Mathematical Biosciences and Engineering

MBE, 20(6): 9759–9780.
DOI: 10.3934/mbe.2023428
Received: 15 January 2023
Revised: 09 March 2023
Accepted: 09 March 2023
Published: 24 March 2023

http://www.aimspress.com/journal/MBE

*Research article*

# DGA-5mC: A 5-methylcytosine site prediction model based on an improved DenseNet and bidirectional GRU method

**Jianhua Jia*, Lulu Qin* and Rufeng Lei**

School of Information Engineering, Jingdezhen Ceramic University, Jingdezhen 333403, China

* **Correspondence:** Email: jjh163yx@163.com, lulu9825@163.com.

**Abstract:** The 5-methylcytosine (5mC) in the promoter region plays a significant role in biological processes and diseases. A few high-throughput sequencing technologies and traditional machine learning algorithms are often used by researchers to detect 5mC modification sites. However, high-throughput identification is laborious, time-consuming and expensive; moreover, the machine learning algorithms are not so advanced. Therefore, there is an urgent need to develop a more efficient computational approach to replace those traditional methods. Since deep learning algorithms are more popular and have powerful computational advantages, we constructed a novel prediction model, called DGA-5mC, to identify 5mC modification sites in promoter regions by using a deep learning algorithm based on an improved densely connected convolutional network (DenseNet) and the bidirectional GRU approach. Furthermore, we added a self-attention module to evaluate the importance of various 5mC features. The deep learning-based DGA-5mC model algorithm automatically handles large proportions of unbalanced data for both positive and negative samples, highlighting the model's reliability and superiority. So far as the authors are aware, this is the first time that the combination of an improved DenseNet and bidirectional GRU methods has been used to predict the 5mC modification sites in promoter regions. It can be seen that the DGA-5mC model, after using a combination of one-hot coding, nucleotide chemical property coding and nucleotide density coding, performed well in terms of sensitivity, specificity, accuracy, the Matthews correlation coefficient (MCC), area under the curve and Gmean in the independent test dataset: 90.19%, 92.74%, 92.54%, 64.64%, 96.43% and 91.46%, respectively. In addition, all datasets and source codes for the DGA-5mC model are freely accessible at https://github.com/lulukoss/DGA-5mC.

**Keywords**: promoter; 5-methylcytosine site identification; DenseNet; BGRU; self-attention; deep learning; ensemble learning

## 1. Introduction

DNA methylation is a dynamic, reversible and heritable form of epigenetic modification that occurs primarily during primordial mammalian germ cell and early embryonic development, and it is widely studied in the context of bioinformatics and disease [1]. DNA methylation, which influences gene expression, genomic imprinting, epigenetic changes and other biological processes without changing the DNA sequence, is crucial for human development. It can modify DNA sequences by attaching to the CpG region of DNA through a highly dynamic and synergistic nuclease network, and it accordingly controls gene expression by changing the regulatory region's functional status without changing the genetic information carried by the DNA sequence [2]. One type of DNA methylation occurs when methyl bonds to cytosine at the cytosine guanine dinucleotide (CpG site). Currently, the three most prevalent forms of DNA methylation in living creatures are n4-methylcytosine (4mC), 5-methylcytosine (5mC) and n6-methyladenine (6mA), which are not regulated by the same mechanisms and functions within an individual [3]. DNA molecules carry a variety of modifications, in which the most prevalent DNA modification in prokaryotes is 6-methyladenine (6mA), whereas the most prevalent DNA modification in eukaryotes is 5-methylcytosine (5mC). The regulation of gene expression is greatly influenced by 5mC, a key mechanism of epigenetic modification that is also a hotly debated subject in the field of epigenetic modification [4,5].

Several disorders [6–10], including cancer, such as liver, lung, kidney, cervical, ovarian and breast cancers, can be brought on by abnormal DNA methylation, which results in dysregulation of gene expression. According to the numerous studies, it also plays a role in the onset of autoimmune rheumatic disorders such as rheumatoid arthritis and systemic lupus erythematosus [11]. Furthermore, cell differentiation, immune system control and the emergence of cancer are all closely related to the DNA methylation of promoters and enhancers [12]. For example, promoter methylation is closely correlated to the incidence of several disorders, including Alzheimer's disease [13], diabetes-related obesity [14,15], Parkinson's disease [16] and malignancies [17]. DNA methylation is crucial for physiological and pathological research, and studies have demonstrated that it can be exploited as a critical biomarker for the early detection and management of diseases [18]. As one of the reversible epigenetic markers in humans, our research on 5mC site identification has not stagnated. Therefore, whether the 5mC site in the DNA promoter can be precisely recognized is significant for promoter methylation in cancer and human genetic diseases.

Numerous academics have proposed computational methods to identify 5mC sites during the past decade. Some high-throughput sequencing techniques [19,20] were frequently used to detect 5mC modification sites in the past, but this method was either expensive or time-consuming. Therefore, finding effective and robust methods to identify 5mC modification sites is urgently needed. However, predicting 5mC sites often relies on conventional machine learning algorithms. For instance, Chai et al. [21] constructed the Staem5 machine learning model based on a stacking ensemble to identify 5mC modification sites. Liu et al. [22] designed a fresh approach for RNA 5mC site prediction based on XGBoost; they named it m5Cpred-XS. Chen et al. [23] proposed a predictive model for RNA 5mC sites by using the model m5CPred-SVM. Recently, it was developed into a commonly utilized technique for categorization learning prediction due to the growing popularity of neural networks in deep learning. Several algorithms-based deep learning has also been used to predict 5mC modification sites. For the case in point, Hasan et al. [24] proposed a hybrid framework for deep learning based on a stacking ensemble to identify 5mC sites. Shi et al. [25] constructed a model

R5hmCFDV based on deep voting and deep feature fusion to predict 5mC modification sites. These approaches are presented to promote the research of 5mC modification sites. More information on methods to predict 5mC modification sites and other modification sites of RNA can be referenced in the reviews [26–28].

Identifying 5mC modification sites in promoter regions is important to reveal DNA methylation modifications. We worked to improve the capability to obtain genome-wide promoter methylation sites in small cell lung cancer (SCLC) for this study because of lung cancer's high morbidity and mortality. Zhang et al. [29] presented the iPromoter-5mC model for the first time. From the Cancer Cell Line Encyclopedia (CCLE) database, they first created an SCLC promoter methylation dataset, and then they processed it with CD-HIT [30] software. The iPromoter-5mC model uses one-hot encoding to extract 5mC sequence characteristics and a straightforward deep neural network (DNN) to predict 5mC modification sites in the promoter region. Then, Nguyen et al. [31] constructed a machine learning-based predictor named 5mC-Pred, and their experimental results using the k-mers embedding feature encoding method on XGBoost outperformed the iPromoter-5mC model. Meanwhile, Qiu et al. [32] proposed a new SCLC promoter methylation dataset based on the model developed by Zhang et al. [29]. They proposed an m5C-HPromoter predictor based on stacking ensemble learning, and the classifier consists of a combination of machine learning algorithms: XGBoost, SVM, LightGBM and DNN algorithm. The experimental comparison with the iPromoter-5mC model was also performed on the same dataset used in our work and it achieved better performance. Subsequently, Cheng et al. [33] constructed a deep learning-based predictor, designated as BiLSTM-5mC, which was encoded with a combination of one-hot and nucleotide chemical properties (NCP). The best performance so far on the BiLSTM DNN was achieved, with the experimental results of both independent tests and five-fold cross-validation above the existing predictors.

With the development of deep learning, densely connected convolutional network (DenseNet) [34] algorithms are becoming more and more popular in the field of bioinformatics. In 2020, Wang et al. [34] used a DenseNet to identify lysine acetylation sites and achieved good experimental results. In 2022, Jia et al. [35,36] successfully implemented the prediction of lysine succinylation sites and lysine glutarylation sites using the DenseNet. In deep learning networks, the traditional convolutional neural network will lead to a gradient disappearance problem as the number of layers rises, while the residual neural network (ResNet) [37] can better solve the gradient disappearance and gradient descent problem than the convolutional neural network. However, the depth of the ResNet determines the number of parameters, and, as the numbers of layers and layer weights increase, the number of parameters will also increase. Furthermore, the performance of the DenseNet has recently been enhanced based on ResNet and subsequently compared with the case of ResNet. The densely connected uniqueness of DenseNet effectively solves the problem of increasing the number of parameters to some extent, and it alleviates the gradient disappearance of the neural network, so there is no issue with training a deeper network.

