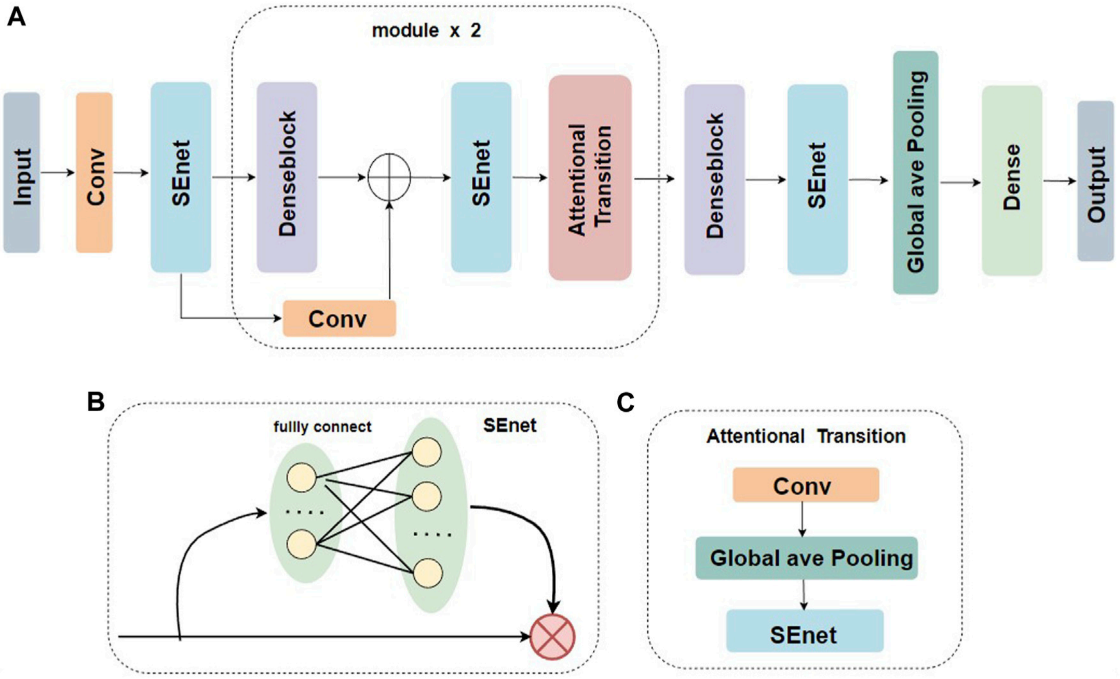

FIGURE 1
Base classifier for EMDL-ac4C. **(A)** Schematic graph of the two-branch residual connection DenseNet model. **(B)** Schematic diagram of the SEnet module. **(C)** Schematic diagram of the Attentional Transition module.

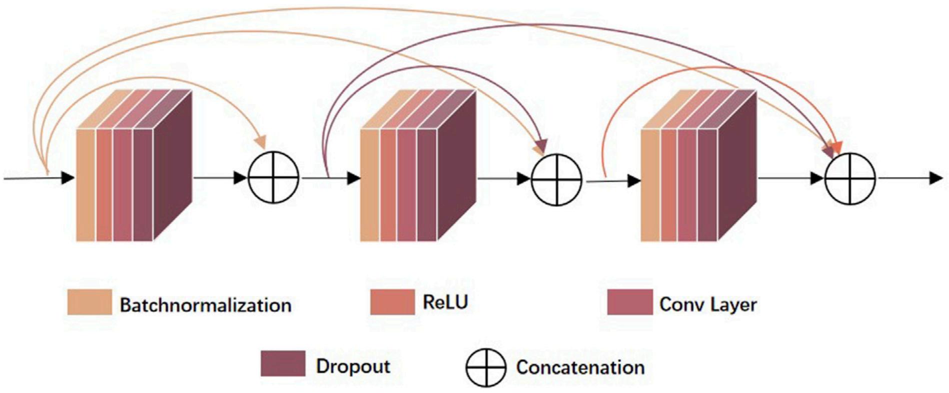

FIGURE 2
The structure of denseblock.

they created ten balanced datasets using the unbalanced training and test sets, respectively, to make it easier to train and test the model. The number of positive samples in each sub-training set and test set remains the same, and negative samples are randomly selected from the corresponding negative dataset, and the number is consistent with the positive sample size. The ten samples are denoted as D1, D2, , D10. Table 1 shows an example of the distribution of data set D1: (the distribution of D2, D3, , D10 is the same).

As in Eq. 1, nucleotide sequences containing potential N4-acetylcytidine sites can normally be read as:

$$f_{\delta}(K) = R_{-\delta}R_{-(\delta-1)}\cdots R_{-2}R_{-1}KR_{+1}R_{+2}\cdots R_{+(\delta-1)}R_{+\delta} \qquad (1)$$

where the center K denotes “N4-acetylcytidine,” $R_{-\delta}$ means the δth upstream nucleotide from the center K, while $R_{+\delta}$ stands for the δth downstream nucleotide from the center K. In this study, δ is 207, that is, the length of a nucleotide sequence is (2δ +1).

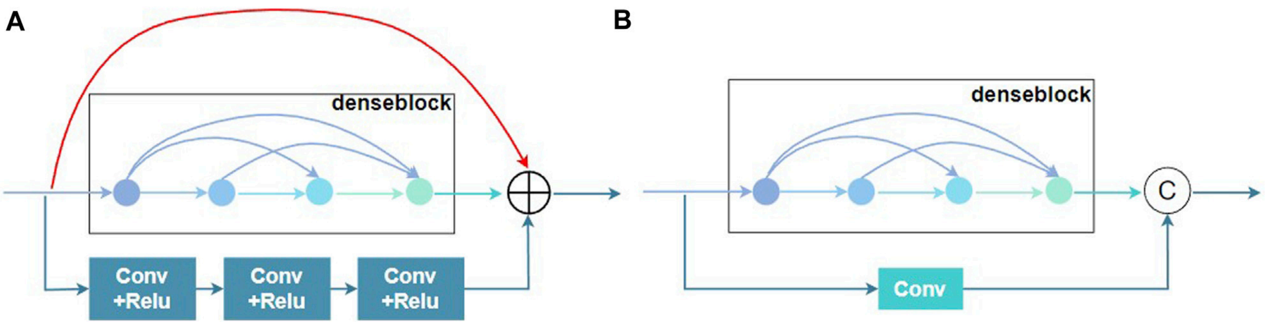

FIGURE 3
Residual connection schematic of the denseblock module. **(A)** RXDNFuse's multi-branch Residual denseblock. **(B)** Two-branch Residual denseblock.

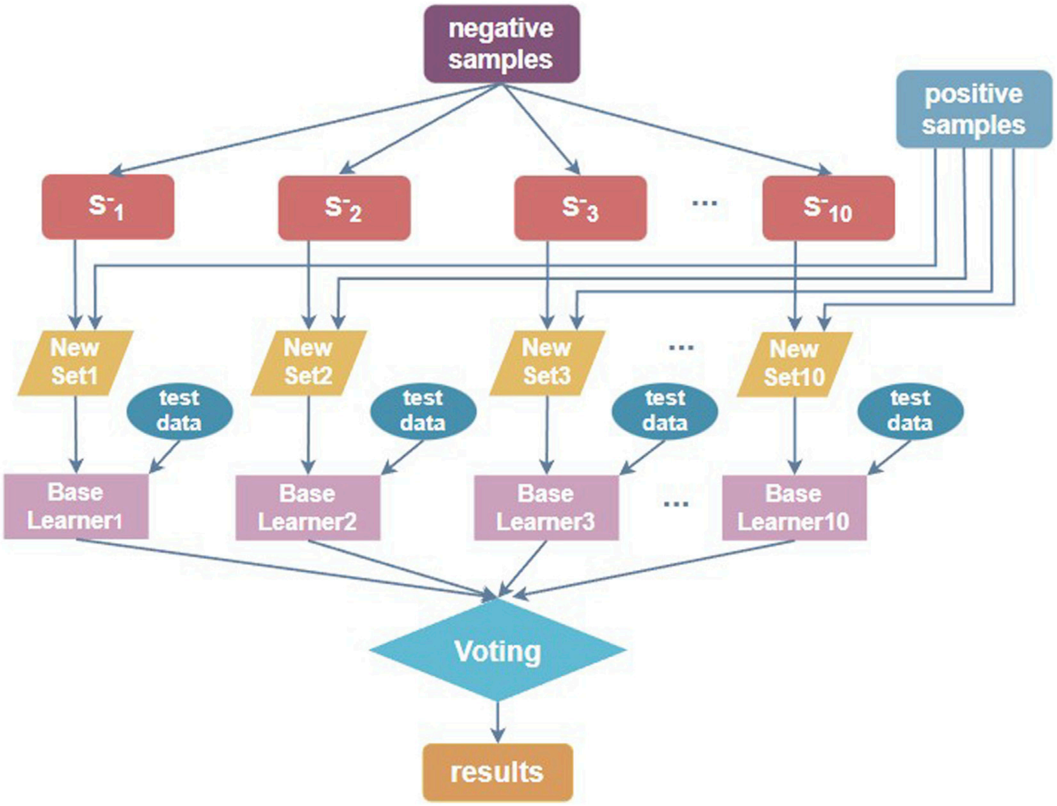

FIGURE 4
Downsampling ensemble classifier construction process.

