scientific reports

Check for updates

OPEN

A robust deep learning approach for identification of RNA 5-methyluridine sites

Md. Shazzad Hossain Shaon[1], Tasmin Karim[1], Md. Mamun Ali[2,3], Kawsar Ahmed[4,5,6]✉, Francis M. Bui[4], Li Chen[4] & Mohammad Ali Moni[7,8]✉

RNA 5-methyluridine (m5U) sites play a significant role in understanding RNA modifications, which influence numerous biological processes such as gene expression and cellular functioning. Consequently, the identification of m5U sites can play a vital role in the integrity, structure, and function of RNA molecules. Therefore, this study introduces GRUpred-m5U, a novel deep learning-based framework based on a gated recurrent unit in mature RNA and full transcript RNA datasets. We used three descriptor groups: nucleic acid composition, pseudo nucleic acid composition, and physicochemical properties, which include five feature extraction methods ENAC, Kmer, DPCP, DPCP type 2, and PseDNC. Initially, we aggregated all the feature extraction methods and created a new merged set. Three hybrid models were developed employing deep-learning methods and evaluated through 10-fold cross-validation with seven evaluation metrics. After a comprehensive evaluation, the GRUpred-m5U model outperformed the other applied models, obtaining 98.41% and 96.70% accuracy on the two datasets, respectively. To our knowledge, the proposed model outperformed all the existing state-of-the-art technology. The proposed supervised machine learning model was evaluated using unsupervised machine learning techniques such as principal component analysis (PCA), and it was observed that the proposed method provided a valid performance for identifying m5U. Considering its multi-layered construction, the GRUpred-m5U model has tremendous potential for future applications in the biological industry. The model, which consisted of neurons processing complicated input, excelled at pattern recognition and produced reliable results. Despite its greater size, the model obtained accurate results, essential in detecting m5U.

Keywords RNA 5-methyluridine, RNA modifications, Physicochemical properties, Deep-learning, Principal component analysis, Transcript RNA

Post-transcriptional regulation is the modification method, where the prime transcriptions RNA is transformed into a balanced RNA called a "mature RNA" in the species. This strategy usually includes molecular compounds to the foundation of RNA, which is a significant phase for gene transcription[1,2]. According to the study, more than 170 post-transcriptional variations have been successfully discovered using high-quality technology[3]. These 170 variations are defined as "epi transcriptomics"[4]. 5- methyluridine termed "m5U" is one of the most crucial components of RNAs among all the epi transcriptomics variants[5], which is mostly found in rRNAs, tRNAs, mRNAs, and incRNAs. So, m5U impacts their functionalities[6–8]. The m5U is obtained in the fifth carbon (C_5) position of uridine catabolized in RNAs[9]. Various major catalysts are involved in the chemical transformation of m5U, which is in *Escherichia coli (E.coli)*: TrmA, RlmC, and RlmD[10,11], in *mammals*: TRMT2A and TRMT2B[12,13], and in *Saccharomyces cerevisiae*: Trm2[14]. In *E. coli tmRNA*, TrmA efficiently methylates the

[1]Department of Computer Science and Informatics, Oakland University, Rochester, MI 48309, USA. [2]Division of Biomedical Engineering, University of Saskatchewan, 57 Campus Drive, Saskatoon, SK S7N 5A9, Canada. [3]Department of Software Engineering, Daffodil Smart City (DSC), Daffodil International University, Birulia, Savar, Dhaka 1216, Bangladesh. [4]Department of Electrical and Computer Engineering, University of Saskatchewan, 57 Campus Drive, Saskatoon, SK S7N 5A9, Canada. [5]Group of Bio-photomatiχ, Department of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Santosh 1902, Tangail, Bangladesh. [6]Health Informatics Research Lab, Department of Computer Science and Engineering, Daffodil International University, Daffodil Smart City, Dhaka 1216, Birulia, Bangladesh. [7]AI & Digital Health Technology, Artificial Intelligence & Cyber Future Institute, Charles Sturt University, Bathurst, NSW 2795, Australia. [8]AI & Digital Health Technology, Rural Health Research Institute, Charles Sturt University, Orange, NSW 2800, Australia. ✉email: k.ahmed.bd@ieee.org; kawsar.ict@mbstu.ac.bd; k.ahmed@usask.ca; mmoni@csu.edu.au

T-loop, which has been observed to modify the bacterial 16 S in *vivo*, but this alteration was not identified in *vivo*[15,16]. The researchers have verified distinct biochemical impacts from this molecule, such as the growth of the brain[17,18], different stages of embryogenesis[19], determinations affecting tumorigenesis[20,21], modulation of plants' evolution, and signaling pathway[22]. Advanced and effective innovations should be introduced for locating m5U sites. Deep learning approaches can be one of the innovative methods to detect these sites. From this perspective, we aim to design a deep-learning method to identify m5U sites.

In recent years, some methods have been introduced to identify m5U sites using both machine learning and deep learning methods. In 2019, Carter et al. presented the Fluorouracil-Induced-Catalytic-Crosslinking-Sequencing method abbreviation of "FICC-Seq" to locate the m5U from *Homo sapiens*[13], where the authors highlighted that this method has been described the characteristics by the molecule m5U sites. However, machine-learning and deep-learning approaches are more suitable and more accessible to obtain the sequences than traditional methods. Jiang et al. proposed a m5U prediction model called "m5Upred" based on a support vector machine (SVM), a machine learning method in 2020[23]. The authors evaluated for performance by using independent tests and five-fold cross-validation approaches. However, there could be a severe possibility of overfitting concerns with redundant information. Feng et al. proposed the "iRNA-m5U" technique based on the SVM algorithm to detect the m5U sites from the RNA sequences in 2021[24]. The researchers stated that they only utilized a limited dataset to train their framework. Besides, the feature selection method was inadequate at acquiring the relevant information to accurately identify the nucleotides, which include the m5U site in all organisms.

Li et al. in 2022, proposed a deep transfer learning neural network framework, called RNADSN, to identify the m5U in mRNA by a similar feature extraction method between mRNA and tRNA[25]. However, the authors did not compare their models with the existing models. They did not present any graphical presentation of how much their model performance is between negative and positive classes. In 2023, Yu et al. developed the "Deepm5U" model based on the deep neural network method[26] using auto BioSeqpy tools[27]. However, the researchers did not address their model's biases or state which deep learning model is best suited to detect the m5U. In the same year, Ao et al. designed a machine learning algorithm called "m5U-SVM", where the authors demonstrated the Multiview feature selection approaches to predict the m5U sites from the RNA sequences[28]. As the m5U sites in RNA are associated with various diseases due to their regulatory roles in RNA stability and function, research indicates their relevance in antifungal and antiviral responses, suggesting potential therapeutic targets for combating these infections such as the study "iAFPs-Mv-BiTCN" presents a technique for predicting antifungal peptides utilizing evolutionary characteristics, bidirectional temporal convolutional networks, and self-attention transformer embedding[29]. The models "Deepstacked-AVPs" and "DeepAVP-TPPred" focus on predicting antiviral peptides, leveraging deep learning architectures to enhance accuracy in identifying peptides with antiviral properties[30,31]. Additionally, the "pAtbP-EnC" framework is tailored to predict anti-tubercular peptides, aiming to support tuberculosis treatment strategies. Other models like "AIPs-SnTCN" are designed for detecting anti-inflammatory peptides[32,33], and the "Deep5HMC" model for 5-hydroxymethylcytosine modification[34]. In another study, Naeem et al. introduced various computational methods, including a deep learning-based model for diabetes prediction[35], in another article, the "DBSCAN" framework was applied for improving clustering techniques[36]. Another paper developed a model for predicting SUMOylating sites[37]. Another "Enet-6 mA" was designed to predict RNA sites[38] and ORI-Explorer to identify origin replication sites[39]. These efforts showcase the wide-ranging applications of computational techniques in peptide research aimed at therapeutic development.

The discussion above indicates a potential scope for developing a more efficient model for identifying m5U sites. Consequently, we performed this study to introduce a well-performed, efficient, and reliable deep learning model to identify m5U sites with higher efficiency. In the current study, we have explored deep-learning approaches to develop the expected model. Our contribution to this study is as follows:

1. The study merged five features from three distinct descriptor categories. This vigilant dataset has 454 features, including both accomplished transcript RNA and mature RNA.
2. This study applied several deep-learning models to detect the active and inactive classes in the sequences and picked the best-fit models for identifying m5U sites.
3. The proposed approach developed based on the Gated Recurrent Unit (GRU) model obtained better performances than the other existing predictors.