The term time-series data refers to the fact that sequence-type data are often temporally correlated. This means that the output of the network at a given moment is related to the input at the current moment, in addition to the output at a previous moment or moments. Recurrent neural networks [38] can handle time-series data; however, they suffer from gradient disappearance or gradient explosion during the learning process, making it difficult to establish dependencies between long distances in long sequences. So, in this study, we introduced a bidirectional gated recurrent unit (BGRU) [39] to capture the long-term dependencies between 5mC features. Due to its distinct advantages, the attention

mechanism [25] has recently been used in a wide range of fields. The core idea is to find the correlation between the original data so that we can ignore the rest of the non-dominant features and focus on the key features. Of course, the attention mechanism is also used by many researchers in the field of bioinformatics. For illustration, a multi-module deep learning system based on the attention mechanism, called DLF-Sul, was proposed by Ning and Li [40] to predict the S-sulfenylation sites in proteins. Zhang et al. [41] designed a classification model iLoc-miRNA to predict miRNA by fusing both attention mechanisms and a bidirectional long short-term memory network (Bi-LSTM). Thus, we also added the attention mechanisms to focus on the vital information between 5mC sequence features in this study.

In this study, we summarized the prediction methods to enable identification of 5mC modification sites on the same dataset and the current progress in 5mC modification site prediction. Although BiLSTM-5mC [33] has made considerable strides, there are still certain shortcomings to overcome. Therefore, we constructed a new deep learning model DGA-5mC to identify promoter 5mC modification sites. In this DGA-5mC model, we first used an improved DenseNet, which is the 8 × 41 feature matrix of the original DNA sequences after one-hot, NCP and nucleotide chemical density (ND) hybrid encoding processing input to the dense block to obtain high-level features. Then, we added a BGRU network after the improved DenseNet to obtain long-term dependencies between high-level features and introduced a self-attention module to evaluate the importance of features. Eventually, the fully connected layer receives these high-level features as input, and the softmax function is used to calculate a probability value between 0 and 1. To make the DGA-5mC model proposed in this work better, we also employed a homogeneous ensemble [42] under three identical web frameworks, and the three probability values were averaged to obtain the final prediction probabilities. If it is larger than 0.5, a 5mC modified site will be identified; otherwise, it is the opposite. The DGA-5mC network architecture is presented in Figure 1.

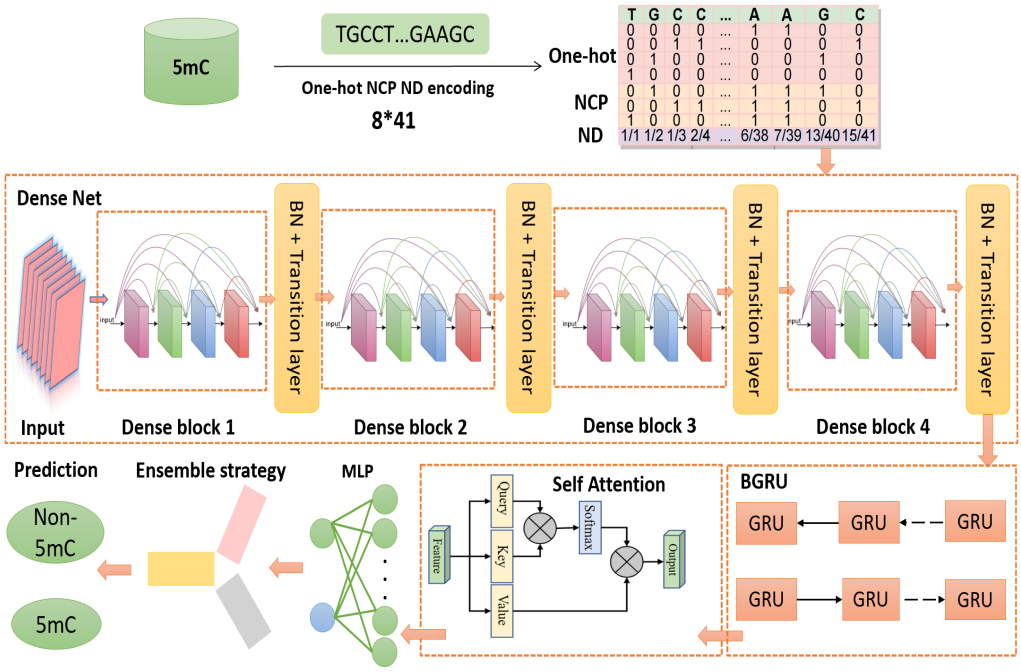


**Figure 1.** Overall flowchart of DGA-5mC.

## 2. Materials and methods

In this study, we developed a method based on deep learning to recognize 5-methylcytosine locations in genome-wide DNA promoters of SCLC. In what follows, we divide the work into three sections: the benchmark dataset, the feature extraction technique and the classification model.

### *2.1. Benchmark dataset*

The source of the benchmark data in this work was a previous study by Zhang et al [29], who used SCLC as a target to examine the distribution of 5mC modification sites in the promoter. They obtained nucleotide sequences with a length of 41 and cytosine at the center from SCLC data from the CCLE database [43] to improve confidence in the data. Subsequently, they used CD-HIT [30] software to eliminate DNA sequences with more than 80% similarity. The benchmark dataset ultimately obtained 893,326 promoter methylation sample sequences from the benchmark dataset, consisting of 69,750 positive samples and 823,576 negative samples. Here, promoter segments with 5mC sites were the positive samples, whereas promoter segments without 5mC sites were the negative samples. Although the ratio of our positive to negative samples was roughly 1:11, this imbalanced data can mirror the distribution of 5mC modification sites in promoters more objectively. The benchmark dataset is shown in Table 1.

**Table 1.** Details of the benchmark dataset.

| Original dataset | Positive sample | Negative sample |
|---|---|---|
| Total | 69,750 | 823,576 |
| Training dataset | 55,800 | 658,861 |
| Testing dataset | 13,950 | 164,715 |

### *2.2. Feature extraction methods*

We used three feature extraction methods in this work, namely, one-hot, ND and NCP encoding to identify 5mC modification sites in promoters, which will be presented in further detail in this section.

#### 2.2.1. One-hot encoding

One-hot encoding [33] is a simple and effective feature extraction method that has been widely used in the field of bioinformatics. It represents the four DNA bases of adenine (A), cytosine (C), guanine (G) and thymine (T) on the nucleotide chain of a DNA molecule as a binary vector consisting of 0 and 1. Specifically, it means that the nucleotides A, C, G and T can be represented by four vectors (1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0) and (0, 0, 0, 1), respectively. For this work, the length of the 5mC site sequence in the promoter was 41bp, so each sequence was transformed into a $4 \times 41$ feature matrix after encoding with this method. The encoding process is shown in Figure 2.

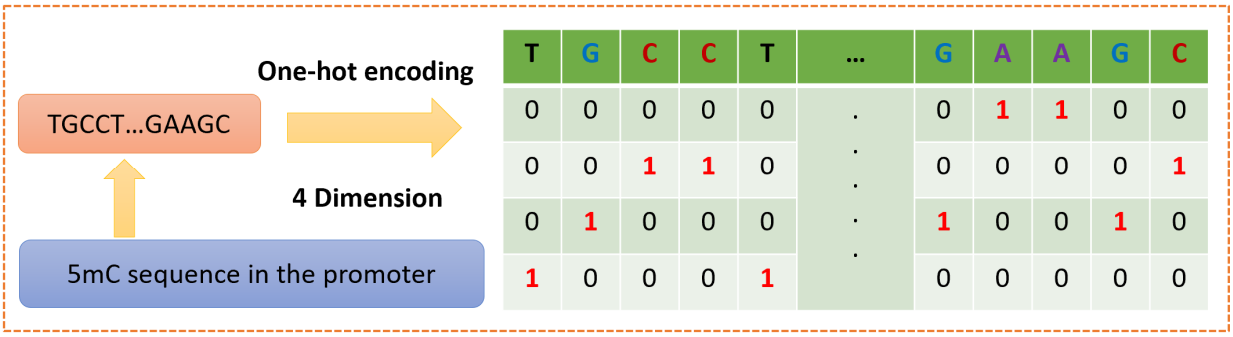


**Figure 2.** One-hot encoding.

### 2.2.2. NCP encoding

Recently, the NCP [44] encoding method has been applied in many studies in bioinformatics. This encoding method is based on three chemical properties, and it is a relatively simple encoding scheme. Different nucleotides have different chemical properties, and their detailed properties are listed below.

1) From the perspective of the functional groups contained in the nucleotides, A and C both contain amino groups, and G and T both contain ketone groups.

2) In terms of the ring structure, A and G contain two ring structures, while G and C have only one ring structure.

3) From the perspective of base complementary pairing, A and T are linked by two hydrogen bonds when paired, while G and C are linked by three hydrogen bonds when paired.