2.2 Feature coding methods

2.2.1 One-hot coding

The RNA sequence used in this study consists of four nucleotides and a "-," where the "-" represents a missing value or an undetected nucleotide in the RNA sequence. One-hot encoding, also known as binary encoding, converts each nucleotide into a numeric vector of 0 and 1, encoding "A" as [1, 0, 0, 0, 0], "C" as [0, 1, 0, 0, 0], "G" as [0, 0, 1, 0, 0], "T" as [0, 0, 0, 1, 0], and "-" is assigned as [0, 0, 0, 0, 0, 1]. One-hot encoding not only simply converts sequence information into digital information for computer processing, but more importantly, it makes the calculation of distances between nucleotides more reasonable. For example, if nucleotides are represented in sequential encoding: 1:A, 2:C, 3:G, 4:T, then the distance between A (adenine ribonucleotide) and C (cytosine ribonucleotide) is smaller than the distance between A (adenine ribonucleotide) and G (guanine ribonucleotide), which is not reasonable. At the same time, one-hot coding in fact means the

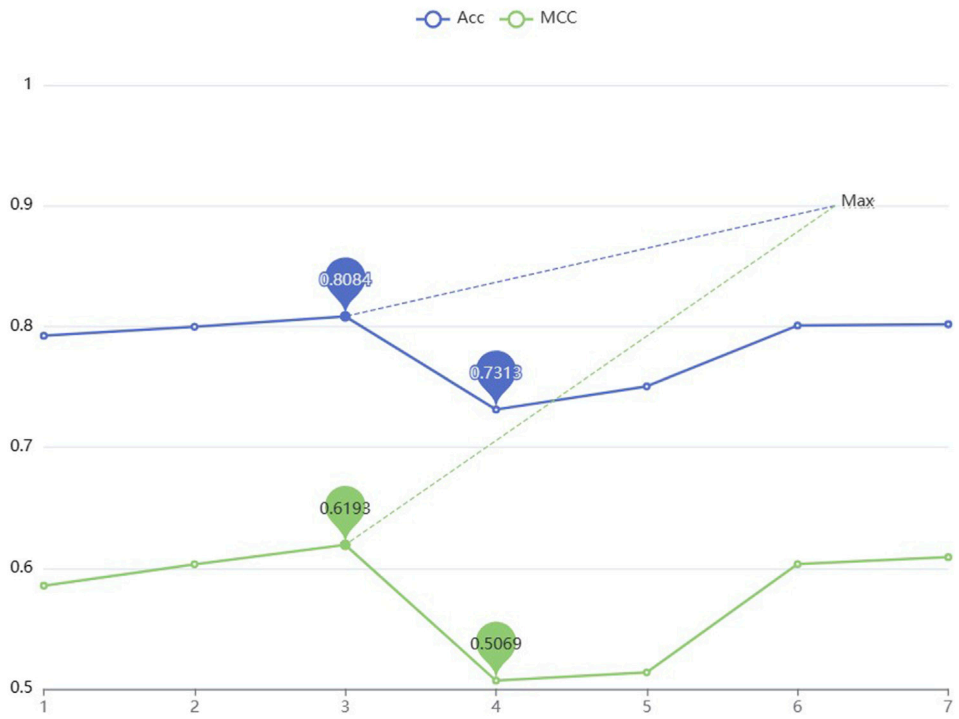

FIGURE 5
The comparison of Acc and MCC values for models with various amounts of denseblocks.

probability distribution of nucleotides, that is, one nucleotide has probability 1 and the others are all 0.

2.3 Classification model

2.3.1 Base learning model

Figure 1 illustrates the model structure of the base learner two-branch residual connection DenseNet for constructing EMDL-ac4C. A): The basic structure of two-branch residual connection DenseNet consists of SENet, denseblock, attentional-transition, residual convolution layer and fully connected layer. B): SENet is composed of two fully connected layers with different number of neurons to fulfill the aim of first dimensionality reduction and then dimensionality increase of the feature map. C): The attentional-transition is composed of 1 × 1 convolution, global average pooling layer and SENet, where the role of convolution and global average pooling is to condense the output feature map of two-branch denseblock and convolutional residual module, decrease the size and dimension of the feature map, and simultaneously can successfully decrease the quantity of denseblock parameters and stop the network to excessive fitting. The purpose of SENet is to extract important features. The final fully connected layer is used as the classification prediction of the model.

Below are explanations of each step in greater detail.

2.3.1.1 DenseNet

Huang et al. (2017) proposed DenseNet in 2017 for the target recognition task to alleviate the gradient disappearance issue that often occurs in deep networks, while feature reuse also enhanced feature propagation with fewer parameters in a network of equal layer depth. As shown in Eq. 2, ResNet only adds features to the input of the latter layer and connects them in a summation manner. DenseNet is an improvement of ResNet in that it combines the features of each layer by concatenation. As shown in Eq. 3, DenseNet connects all the previous layers as the input of the next layer, obtaining better performance than ResNet with fewer parameters and computational cost.

$$x_l = H_l(x_{l-1}) + x_{l-1} \tag{2}$$

$$x_l = H_l([x_0, x_1, \ldots, x_{l-1}]) \tag{3}$$

The $H_l(\bullet)$ in Eqs 2, 3 represents the non-linear transformation function, which is a combined operation that may include a series of BN (Batch Normalization), ReLU(Rectified Linear Unit), POOL (Pooling) and Conv (Convolution) operations. The non-linear transformation function in this paper adopted the structure of BN+ReLU+1*1 Conv +3*3 Conv, which were created through a preactivation strategy to facilitate network training as well as enhanced the efficiency of generalization, and 1 × 1 Conv served to reduce the number of features, thus reduced computational workload and improved computational efficiency. In addition, 3 × 3 Conv offered a larger receptive field.

As shown in Figure 2, the denseblock is a module containing many layers, each layer has the same feature map size, and the layers are closely connected to each other, while the Transition module connects two neighboring denseblocks and reduces the feature map size by Pooling. In this paper, we used a new approach Attentional

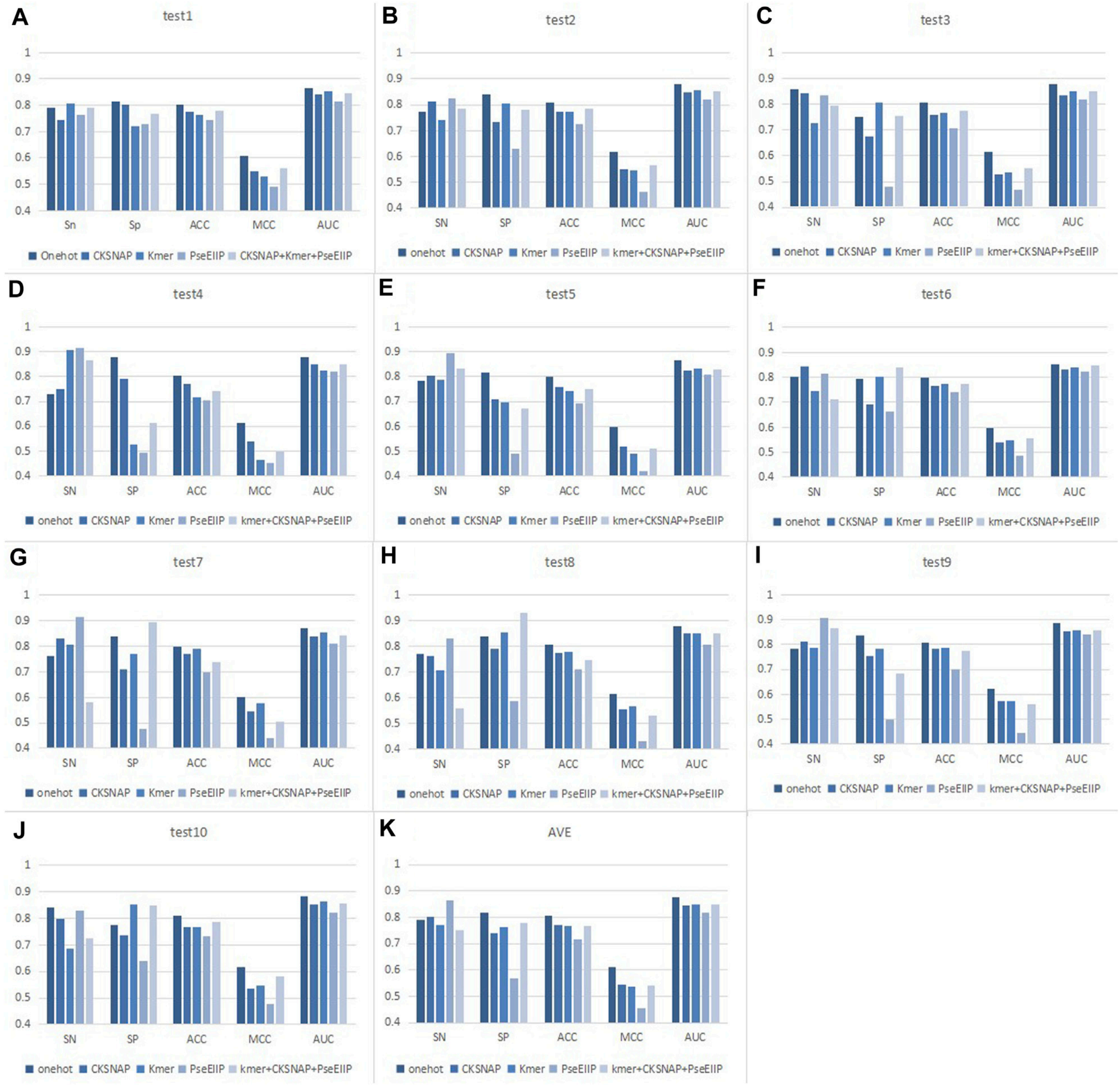

FIGURE 6
Performance comparison of different encoding methods on ten test sets. **(A–J)** Results for different test sets. **(K)** Average results of 10 test sets.

Transition instead of transition, whose construct is depicted in Figure 1C.

2.3.1.2 Residual connection

In order to effectively spread shallow information to deep layers, the output of residual connection is represented by a linear superposition of the inputs and their nonlinear transformations. Deep learning propagates the gradient (derivative) of the loss function step by step from the back to the front in the back propagation process, and the gradient is less than 1 at each level. Due to the cumulative multiplicative effect, the gradient may be too small and cause the network to stop training optimization. Therefore, the addition of residual connections to the network can address the issue of network degradation and enhance network functionality.