Our work follows Chou's five-point guidelines[40,41], to this end, we formulate the RNA modification prediction problem precisely, build a complete model with Conv1D and GRU layers and all required preprocessing steps, create a robust algorithm that is validated by 10-fold cross-validation, and offer thorough documentation to guarantee repeatability. This systematic process guarantees our research's rigor, dependability, and clarity.

Materials and methods

Data collection

The current study collected 3262 positive and 4825 negative samples from the Ao et al.[28] articles, where the datasets had already been preprocessed with a 0.8 threshold in order to remove redundancy by employing the Cluster Database at High Identity with Tolerance (CD-HIT) method[42]. The full transcript dataset has 2034 active m5U sequences and 3593 inactive m5U sequences. The mature dataset has 1228 active and inactive sequences. We collected another dataset from a study conducted by Jiang et al., where the authors obtained 4928 positive and 4928 negative redundant datasets[23].

Data preprocessing phase

This study proposes a robust deep learning-based prediction approach, and the corresponding flowchart of the study is represented in Fig. 1. After collecting the dataset, we extracted features using five methods from three descriptor groups. This step is essential because sequential data must be numerically represented for model development. Machine learning models, especially GRUs or LSTMs, require numerical inputs to learn patterns and perform mathematical operations for training. This transformation enables effective model optimization. Therefore, from nucleic acid composition groups, we have considered enhanced nucleic acid composition (ENAC)[43], and basic Kmer (Kmer)[44], from physicochemical property groups, we have applied dinucleotide physicochemical properties (DPCP) and (DPCP Type 2)[45], and from pseudo nucleic acid composition groups, we considered Pseudo dinucleotide composition (PseDNC) descriptor[46,47]. All the features were generated by using the ilearnplus tool[48]. We combined every characteristic to produce a new dataset with 454 properties. To create the best-fit model, we used independent testing, individual assessments, and 10-fold cross-validation on this dataset. Our study design was motivated by the analysis's conclusions that a merged-based cross-validation technique would be more acceptable.

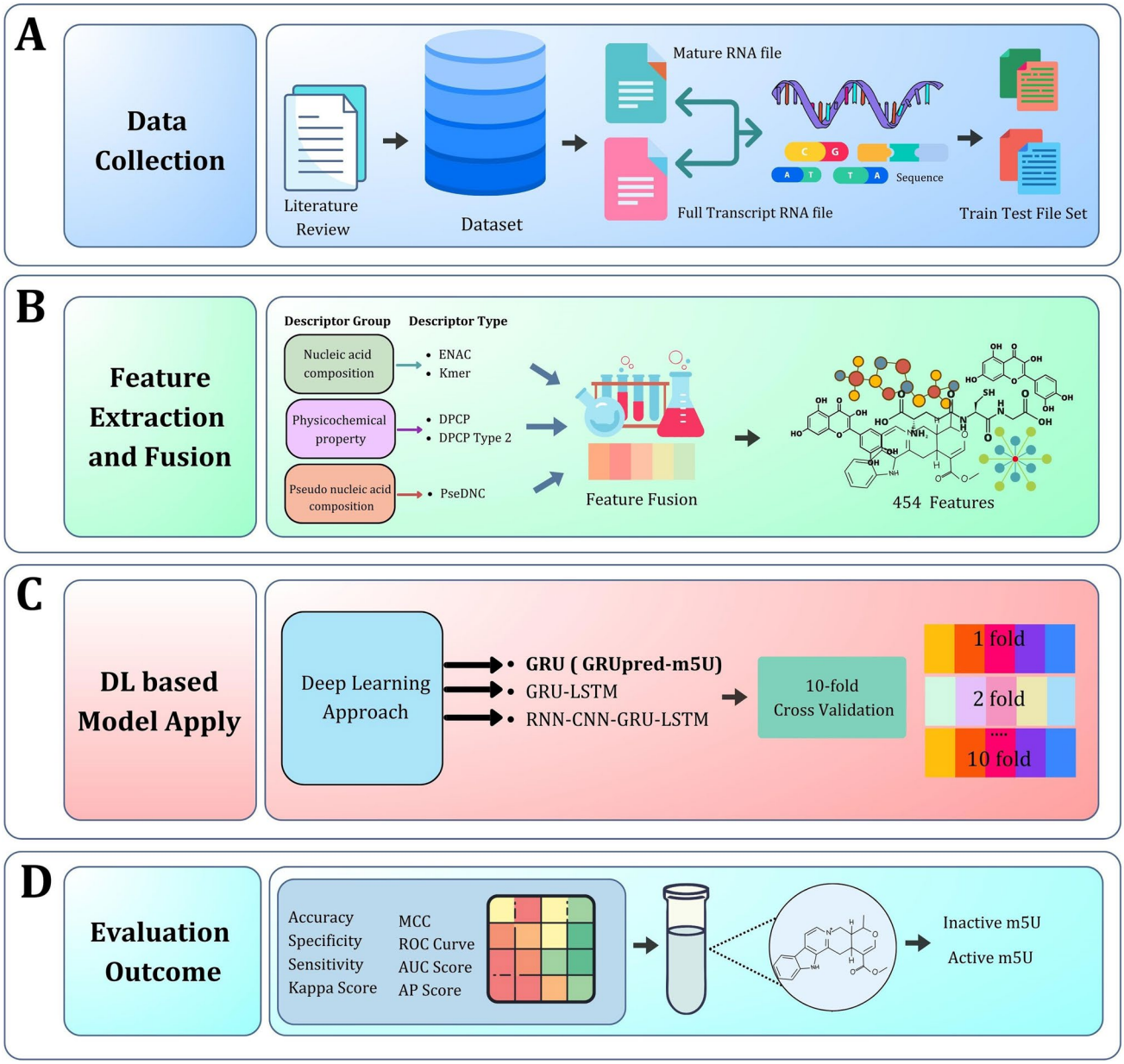

Fig. 1. Overview of the m5U analysis workflow. (**A**) collection of the two types of datasets: the full transcript and mature mRNA, (**B**) feature extraction and fusion method to create a two-stage feature selection approach, (**C**) applying deep learning model on cross-validation evaluation method, (**D**) performance assessment with the evaluation metrics and finally discovering the active class from the datasets. (Icons collection and diagram design URLs: https://www.flaticon.com/icons, and https://app.diagrams.net/).

Description of the proposed model construction

In this study, different deep learning models were applied, and we considered three individual models to present in the article to compare their performances. These three hybridization models are GRU with Long Short-Term Memory (LSTM) model, GRU with Recurrent neural network (RNN), convolutional neural networks (CNN), and LSTM layer, and the proposed model with GRU and CNN layer. All the hyperparameters used in this study are represented in Table 1. All the layers consist of convolutional layers (Conv1D) followed by Batch Normalization, activation functions (sigmoid and ReLU), max pooling, and dropout layers. The models were built using the Adam optimizer and binary cross-entropy loss, with accuracy and the Area Under the ROC Curve (AUC) used as evaluation metrics. In this section, we describe our proposed construction procedure approach in a proper mathematical way. Accordingly, the other model's architecture is found in supplementary files Figure S1 and Figure S2 with descriptions.

Construction of the GRUpred-m5U framework

The proposed GRUpred-m5U model was developed using the GRU and CNN layer. Figure 2 represents the proposed model, where we developed the proposed model following nine steps, using various layers such as convolution layers, batch normalization layers, activation layers, max-pooling layers, GRU layers, dense layers, and dropout layers. Every stage has particular characteristics and functions. The following steps represent the GRUpred-m5U model:

$$Conv1_{x(j)} = \varphi\left(w1 \times\ x[j : j+3] +\ k_1\right) \tag{1}$$

$$max1_j = \left(Conv1_{x(2\times j),} Conv1_{(j)[2j+1]}\right) \tag{2}$$

$$bn1_j = \frac{\theta\left(j - m(j)\right.}{sqrt\left(v\left(j\right) + \epsilon\right) + \lambda} \tag{3}$$