**Table 2.** NCPs.

| Chemical properties | Classification | Nucleotide | Encoding form |
|---|---|---|---|
| Ring structure | Purine (two) | A, G | 1 |
| | Pyrimidines (three) | C, T | 0 |
| Functional group | Amine | A, C | 1 |
| | Ketone group | G, T | 0 |
| Hydrogen bonding | High stability | C, G | 0 |
| | Weak stability | A, T | 1 |

For the 5mC site sequence sample of the promoter in this paper, each nucleotide can be represented as a three-dimensional vector according to the NCP encoding form. Thus, the four nucleotides A, C, G and T are represented by (1, 1, 1), (0, 1, 0), (1, 0, 0) and (0, 0, 1), respectively. Table 2 gives the specific chemical properties and encoding representation among the nucleotides.

### 2.2.3. ND encoding

The ND [42] encoding method is also one of the DNA sequence encoding methods, and it is often used in combination with other encoding methods. The main principle is to take one or several bases in a DNA sequence sample as an element and calculate the frequency of this element occurring in the

sample where it is located.

Assume that the DNA sequence samples are composed of $l$ nucleotides, where $R_i$ is one of the four nucleotides. Then, the DNA sequence samples can be expressed in the form of Eq (1).

$$Y = R_1R_2R_3R_4 \ldots R_i \ldots R_l \tag{1}$$

Take the calculation of single ND as an example, where $P_m$ is the density of the occurrence of nucleotide $R_i$ at position $i$ in the DNA sequence sample. The calculation method is shown in Eq (2).

$$P_m = \frac{\sum_{i=1}^{m} f(R_i)}{m} \tag{2}$$

where $f(R_i)$ is calculated as shown in Eq (3), and $R_m$ represents the $m$th nucleotide.

$$f(R_i) = \begin{cases} 1, & R_i = R_m \\ 0, & \text{others} \end{cases} \tag{3}$$

Each nucleotide can be represented as a one-dimensional vector using the ND encoding method. We take a 41bp long sequence "TGCCT...GAAGC" in promoter 5mC as an example: "A" at positions 18, 19, ..., 38 and 39 with densities of 1/18, 2/19, ..., 6/38 and 7/39, respectively; "C" at positions 3, 4, ..., 36 and 41 with densities of 1/3, 2/4...14/36, ..., and 15/41, respectively; "G" at positions 2, 6, ..., 37 and 40 with densities of 1/2, 2/6, ..., 12/37, ..., and 13/40, respectively; and "T" at positions 1, 5, ..., 30 and 32 with the densities of 1/1, 2/5, ..., 5/30 and 6/32, respectively. Eventually, this sequence can be represented as a 41-dimensional vector.

We combined NCP encoding and ND encoding, in which each nucleotide can then be represented by a four-dimensional vector. Therefore, a promoter 5mC sequence can be represented as a 4 × 41 feature matrix, as illustrated in Figure 3.

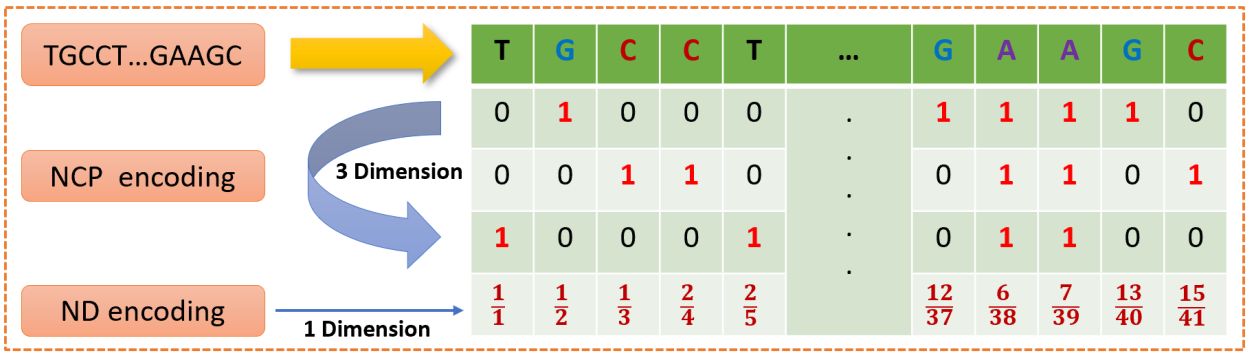


**Figure 3.** NCP encoding and ND encoding.

### *2.3. Classification model*

#### 2.3.1. DenseNet

The DenseNet [34] is a neural network framework based on the ResNet, which is composed of

three layers: the convolutional layer, the dense block layer and the transition layer. The original features are first convolved with the convolutional layer and then combined with several densely connected dense blocks and transition layers to obtain the high-level features of the sequence. The specific network structure is represented in Figure 4.

To ensure maximum information transfer to and from layers, the most unique aspect of DenseNet is the proposed dense connection mechanism, which ensures that all layers are interconnected. Specifically, each layer is used as input to the subsequent layer, and it is to be connected to all of the previous layers in the channel dimension. This implies that each layer is connected to all previous layers in the channel dimension and serves as input to the next layer. There are $(L\ (L+1))/2$ connections total for an $L$-layer DenseNet. It can be seen that this is a more dense way of connection. Moreover, DenseNet directly concatenates feature maps of different layers, which can realize feature reuse and improve efficiency.

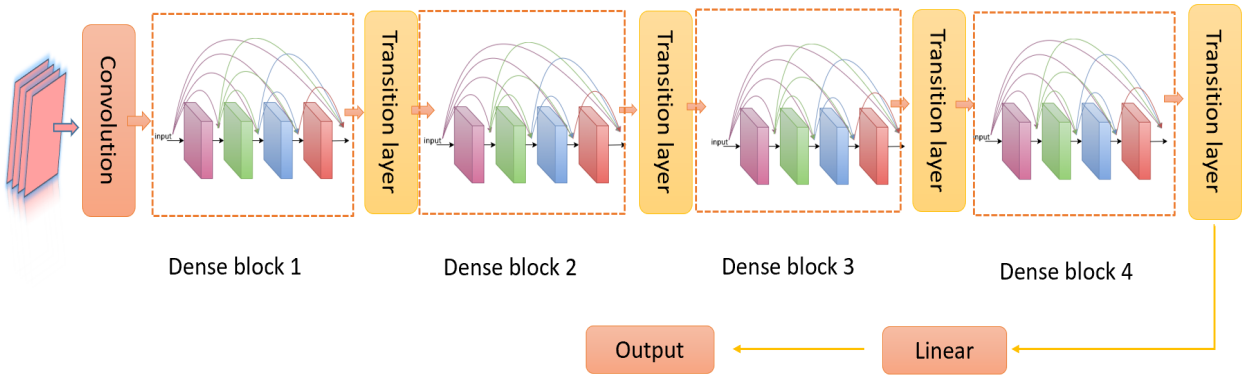


**Figure 4.** Original structure of DenseNet.

A dense block consists of an $L$-layer network structure with a nonlinear transform function, and the specific network structure is represented in Figure 5. The nonlinear transformation function consists of a normalization function batch normalization (BN), rectified linear unit (ReLU) and a 3 × 3 convolution kernel. The $L$th layer of DenseNet will have $L$ inputs, which means that the $L$th layer receives all feature map outputs from the previous $L$–1 layers. Its output is calculated as

$$x_L = H_L([x_0, x_1, \ldots, x_{L-1}]) \tag{4}$$

where $[x_0, x_1, \ldots, x_{L-1}]$ denotes the feature maps from layer 0 to layer $L$–1, which are concatenated, $L$ denotes the layer, $x_L$ denotes the output of layer $L$ and $H_L$ denotes a nonlinear transformation.

The transition layer mainly connects two adjacent dense blocks and reduces the feature map size. The transition layer consists of a 1 × 1 convolution and 2 × 2 AvgPooling with the structure of BN+ReLU+1 × 1 Conv + 2 × 2 AvgPooling, which can lead to the features' dimensional reduction and show the result of compressing the model.