Long et al. (2021) proposed the aggregated residual dense network (RXDNFuse) for the mix problem of IR and visible images, which combined the residual and convolutional residuals into parallel dense blocks to extract multi-level features, as depicted in Figure 3A. The results also demonstrated that RXDNFuse can effectively retain the important thermal radiation information in the feature map, from which this work was inspired to propose Two-branch residual connection dense network, which consisted of a denseblock and a convolutional residual connection to form two channels. As illustrated in Figure 3B, compared with RXDNFuse, two-

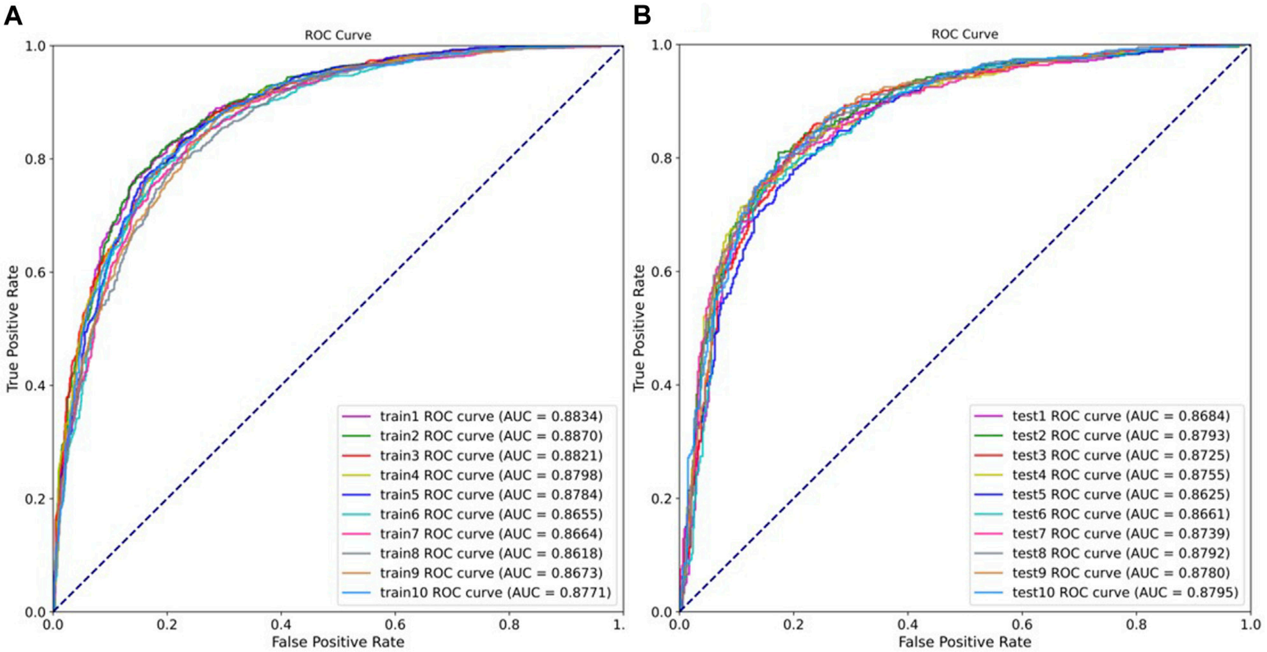

FIGURE 7
ROC curve of two-branch residual connection DenseNet on each dataset. **(A)** Results on training sets. **(B)** Results on testing sets.

branch residual connection DenseNet has three differences, firstly, it reduces one layer of residual connection, secondly, it changes convolutional residual connection part from 3 layers of Conv+ReLU to one layer of Conv to extract features, and for the last, it changes the residual connection method from summation to concatenation. In this way, we can improve the diversity of feature extraction and achieve the effect of two-branch feature extraction, while reducing the complexity of the model.

2.3.1.3 SENet

Introducing attention mechanism in the predictor can make model more efficient in learning the interrelationships between feature information (Vaswani et al., 2017), while focusing on useful information. (Wang et al., 2020; Wang et al., 2021; Jia et al., 2022; Jia et al., 2023). The use of the attention mechanism in both image processing and natural language learning has demonstrated its value in enhancing the model's capacity for recognition, and our study again confirms this point that the attention mechanism can help suppress useless information, pay attention to critical information, and improve the model performance.

Researchers have developed various attention mechanisms, commonly used including self-attention mechanisms, spatial attention mechanism, channel attention mechanism, and so on. Squeeze-and-Excitation Networks (SENet) (Hu et al., 2018) is one of the channel attention mechanisms. We investigated the use of attention mechanisms in our model in two ways, one was to choose the attention mechanism that best matches the model and the data characteristics, and try to use the four attention mechanisms individually or in combination. The second was where to place the attention mechanism in the model. There are various options for where to insert attention into the model, including introducing it in densecells (Bastings and Filippova, 2020), adding it in denseblocks (Wei et al., 2019; Zhou et al., 2019; Wang et al., 2021), inserting it between denseblocks and transitions (Jia et al., 2022; Jia et al., 2023), bring in the attention mechanism in the transition layer (Song et al., 2021), or attaching it before the data enters DenseNet or at the end of the model prediction.

The combined use of SENet and DenseNet has been repeatedly shown to boost network detection and site prediction performance (Yan et al., 2019; Wang et al., 2021; Shi et al., 2021; Jia et al., 2023). We had also found after numerous ablation experiments that SENet alone works best, while it was advisable to place SENet before DenseNet, between denseblock and Attentional Transition, and before the final global average pooling, as seen in Figure 1A. Adding the attentional layer before the initial feature map enters the DenseNet helps the model not miss important information in the original feature map, while the attentional layer behind denseblock aims to repress redundant features and strengthen propagation of important features.

The construction of SENet is to first perform a global pooling operation on the feature map with input h*w*c, which is a spatial compression process that makes the feature map 1*1*c in size. Next there are two fully connected layers. The first full connection has c/16 neurons, which is a dimensionality reduction procedure, and the second fully connected is ascending to c neurons. The significance of dimensionality reduction and then dimensionality increase is to discover the correlation between channels. The final step is to multiply the original h*w*c feature map with the 1*1*c feature map after dimensionality down and dimensionality up to obtain a feature map with the importance levels of different channels.

It is worth noting that the SENet used in this paper removes the global pooling operation at the beginning, the reason is that the global pooling operation will lose some location information.

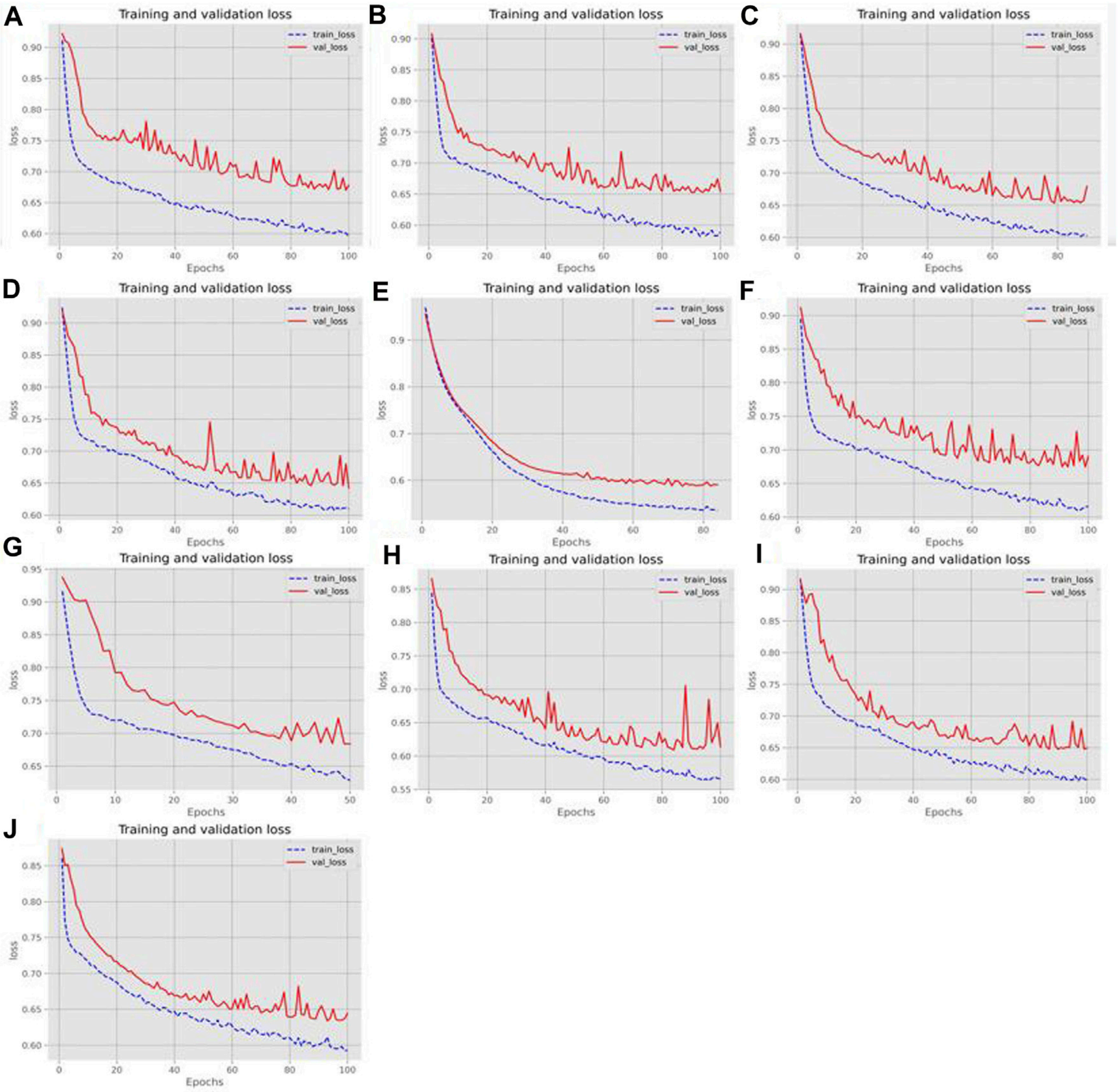

FIGURE 8
The loss change of the training and validation sets. **(A–J)** Loss changes in ten different training and validation sets.

2.3.1.4 Attentional transition

For the transition layer, it mostly connects two neighboring denseblocks and decreases the size of the feature map. Yang et al. (Yang et al., 2018) proposed CliqueNet to alleviate the training challenge of deep networks, which introduced a channel-based attentional mechanism in the transition layer. Following their approach, we also introduce the attentional mechanism in transition layers to ensure high-quality propagation of features between dense blocks. As Figure 1C shows, after convolution, the feature map is then brought into SENet, which means that the feature map is first global average pooled, followed by two fully connected layers (FC) to complete the descending and ascending operations.

2.3.2 Ensemble learning

When the number of positive samples in the benchmark dataset is significantly lower than the number of negative sample data, the dataset is unbalanced. For such a non-equilibrium dataset, using a simple model is not friendly to identify positive samples, while for us, the information of N4-acetylcytidine loci is the most critical and the most necessary to be identified. The ensemble learning method can be used to downsample the non-balanced data, meaning that for the majority of negative samples, downsampling is performed each time, and the same number of subsets as positive samples are extracted, and the two constitute a balanced dataset for training. Multiple sub-classifiers are thus constructed, and then the training models are validated by cross-validation and the model training effect is verified by independent test sets, respectively. Such use of the ensemble classifier dramatically improves the accuracy of loci prediction (Jia et al., 2016). The process of constructing the downsampling ensemble classifier is shown in Figure 4.