Models	Parameters
GRU-LSTM	Conv1D (1): filters = 64, kernel_size = 3, Conv1D (2): filters = 64, kernel_size = 3, BatchNorm (1): Yes, Activation (1): sigmoid, MaxPool (1): pool_size = 2, Dropout (1): rate = 0.5 GRU (1): units = 256, return_sequences = True LSTM (1): units = 128, return_sequences = True GRU (2): units = 64, return_sequences = True LSTM (2): units = 32, return_sequences = True Flatten: Yes Dense (1): units = 64 BatchNorm (2): Yes Activation (2): relu Dropout (2): rate = 0.5 Dense (2): units = 1, activation = sigmoid
RNN-CNN-GRU-LSTM	Conv1D (1): filters = 64, kernel_size = 3 BatchNorm (1): Yes Activation (1): relu MaxPool (1): pool_size = 2 Dropout (1): rate = 0.5 SimpleRNN: units = 64, return_sequences = True BatchNorm (2): Yes LSTM: units = 64, return_sequences = True GRU: units = 64 Dropout (2): rate = 0.5 Flatten: Yes Dense: units = 64 Activation: relu Dense: units = 1, activation = sigmoid
GRUpred-m5U	Conv1D (1): filters = 64, kernel_size = 3 BatchNorm (1): Yes Activation (1): sigmoid MaxPool (1): pool_size = 2 Conv1D (2): filters = 128, kernel_size = 3 BatchNorm (2): Yes Activation (2): relu MaxPool (2): pool_size = 2 Conv1D (3): filters = 64, kernel_size = 3 BatchNorm (3): Yes Activation (3): relu MaxPool (3): pool_size = 2 GRU (1): units = 128, return_sequences = True GRU (2): units = 64 Dense (1): units = 64 BatchNorm (4): Yes Activation (4): relu Dropout: rate = 0.5 Dense (2): units = 1, activation = sigmoid

Table 1. Hyperparameters of applied three hybrid models.

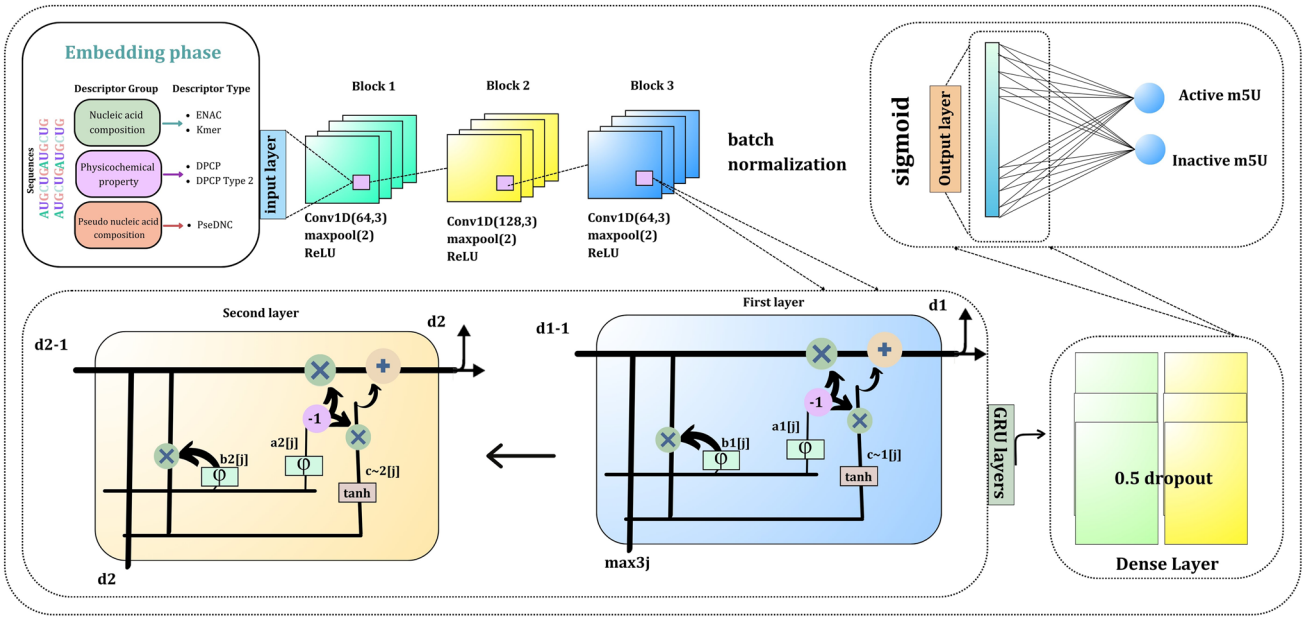

Fig. 2. An Overview of the GRUpred-m5U Architecture. (Icons collection and diagram design URLs: https://www.flaticon.com/icons, and https://app.diagrams.net/).

At first, we took 454 features as input shapes, whereas in the first convolutional layers, we used 64 filters with 3 kernel sizes. To obtain an accurate performance, we used the ReLU activation function, and to reduce the variance, we used a max pooling size of 2. Here, φ denotes the activation function, k_1 is the bias vector, θ is the scale parameter, λ is the shift parameter, $m(j)$ denotes the mean of j, and ϵ reduces the division by 0, whereas the epsilon prevents generally positive numbers. $Conv1_{x(j)}$ means the convolutional layer, $max1_j$ is the max-pool, and $bn1_j$ is the batch normalization. After the first layer, we have created a second convolutional layer, as represented as:

$$Conv2_{max1_j} = \varphi\ (w2 \times\ max1_j\,[j : j+3] +\ k_2) \tag{4}$$

$$max2_j = \left(Conv2_{max1_j} Conv2_{2(max1_j)[2j+1]}\right) \tag{5}$$

We used 128 filters with 3 kernel sizes and applied similar activation functions and max-pool layers as used in the first convolution layers. In the second layer, we took the first layer output as an input and generated another result that passed through the third convolution layer. The third layer is denoted as:

$$Conv3_{max2_j} = \varphi\,(w3 \times\ max2_j[j : j+3] +\ k_3) \tag{6}$$

$$max3_j = max\left(Conv3_{max2_j} Conv3_{2(max2_j)[2j+1]}\right) \tag{7}$$

where we applied 64 filters, 3 kernel sizes, ReLU activation, and 2 max-pooling rare, accordingly. Afterward, we applied GRU layers, where in the first layer we took 128 unit numbers, and in the next layer we took 64 numbers of units. The GRU layers can be stated as:

$$a1\,[j] = \varphi\,(L_{a1}\,.max3_j + U_{a1}.d1[j-1] + g_{a1} \tag{8}$$

$$b1\,[j] = \varphi\,(L_{b1}\,.max3_j + U_{b1}.d1[j-1] + g_{b1} \tag{9}$$

$$c\sim\ 1\,[j] = \varphi\,(L_{c1}\,.max3_j + U\sim\ 1(b1\,[j]\,.d1[j-1] + g_{d1} \tag{10}$$

$$d1\,[j] = (1 - a1[j].d1[j-1] + a1[j].c\sim\ 1[j] \tag{11}$$

After the first GRU layer output can be through to the another GRU layer and the second GRU layer as stated:

$$a2\,[j] = \varphi\,(L_{a2}\,.d[j] + U_{a2}.d2[j-1] + g_{a2} \tag{12}$$

$$b2\,[j] = \varphi\ \ (L_{b2}\,d1[\,j] + U_{b2}.d2[j-1] + g_{b2} \tag{13}$$

$$c\sim\ 2\,[j] = \varphi\,(L_{c2}\,.d1[j] + U\sim\ 2(b2\,[j]\,.d2[j-1] + g_{d1} \tag{14}$$

$$d2\,[j] = (1 - a2[j].d2[j-1] + a2[j].c\sim\ 2[j] \tag{15}$$

where φ the activation function, *j* is the weighted sum of time steps, $j-1$is the hidden state at time steps, $d1[j-1]$, $d2[j-1]$ control the previous hidden states, g_{a1}. $g_{d1,}g_{a2}\cdots g_{d1}$ is the bias vector. $a1\left[j\right], a2\left[j\right]$ is the updated gate, $b1\left[j\right], b2\left[j\right]$ is the reset gate, $c\sim\ 1\left[j\right], c\sim\ 2\left[j\right].$ is the candidate activation, $d1\left[j\right], d2\left[j\right]$ is the updated hidden state. Therefore, we added the dense layer with the batch normalization, The ReLU activation layer can be stated as:

$$dense = W.d2\left[j\right] + D_{res} \tag{16}$$

Finally, we have created the final classifier with a sigmoid activation function and 0.5 dropout layer, the output layer expressed as:

$$res = W.dr + D_{res} \tag{17}$$

where W represents the weight matrix, dr denotes the 50% dropout and D_{res} denotes the bias vector.