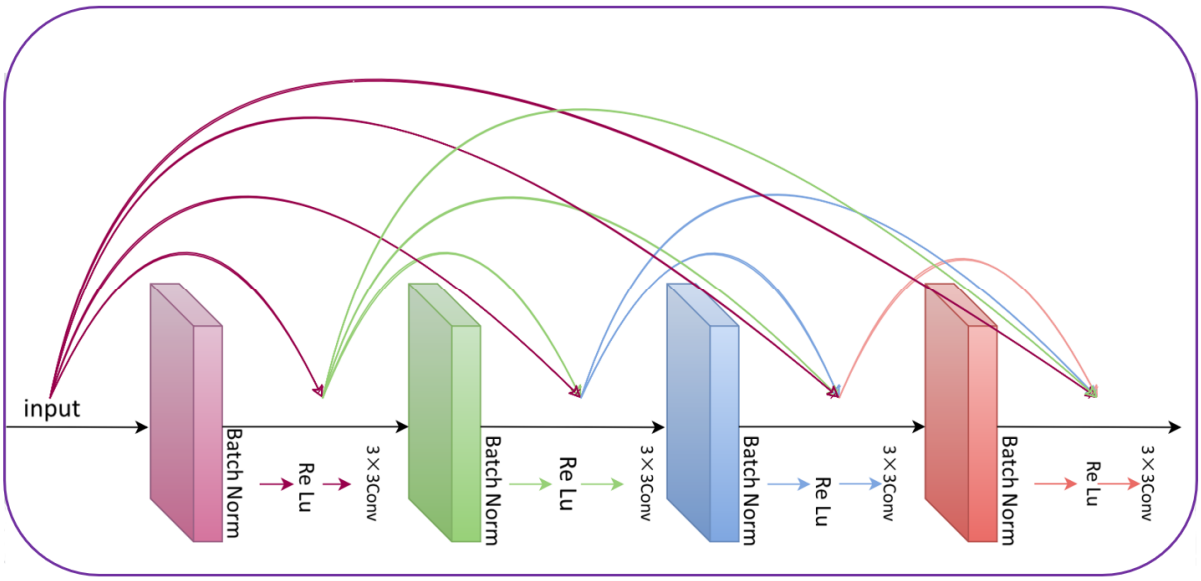


**Figure 5.** Structure of a dense block.

In this study, we have modified the original network framework of DenseNet. In detail, we removed the first convolutional layer and inputted the original features of the one-hot encoded promoter 5mC sequence directly to the dense block. Furthermore, we added a BN layer between the dense block and the transition layer, which extracted the original feature information at a deeper level and improved the generalization ability of the model. The BN is given as

$$\widetilde{x_i} = \frac{x_i - \mu}{\sqrt{\sigma^2 - \epsilon}} \tag{5}$$

$$y_i = \gamma \widetilde{x_i} + \beta \tag{6}$$

where $\mu$ is the mean of the feature dataset; $\sigma^2$ is the variance of the feature dataset; $\gamma$ and $\beta$ are the trainable parameters.

By repeating the experiment, we adjusted the network parameters and selected the four-layer dense block that can obtain the optimal prediction results. The improved DenseNet structure for this work was shown in Figure 1.

### 2.3.2. BGRU

To obtain the long-term dependence between the 5mC features, we added two layers of BGRUs [45,46] after the DenseNet to extract deeper features. The network structure is shown in Figure 6. The feature maps generated after the DenseNet are fed into a two-layer BGRU, which has 500 neurons per layer, to extract higher-level features.

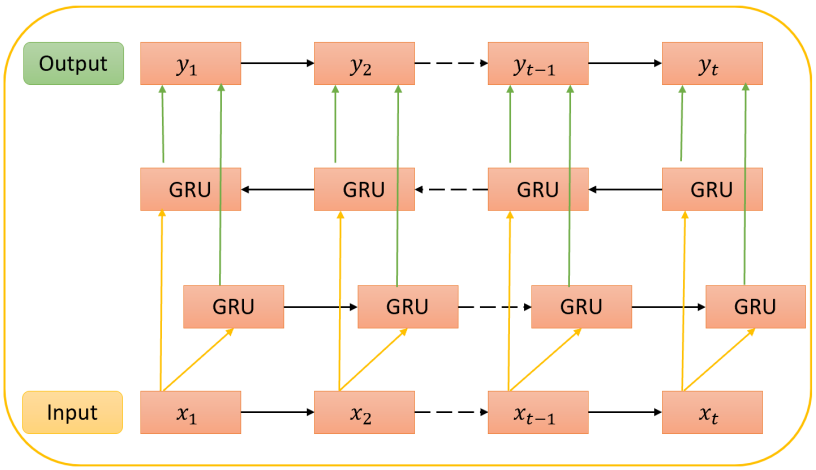


**Figure 6.** Structure of BGRU.

The BGRU consists of two GRUs, including a forward GRU model that accepts forward inputs and a reverse GRU model that learns reverse inputs. The BGRU performance is similar to that of a Bi-LSTM [33], which is essentially a Bi-LSTM without output gates, but with fewer parameters and lower computational complexity. The network structure of a BGRU is relatively simple, as it consists of only update gates and reset gates. Figure 7 represents a standard GRU architecture. The update gate indicates the state of a cell at a certain time, and a larger value indicates that more information about the previous state is remembered. The reset gate is applied to regulate the degree of forgetting the state information of the previous moment, and a smaller value means that more is forgotten. The GRU can be calculated as

$$\begin{aligned} r_t &= \sigma(W_t x + U_t h_{t-1}) \\ z_t &= \sigma(W_z x_t + U_z h_{t-1}) \\ \widetilde{h_t} &= tanh(W x_t + U(r_t \odot h_{t-1})) \\ h_t &= (1 - Z_t) \odot h_{t-1} + Z_t \odot \widetilde{h_t} \end{aligned} \tag{7}$$

where $\sigma$ is the sigmoid function, W and U are the weight matrices, $\odot$ denotes the element multiplication, $h_{t-1}$ denotes the hidden state at the previous moment, $h_t$ denotes the hidden state at the current moment, $x_t$ denotes the input sequence information, $r_t$ denotes the reset gate and $z_t$ denotes the update gate.

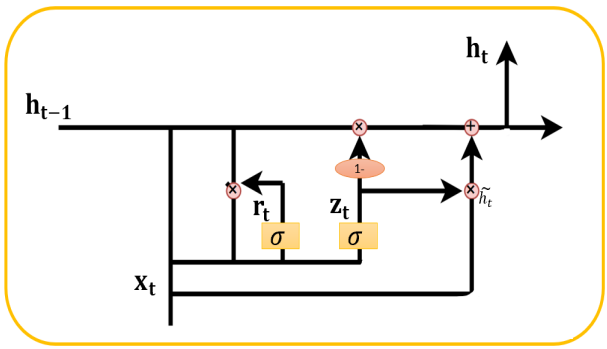


**Figure 7.** Inner structure of a GRU.

### 2.3.3. Self-attention

After processing through the BGRU, we introduced another self-attention module [25] to learn the importance of the features, as shown in Figure 8. We have taken the 5mC high-level feature output after the self-attention module and then input them into a two-layer fully connected layer. The first layer consists of 240 neurons with a dropout mechanism with a 50% random deletion rate, and the second layer consists of 40 neurons with a random deletion rate of 20% for the dropout mechanism. We chose softmax [25] as the activation function for the DGA-5mC model to obtain the predicted probability of the 5mC sites in the promoter.

The self-attention mechanism module converts the input data into three vectors: $q_i$, $k_i$ and $v_i$. The output vector is a weighted sum of each value vector, and it is obtained by querying the correlation of the vector with the corresponding vector to calculate the weight of each value vector. The calculation method is shown as

$$\begin{aligned} q_i &= W_q b_i \\ k_i &= W_k b_i \\ v_i &= W_v b_i \\ w_{ti} &= \frac{\exp\left(similarity(h_i,h_j)\right)}{\sum_{i=1}^{t}\exp\left(similarity(h_i,h_j)\right)} \end{aligned} \tag{8}$$

where $W_q$, $W_k$ and $W_v$ are the parameter matrices; $q_i$, $k_i$ and $v_i$ stand for the query, key and value vectors, respectively; $w_{ti}$ is the weight assignment to the input vectors.

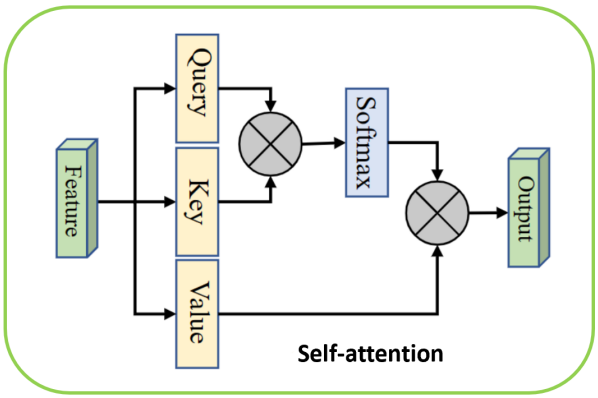


**Figure 8.** Structure of self-attention.

### 2.3.4. Ensemble learning

In machine learning, we input independent test datasets to several of the same or different models, and then calculate several predictions and average them. This ensemble learning method is known as model averaging. The advantage of model averaging is that different models do not usually produce identical errors on the independent test dataset, and it is a very powerful method for reducing generalization errors. In this study, the homogeneous ensemble algorithm refers to the use of the same feature extraction and model framing methods for the same training dataset. It makes use of the very

idea of model averaging introduced above. In this study, we used five-fold cross-validation. The training dataset was divided into five parts, four of which were used for training and one for validation. For the training dataset, we put the validation set into three models in each fold, through which three predictions were obtained. And, the three predictions were averaged to get the validation results for each fold. For the independent test dataset, the same method was used as that for the training dataset. The exact structure of the ensemble learning algorithm is shown in Figure 9. It is worth noting here that Models 1, 2 and 3 in Figure 9 are the same model framework, which is all network frameworks in the DGA-5mC model without the homogeneous ensemble learning. In detail, the promoter 5mC sequence is processed by one-hot, NCP, and ND hybrid coding to obtain an 8 × 41 feature matrix. We first inputted this 8 × 41 matrix into the improved DenseNet to obtain the high-level features. Next, we added a BGRU network to obtain the long-term dependencies between high-level features. Subsequently, a self-attention module was introduced to evaluate the importance of the features. Finally, the high-level features were fed into the fully connected layer and a probability value between 0 and 1 was derived using softmax.