Wang et al. (Wang et al., 2021) constructed the dataset in a similar manner to Figure 4, with ten random downsampling of negative samples to build ten balanced datasets. Therefore, we used 10 balanced training sets to build 10 sub-classifiers, trained the model by cross-validation, and then validated the model training effect using ten independent test sets

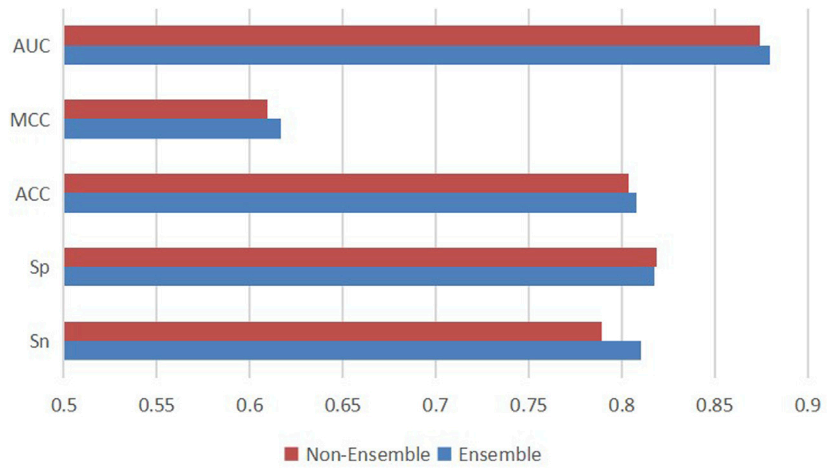

FIGURE 9
Performance comparison of single two-branch residual connection DenseNet and downsampling ensemble model EMDL-ac4C.

one by one. The prediction results obtained from multiple sub-classifiers were soft-voted to obtain the final ensemble results. The base sub-classifiers are implemented with the two-branch residual connection DenseNet shown in Figure 1A.

2.3.3 Performance evaluation

Five metrics are typically used to evaluate models in such studies: Accuracy (Acc), Sensitivity (Sn), Specificity (Sp), Area under curve (AUC) and Matthew's correlation coefficient (MCC), which are computed as follows in Eq. 4:

$$Sn = \frac{TP}{TP + FN}$$

$$Sp = \frac{TN}{TN + FP}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TN + FN) \times (TP + FP) \times (TN + FP)}} \tag{4}$$

TP is a correctly identified positive ac4C site, FN is a misidentified positive ac4C site, and TN and FP are correctly and incorrectly predicted negative ac4C sites, respectively. Researchers often use ROC curves to indicate the performance of classifiers, and the area under the ROC curve is measured as the AUC value, and a bigger value indicates greater performance.

In this study, we used these five prevalent evaluation metrics to assess the performance of EMDL-ac4C.

3 Results and discussion

3.1 Comparison of models with different denseblocks

To improve the performance of the predictor, the parameters are optimized in two-branch residual connection DenseNet. In this section, different numbers of denseblocks are set and the Acc and MCC values are used to compare the model performance for different numbers of denseblocks. According to Figure 5, selecting a model with three denseblocks yields the highest performance. The trough is reached when the number is 4. After the number is 5, the Acc and MCC values increase as the amount of blocks grows, but considering that when the amount of denseblocks reaches 8, the maximum feature map scale in the model run has reached [211464,768], which is easy to cause insufficient server memory. Therefore, based on the consideration of computational complexity, the denseblock = 3 is selected.

3.2 Comparison of models with various encoding methods

For the choice of feature coding methods, we considered four traditional coding methods, namely,: One-hot, composition of k-spaced nucleic acid pairs (CKSNAP), Kmer, and electron–ion interaction pseudopotentials of trinucleotide (PseEIIP). Also included: this coding scheme of CKSNAP + Kmer + PseEIIP combination. These coding methods have been applied in many studies (Li et al., 2020; Wang et al., 2021; Chen et al., 2021; El Allali et al., 2021; Le et al., 2022; Wang et al., 2023) and are not described in detail here. In addition to these traditional coding approaches, there are some new coding approaches including: Gene2vec (Zou et al., 2019), Geo2vec (Huang et al., 2022), Genomics features (Chen et al., 2019), Chemical property (Chen et al., 2017), Heuristic nucleotide physical-chemical properties reduction (Dou et al., 2020) also gave us a lot of inspiration on sequence encoding. We compared these coding schemes on ten balanced test datasets, as indicated in Figure 6. In the ten experiments, we tested univariate the encoding style suitable for the model using a unified classifier: two-branch residual connection DenseNet.

As seen in Figure 6, we can find that for each test set, the values of the indicators corresponding to the same coding method do not differ much. For example, the AUC value of the model with one-hot

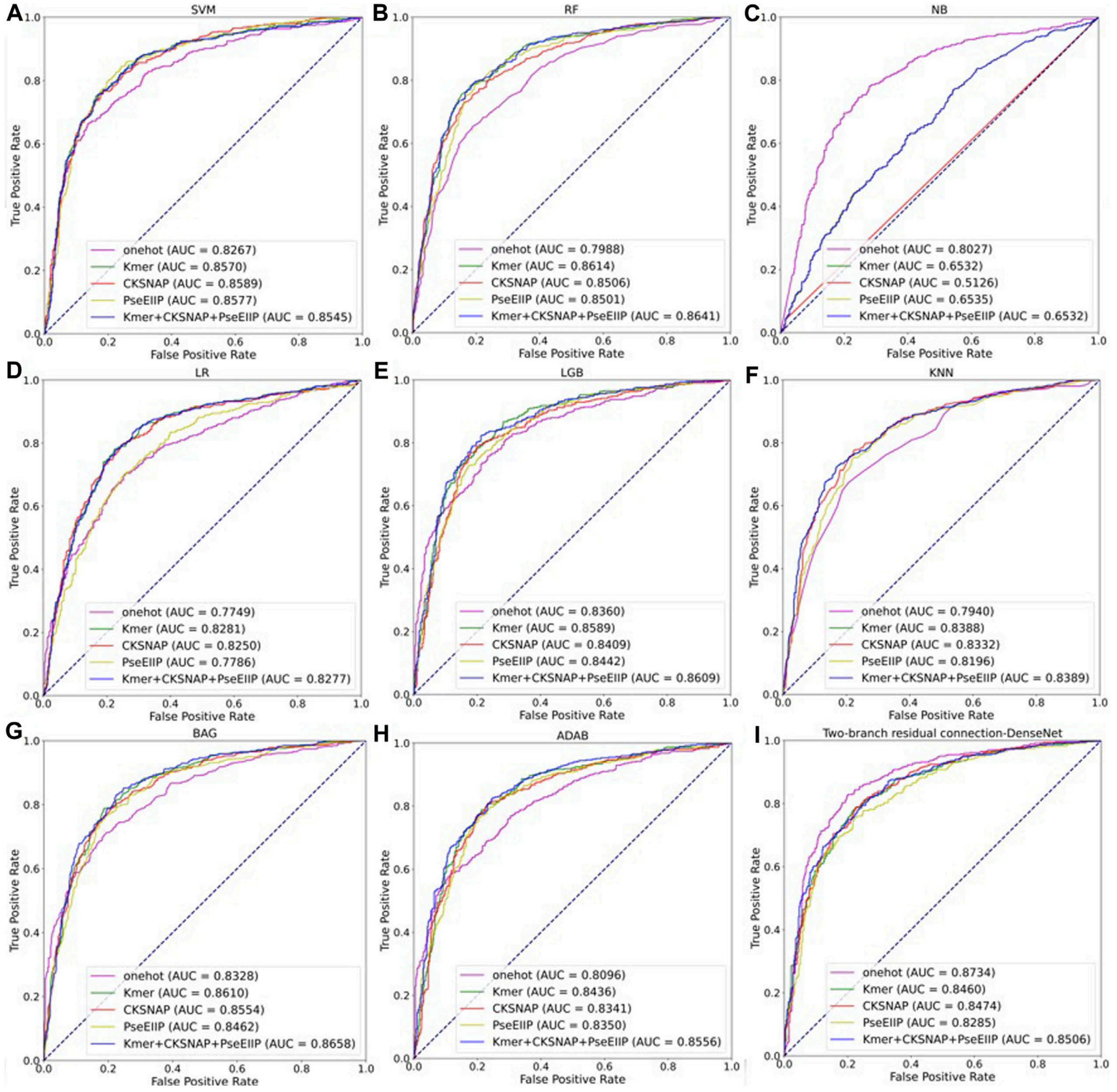

FIGURE 10
Multiple model ROC curves based on test sets with different encoding methods. **(A–H)** Results of machine learning. **(I)** Results of two-branch residual connection DenseNet.

encoding in each dataset is about 0.87. The detailed performance of the different encoding methods of EMDL-ac4C on ten independent test sets is shown in Supplementary Tables S1–S10. This indicates that two-branch residual connection DenseNet has good generalizability and does not show excessive differences depending on the dataset. In addition, it is also obvious from Figure 6 that the model with one-hot coding has significantly higher Acc, MCC and AUC values than the other coding methods. After comprehensive consideration, we concluded that one-hot encoding is more suitable for two-branch residual connection DenseNet model for predicting ac4C sites.

3.3 Comparison of ensemble and non-ensemble models

To illustrate the effectiveness of the two-branch residual connection DenseNet more intuitively, we further tested its ROC curves on ten training and test sets, as shown in Figure 7. Also, Figure 8 showed the loss changes on the training and validation sets. The model demonstrated excellent performance and balanced results on ten training sets and ten testing sets. The difference in indicators on each dataset done not exceed 2.52%, indicating that the two branch residual connection DenseNet can stably predict the ac4C sites.

TABLE 2 Average performance comparison of EMDL-ac4C and additional machine learning models on ten independent test sets.

Model	Sn	Sp	MCC	AUC	Auprc
SVM	0.7610	0.7668	0.5279	0.8267	0.8301
RF	0.7248	0.7602	0.4854	0.7988	0.6824
NB	0.8046	0.6634	0.4956	0.8027	0.7904
LR	0.6998	0.6893	0.3892	0.7749	0.7626
LGB	0.7030	0.8081	0.5141	0.8360	0.8538
KNN	0.7786	0.6294	0.5028	0.7940	0.7873
DT	0.6998	0.6893	0.3892	0.625040	0.5783
BAG	0.7717	0.740814	0.5427	0.8328	0.8557
ADAB	0.6833	0.7636	0.4486	0.8096	0.8453
EMDL-ac4C	**0.8104**	**0.8173**	**0.6169**	**0.8734**	**0.8643**

The best outcomes are in bold.

TABLE 3 Average performance of several advanced models on ten test sets.