Description of evaluation metrics

We used a wide range of evaluation metrics to evaluate the model's performance. We applied accuracy (ACC) to measure the accurate classification of all the instances. Cohen's kappa score (KP) evaluates the extent of compatibility between the envisioned and actual categorization. Matthew's coefficient correlation (MCC) is a method for analyzing the reliability of dichotomous categories, particularly those derived from imbalanced data sets. Sensitivity (Sn), and specificity (Sp) are used to detect the actual positive class and specific classification from the RNA samples. To confirm the model's ability to distinguish between the active and inactive sequences, we applied the area under curve score (AUC) and average precision score (AP). All the evaluation metrics can be stated as:

$$ACC = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{18}$$

$$MCC = \frac{TP*TN - FP*FN}{\sqrt{TP+FP*(TP+FN)*(TN+FP)}} \tag{19}$$

$$Kp = \frac{2*(TP*TN - FP*FN)}{(TP+FP)*(FP+TN)+(TP+FN)*(FN+TN)} \tag{20}$$

$$Sp = \frac{TN}{(TN+FP)} \tag{21}$$

$$Sn = \frac{TP}{(TP+FN)} \tag{22}$$

$$AP = \frac{1}{m}\Sigma\,{}^{m}_{k=1}\left(p(\,V_k) - p(V_{k-1}) \times\ V_k\right. \tag{23}$$

where m is the total value of the recall on the precision curve, $p\left(V_k\right)$ denotes the interpolated value at the *kth* recall level, V_k is the *kth* recall value. *TP* denotes the true positive, *TN* denotes the true negative, *FP* is the false positive, and *FN* is the false negative value[49–51].

Experimental results

This study employed several deep-learning-based methods to find the best-performing classification method to identify m5U sites. Finally, the best-performing three models were chosen for further analysis to develop the proposed method. We used GRU-based hybridization methods such as RNN-CNN-GRU-LSTM, GRU with LSTM, and GRU with CNN, which is our proposed approach named GRUpred-m5U. In the result sections, we summarize the overall performance of our proposed models compared to other applied models and state-of-the-art technology performance. Table 2 shows the three applied models' performance on a 10-fold cross-validation approach for the two datasets. In the supplementary file, Table S1 presents the independent test results and Table S2 represents the performances of the applied models for individual feature extraction methods.

Mode	Classifier	ACC (%)	MCC (%)	Kp (%)	Sp (%)	Sn (%)	AUC (%)	AP (%)
Full_transcript RNA	RNN-CNN-GRU-LSTM	75.80	42.52	37.75	96.87	37.92	89.09	82.03
	GRU-LSTM	90.85	80.55	80.30	91.63	89.49	95.94	93.41
	GRUpred-m5U	**96.70**	**92.77**	**92.76**	**97.41**	**95.37**	**98.89**	**98.18**
Mature RNA	RNN-CNN-GRU-LSTM	82.24	65.92	64.30	96.08	68.08	94.11	93.73
	GRU-LSTM	67.08	43.17	34.64	35.05	99.59	96.42	96.77
	GRUpred-m5U	**98.41**	**96.87**	**96.81**	**99.18**	**97.65**	**99.83**	**99.86**

Table 2. Performance analysis on 10-fold cross-validation approach in two modes.

Table 2 demonstrates that the GRUpred-m5U model obtained the most compatible performance on every evaluation metric. In full transcript mode, it provided 96.70% accuracy, which is higher than the other models. GRU-LSTM obtained 6% and RNN-CNN-GRU-LSTM achieved 15% less accuracy, compared to the proposed model. These evaluation metrics state that the proposed model can identify the active and inactive m5U sites from the input sequences. Moreover, in terms of AUC, sensitivity, and specificity, the model achieved 98.89%, 95.37%, and 97.41%, respectively, which denotes that the proposed model identified the positive sequences and the negative sequences correctly more than 95% accurately, which denotes a successful performance. Accordingly, the other evaluation metrics demonstrated excellent performances. The GRU-LSTM achieved 90.85% accuracy, but this model could not perform better than the proposed model in evaluating the other metrics. Similarly, the RNN-CNN-GRU-LSTM performed lower than the other applied two models. The GRUpred-m5U model has successfully demonstrated acceptable performance on every platform in a 10-fold cross-validation method to compare them in full transcript RNA mode.

In mature RNA, GRUpred-m5U achieved higher accuracy with 98.41% in every aspect of the evaluation metrics. The proposed model accurately predicted the m5U sites, where this model gained 99.83%, 97.65%, and 99.18% AUC scores, sensitivity, and specificity, which denotes a remarkable performance of the proposed model. The other applied models, GRU-LSTM and RNN-CNN-GRU-LSTM achieved lower performance in all the evaluation metrics. It is seen that the GRU-LSTM method obtained 99.59% sensitivity, which denotes a higher performance in identifying positive classes. However, the specificity of the GRU-LSTM model could be better, which indicates that the model cannot identify negative sequences. Overall performances of the applied models on two datasets indicate that the proposed model, GRUpred-m5U, is highly capable of detecting the m5U sites from the RNA sequences in both methods.

Figure 3 exhibits the ROC curve in mature and complete transcript RNA modes. The ROC curve and measuring the AUC give an exhaustive understanding of a binary classification model's effectiveness, promoting a well-informed comparative analysis. In these subplots, it is stated that our models play an important role in detecting the sites, where GRUpred-m5U obtained 99.89 and 99.83% of AUC scores in two datasets.

Figure 4 compares the performance of the GRUpred-m5U in two modes using two confusion matrices, labeled (A) and (B). There are 487 true negatives, 13 false positives, 23 false negatives, and 477 genuine positives in matrix (A). The model performs better with 488 true positives, 12 false negatives, 4 false positives, and 496 true negatives in matrix (B). The number of samples is represented by a blue color gradient in both matrices; deeper shades indicate larger counts. Compared to matrix (A), matrix (B) shows a more accurate model with fewer classification mistakes.

Principal component analysis

In this study, we used unsupervised machine learning methods on two training datasets such as mature and full transcript RNA. We applied the principal component analysis (PCA) on two of the datasets to show the data distribution and the pattern of the data before and after the cluster to justify those two classes of positive and negative separating the processed data. In Fig. 5, A and C represent the original train sets of the two datasets, and B and D visualize the clustering result of the applied two datasets. We used the Gaussian matrix algorithm to cluster the datasets.

According to Fig. 5(A) and 5(C), the positive and negative classes are not differentiable from the raw data since they overlap in both datasets. However, the two classes are completely separate in both training datasets after preprocessing, according to Fig. 5(B) and 5(D). Moreover, the unsupervised machine learning results indicate that the m5U sites can be separated and identified from RNA sequences since they are linearly differentiable. The results also validate that the supervised machine learning performances are valid since the processed dataset shows a perfect distribution. The red circle marked data points are supposed to be in one cluster but belong to another. Consequently, it proves that these data points have some other characteristics that require further study. So, the unsupervised machine learning results indicate that the proposed model's performance is highly reliable, and the model is highly efficient in identifying m5U sites.

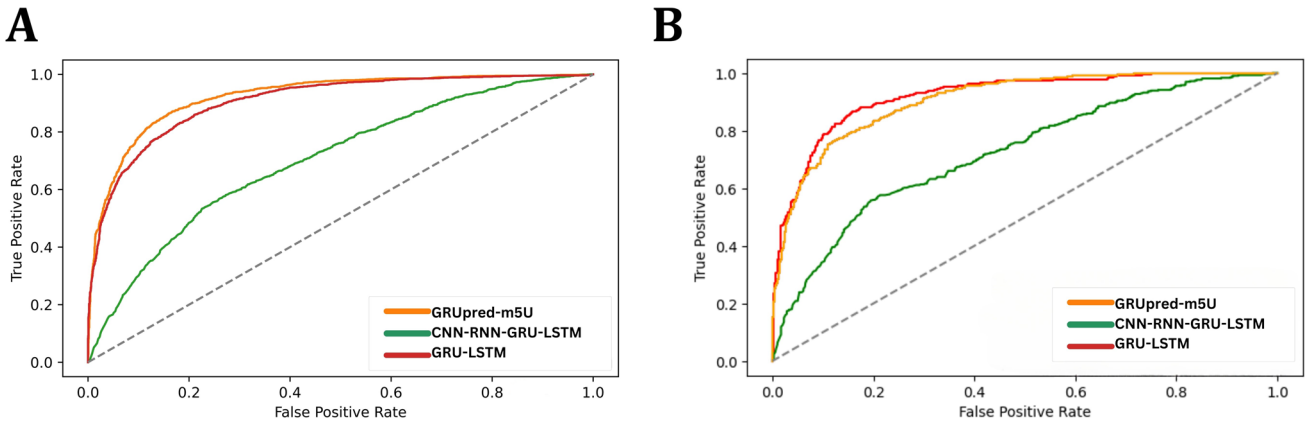

Fig. 3. ROC curve analysis on full transcript RNA and mature RNA. (**A**) full-transcript RNA; (**B**) mature RNA.