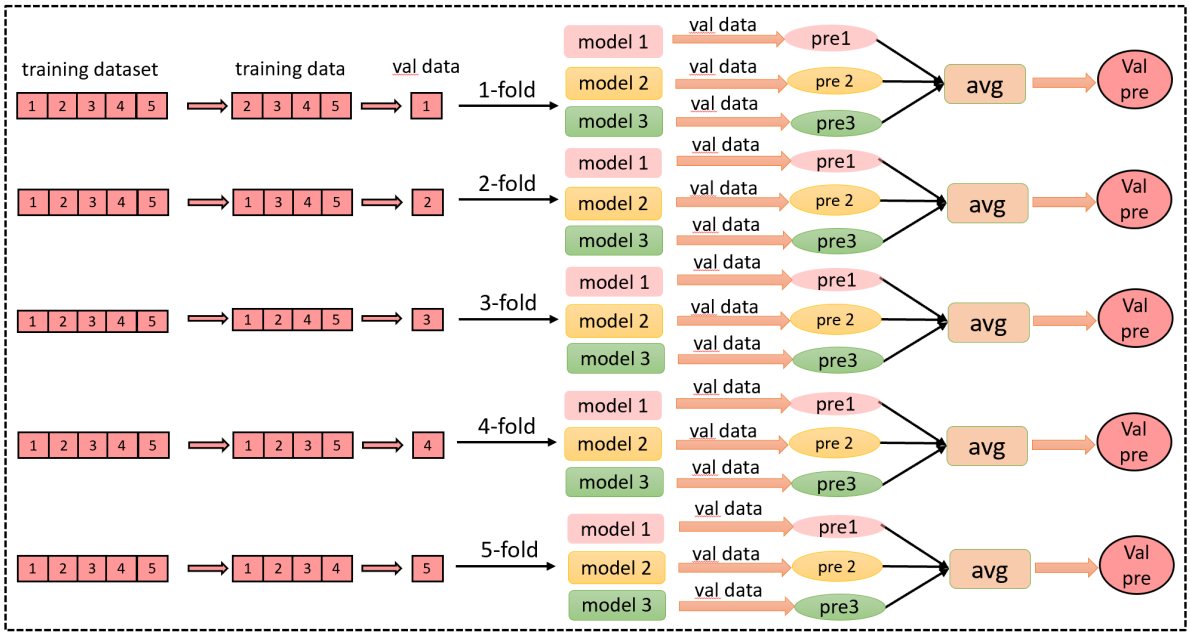


**Figure 9.** Ensemble learning figure of 5-fold cross-validation.

### 2.3.5. Performance evaluation

In this study, we selected five commonly used classifier evaluation metrics to evaluate the prediction performance of the DGA-5mC model, including the sensitivity (Sn), specificity (SP), Gmean, accuracy (ACC) and Matthews correlation coefficient (MCC); the specific definitions are respectively given as

$$
\begin{aligned}
Sn &= \frac{TP}{TP+FN} \\
SP &= \frac{TN}{TN+FP} \\
Gmean &= \sqrt{Sn \times SP} \\
ACC &= \frac{TP+TN}{TP+TN+FP+FN} \\
MCC &= \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN) \times (TN+FN) \times (TP+FP) \times (TN+FP)}}
\end{aligned} \tag{9}
$$

where TP refers to true positives, computing the number of positive samples that are truly predicted as positive. FP refers to false positives, i.e., the number of negative samples that are incorrectly classified as positive. TN indicates true negatives, corresponding to the number of negative samples that are correctly classified. FN denotes false negatives, or the number of positive samples that are incorrectly classified as negative. Sn and SP respectively represent the proportion of positive and negative samples predicted correctly, and G-mean is a composite of positive sample accuracy and negative sample accuracy. ACC represents the proportion of the whole sample predicted correctly, and MCC can accurately evaluate the performance of the model.

In addition, we have also introduced the receiver operating characteristic (ROC) curve [47] and calculated the area under the ROC curve (AUC) to evaluate the overall performance of the predictor. The value of AUC is in the range of [0, 1], and its value is positively correlated with the prediction performance, where the larger the value of AUC, the better the overall performance of the predictor.

## 3. Results and discussion

In this study, we first evaluated the 5mC sequence feature extraction encoding method and conducted ablation experiments on the network architecture of the DGA-5mC deep learning model. Then, we evaluated the performance of the DGA-5mC predictor and obtained the experimental results for five-fold cross-validation and independent testing. Finally, we also performed a comparison with existing predictors, which showed superior performance in the Sn, MCC and AUC metrics.

### *3.1. Contrasting various feature extraction techniques*

In the DGA-5mC network framework proposed in this paper, we compared the performance of three different feature encoding methods, including one-hot encoding, NCP and ND encoding (NPF+ND) and their hybrid encoding (one-hot+NPF+ND). As described in Section 2.2 introduced earlier, One-hot encoding and NPF+ND encoding can encode the nucleotide sequences of 5mC or non-5mC sites into a matrix of size 4 × 41, respectively. One-hot+NPF+ND hybrid encoding can encode the nucleotide sequences of 5mC or non-5mC sites into a matrix of size 8 × 41.

We inputted the feature matrices generated by these three encoding methods into the DGA-5mC network framework respectively; the experimental results on the training dataset and the independent test dataset are shown in Tables 3 and 4. It was easy to see that one-hot+NPF+ND hybrid coding was the best in terms of five performance evaluation metrics. Therefore, we adopted one-hot+NPF+ND hybrid encoding as the final encoding method in this study.

**Table 3.** Feature encoding methods based on 5-fold cross-validation on training dataset.

| Encoding | Sn | SP | ACC | MCC | AUC | Gmean |
|---|---|---|---|---|---|---|
| One-hot | 0.9027 | 0.9245 | 0.9228 | 0.6411 | 0.9634 | 0.9135 |
| NCP+ND | 0.8990 | **0.9265** | **0.9244** | 0.6407 | 0.9626 | 0.9126 |
| One-hot+NCP+ND | **0.9044** | 0.9250 | 0.9234 | **0.6417** | **0.9639** | **0.9146** |

**Table 4.** Feature encoding methods based on independent test dataset.

| Encoding | Sn | SP | ACC | MCC | AUC | Gmean |
|---|---|---|---|---|---|---|
| One-hot | 0.9005 | 0.9255 | 0.9235 | 0.6407 | 0.9638 | 0.9129 |
| NCP+ND | 0.8991 | 0.9264 | 0.9243 | 0.6421 | 0.9636 | 0.9126 |
| One-hot+NCP+ND | **0.9019** | **0.9274** | **0.9254** | **0.6464** | **0.9644** | **0.9146** |

### *3.2. Ablation experiment for model architecture*

We conducted ablation experiments on the network framework to determine which combination of the four methods was most suitable as the network framework for the model. The results of the ablation experiments based on the five-fold cross-validation of the training dataset are presented in Table 5. The results of the ablation experiments based on the independent test dataset are shown in Table 6. In Tables 5 and 6, the results of seven experimental combinations are shown, and the results of the best combination are indicated in bold. If there is a marker "√" in the corresponding row of each network method, it means that the method was selected for this experiment; if not, it means that the method was not selected.

This shows that the network framework with a combination of the four methods works best. This network framework extracted more advanced features compared to other network frameworks, and it was the best in terms of the MCC metric. Therefore, we finally chose a combination of four methods as our network framework model for DGA-5mC.