Model	Sn	Sp	MCC	Acc	AUC
VGG-16	**0.9227**	0.2921	0.2533	0.6074	0.6850
ResNet	0.6673	0.6285	0.3008	0.6479	0.7086
CSPNet	0.8267	0.7349	0.5683	0.7808	0.8576
VGG-19	0.7537	0.6417	0.4304	0.6977	0.8231
Inception V3	0.7934	0.3488	0.1770	0.5711	0.6596
EMDL-ac4C	0.8104	**0.8173**	**0.6169**	**0.8080**	**0.8794**

The best outcomes are in bold.

In addition, to research whether downsampling ensemble has better prediction ability, we made a comparison between the downsampling ensemble model and single two-branch residual connection DenseNet. On ten independent test sets, we used the ensemble classifier and the non-ensemble classifier to predict ac4C sites based on one-hot coding, and the average results were shown in Figure 9. The ensemble model EMDL-ac4C outperforms the non-ensemble model in Sn, Acc, MCC and AUC by 2.13%, 0.42%, 0.74% and 0.53%, respectively. In contrast, Sp is 0.14% lower, and collectively, the performance of the ensemble model is better than that of the non-ensemble model.

3.4 Comparison with other machine learning models

To further evaluate the performance of the two-branch residual connection DenseNet model with that of other machine learning

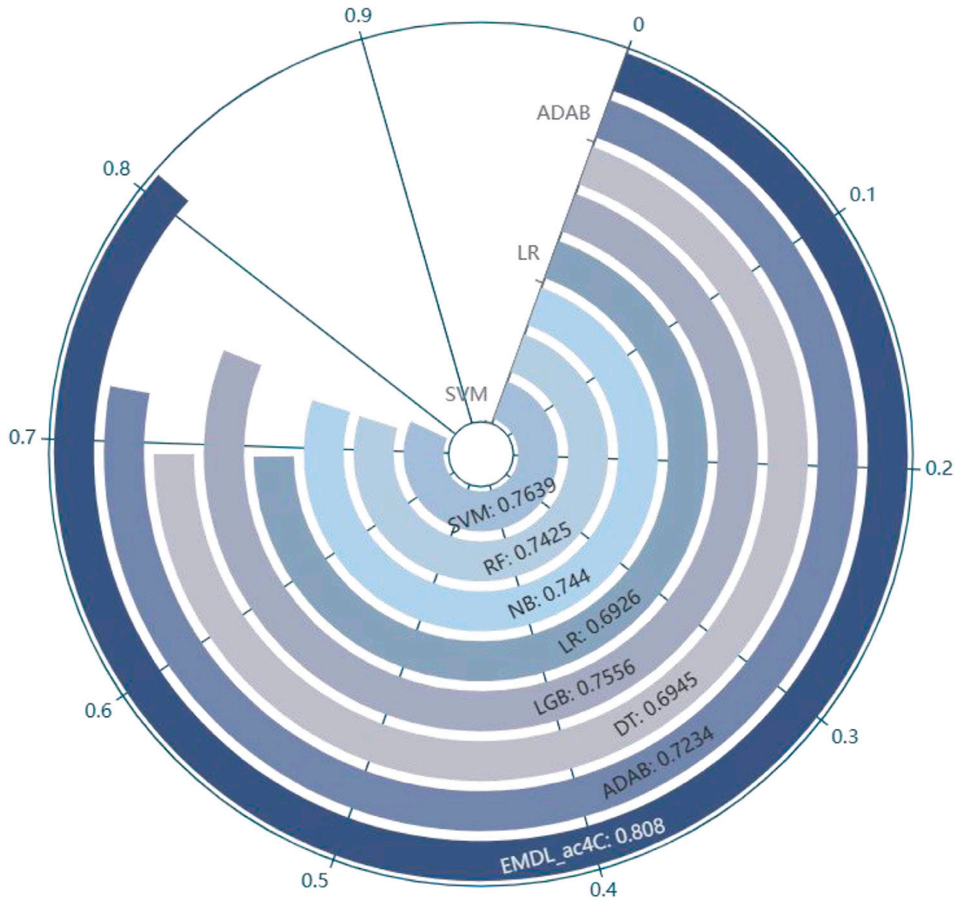

FIGURE 11
Comparison of Average Accuracy of EMDL-ac4C and additional machine learning models on ten independent test sets.

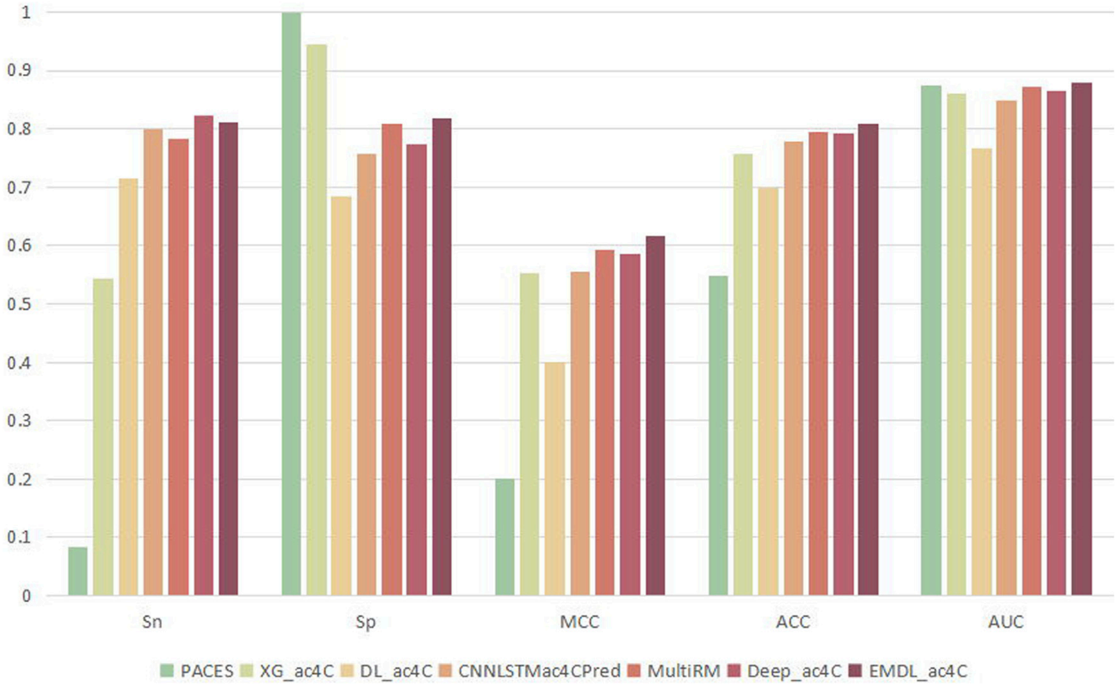

FIGURE 12
Comparison of the average results of ten test sets on various predictors.

models, we performed a comparison of the average results of various coding methods and distinct classifiers on ten test datasets. The ROC curves comparison are shown in Figure 10, where, A-H are the machine learning models: support vector machine (SVM), random forest (RF), naive Bayesian (NB), logistic regression (LR), light gradient boosting machine (LGB), k-nearest neighbor (KNN), bagging (BAG) and adaboost classifier (ADAB). I is the result of two-branch residual connection DenseNet. It is clear from Figure 10 that among the machine learning models, SVM, LGB, and BAG perform well and NB performs the worst. Of course, the combination of the two-branch residual connection DenseNet and one-hot coding has the best performance. Table 2 shows the average performance metrics of EMDL-ac4C and additional machine learning models using one-hot coding on ten independent test sets. All four metrics of EMDL-ac4C return higher values than the other traditional machine learning models.

In balanced datasets, accuracy (Acc) is one of the important metrics to evaluate the classifier performance. Figure 11 visualizes that EMDL-ac4C is more effective than other machine learning models.

3.5 Comparison with other advanced models

To assess the prediction performance of EMDL-ac4C, we compared EMDL-ac4C with several advanced models for analysis, including: Cross Stage Partial DenseNet (CSPNet) (Wang et al., 2020), VGG-16 (Guan et al., 2019), ResNet (He et al., 2016), VGG-19 (Xiao et al., 2020), Inception V3 (Yu et al., 2017). These models performed differently on the ten balanced test sets, as shown in Supplementary Tables S12–S17, while the average prediction results on the ten datasets are shown in Table 3. VGG-16 scored the highest in Sn, but the score of Sp was too low, and the balance between Sn and Sp was lost, with too much deviation, and the prediction accuracy was not high. In contrast, our model EMDL-ac4C reached a balance between Sn and Sp with less than 1% deviation, and obtained the highest Sp, MCC, Acc, and AUC scores among several models, getting the greatest prediction results.

To further evaluate the predictive performance of EMDL-ac4C, we collected three ac4C sites identified by Oxford Nanopore Technology (ONT) from the large public database DirectRMDB (Zhang. et al., 2022) as an additional test set, these three ac4C loci are located at positions 453630, 455959, and 456452 of chromosome NC_001144.5, respectively. After testing, all three ac4C loci were correctly predicted, and the probabilities of predicting positive samples were 0.7923848, 0.9826068, and 0.9666126, respectively. Therefore, we can consider EMDL-ac4C as a high-performance ac4C classifier.

3.6 Comparison of different classifiers

To prove the validity of EMDL-ac4C, we found six models that can be used to predict ac4C loci for comparison, including PACES (Zhao et al., 2019a), XG_ac4C (Alam et al., 2020b), DL_ac4C (Iqbal et al., 2022), CNNLSTMac4CPred(Zhang et al., 2022), MultiRM (Song et al., 2021) and DeepAc4C (Wang et al., 2021). Among these six predictors, PACES and XG_ac4C used machine learning approaches: random forest, XGboost. While DL_ac4C,

TABLE 4 The performance of DeepAc4C and EMDL-ac4C.