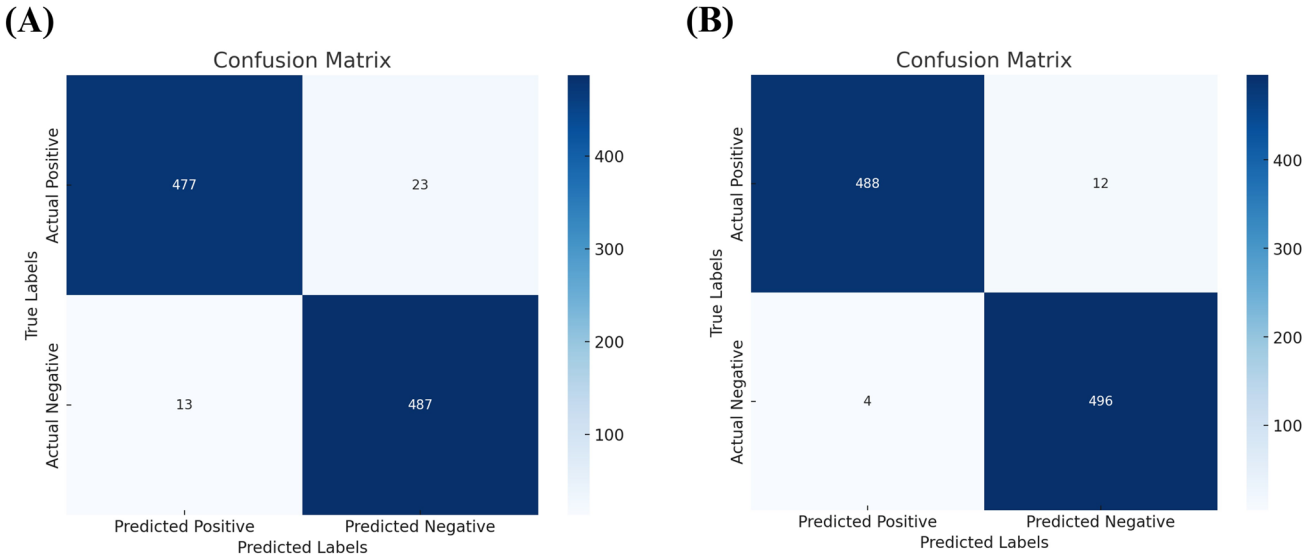

Fig. 4. Confusion matrix visualization of the proposed models, where A is the Full transcript RNA, and B is the mature RNA.

Discussion

5-methyluridine (m5U) findings become essential for numerous species, including humans, mammals, and plants. m5U has significance in RNA modification, especially in transfer RNA (tRNA). This alteration is required to preserve the consistency and functionality of tRNA compounds, which contribute to protein synthesis. Protein synthesis is critical for cells' regular functioning and survival in all living creatures. As a result, identification of m5U proves essential to comprehending and ensuring the strength of processes in cells in many species, eventually affecting the health and well-being of creatures throughout the biological spectrum. The proposed model identifies the m5U precisely, where we developed a GRU-based model with several functions with multiple GRU and convolutional layers. The model could acquire hierarchical representations of the input data using the initial Conv1D layers with variable filter sizes and activation algorithms. The Batch Normalization layers assist in training stability and effectiveness by minimizing internal covariate alteration. The MaxPooling1D layers further mitigate computational strain by downsampling the learned features. The addition of GRU layers to the data adds the capacity to capture sequential patterns and long-term relationships, which is crucial for jobs with spatiotemporal periods. A selective choice of hyperparameters and optimization methods generates robust performance, leading to highly precise results. The proposed approach uses its capacity to automate feature extraction from data to improve classification performance over conventional feature selection techniques. Conventional techniques frequently depend on heuristic- or manual-based feature selection, which may miss intricate connections or interactions in the data. However, the suggested model—probably based on deep learning techniques—can better manage big, high-dimensional datasets. It improves classification accuracy and resilience by detecting complex patterns and connections that conventional approaches might overlook. An imbalanced dataset was employed int this study because, in the real world, the possibility of getting balanced dataset is very poor. In most cases, all the datasets are imbalanced. Consequently, we invested our efforts to build a model with imbalanced data with higher efficiency and reliability, so that the model can be used in the real world with any sort of balanced or imbalanced dataset. To ensure the reliability of the model with imbalanced dataset, we have considered some performance evaluation metrics especially, sensitivity, specificity, and AUROC curve. Our proposed model gained 97.65% and 95.37% sensitivity for Mature RNA and Full_transcript RNA respectively, which ensures that the proposed model is highly capable of identifying positive data. At the same time, the proposed model gained 97.41% specificity for Full_transcript RNA and 99.18% specificity for Mature RNA, which ensures that the model is highly capable of identifying negative. In addition to that we have drawn AUROC and found that the model is not overfitted. So, the overall result shows that imbalanced dataset did not impact the reliability of the proposed model. So, The proposed model is highly efficient and reliable.

Comparison of GRUpred-m5U with the state-of-the-art method

According to the literature search, it is found that our approach is the most successful compared to other methods. We compared the proposed model with the existing state-of-the-art models based on machine learning and deep learning-based approaches. Figure 6 compares the proposed model's performance with other existing models. We have compared in terms of ACC, Sn, Sp, and AUC scores. Figure 5(A) denotes the full transcript RNA, and Fig. 5(B) denotes mature RNA. It is clearly shown that our proposed model obtained a remarkable performance compared to the other existing models. In the full transcript mode, our model achieved a higher level of performance, especially in the sensitivity our models distinguished the positive class more accurately. In the mature RNA, we can observe that our model acquired a higher performance in sensitivity and specificity.