**Table 5.** Ablation experiments based on 5-fold cross-validation on training dataset.

| DenseNet | √ | √ | | √ | √ | √ | √ |
|---|---|---|---|---|---|---|---|
| BGRU | | √ | √ | | √ | √ | √ |
| Self-Attention | | | √ | √ | √ | | √ |
| Ensemble | | | | | | √ | √ |
| Sn | 0.9515 | 0.8985 | **0.9558** | 0.9504 | 0.8899 | 0.9108 | 0.9044 |
| SP | 0.8996 | 0.9228 | 0.8680 | 0.8595 | **0.9252** | 0.9223 | 0.9250 |
| ACC | 0.9037 | 0.9209 | 0.8749 | 0.8666 | 0.9224 | 0.9214 | **0.9234** |
| MCC | 0.6125 | 0.6330 | 0.5566 | 0.5402 | 0.6327 | 0.6406 | **0.6417** |
| AUC | 0.9637 | 0.9104 | 0.9572 | 0.9524 | 0.9585 | 0.9627 | **0.9639** |
| Gmean | **0.9252** | 0.9104 | 0.9108 | 0.9038 | 0.9069 | 0.9165 | 0.9146 |

**Table 6.** Ablation experiments based on independent test dataset.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| DenseNet | √ | √ | | √ | √ | √ | √ |
| BGRU | | √ | √ | | √ | √ | √ |
| Self-Attention | | | √ | √ | √ | | √ |
| Ensemble | | | | | | √ | √ |
| Sn | 0.9607 | 0.8776 | 0.9468 | **0.9615** | 0.8826 | 0.8874 | 0.9019 |
| SP | 0.8950 | 0.9301 | 0.8753 | 0.8337 | 0.9279 | **0.9299** | 0.9274 |
| ACC | 0.9001 | 0.9260 | 0.8808 | 0.8437 | 0.9244 | **0.9266** | 0.9254 |
| MCC | 0.6086 | 0.6381 | 0.5634 | 0.5082 | 0.6357 | 0.6439 | **0.6464** |
| AUC | **0.9646** | 0.9603 | 0.9571 | 0.9488 | 0.9592 | 0.9640 | 0.9644 |
| Gmean | **0.9262** | 0.9089 | 0.9102 | 0.9115 | 0.9214 | 0.9106 | 0.9146 |

### *3.3. Performance of DGA-5mC on the training dataset*

To analyze the performance of DGA-5mC, we performed a five-fold cross-validation on the training dataset. The five-fold cross-validation performance of the DGA-5mC model on the training dataset is shown in Figure 10. It is not difficult to find that the values of the Sn, SP, ACC, MCC, AUC and Gmean metrics on the training dataset are very stable and fluctuate relatively little, effectively avoiding the problem of overfitting. Therefore, this indicates that the DGA-5mC model had good performance on the training dataset.

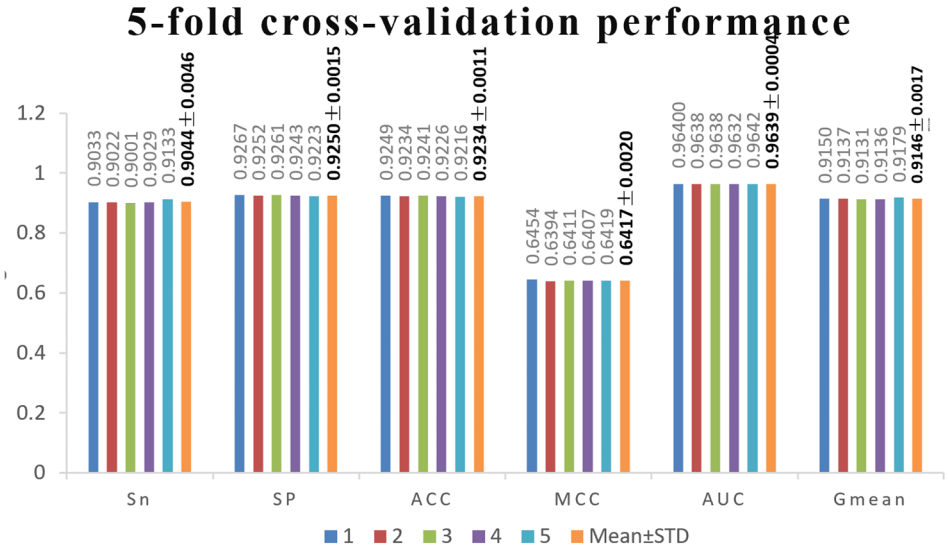


**Figure10.** Performance of DGA-5mC on the training dataset.

The ROC curve of the DGA-5mC model on the training dataset is shown in Figure 11, and the mean value of AUC was 0.9639. The AUC values for each fold of cross-validation were greater than 0.96. This indicates that our proposed DGA-5mC model has good stability.

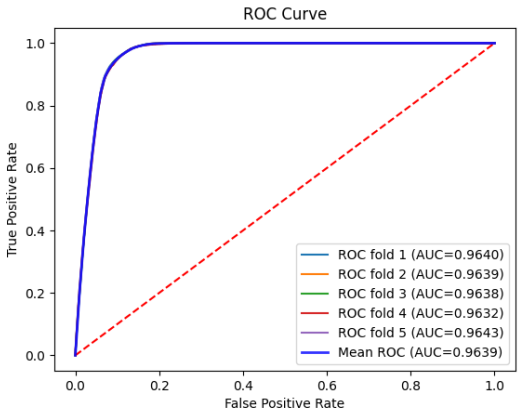


**Figure 11.** ROC curve for DGA-5mC on the training dataset.

### *3.4. Comparison with existing predictors*

The five-fold cross-validation performance of the DGA-5mC model compared to three existing prediction methods on the same dataset is shown in Table 7. It is worth noting that this work has an unbalanced ratio of positive and negative samples in the training dataset. However, we did not train the DGA-5mC model by processing the unbalanced data into balanced data as in the case of the BiLSTM-5mC model. Therefore, this made the training process biased toward the identification of majority class samples (promoter fragments without 5mC sites). As can be seen in Table 7, the DGA-5mC model had a higher Sn than the other three prediction methods, including being 9.48% higher than the model BiLSTM-5mC, which is the latest prediction method, at the cost of a 1.54% decrease in SP and a 0.68% decrease in ACC. Since our dataset was a large amount of unbalanced data, comparing ACC is not meaningful, and MCC is an important measure of unbalanced data. It can be seen that, although SP and ACC were slightly reduced, the MCC assessment index was 1.82% higher than its BiLSTM-5mC model counterpart, and the AUC was basically the same as that for the BiLSTM-5mC model, proving that the predictive model DGA-5mC proposed in this paper focuses more on the accurate identification of minority class samples (promoter fragments with 5mC sites) than the other three prediction methods. Taking into consideration both positive and negative sample precision, we achieved a value of 91.46% by using the Gmean metric for both the five-fold cross-validation and independent test datasets.

The independent test performance of the DGA-5mC model and other prediction methods on the same dataset is shown in Table 8. It can be seen that the DGA-5mC model developed in this research achieved the best performance and outperformed the other three models in terms of the Sn, MCC and AUC metrics. The Sn metric was 0.9019, which was 3.58% higher than that for the BiLSTM-5mC model of the latest prediction method. However, the SP and ACC values from our method were slightly lower than those for the BiLSTM-5mC model, but they were higher than those for the other two models, iPromoter-5mC and 5mC-Pred. This may be due to the very low Sn for the BiLSTM-5mC model, resulting in an elevated SP, with a large amount of imbalanced data biasing the identification of most class samples (promoter fragments without 5mC sites) during training. The BiLSTM-5mC model reuses the number of positive samples 10 times and divides the number of negative samples into 11 copies. Thus, the 1:11 unbalanced data was converted into 1:1 balanced data, resulting in 11 sub-models

of balanced data. Each model needed to be trained, and eventually 11 sub-models were trained. In contrast, the DGA-5mC model has the advantage that it can feed unbalanced data directly into that model, requiring only one model to be trained. In terms of computational effort, the DGA-5mC model has fewer parameters than the BiLSTM-5mC model, so it requires less time. Therefore, the DGA-5mC model is preferred over the other models.

The ROC curve of DGA-5mC on the independent test dataset is shown in Figure 12, with a value of 0.9644 for AUC. In summary, the DGA-5mC model achieved excellent performance on both the training dataset and the independent test dataset, outperforming the other existing predictors. These comparative results indicated that the DGA-5mC model has better generalization ability and stronger prediction ability to accurately identify the potential 5mC sites.