Dataset	Model	Training Acc	Validation Acc	Test Acc	Test MCC	Test AUC
D1	DeepAc4C	0.8093	0.8043	0.7934	0.5871	0.8620
	EMDL-ac4C	**0.8911**	**0.8092**	**0.8104**	**0.6217**	**0.8821**
D2	DeepAc4C	0.8195	0.8000	0.7943	0.5921	0.8660
	EMDL-ac4C	**0.8967**	**0.8075**	**0.8126**	**0.6267**	**0.8801**
D3	DeepAc4C	0.8209	0.8043	0.7874	0.5777	0.8641
	EMDL-ac4C	**0.8985**	**0.8097**	**0.8115**	**0.6244**	**0.8832**
D4	DeepAc4C	0.8490	**0.8348**	0.7969	0.5945	0.8658
	EMDL-ac4C	**0.8954**	0.8096	**0.8040**	**0.6102**	**0.8819**
D5	DeepAc4C	0.8403	0.8043	0.7950	0.5902	0.8646
	EMDL-ac4C	**0.8919**	**0.8063**	**0.8019**	**0.6041**	**0.8647**
D6	DeepAc4C	0.7996	0.7913	0.7938	0.5897	0.8645
	EMDL-ac4C	**0.8871**	**0.8076**	**0.7997**	**0.6000**	**0.8665**
D7	DeepAc4C	0.8209	**0.8609**	0.7911	0.5836	0.8671
	EMDL-ac4C	**0.8915**	0.8179	**0.7976**	**0.5956**	**0.8749**
D8	DeepAc4C	0.8475	0.8043	0.7978	0.5959	0.8657
	EMDL-ac4C	**0.8880**	**0.8056**	**0.8147**	**0.6306**	**0.8809**
D9	DeepAc4C	0.8078	0.8130	0.7846	0.5703	0.8615
	EMDL-ac4C	**0.8308**	**0.8177**	**0.8169**	**0.6340**	**0.8908**
D10	DeepAc4C	0.8277	**0.8217**	0.7850	0.5725	0.8680
	EMDL-ac4C	**0.8924**	0.8201	**0.8147**	**0.6295**	**0.8913**
Average	DeepAc4C	0.8242	**0.8139**	0.7919	0.5857	0.8649
	EMDL-ac4C	**0.8927**	0.8111	**0.8084**	**0.6177**	**0.8794**

The best outcomes are in bold.

TABLE 5 The AUC values and Aupr values of PACES, XG-ac4C and Two-branch residual connection DenseNet.

Dataset	Methods	AUC	Aupr
Cross-validation	PACES	0.885	0.559
	XG-ac4C	**0.910**	**0.653**
	Two-branch residual connection DenseNet	0.904	0.615
Independent-test	PACES	0.874	0.485
	XG-ac4C	0.889	0.581
	Two-branch residual connection DenseNet	**0.901**	**0.594**

The best outcomes are in bold.

CNNLSTMac4CPred, MultiRMand DeepAc4C used the deep learning approach. For a fair comparison, all seven predictors were tested using the same training and testing sets, and the predicted results were compared to determine their performance. Figure 12 had shown the average result of the seven predictors on ten test sets. Among them, EMDL-ac4C had the best comprehensive performance, followed by the deep learning model MultiRM, and then DeepAc4C, CNNLSTMac4CPred, XG_ac4C and DL_ac4C. PACES had the worst performance. Compared to MultiRM, EMDL-ac4C was 2.88%, 0.79%, 2.33%, 1.25% and 0.85% higher for Sn, Sp, MCC, ACC, and AUC, respectively. Meanwhile, compared to DeepAc4C, EMDL-ac4C was 4.39%, 3.12%, 1.61%, and 1.45% higher in Sp, MCC, Acc, and AUC, respectively. In addition, EMDL-ac4C was also higher than CNNLSTMac4CPred in all metrics, and the average is 3.93% higher. For Sp, PACES and XG_ac4C had higher return values than EMDL-ac4C. Nevertheless, in particular, the difference between the Sn and Sp values of PACES and XG_ac4C was too large, even reaching 91.52% for PACES, and too low Sn indicated that few positive samples were identified, which was not a good phenomenon. For the balanced dataset, we pay more attention to the return value of Acc, and the larger the Acc, the better the model performance. Our model EMDL-ac4C had higher Acc metrics than DL_ac4C, XG_ac4C and PACES: 10.89%, 5.08% and 26.08% higher, respectively. DeepAc4C is the most advanced model at present, so the detailed analysis of EMDL-ac4C and DeepAc4C is compared, see Table 4.

Some of the classification performance metrics of EMDL-ac4C and DeepAc4C on the training set, validation set, and test set were listed in Table 4 with comparative analysis of the two predictors. EMDL-ac4C

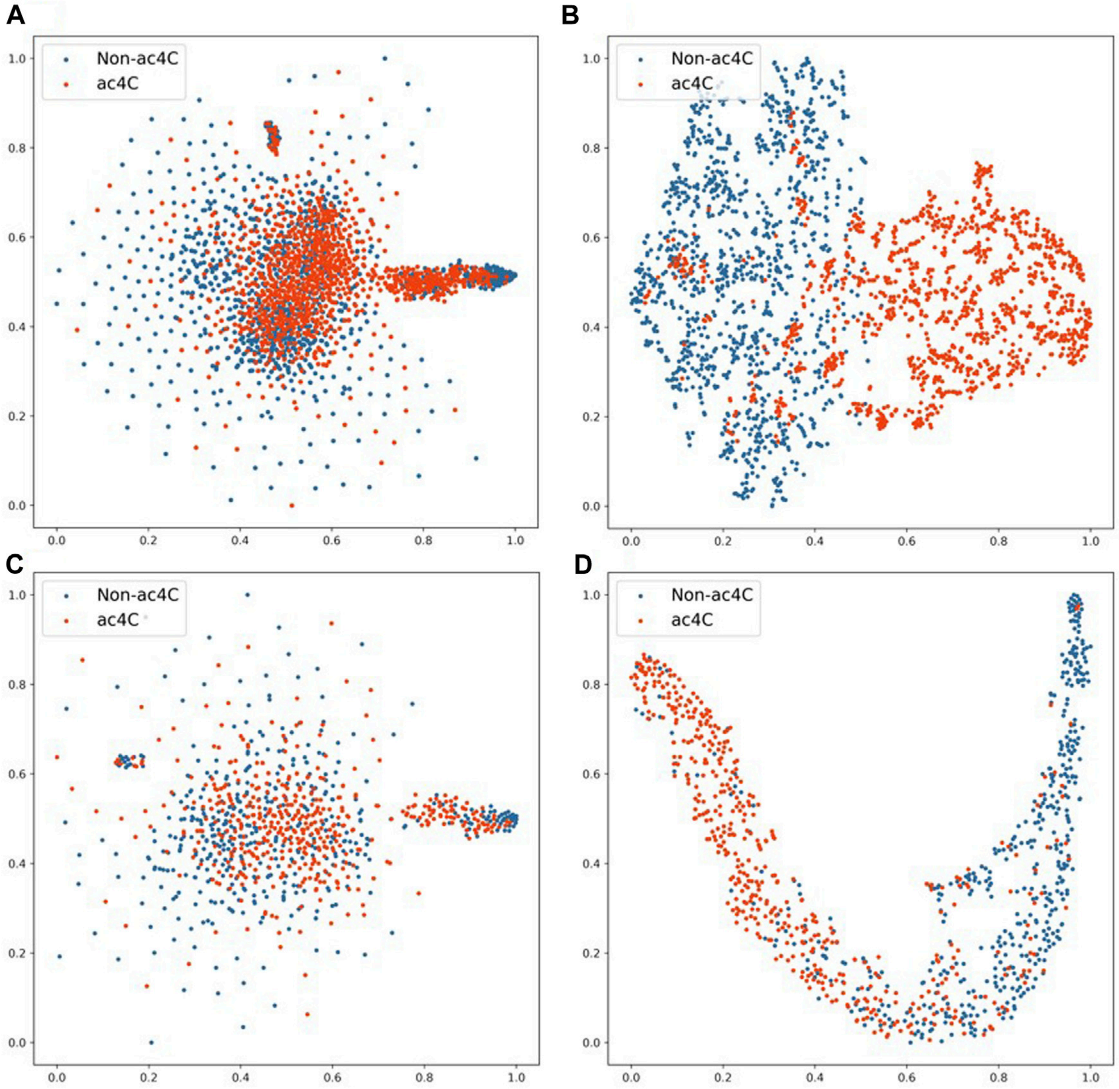

FIGURE 13
2D t-SNE visualization of the training and testing sets. **(A)** 2D t-SNE visualization of the training set with one-hot encoding. **(B)** 2D t-SNE visualization of features learned from the training set by EMDL-ac4C. **(C)** 2D t-SNE visualization of the testing set with one-hot encoding. **(D)** 2D t-SNE visualization of features learned from the testing set by EMDL-ac4C.

had a greater accuracy (Acc) than DeepAc4C in all 10 balanced datasets used for training. Among them, the best performance was on D6, which was 8.75% higher. On the validation dataset, EMDL-ac4C performed relatively poorly on the D4, D7 datasets, and it also performed slightly worse than DeepAc4C on D10. In this regard, we believe that EMDL-ac4C is unable to learn enough information and features from the data due to the small number of data in the validation set, which affects the performance of the model. For the datasets in this paper, the samples used for training in each balanced dataset were 2066, while the amounts of samples used for validation and testing were 230 and 934, correspondingly.

For the ten balanced test sets, EMDL-ac4C obtained good prediction results, with the Acc, MCC and AUC values of each test subset exceeding the corresponding metrics of DeepAc4C. The average Acc of the ten test sets of EMDL-ac4C was 1.61% higher than the average Acc of DeepAc4C, and the average MCC and average AUC values were 3.12% and 1.45% higher, respectively. This showed that the EMDL-ac4C model is effective and also has good generalization. In addition, it can also be seen from Table 4 that the difference of each metric corresponding to the ten balanced data sets is small, which proven that EMDL-ac4C is a stable and reliable model. By comparison, see Supplementary Tables S14, S15, the MCC values of VGG16 and VGG 19 models on different equilibrium datasets vary widely. VGG 16 had a MCC value of 0.99% on D7, and the MCC value on D1 reaches 52.75%, with a variance of 51.76%. The maximum difference in MCC values for VGG19 was also as high as 42.63%, which indicated that the VGG model is not stable on the data sets of this paper.

To evaluate the performance of our model on the unbalanced dataset, we use the dataset downloaded from the PACES website (http://www.rnanut.net/paces/) for testing. In this case, the training set contains 1,160 positive and 10,855 negative samples, respectively, while the test set contains 469 positive and 4,343 negative samples.

The results of the 5-fold cross-validation and independent tests of our base model Two-branch residual connection DenseNet and two other predictors PACES (Zhao et al., 2019b) and XG-ac4C (Alam et al., 2020a) are shown in Table 5.

On cross-validation, the AUC value of Two-branch residual connection DenseNet is slightly lower than XG-ac4C, but exceeds PACES, and the same is true for the Aupr value, which may be due to the fact that Two-branch residual connection DenseNet is only the base model, without unbalanced preprocessing of the dataset and without using the ensemble method. In the independent test, Two-branch residual connection DenseNet is better than the other two models, and the difference between the independent test and cross-validation results is small, not more than 0.021, and there is no overfitting, which just shows that Two-branch residual connection DenseNet has a strong generalization ability.