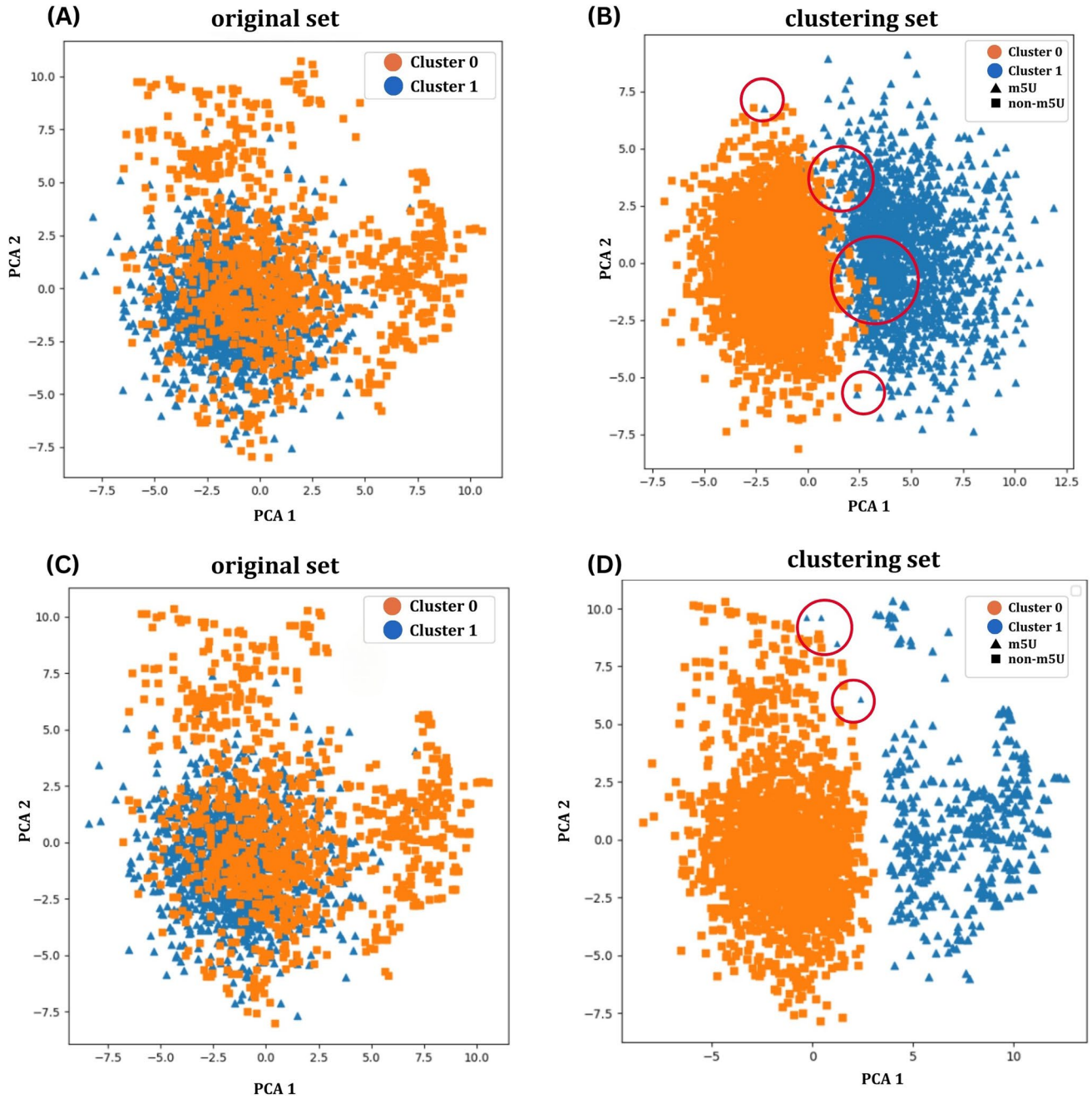

Fig. 5. PCA analysis on the training sets. (**A**) original set of the full transcript RNA (**B**) clustering of the training set. (**C**) original set of the mature RNA, (**D**) clustering set of the training set of mature RNA.

Overall, we can conclude that the proposed model was successfully equipped with balancing performance to detect the m5U site more accurately.

In Table 3, we summarized the overall performance of our models with the other existing models for the full transcript and mature RNA datasets. We compared the proposed model with the recent studies. m5U-SVM model achieved 88.87% accuracy in full transcription with 95.53% AUC score with a good level of specificity, but they achieved 81.22% sensitivity, which indicates the model could not perform better to differentiate the positive sequences. Hence the model might count the positive sequences as unfavorable sometimes, which is a serious flaw in the bioinformatics field. Another model, m5UPred has similar issues with sensitivity. The model achieved a 72.81% sensitivity score, indicating that the m5UPred model detected negative sequences more precisely than positive ones. The DeepmU model has excellent accuracy and AUC scores, but we have a balanced performance in all the evaluation metrics. GRUpred-m5U has 95.37% sensitivity, which denotes the model differentiates the positive sequences more precisely than the others, which is improved by more than 10%.

Moreover, 98.89% of AUC scores indicate that the model can distinguish between the negative and positive classes of m5U sites. In the other dataset, in terms of mature RNA, our model obtained 98.41% accuracy, which

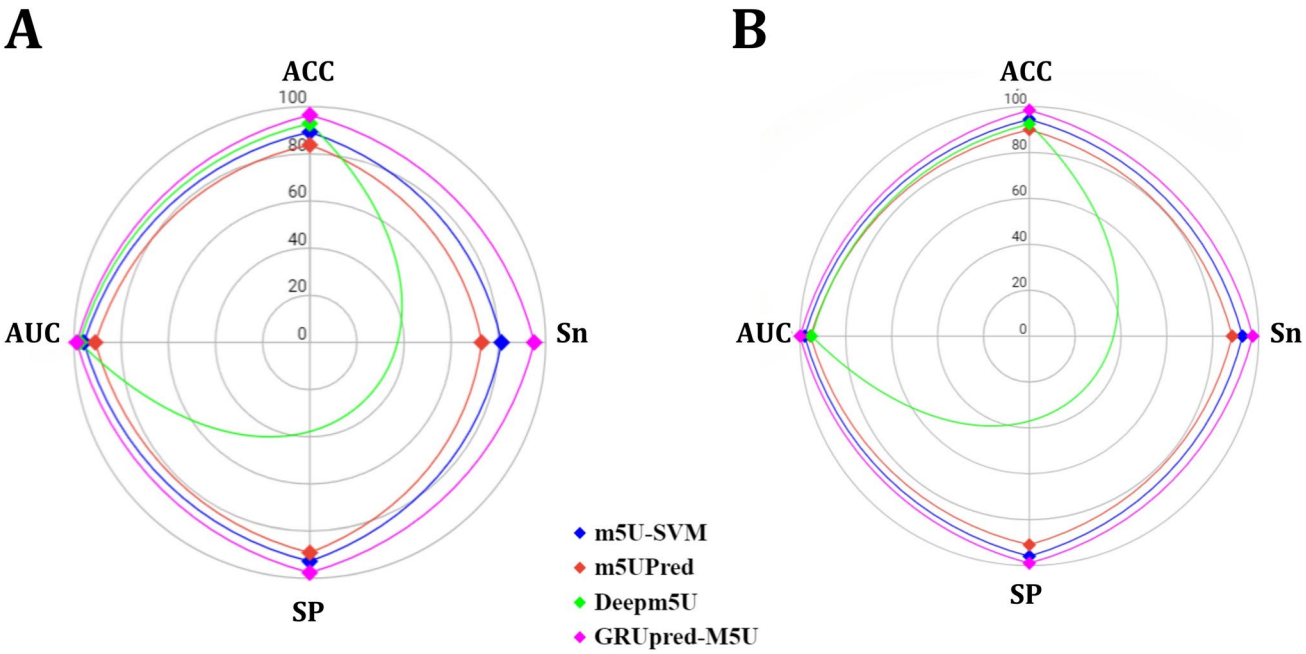

Fig. 6. Performance comparison of GRUpred-m5U with other existing approaches. (**A**) full-transcript RNA (**B**) mature RNA.

Mode	Models	ACC (%)	Sn (%)	Sp (%)	AUC (%)
Full_transcipt RNA	m5U-SVM[28]	88.87	81.22	92.97	95.53
	m5UPred[23]	83.59	72.81	89.37	91.09
	Deepm5U[26]	92.91	-	-	97.73
	GRUpred-M5U	**96.70**	**95.37**	**97.41**	**98.89**
Mature RNA	m5U-SVM[28]	94.35	92.98	95.73	98.04
	m5UPred[23]	89.91	88.64	91.18	95.60
	Deepm5U[26]	92.48	-	-	95.11
	GRUpred-M5U	**98.41**	**97.65**	**99.18**	**99.83**

Table 3. Performance comparison of the state-of-the-art with GRUpred-m5U.

is more improved than the different models, where this model obtained < 4% of the recent studies. The model could be performed to specify the positive and negative sequences of more than 5%, Accordingly, in the AUC scores, our suggested model proved that it could access the sites from the RNA sequences. According to the study, we have chosen these features because they are highly effective in discovering drugs, which may have various implications for human metabolism and diseases. This comprehension contributes to the development of medications and oligonucleotide-based therapies. Xu et al., Wang et al., and Ao et al.[28,52–54] validate that these feature groups contain valuable information for the potential drug targets and deployment of bioinformatics tools. A brief description has been included in the Supplementary file to help with comprehension of all the suitable feature groups[55,56,57,58,59,60,61,62,63].

Conclusion

The current study has proposed a deep learning-based model named GRUped-m5U to identify the m5U sites from the FASTA format RNA sequences. Three descriptor groups were employed to extract features from the RNA sequences, including five feature extraction methods: ENAC, PseDNC, DPCP, DPCP type 1, and Kmer. The proposed model achieved outstanding performance compared to the existing state-of-the-art model. A 10-fold cross-validation approach with several evaluation metrics was used to evaluate the proposed model. The model accurately predicted the m5U sites from the multi-view feature extraction methods. In addition, our methodology for identifying m5U sites has broad significance for improving the understanding of RNA structure, including in the area of epi transcriptomics, potentially offering applications in biomarker identification and therapeutic development. Although this study has some limitations in model development, particularly with the use of the GRU model, GRUs can be unnecessarily complex and computationally expensive compared to simpler models, especially when dealing with smaller datasets. In the future, we will work with larger datasets and various species. Moreover, we will work on more feature extractions, reducing the model's complexity. As the field continues to

evolve, the accurate prediction of RNA modifications becomes increasingly vital for unraveling the complexities of cellular processes and disease mechanisms.

Data availability

The dataset and the source code have been available for this study is here. https://github.com/Shazzad-Shaon3404/m5U_detection-.git.