**Table 7.** 5-fold cross-validation performance of DGA-5mC and other predictors.

| Predictor | Sn | SP | ACC | MCC | AUC | Gmean |
|---|---|---|---|---|---|---|
| iPromoter-5mC | 0.8746 | 0.9039 | 0.9016 | 0.5743 | 0.9566 | – |
| 5mC-Pred | 0.8990 | 0.9200 | 0.9180 | 0.6260 | 0.9620 | – |
| BiLSTM-5mC | 0.8096 | 0.9404 | 0.9302 | 0.6235 | 0.9644 | – |
| **DGA-5mC** | **0.9044** | 0.9250 | 0.9234 | **0.6417** | 0.9639 | 0.9146 |

**Table 8.** Independent test dataset performance of DGA-5mC and other predictors.

| Predictor | Sn | SP | ACC | MCC | AUC | Gmean |
|---|---|---|---|---|---|---|
| iPromoter-5mC | 0.8777 | 0.9042 | 0.9022 | 0.5771 | 0.9570 | – |
| 5mC-Pred | 0.8950 | 0.9200 | 0.9180 | 0.6250 | 0.9620 | – |
| BiLSTM-5mC | 0.8661 | 0.9374 | 0.9303 | 0.6384 | 0.9635 | – |
| **DGA-5mC** | **0.9019** | 0.9274 | 0.9254 | **0.6464** | **0.9644** | 0.9146 |

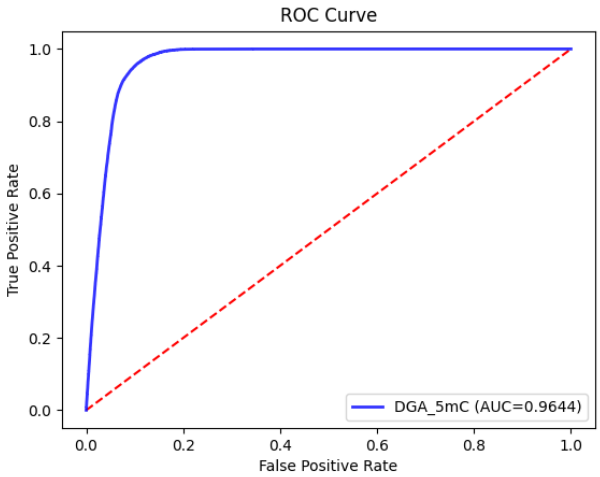


**Figure 12.** ROC curve for DGA-5mC on the independent testing dataset.

## 4. Conclusions

We developed a novel deep learning model, named DGA-5mC, to accurately identify 5mC modification sites in the genome-wide promoter region of SCLC cell lines in this research. The following three features of our model are wherein the significant novelties lie. First, we added ND encoding to the BiLSTM-5mC model encoding method and used their hybrid encoding to extract the original features of DNA sequences. Second, the ratio of positive to negative samples in the dataset for this work was unbalanced at 1:11. The DGA-5mC model algorithm automatically handles large proportions of unbalanced data for both positive and negative samples and does not require manual processing into balanced data, which highlights the reliability and superiority of the model. Lastly, we investigated the improved network framework of DenseNet and BGRU methods based on deep learning methods. We added a self-attentive module and classified it with fully connected layers, using a homogeneous ensemble. The experimental findings demonstrated that our proposed DGA-5mC model shows better prediction and generalization ability than the existing advanced models.

The accomplishment of the model will assist researchers in better identifying 5mC modification sites in the promoter region. In the future, we will extend this work by trying to build a network of servers, which will provide numerous conveniences. Furthermore, all datasets and the source code of the DGA-5mC model can be accessed for free at https://github.com/lulukoss/DGA-5mC.

## Acknowledgments

The authors are grateful for the constructive comments and suggestions made by the reviewers. 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 (No. GJJ190695). These funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

## Conflict of interest

The authors claim to have no conflict of interest.

## References

1. Y. Assenov, F. Muller, P. Lutsik, J. Walter, T. Lengauer, C. Bock, Comprehensive analysis of DNA methylation data with RnBeads, *Nat. Methods*, **11** (2014), 1138–1140. https://doi.org/10.1038/nmeth.3115
2. M. Beetch, C. Boycott, S. Harandi-Zadeh, T. Yang, B. J. E. Martin, T. Dixon-McDougall, et al., Pterostilbene leads to DNMT3B-mediated DNA methylation and silencing of OCT1-targeted oncogenes in breast cancer cells, *J. Nutr. Biochem.*, **98** (2021), 108815. https://doi.org/10.1016/j.jnutbio.2021.108815

3. H. Lv, F. Y. Dao, D. Zhang, H. Yang, H. Lin, Advances in mapping the epigenetic modifications of 5-methylcytosine (5mC), N6-methyladenine (6mA), and N4-methylcytosine (4mC), *Biotechnol. Bioeng.*, **118** (2021), 4204–4216. https://doi.org/10.1002/bit.27911
4. J. Karanthamalai, A. Chodon, S. Chauhan, G. Pandi, DNA $N^6$-methyladenine modification in plant genomes-A glimpse into emerging epigenetic code, *Plants*, **9** (2020). https://doi.org/10.3390/plants9020247
5. J. Xiong, K. K. Chen, N. B. Xie, T. T. Ji, S. Y. Yu, F. Tang, et al., Bisulfite-free and single-base resolution detection of epigenetic DNA modification of 5-Methylcytosine by methyltransferase-directed labeling with APOBEC3A deamination sequencing, *Anal. Chem.*, **94** (2022), 15489–15498. https://doi.org/10.1021/acs.analchem.2c03808
6. Q. Zhang, Y. Wu, Q. Xu, F. Ma, C. Y. Zhang, Recent advances in biosensors for in vitro detection and in vivo imaging of DNA methylation, *Biosens. Bioelectron.*, **171** (2021), 112712. https://doi.org/10.1016/j.bios.2020.112712
7. D. K. Vanaja, M. Ehrich, D. Van den Boom, J. C. Cheville, R. J. Karnes, D. J. Tindall, et al., Hypermethylation of genes for diagnosis and risk stratification of prostate cancer, *Cancer Invest.*, **27** (2009), 549–560. https://doi.org/10.1080/07357900802620794
8. K. Chen, J. Zhang, Z. Guo, Q. Ma, Z. Xu, Y. Zhou, et al., Loss of 5-hydroxymethylcytosine is linked to gene body hypermethylation in kidney cancer, *Cell Res.*, **26** (2016), 103–118. https://doi.org/10.1038/cr.2015.150
9. D. W. Tucker, C. R. Getchell, E. T. McCarthy, A. W. Ohman, N. Sasamoto, S. Xu, et al., Epigenetic reprogramming strategies to reverse global loss of 5-Hydroxymethylcytosine, a prognostic factor for poor survival in high-grade serous ovarian cancer, *Clin. Cancer Res.*, **24** (2018), 1389–1401. https://doi.org/10.1158/1078-0432.CCR-17-1958
10. P. Devi, S. Ota, T. Punga, A. Bergqvist, Hepatitis C virus core protein down-regulates expression of src-homology 2 domain containing protein tyrosine phosphatase by modulating promoter DNA methylation, *Viruses*, **13** (2021). https://doi.org/10.3390/v13122514
11. J. Rodriguez-Ubreva, C. de la Calle-Fabregat, T. Li, L. Ciudad, M. L. Ballestar, F. Catala-Moll, et al., Inflammatory cytokines shape a changing DNA methylome in monocytes mirroring disease activity in rheumatoid arthritis, *Ann. Rheum. Dis.*, **78** (2019), 1505–1516. https://doi.org/10.1136/annrheumdis-2019-215355
12. L. Wei, R. Su, S. Luan, Z. Liao, B. Manavalan, Q. Zou, et al., Iterative feature representations improve N4-methylcytosine site prediction, *Bioinformatics*, **35** (2019), 4930–4937. https://doi.org/10.1093/bioinformatics/btz408
13. S. Shinagawa, N. Kobayashi, T. Nagata, A. Kusaka, H. Yamada, K. Kondo, et al., DNA methylation in the NCAPH2/LMF2 promoter region is associated with hippocampal atrophy in Alzheimer's disease and amnesic mild cognitive impairment patients, *Neurosci. Lett.*, **629** (2016), 33–37. https://doi.org/10.1016/j.neulet.2016.06.055
14. L. Zhang, Y. Z. Xu, X. F. Xiao, J. Chen, X. Q. Zhou, W. Y. Zhu, et al., Development of techniques for DNA-methylation analysis, *TrAC, Trends Anal. Chem.*, **72** (2015), 114–122. https://doi.org/10.1016/j.trac.2015.03.025
15. M. Lecorguille, F. M. McAuliffe, P. J. Twomey, K. Viljoen, J. Mehegan, C. C. Kelleher, et al., Maternal glycaemic and insulinemic status and newborn DNA methylation: findings in women with overweight and obesity, *J. Clin. Endocrinol. Metab.*, **108** (2023), 85–98. https://doi.org/10.1210/clinem/dgac553