3.7 Visualization of the classification ability of EMDL-ac4C

To test the classification performance of EMDL-ac4C, we selected D1 in ten balanced datasets for validation. After encoding the sequence data with one-hot, We reduced the encoding vector to two-dimensional using t-distribution random neighbor embedding (t-SNE) (van der Maaten and Hinton, 2008) method, as shown in Figures 13A, C, where (A) and (C) were the classification effects of the training and testing sets after one-hot encoding using t-SNE downscaling, respectively. In order to compare with (A) (C) in Figure 13, we first extracted the important features from the training and testing data after one-hot encoding using EMDL-ac4C, and then downscaled them by t-SNE, and finally displayed the classification effects as shown in Figures 13B, D. The red circles in Figure 13 stand for the positive class samples, whereas the blue circles for the negative class samples. From Figure 13, we can clearly see that there are more overlapping clusters generated after one-hot coding, which indicates that the quality of one-hot coding is imperfect. In contrast, after EMDL-ac4C extracts the features, fewer overlapping clusters are generated, especially the two classes of clusters generated in the testing set have reached a highly disjoint classification effect, which proves the efficiency of EMDL-ac4C in terms of extracting features, that is, EMDL-ac4C has powerful classification ability.

4 Conclusion

In this work, we built a downsampling ensemble learning model called EMDL-ac4C, which aimed to predict ac4C sites from sequence fragments of RNA. To effectively identify the ac4C locus, we had done a lot of work at both sequence encoding and feature extraction levels. Firstly, we had compared five commonly used feature encoding schemes and found that the combination of simple one-hot encoding and deep learning models can identify ac4C loci more efficiently. Second, we proposed the ensemble learning model EMDL-ac4C to extract features and predict sites, whose underlying learner was two-branch residual connection DenseNet. Compared with other advanced models and predictors for identifying ac4C, EMDL-ac4C obtained superior performance in independent tests, which proved EMDL-ac4C's powerful feature learning capability and predictive power. We will develop the model and increase its prediction power in subsequent studies. For instance, we will be able to anticipate multi-class sites simultaneously, such as 6ma, 4mc, *etc.*

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding authors.

Author contributions

The studies were created and planned by JJ and ZW. ZW carried out the feature extraction, model building, deep learning, and performance assessment. The manuscript was written by ZW and revised by JJ and XC. This work was supervised by JJ and XC. All authors contributed to the article and approved the submitted version.

Funding

This work was partially supported by the National Natural Science Foundation of China (Nos. 61761023, 62162032, and 31760315), the Natural Science Foundation of Jiangxi Province, China (Nos. 20202BABL202004 and 20202BAB202007), the Scientific Research Plan of the Department of Education of Jiangxi Province, China (GJJ190695 and GJJ212419).

Acknowledgments

The authors are grateful for the constructive comments and suggestions made by the reviewers.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1232038/full#supplementary-material

References

Alam, W., Tayara, H., and Chong, K. T. (2020). XG-ac4C: Identification of N4-acetylcytidine (ac4C) in mRNA using eXtreme gradient boosting with electron-ion interaction pseudopotentials. *Sci. Rep.* 10, 20942. doi:10.1038/s41598-020-77824-2

Arango, D., Sturgill, D., Alhusaini, N., Dillman, A. A., Sweet, T. J., Hanson, G., et al. (2018). Acetylation of cytidine in mRNA promotes translation efficiency. *Cell.* 175, 1872–1886 e24. doi:10.1016/j.cell.2018.10.030

Basith, S., Manavalan, B., Shin, T. H., and Lee, G. (2019). SDM6A: A web-based integrative machine-learning framework for predicting 6mA sites in the rice genome. *Nucleic acids.* 18, 131–141. doi:10.1016/j.omtn.2019.08.011

Bastings, J., and Filippova, K. (2020). "The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? " in Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, November 2020, South Carolina, pp. 149–155.

Boccaletto, P., Machnicka, M. A., Purta, E., Piątkowski, P., Bagiński, B., Wirecki, T. K., et al. (2018). Modomics: A database of RNA modification pathways 2017 update. *Nucleic Acids Res.* 46, D303–D307. doi:10.1093/nar/gkx1030

Chen, K., Wei, Z., Zhang, Q., Wu, X., Rong, R., Lu, Z., et al. (2019). Whistle: A high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach. *Nucleic Acids Res.* 47, e41. doi:10.1093/nar/gkz074

Chen, L., Wang, W. J., Liu, Q., Wu, Y. K., Wu, Y. W., Jiang, Y., et al. (2022). NAT10-mediated N4-acetylcytidine modification is required for meiosis entry and progression in male germ cells. *Nucleic Acids Res.* 50, 10896–10913. doi:10.1093/nar/gkac594

Chen, W., Tang, H., and Lin, H. (2017). MethyRNA: A web server for identification of N6-methyladenosine sites. *J. Biomol. Struct. Dyn.* 35, 683–687. doi:10.1080/07391102.2016.1157761

Chen, Z., Zhao, P., Li, C., Li, F., Xiang, D., Chen, Y.-Z., et al. (2021). iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. *Nucleic Acids Res.* 49, e60. doi:10.1093/nar/gkab122

Dominissini, D., and Rechavi, G. (2018). N(4)-acetylation of cytidine in mRNA by NAT10 regulates stability and translation. *Cell.* 175, 1725–1727. doi:10.1016/j.cell.2018.11.037

Dou, L., Li, X., Ding, H., Xu, L., and Xiang, H. (2020). Prediction of m5C modifications in RNA sequences by combining multiple sequence features. *Nucleic acids.* 21, 332–342. doi:10.1016/j.omtn.2020.06.004

El Allali, A., Elhamraoui, Z., and Daoud, R. (2021). Machine learning applications in RNA modification sites prediction. *Comput. Struct. Biotechnol. J.* 19, 5510–5524. doi:10.1016/j.csbj.2021.09.025

Feng, Z., Li, K., Qin, K., Liang, J., Shi, M., Ma, Y., et al. (2022). The LINC00623/NAT10 signaling axis promotes pancreatic cancer progression by remodeling ac4C modification of mRNA. *J. Hematol. Oncol.* 15, 112. doi:10.1186/s13045-022-01338-9

Gamage, S. T., Sas-Chen, A., Schwartz, S., and Meier, J. L. (2021). Quantitative nucleotide resolution profiling of RNA cytidine acetylation by ac4C-seq. *Nat. Protoc.* 16, 2286–2307. doi:10.1038/s41596-021-00501-9

Guan, Q., Wang, Y., Ping, B., Li, D., Du, J., Qin, Y., et al. (2019). Deep convolutional neural network VGG-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: A pilot study. *J. Cancer* 10, 4876–4882. doi:10.7150/jca.28769

Hao, H., Liu, W., Miao, Y., Ma, L., Yu, B., Liu, L., et al. (2022). N4-acetylcytidine regulates the replication and pathogenicity of enterovirus 71. *Nucleic Acids Res.* 50, 9339–9354. doi:10.1093/nar/gkac675

Hasan, M. M., Basith, S., Khatun, M. S., Lee, G., Manavalan, B., and Kurata, H. (2021). Meta-i6mA: An interspecies predictor for identifying DNA N6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework. *Briefings Bioinforma.* 22, bbaa202. doi:10.1093/bib/bbaa202

Hasan, M. M., Tsukiyama, S., Cho, J. Y., Kurata, H., Alam, M. A., Liu, X., et al. (2022). Deepm5C: A deep-learning-based hybrid framework for identifying human RNA N5-methylcytosine sites using a stacking strategy. *Mol. Ther. J. Am. Soc. Gene Ther.* 30, 2856–2867. doi:10.1016/j.ymthe.2022.05.001

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition." in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, June 2016, pp. 770–778.

Hu, J., Shen, L., and Sun, G. (2018). "Squeeze-and-Excitation networks." in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, June 2018, pp. 7132–7141.

Huang, D., Chen, K., Song, B., Wei, Z., Su, J., Coenen, F., et al. (2022). Geographic encoding of transcripts enabled high-accuracy and isoform-aware deep learning of RNA methylation. *Nucleic Acids Res.* 50, 10290–10310. doi:10.1093/nar/gkac830

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (2017). "Densely connected convolutional networks." in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, June 2017, pp. 4700–4708.

Iqbal, M. S., Abbasi, R., Bin Heyat, M. B., Akhtar, F., Abdelgeliel, A. S., Albogami, S., et al. (2022). Recognition of mRNA N4 acetylcytidine (ac4C) by using non-deep vs. Deep learning. *Deep Learn. Appl. Sci.* 12, 1344. doi:10.3390/app12031344

Ito, S., Akamatsu, Y., Noma, A., Kimura, S., Miyauchi, K., Ikeuchi, Y., et al. (2014). A single acetylation of 18 S rRNA is essential for biogenesis of the small ribosomal subunit in *Saccharomyces cerevisiae*. *J. Biol. Chem.* 289, 26201–26212. doi:10.1074/jbc.M114.593996

Jia, J., Liu, Z., Xiao, X., Liu, B., and Chou, K. C. (2016). pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. *J. Theor. Biol.* 394, 223–230. doi:10.1016/j.jtbi.2016.01.020

Jia, J., Sun, M., , et al.Jia, J., Sun, M., Wu, G., Qiu, W. (2023). DeepDN_iGlu: Prediction of lysine glutarylation sites based on attention residual learning method and DenseNet. *Math. Biosci. Eng.* 20, 2815–2830. doi:10.3934/mbe.2023132

Jia, J., Wu, G., Li, M., and Qiu, W. (2022). pSuc-EDBAM: Predicting lysine succinylation sites in proteins based on ensemble dense blocks and an attention module. *BMC Bioinforma.* 23 (1), 450. doi:10.1186/s12859-022-05001-5

Jin, G., Xu, M., Zou, M., and Duan, S. (2020). The processing, gene regulation, biological functions, and clinical relevance of N4-acetylcytidine on RNA: A systematic review. *Mol. Ther. Nucleic Acids* 20, 13–24. doi:10.1016/j.omtn.2020.01.037

Kumbhar, B. V., Kamble, A. D., and Sonawane, K. D. (2013). Conformational preferences of modified nucleoside N(4)-acetylcytidine, ac4C occur at "wobble" 34th position in the anticodon loop of tRNA. *Cell. Biochem. Biophys.* 66, 797–816. doi:10.1007/s12013-013-9525-8

Le, N. Q. K., Ho, Q.-T., Nguyen, V.-N., and Chang, J.-S. (2022). BERT-Promoter: An improved sequence-based predictor of DNA promoter using BERT pre-trained model and SHAP feature selection. *Comput. Biol. Chem.* 99, 107732. doi:10.1016/j.compbiolchem.2022.107732