Received: 16 May 2024; Accepted: 10 October 2024
Published online: 28 October 2024

References

1. Carlile, T. M., Rojas-Duran, M. F. & Gilbert, W. V. Pseudo-Seq: genome-wide detection of pseudouridine modifications in RNA. In *Methods in enzymology* (Vol. 560, pp. 219–245). Academic Press. (2015). https://doi.org/10.1016/bs.mie.2015.03.011
2. Li, S. & Mason, C. E. The pivotal regulatory landscape of RNA modifications. *Annu. Rev. Genom. Hum. Genet.* **15**, 127–150. https://doi.org/10.1146/annurev-genom-090413-025405 (2014).
3. Boccaletto, P. et al. MODOMICS: a database of RNA modification pathways. 2021 update. *Nucleic Acids Res.* **50** (D1). https://doi.org/10.1093/nar/gkab1083 (2022).
4. Saletore, Y. et al. The birth of the Epitranscriptome: deciphering the function of RNA modifications. *Genome Biol.* **13**, 1–12. https://doi.org/10.1186/gb-2012-13-10-175 (2012).
5. Xiong, Q. & Zhang, Y. Small RNA modifications: regulatory molecules and potential applications. *J Hematol Oncol.* **16**(1), 1–24. https://doi.org/10.1186/s13045-023-01466-w (2023).
6. Desrosiers, R., Friderici, K. & Rottman, F. Identification of methylated nucleosides in messenger RNA from Novikoff hepatoma cells. *Proc Natl Acad Sci.* **71**(10), 3971–3975. https://doi.org/10.1073/pnas.71.10.3971 (1974).
7. Oerum, S., Meynier, V., Catala, M. & Tisné, C. A comprehensive review of m6A/m6Am RNA methyltransferase structures. *Nucleic Acids Res.* **49** (13), 7239–7255. https://doi.org/10.1093/nar/gkab378 (2021).
8. Alarcón, C. R., Lee, H., Goodarzi, H., Halberg, N. & Tavazoie, S. F. N 6-methyladenosine marks primary microRNAs for processing. *Nature*. **519** (7544), 482–485. https://doi.org/10.1038/nature14281 (2015).
9. Bujnicki, J. M., Feder, M., Ayres, C. L. & Redman, K. L. Sequence–structure–function studies of tRNA: m5C methyltransferase Trm4p and its relationship to DNA: m5C and RNA: m5U methyltransferases. *Nucleic Acids Res.* **32** (8), 2453–2463. https://doi.org/10.1093/nar/gkh564 (2004).
10. Urbonavičius, J., Jäger, G. & Björk, G. R. Amino acid residues of the Escherichia coli tRNA (m5U54) methyltransferase (TrmA) critical for stability, covalent binding of tRNA and enzymatic activity. *Nucleic Acids Res.* **35** (10), 3297–3305. https://doi.org/10.1093/nar/gkm205 (2007).
11. Powell, C. A. & Minczuk, M. TRMT2B is responsible for both tRNA and rRNA m5U-methylation in human mitochondria. *RNA Biol.* **17** (4), 451–462. https://doi.org/10.1080/15476286.2020.1712544 (2020).
12. Pereira, M. et al. m5U54 tRNA hypomodification by lack of TRMT2A drives the generation of tRNA-derived small RNAs. *Int. J. Mol. Sci.* **22** (6), 2941. https://doi.org/10.3390/ijms22062941 (2021).
13. Carter, J. M. et al. FICC-Seq: a method for enzyme-specified profiling of methyl-5-uridine in cellular RNA. *Nucleic Acids Res.* **47** (19). https://doi.org/10.1093/nar/gkz658 (2019).
14. Nordlund, M. E., JOHANSSON, J. M., von Pawel-Rammingen, U. & BYSTROeM, A. S. Identification of the TRM2 gene encoding the tRNA (m5U54) methyltransferase of Saccharomyces cerevisiae. *Rna*. **6** (6), 844–860. https://doi.org/10.1017/S1355838200992422 (2000).
15. Ranaei-Siadat, E. et al. RNA-methyltransferase TrmA is a dual-specific enzyme responsible for C5-methylation of uridine in both tmRNA and tRNA. *RNA Biol.* **10** (4), 572–578. https://doi.org/10.4161/rna.24327 (2013).
16. Gu, X., Ofengand, J. & Santi, D. V. In vitro methylation of Escherichia coli 16S rRNA by tRNA (m5U54)-methyltransferase. *Biochemistry*. **33** (8), 2255–2261. https://doi.org/10.1021/bi00174a036 (1994).
17. Mathoux, J., Henshall, D. C. & Brennan, G. P. Regulatory mechanisms of the RNA modification m6A and significance in brain function in health and disease. *Front. Cell. Neurosci.* **15**, 671932. https://doi.org/10.3389/fncel.2021.671932 (2021).
18. Livneh, I., Moshitch-Moshkovitz, S., Amariglio, N., Rechavi, G. & Dominissini, D. The m6A epitranscriptome: transcriptome plasticity in brain development and function. *Nat. Rev. Neurosci.* **21** (1), 36–51. https://doi.org/10.1038/s41583-019-0244-z (2020).
19. Zhang, M., Zhai, Y., Zhang, S., Dai, X. & Li, Z. Roles of N6-Methyladenosine (m6A) in stem cell fate decisions and early embryonic development in mammals. *Front. Cell. Dev. Biology*. **8**, 782. https://doi.org/10.3389/fcell.2020.00782 (2020).
20. Delaunay, S. & Frye, M. RNA modifications regulating cell fate in cancer. *Nat. Cell Biol.* **21** (5), 552–559. https://doi.org/10.1038/s41556-019-0319-0 (2019).
21. Liang, W., Lin, Z., Du, C., Qiu, D. & Zhang, Q. mRNA modification orchestrates cancer stem cell fate decisions. *Mol. Cancer*. **19(1)**, 1–12. https://doi.org/10.1186/s12943-020-01166-w (2020).
22. Wang, Y. et al. Identification of tRNA nucleoside modification genes critical for stress response and development in rice and Arabidopsis. *BMC Plant Biol.* **17** (1), 1–15. https://doi.org/10.1186/s12870-017-1206-0 (2017).
23. Jiang, J. et al. m5UPred: a web server for the prediction of RNA 5-methyluridine sites from sequences. *Mol. Therapy-Nucleic Acids*. **22**, 742–747. https://doi.org/10.1016/j.omtn.2020.09.031 (2020).
24. Feng, P. & Chen, W. iRNA-m5U: a sequence based predictor for identifying 5-methyluridine modification sites in saccharomyces cerevisiae. *Methods*. **203**, 28–31. https://doi.org/10.1016/j.ymeth.2021.04.013 (2022).
25. Li, Z., Mao, J., Huang, D., Song, B. & Meng, J. RNADSN: transfer-learning 5-Methyluridine (m5U) modification on mRNAs from common features of tRNA. *Int. J. Mol. Sci.* **23** (21), 13493. https://doi.org/10.3390/ijms232113493 (2022).
26. Yu, L. et al. Evaluation and development of deep neural networks for RNA 5-Methyluridine classifications using autoBioSeqpy. *Front. Microbiol.* **14**, 1175925. https://doi.org/10.3389/fmicb.2023.1175925 (2023).
27. Jing, R. et al. autoBioSeqpy: a deep learning tool for the classification of biological sequences. *J. Chem. Inf. Model.* **60** (8), 3755–3764. https://doi.org/10.1021/acs.jcim.0c00409 (2020).
28. Ao, C., Ye, X., Sakurai, T., Zou, Q. & Yu, L. m5U-SVM: identification of RNA 5-methyluridine modification sites based on multi-view features of physicochemical features and distributed representation. *BMC Biol.* **21** (1). https://doi.org/10.1186/s12915-023-01596-0 (2023).
29. Akbar, S., Zou, Q., Raza, A. & Alarfaj, F. K. iAFPs-Mv-BiTCN: Predicting antifungal peptides using self-attention transformer embedding and transform evolutionary based multi-view features with bidirectional temporal convolutional networks. *Art Intell Med.* **151**, 102860 (2024).
30. Akbar, S., Raza, A. & Zou, Q. Deepstacked-AVPs: predicting antiviral peptides using tri-segment evolutionary profile and word embedding based multi-perspective features with deep stacking model. *BMC Bioinform.* **25** (1), 102. https://doi.org/10.1186/s12859-024-05726-5 (2024).