16. X. Su, Y. Chu, J. H. Kordower, B. Li, H. Cao, L. Huang, et al., PGC-1α promoter methylation in Parkinson's disease, *PLoS One*, **10** (2015), e0134087. https://doi.org/10.1371/journal.pone.0134087
17. L. Yang, Y. Chen, N. Liu, Y. Lu, X. Li, W. Ma, et al., 5mC and H3K9me3 of TRAF3IP2 promoter region accelerates the progression of translocation renal cell carcinoma, *Biomarker Res.*, **10** (2022). https://doi.org/10.1186/s40364-022-00402-3
18. F. Nassiri, A. Chakravarthy, S. Feng, S. Y. Shen, R. Nejad, J. A. Zuccato, et al., Detection and discrimination of intracranial tumors using plasma cell-free DNA methylomes, *Nat. Med.*, **26** (2020), 1044–1047. https://doi.org/10.1038/s41591-020-0932-2
19. M. J. Booth, T. W. Ost, D. Beraldi, N. M. Bell, M. R. Branco, W. Reik, et al., Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine, *Nat. Protoc.*, **8** (2013), 1841–1851. https://doi.org/10.1038/nprot.2013.115
20. Y. Li, T. O. Tollefsbol, DNA methylation detection: bisulfite genomic sequencing analysis, in *Epigenetics Protocols*, Humana Press, **791** (2011), 11–21. https://doi.org/10.1007/978-1- 61779-316-5_2
21. D. Chai, C. Jia, J. Zheng, Q. Zou, F. Li, Staem5: A novel computational approachfor accurate prediction of m5C site, *Mol. Ther. Nucleic Acids*, **26** (2021), 1027–1034. https://doi.org/10.1016/j.omtn.2021.10.012
22. Y. Liu, Y. Shen, H. Wang, Y. Zhang, X. Zhu, m5Cpred-XS: A new method for predicting RNA m5C sites based on XGBoost and SHAP, *Front. Genet.*, **13** (2022). https://doi.org/10.3389/fgene.2022.853258
23. X. Chen, Y. Xiong, Y. Liu, Y. Chen, S. Bi, X. Zhu, m5CPred-SVM: a novel method for predicting m5C sites of RNA, *BMC Bioinf.*, **21** (2020). https://doi.org/10.1186/s12859-020-03828-4
24. M. M. Hasan, S. Tsukiyama, J. Y. Cho, H. Kurata, M. A. Alam, X. Liu, et al., Deepm5C: A deep-learning-based hybrid framework for identifying human RNA N5-methylcytosine sites using a stacking strategy, *Mol. Ther.*, **30** (2022), 2856–2867. https://doi.org/10.1016/j.ymthe.2022.05.001
25. H. Shi, S. Zhang, X. Li, R5hmCFDV: computational identification of RNA 5-hydroxymethylcytosine based on deep feature fusion and deep voting, *Briefings Bioinf.*, **23** (2022). https://doi.org/10.1093/bib/bbac341
26. H. Wang, S. Wang, Y. Zhang, S. Bi, X. Zhu, A brief review of machine learning methods for RNA methylation sites prediction, *Methods*, **203** (2022), 399–421. https://doi.org/10.1016/j.ymeth.2022.03.001
27. G. Guo, K. Pan, S. Fang, L. Ye, X. Tong, Z. Wang, et al., Advances in mRNA 5-methylcytosine modifications: Detection, effectors, biological functions, and clinical relevance, *Mol. Ther. Nucleic Acids*, **26** (2021), 575–593. https://doi.org/10.1016/j.omtn.2021.08.020
28. A. El Allali, Z. Elhamraoui, R. Daoud, Machine learning applications in RNA modification sites prediction, *Comput. Struct. Biotechnol. J.*, **19** (2021), 5510–5524. https://doi.org/10.1016/j.csbj.2021.09.025
29. L. Zhang, X. Xiao, Z. C. Xu, iPromoter-5mC: A novel fusion decision predictor for the identification of 5-Methylcytosine sites in genome-wide DNA promoters, *Front. Cell Dev. Biol.*, **8** (2020). https://doi.org/10.3389/fcell.2020.00614
30. L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data, *Bioinformatics*, **28** (2012), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

31. T. T. Nguyen, T. A. Tran, N. Q. Le, D. M. Pham, Y. Y. Ou, An extensive examination of discovering 5-Methylcytosine sites in genome-wide DNA promoters using machine learning based approaches, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **19** (2022), 87–94. https://doi.org/10.1109/TCBB.2021.3082184
32. W. R. Qiu, X. Xiao, Y. T. Shao, Z. T. Luo, m5C-HPromoter: An ensemble deep learning predictor for identifying 5-methylcytosine sites in human promoters, *Curr. Bioinf.*, **17** (2022), 452–461. https://doi.org/10.2174/1574893617666220330150259
33. X. Cheng, J. Wang, Q. Li, T. Liu, BiLSTM-5mC: A bidirectional long short-term memory-based approach for predicting 5-Methylcytosine sites in genome-wide DNA promoters, *Molecules*, **26** (2021). https://doi.org/10.3390/molecules26247414
34. H. Wang, Z. Yan, D. Liu, H. Zhao, J. Zhao, MDC-Kace: A model for predicting Lysine acetylation sites based on modular densely Connected Convolutional Networks, *IEEE Access*, **8** (2020), 214469–214480. https://doi.org/10.1109/access.2020.3041044
35. J. Jia, G. Wu, M. Li, W. Qiu, pSuc-EDBAM: Predicting lysine succinylation sites in proteins based on ensemble dense blocks and an attention module, *BMC Bioinf.*, **23** (2022), 450. https://doi.org/10.1186/s12859-022-05001-5
36. J. Jia, M. Sun, G. Wu, W. Qiu, DeepDN_iGlu: prediction of lysine glutarylation sites based on attention residual learning method and DenseNet, *Math. Biosci. Eng.*, **20** (2022), 2815–2830. https://doi.org/10.3934/mbe.2023132
37. X. Li, S. Zhang, H. Shi, An improved residual network using deep fusion for identifying RNA 5-methylcytosine sites, *Bioinformatics*, **38** (2022), 4271–4277. https://doi.org/10.1093/bioinformatics/btac532
38. S. Min, B. Lee, S. Yoon, Deep learning in bioinformatics, *Briefings Bioinf.*, **18** (2017), 851–869. https://doi.org/10.1093/bib/bbw068
39. J. Jin, Y. Yu, L. Wei, Mouse4mC-BGRU: Deep learning for predicting DNA N4-methylcytosine sites in mouse genome, *Methods*, **204** (2022), 258–262. https://doi.org/10.1016/j.ymeth.2022.01.009
40. Q. Ning, J. Li, DLF-Sul: a multi-module deep learning framework for prediction of S-sulfinylation sites in proteins, *Briefings Bioinf.*, **23** (2022). https://doi.org/10.1093/bib/bbac323
41. Z. Y. Zhang, L. Ning, X. Ye, Y. H. Yang, Y. Futamura, T. Sakurai, et al., iLoc-miRNA: extracellular/intracellular miRNA prediction using deep BiLSTM with attention mechanism, *Briefings Bioinf.*, **23** (2022). https://doi.org/10.1093/bib/bbac395
42. Z. Luo, W. Su, L. Lou, W. Qiu, X. Xiao, Z. Xu, DLm6Am: A Deep-Learning-Based Tool for Identifying N6, 2'-O-Dimethyladenosine Sites in RNA Sequences, *Int. J. Mol. Sci.*, **23** (2022). https://doi.org/10.3390/ijms231911026
43. H. Li, S. Ning, M. Ghandi, G. V. Kryukov, S. Gopal, A. Deik, et al., The landscape of cancer cell line metabolism, *Nat. Med.*, **25** (2019), 850–860. https://doi.org/10.1038/s41591-019-0404-8
44. T. H. Nguyen-Vo, Q. H. Nguyen, T. T. T. Do, T. N. Nguyen, S. Rahardja, B. P. Nguyen, iPseU-NCP: Identifying RNA pseudouridine sites using random forest and NCP-encoded features, *BMC Genomics.*, **20** (2019), 971. https://doi.org/10.1186/s12864-019-6357-y
45. Z. Cui, L. Kang, L. Li, L. Wang, K. Wang, A combined state-of-charge estimation method for lithium-ion battery using an improved BGRU network and UKF, *Energy*, **259** (2022). https://doi.org/10.1016/j.energy.2022.124933

46. Y. Shang, X. Tang, G. Zhao, P. Jiang, T. Ran Lin, A remaining life prediction of rolling element bearings based on a bidirectional gate recurrent unit and convolution neural network, *Measurement*, **202** (2022). https://doi.org/10.1016/j.measurement.2022.111893
47. S. Yang, G. Berdine, The receiver operating characteristic (ROC) curve, *Southwest Respir. Crit. Care Chron.*, **5** (2017), 34–36. https://doi.org/10.12746/swrccc.v5i19.391


AIMS Press