Li, Q., Xu, L., Li, Q., and Zhang, L. (2020). Identification and classification of enhancers using dimension reduction technique and recurrent neural network. *Comput. Math. Methods Med.* 2020, 8852258. doi:10.1155/2020/8852258

Liu, Q., Chen, J., Wang, Y., Li, S., Jia, C., Song, J., et al. (2020). DeepTorrent: A deep learning-based approach for predicting DNA N4-methylcytosine sites. *Briefings Bioinforma.* 22, bbaa124. doi:10.1093/bib/bbaa124

Long, Y., Jia, H., Zhong, Y., Jiang, Y., and Jia, Y. (2021). RXDNFuse: A aggregated residual dense network for infrared and visible image fusion. *Inf. Fusion* 69, 128–141. doi:10.1016/j.inffus.2020.11.009

Lv, H., Zhang, Z.-M., Li, S.-H., Tan, J.-X., Chen, W., and Lin, H. (2020). Evaluation of different computational methods on 5-methylcytosine sites identification. *Briefings Bioinforma.* 21, 982–995. doi:10.1093/bib/bbz048

Mezzar, S., de Schryver, E., and Van Veldhoven, P. P. (2014). RP-HPLC-fluorescence analysis of aliphatic aldehydes: Application to aldehyde-generating enzymes HACL1 and SGPL1. *J. Lipid Res.* 55, 573–582. doi:10.1194/jlr.D044230

Rehman, M. U., Tayara, H., and Chong, K. T. (2021). DCNN-4mC: Densely connected neural network based N4-methylcytosine site prediction in multiple species. *Comput. Struct. Biotechnol. J.* 19, 6009–6019. doi:10.1016/j.csbj.2021.10.034

Sharma, S., Langhendries, J. L., Watzinger, P., Kotter, P., Entian, K. D., and Lafontaine, D. L. (2015). Yeast Kre33 and human NAT10 are conserved 18S rRNA cytosine acetyltransferases that modify tRNAs assisted by the adaptor Tan1/THUMPD1. *Nucleic Acids Res.* 43, 2242–2258. doi:10.1093/nar/gkv075

Sharma, S., Marchand, V., Motorin, Y., and Lafontaine, D. L. J. (2017). Identification of sites of 2′-O-methylation vulnerability in human ribosomal RNAs by systematic mapping. *Sci. Rep.* 7, 11490. doi:10.1038/s41598-017-09734-9

Shi, Z., Wang, T., Huang, Z., Xie, F., and Song, G. (2021). A method for the automatic detection of myopia in Optos fundus images based on deep learning. *Int. J. Numer. Methods Biomed. Eng.* 37, e3460. doi:10.1002/cnm.3460

Song, B., Wang, X., Liang, Z., Ma, J., Huang, D., Wang, Y., et al. (2023). RMDisease V2.0: An updated database of genetic variants that affect RNA modifications with disease and trait implication. *Nucleic Acids Res.* 51, D1388–D1396. doi:10.1093/nar/gkac750

Song, Z., Huang, D., Song, B., Chen, K., Song, Y., Liu, G., et al. (2021). Attention-based multi-label neural networks for integrated prediction and interpretation of twelve widely occurring RNA modifications. *Nat. Commun.* 12, 4011. doi:10.1038/s41467-021-24313-3

Sturgill, D., Arango, D., and Oberdoerffer, S. (2022). Protocol for base resolution mapping of ac4C using RedaC:T-seq. *Star. Protoc.* 3, 101858. doi:10.1016/j.xpro.2022.101858

Tardu, M., Jones, J. D., Kennedy, R. T., Lin, Q., and Koutmou, K. S. (2019). Identification and quantification of modified nucleosides in *Saccharomyces cerevisiae* mRNAs. *ACS Chem. Biol.* 14, 1403–1409. doi:10.1021/acschembio.9b00369

Thomas, J. M., Bryson, K. M., and Meier, J. L. (2019). Nucleotide resolution sequencing of N4-acetylcytidine in RNA. *Methods Enzym.* 621, 31–51. doi:10.1016/bs.mie.2019.02.022

Tsai, K., Jaguva Vasudevan, A. A., Martinez Campos, C., Emery, A., Swanstrom, R., and Cullen, B. R. (2020). Acetylation of cytidine residues boosts HIV-1 gene expression by increasing viral RNA stability. *Cell. Host Microbe* 28, 306–312 e6. doi:10.1016/j.chom.2020.05.011

Tsukiyama, S., Hasan, M. M., Deng, H.-W., and Kurata, H. (2022). BERT6mA: Prediction of DNA N6-methyladenine site using deep learning-based approaches. *Briefings Bioinforma.* 23, bbac053. doi:10.1093/bib/bbac053

van der Maaten, L., and Hinton, G. (2008). Viualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems 30 (NIPS 2017). Long Beach, CA, December 2017.

Wang, C.-Y., Liao, H.-Y. M., Wu, Y.-H., Chen, P.-Y., Hsieh, J.-W., and Yeh, I.-H. (2020a). "CSPNet: A new backbone that can enhance learning capability of CNN." in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, June 2020, pp. 390–391.

Wang, C., Ju, Y., Zou, Q., and Lin, C. (2021a). DeepAc4C: A convolutional neural network model with hybrid features composed of physicochemical patterns and distributed representation information for identification of N4-acetylcytidine in mRNA. *Bioinformatics* 38, 52–57. doi:10.1093/bioinformatics/btab611

Wang, C., Zou, Q., Ju, Y., and Shi, H. (2023). Enhancer-FRL: Improved and robust identification of enhancers and their activities using feature representation learning. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 20, 967–975. doi:10.1109/TCBB.2022.3204365

Wang, H., Yan, Z., Liu, D., Zhao, H., and Zhao, J. (2020b). MDC-kace: A model for predicting lysine acetylation sites based on modular densely connected convolutional networks. *IEEE Access* 8, 214469–214480. doi:10.1109/ACCESS.2020.3041044

Wang, H., Zhao, H., Yan, Z., Zhao, J., and Han, J. (2021b). MDCAN-lys: A model for predicting succinylation sites based on multilane dense convolutional attention network. *Biomolecules* 11, 872. doi:10.3390/biom11060872

Wang, L., Li, Y., Peng, S., Tang, X., and Yin, B. (2021c). Multi-level feature fusion network for crowd counting. *IET Comput. Vis.* 15, 60–72. doi:10.1049/cvi2.12012

Wei, L., Luan, S., Nagai, L. A. E., Su, R., Zou, Q., and Hancock, J. (2019a). Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. *Bioinformatics* 35, 1326–1333. doi:10.1093/bioinformatics/bty824

Wei, Z., Song, H., Chen, L., Li, Q., and Han, G. (2019b). Attention-based DenseUnet network with adversarial training for skin lesion segmentation. *IEEE Access* 7, 136616–136629. doi:10.1109/ACCESS.2019.2940794

Weizhong, L., and Godzik, A. (2006). Cd-Hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22 (13), 1658–1659. doi:10.1093/bioinformatics/btl158

Xiao, J., Wang, J., Cao, S., and Li, B. (2020). Application of a novel and improved VGG-19 network in the detection of workers wearing masks. *J. Phys. Conf. Ser.* 1518, 012041. doi:10.1088/1742-6596/1518/1/012041

Yan, Y., Ni, B., Liu, J., and Yang, X. (2019). Multi-level attention model for person re-identification. *Pattern Recognit. Lett.* 127, 156–164. doi:10.1016/j.patrec.2018.08.024

Yang, S., Yang, Z., and Yang, J. (2023). 4mCBERT: A computing tool for the identification of dna N4-methylcytosine sites by sequence- and chemical-derived information based on ensemble learning strategies. *Int. J. Biol. Macromol.* 231, 123180. doi:10.1016/j.ijbiomac.2023.123180

Yang, W., Li, H. Y., Wu, Y. F., Mi, R. J., Liu, W. Z., Shen, X., et al. (2021). ac4C acetylation of RUNX2 catalyzed by NAT10 spurs osteogenesis of BMSCs and prevents ovariectomy-induced bone loss. *Mol. Ther. Nucleic Acids* 26, 135–147. doi:10.1016/j.omtn.2021.06.022

Yang, Y., Zhong, Z., Shen, T., and Lin, Z. (2018). "Convolutional neural networks with alternately updated clique." in Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, June 2018, pp. 2413–2422.

Yu, H., and Dai, Z. (2019). SNNRice6mA: A deep learning method for predicting dna N6-methyladenine sites in rice genome. *Front. Genet.* 10, 1071. doi:10.3389/fgene.2019.01071

Yu, W., Chang, J., Yang, C., Zhang, L., Shen, H., Xia, Y., et al. (2017). "Automatic classification of leukocytes using deep neural network." in Proceedings of the 2017 IEEE 12th International Conference on ASIC (ASICON), Guiyang, China, October 2017, pp. 1041–1044.

Zhang, G., Luo, W., Lyu, J., Yu, Z. G., and Huang, G. (2022a). CNNLSTMac4CPred: A hybrid model for N4-acetylcytidine prediction. *Interdiscip. Sci.* 14, 439–451. doi:10.1007/s12539-021-00500-0

Zhang, Y., Jiang, J., Ma, J., Wei, Z., Wang, Y., Song, B., et al. (2022b). DirectRMDB: A database of post-transcriptional RNA modifications unveiled from direct RNA sequencing technology. *Nucleic Acids Res.* 51, D106–D116. doi:10.1093/nar/gkac1061

Zhao, W., Zhou, Y., Cui, Q., and Zhou, Y. (2019). Paces: Prediction of N4-acetylcytidine (ac4C) modification sites in mRNA. *Sci. Rep.* 9, 11112. doi:10.1038/s41598-019-47594-7

Zhou, Q., Zhou, Z., Chen, C., Fan, G., Chen, G., Heng, H., et al. (2019). Grading of hepatocellular carcinoma using 3D SE-DenseNet in dynamic enhanced MR images. *Comput. Biol. Med.* 107, 47–57. doi:10.1016/j.compbiomed.2019.01.026

Zhou, Y., Zeng, P., Li, Y.-H., Zhang, Z., and Cui, Q. (2016). Sramp: Prediction of mammalian N ^{6} -methyladenosine (m ^{6} A) sites based on sequence-derived features. *Nucleic Acids Res.* 44, e91. doi:10.1093/nar/gkw104

Zou, Q., Xing, P., Wei, L., and Liu, B. (2019). Gene2vec: Gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA. *RNA (New York, N.Y.)* 25, 205–218. doi:10.1261/rna.069112.118