31. Ullah, M., Akbar, S., Raza, A. & Zou, Q. DeepAVP-TPPred: identification of antiviral peptides using transformed image-based localized descriptors and binary tree growth algorithm. *Bioinformatics*. **40** (5), btae305. https://doi.org/10.1093/bioinformatics/btae305 (2024).
32. Raza, A. et al. AIPs-SnTCN: Predicting anti-inflammatory peptides using fastText and transformer encoder-based hybrid word embedding with self-normalized temporal convolutional networks. *J. Chem. Inf. Model.* **63** (21), 6537–6554. https://doi.org/10.1021/acs.jcim.3c01563 (2023).
33. Akbar, S. et al. pAtbP-EnC: identifying anti-tubercular peptides using multi-feature representation and genetic algorithm based deep ensemble model. *IEEE Access.* **11**, 137099–137114. https://doi.org/10.1109/ACCESS.2023.3321100 (2023).
34. Khan, S. et al. Sequence based model using deep neural network and hybrid features for identification of 5-hydroxymethylcytosine modification. *Sci. Rep.* **14** (1), 9116. https://doi.org/10.1038/s41598-024-59777-y (2024).
35. Naeem, M. & Qiyas, M. Deep intelligent predictive model for the identification of diabetes. *AIMS Math.***8** (7), 16446–16462. https://doi.org/10.3934/math.2023840 (2023).
36. Aurangzeb, K. DBSCAN-based energy users clustering for performance enhancement of deep learning model. *J. Intell. Fuzzy Syst.* **46** (3), 5555–5573. https://doi.org/10.3233/JIFS-235873 (2024).
37. Khan, S. et al. Enhancing sumoylation site prediction: A deep neural network with discriminative features. *Life.* **13**(11), 2153 (2023).
38. Abbas, Z., Tayara, H. & Chong, K. T. ENet-6 mA: identification of 6 mA modification sites in plant genomes using ElasticNet and neural networks. *Int. J. Mol. Sci.* **23** (15), 8314. https://doi.org/10.3390/ijms23158314 (2022).
39. Abbas, Z., Rehman, M. U., Tayara, H. & Chong, K. T. ORI-Explorer: a unified cell-specific tool for origin of replication sites prediction by feature fusion. *Bioinformatics*. **39** (11), btad664. https://doi.org/10.1093/bioinformatics/btad664 (2023).
40. Khan, S., Khan, M., Iqbal, N., Khan, S. A. & Chou, K. C. Prediction of piRNAs and their function based on discriminative intelligent model using hybrid features into Chou's PseKNC. *Chemometr. Intell. Lab. Syst.* **203**, 104056. https://doi.org/10.1016/j.chemolab.2020.104056 (2020).
41. Khan, S., Khan, M., Iqbal, N., Rahman, M. A. A. & Karim, M. K. A. Deep-PiRNA: bi-layered prediction model for PIWI-interacting RNA using discriminative features. *Comput. Mater. Contin.* **72**, 2243–2258 (2022).
42. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. **28** (23), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565 (2012).
43. Chen, Z. et al. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Brief. Bioinform.* **21** (3), 1047–1057. https://doi.org/10.1093/bib/bbz041 (2020).
44. Lee, D., Karchin, R. & Beer, M. A. Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res.***21** (12), 2167–2180. https://doi.org/10.1101/gr.121905.111 (2011). http://www.genome.org/cgi/doi/
45. Manavalan, B. et al. 4mCpred-EL: an ensemble learning framework for identification of DNA N4-methylcytosine sites in the mouse genome. *Cells*. **8** (11), 1332. https://doi.org/10.3390/cells8111332 (2019).
46. Liu, B., Liu, F., Fang, L., Wang, X. & Chou, K. C. repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. *Bioinformatics*. **31** (8), 1307–1309. https://doi.org/10.1093/bioinformatics/btu820 (2015).
47. Liu, B. et al. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.* **43** (W1), W65–W71. https://doi.org/10.1093/nar/gkv458 (2015).
48. Chen, Z. et al. iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. *Nucleic Acids Res.***49** (10), e60–e60. https://doi.org/10.1093/nar/gkab122 (2021).
49. Umakantha, N. A New Approach to Probability Theory with reference to statistics and statistical physics. *J. Mod. Phys.* **7** (09), 989. https://doi.org/10.4236/jmp.2016.79090 (2016).
50. Radhika, C. & Priya, N. Prediction of learning disability of the children using adaptive effective feature Engineering techniques. *J. Posit. School Psychol.* **6** (5), 2768–2783. https://doi.org/10.1002/9781118445112.stat00365.pub2 (2022).
51. Basith, S., Manavalan, B., Shin, H., Lee, G. & T. and Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening. *Med. Res. Rev.* **40** (4), 1276–1314. https://doi.org/10.1002/med.21658 (2020).
52. Xu, Z., Wang, X., Meng, J., Zhang, L. & Song, B. m5U-GEPred: prediction of RNA 5-methyluridine sites based on sequence-derived and graph embedding features. *Front Microbiol*. 14. https://doi.org/10.3389/fmicb.2023.1277099 (2023).
53. Wang, Y. et al. RNAincoder: a deep learning-based encoder for RNA and RNA-associated interaction. *Nucleic Acids Res.* **51** (W1), W509–W519. https://doi.org/10.1093/nar/gkad404 (2023).
54. Khan, S., AlQahtani, S. A., Noor, S. & Ahmad, N. PSSM-Sumo: deep learning based intelligent model for prediction of sumoylation sites using discriminative features. *BMC Bioinform.* **25** (1), 284. https://doi.org/10.1186/s12859-024-05917-0 (2024).
55. Huang, Y., He, N., Chen, Y., Chen, Z. & Li, L. BERMP: a cross-species classifier for predicting m6A sites by integrating a deep learning algorithm and a random forest approach. *Int J Biol Sci.* **14**(12), 1669. https://doi.org/10.7150/ijbs.27819 (2018).
56. El Allali, A., Elhamraoui, Z. & Daoud, R. Machine learning applications in RNA modification sites prediction. *Comput Struct Biotechnol J.* **19**, 5510–5524. https://doi.org/10.1016/j.csbj.2021.09.025 (2021).
57. Orozco-Arias, S. et al. K-mer-based machine learning method to classify LTR-retrotransposons in plant genomes. *PeerJ*. **9**,e11456. https://doi.org/10.7717/peerj.11456 (2021).
58. Ferreira, L. M., Sáfadi, T. & Ferreira, J. L. K-mer applied in Mycobacterium tuberculosis genome cluster analysis. *Brazilian J. Biology*. **84**, e258258. https://doi.org/10.1590/1519-6984.258258 (2022).
59. Teng, Z. et al. i6mA-Vote: cross-species identification of DNA N6-methyladenine sites in plant genomes based on ensemble learning with voting. *Front. Plant Sci.* **13**, 845835. https://doi.org/10.3389/fpls.2022.845835 (2022).
60. Chen, R. et al. ATTIC is an integrated approach for predicting A-to-I RNA editing sites in three species. *Brief. Bioinform.* **24** (3), 170. https://doi.org/10.1093/bib/bbad170 (2023).
61. Chen, Z. et al. iFeatureOmega: an integrative platform for engineering, visualization and analysis of features from molecular sequences, structural and ligand data sets. *Nucleic Acids Res.* **50** (W1), W434–W447. https://doi.org/10.1093/nar/gkac351 (2022).
62. Chen, W., Feng, P. M., Lin, H. & Chou, K. C. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res.* **41**(6), e68. https://doi.org/10.1093/nar/gks1450 (2013).
63. Zheng, L. et al. RAACBook: a web server of reduced amino acid alphabet for sequence-dependent inference by using Chou's five-step rule. *Database.* **2019**, baz131. https://doi.org/10.1093/database/baz131 (2019).

Acknowledgements

This work was supported in part by funding from the Natural Sciences and Engineering Research Council of Canada (NSERC).

Authors' contributions

Conceptualization, M.M. Ali, K. Ahmed; Data curation, Formal analysis, Investigation, M.M. Ali, M.S.H. Shaon, T. Karim; Methodology, M.M. Ali, K. Ahmed, F.M. Bui, M.A. Moni; Project administration, M.M. Ali, K. Ahmed, F.M. Bui; Resources, Software, M.M. Ali, K. Ahmed; Supervision, Validation, M.M. Ali, K. Ahmed, F.M. Bui; Visualization, M.M. Ali, K. Ahmed; Funding, F.M. Bui, L. Chen; Writing - original draft, Writing - review

editing, M.S.H. Shaon, T. Karim, M.M. Ali, K. Ahmed, F.M. Bui, L. Chen, M.A. Moni;. All authors have read and approved the final version of the manuscript.

Funding
This work was supported in part by funding from the Natural Sciences and Engineering Research Council of Canada (NSERC).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-76148-9.

Correspondence and requests for materials should be addressed to K.A. or M.A.M.

Reprints and permissions information is available at www.nature.com/reprints.

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