BMC Bioinformatics

RESEARCH Open Access

Deep-m5U: a deep learning-based approach for RNA 5-methyluridine modi□cation prediction using optimized feature integration

Sumaiya Noor[1], Afshan Naseem[2], Hamid Hussain Awan[3], Wasiq Aslam[3], Salman Khan[4], Salman A. AlQahtani[4] and Nijad Ahmad[5*]

*Correspondence:
Nijad@khurasan.edu.af

[1] Business and Management Sciences Department, Purdue University, West Lafayette, IN, USA
[2] Institute of Oceanography and Environment (INOS), Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia
[3] Department of Computer Science, Muslim Youth University, Islamabad, Pakistan
[4] New Emerging Technologies and 5G Network and Beyond Research Chair, Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia
[5] Department of Computer Science, Khurasan University, Jalalabad, Afghanistan

Abstract

Background: RNA 5-methyluridine (m5U) modifications play a crucial role in biological processes, making their accurate identification a key focus in computational biology. This paper introduces Deep-m5U, a robust predictor designed to enhance the prediction of m5U modifications. The proposed method, named Deep-m5U, utilizes a hybrid pseudo-K-tuple nucleotide composition (PseKNC) for sequence formulation, a Shapley Additive exPlanations (SHAP) algorithm for discriminant feature selection, and a deep neural network (DNN) as the classifier.

Results: The model was evaluated using two benchmark datasets, i.e., Full Transcript and Mature mRNA. Deep-m5U achieved overall accuracies of 91.47% and 95.86% for the Full Transcript and Mature mRNA datasets with 10-fold cross-validation, and for independent samples, the model attained 92.94% and 95.17% accuracy.

Conclusion: Compared to existing models, Deep-m5U showed approximately 5.23% and 3.73% higher accuracy on the training data and 3.95% and 3.26% higher accuracy on independent samples for the Full Transcript and Mature mRNA datasets, respectively. The reliability and e□ectiveness of Deep-m5U make it a valuable tool for scientists and a potential asset in pharmaceutical design and research.

Keywords: RNA 5-methyluridine, PseKNC, Deep learning, Sequence-derived features, SHAP

Introduction

In recent years, the achievement of measurement technologies mapping for RNA has significantly propelled research into RNA epigenetic modifications. Currently, over 170 chemicals identified have been modifications in cellular RNA, with notable examples including 5-methylcytosine (m5C), 5-hydroxymethylcytosine (5hmC), N6-methyladenosine (m6A), N1-methyladenosine (m1A), 2'-O-methylation of ribose (2'-O-Me) and pseudouridine (Ψ) [1]. Lesser-known modifications such as 7-methylguanosine (m7G), adenosine-to-inosine (A-to-I), dihydrouridine (D), N2-methylguanosine (m2G), and N4-acetylcysteine have also been identified [2]. ese modifications can a ect all four

RNA bases, adenine (A), cytosine (C), guanine (G), and uracil (U), as well as the ribose sugar. While nearly all RNA species undergo modification, transfer ribosomal RNA (rRNA) and RNA (tRNA) are the most heavily modified [3]. More than 100 mutations in RNA-modifying enzymes have been linked to human illnesses, underscoring these modifications' critical role in regulating gene expression and protein production [4]. Moreover, RNA modifications, like post-translational protein modifications, serve dynamic cellular functions by modulating cell-specific activities. Current developments in detecting RNA modifications have been simplified by two principal methods: next-generation sequencing (NGS) and liquid chromatography coupled with mass spectrometry (LC–MS). LC–MS is highly sensitive and specific but cannot provide sequence context [5]. In contrast, NGS o ers comprehensive sequence information but struggles to detect modifications directly. is limitation arises because RNA modifications often interfere with reverse transcription, introducing errors or blockages during sequencing [6]. ese technological advancements are crucial for further understanding RNA modifications' complexity and biological significance [7, 8].

e 5-methyluridine (m5U) modification is a significant epigenetic mark that has drawn global attention from researchers. It is commonly found in cytosolic tRNAs and other non-coding RNAs like mRNA and rRNA [9]. Various enzymes, such as TrmA in *E.* coli, *Trm2* in S. cerevisiae, and TRMT2A and TRMT2B, catalyze m5U modifications [10]. is modification is critical to RNA structures, but the conserved T-loop motif is essential in alleviating the secondary structure of RNAs [10, 11]. Despite being among the most prevalent RNA changes, there is still limited research on identifying and understanding the functions of m5U. Methylation of uridine at its fifth carbon, carried out by specific enzymes, may have been among the first pyrimidine methyltransferases to evolve. m5U is linked to diseases like breast cancer and lupus, as well as plant development and stress response. Accurately pinpointing m5U sites is crucial for understanding its biological role. Still, experimental methods like miCLIP-Seq and iCLIP are costly and time-consuming, often yielding limited data due to antibody specificity issues [12].

To address the challenges of detecting RNA modifications, researchers have proposed using machine learning (ML) algorithms [13–15]. Experimentally identifying all RNA modifications is expensive and complicated, so computational methods are now widely used. Various ML-based tools have been developed to predict RNA modifications, such as (WHISTLE [16], SRAMP [17], iRNA-Methyl [18], RNA modification [19], m7GHub V2.0 [20], DirectRMDB [21], MODOMICS [22], ConsRM [23]), m5C site predictors (iRNA-m5C [24], RNADSN [25], pseudouridine predictors (iRNA-PseU, PPUS [26] and RNAm5Cfinder [27]). A few computational and experimental techniques have been created to improve the identification of m5U-modified sites and to handle the complexity of RNA modifications. One of the prominent tools is m5UPred, introduced by Jiang et al. [28], which uses an SVM algorithm. is method uses sequence-derived features like nucleotide concentration and chemistry to predict m5U sites in human RNA. While m5UPred achieved a respectable accuracy of 83.60% for Full Transcripts and 89.91% for Mature mRNA in cross-validation tests, it faced challenges like overfitting. It was limited by the quality of its dataset, which wasn't redundantly processed.

Ao et al. [29] proposed the m5U-SVM model to address these limitations, enhancing prediction by merging distributed representation characteristics with traditional

physicochemical features. is model was also based on SVM and used a multi-view feature approach, enhancing its ability to di erentiate between m5U and non-m5U sites. Under tenfold cross-validation, m5U-SVM showed improved performance, with an estimated average accuracy of 88.876% for Full Transcripts and 94.358% for Mature mRNA. Despite these promising results, both models still rely on conventional learning methods that struggle to accurately predict m5U-modified sites, primarily due to the resemblance among m5U and non-m5U sites.

is paper presents a robust computational predictor named Deep-m5U, designed to accurately identify 5-methyluridine (m5U) modifications, leveraging e ective feature extraction techniques. e model utilizes pseudo-k-tuple nucleotide compositions, which are grouped into parts such as single nucleotide composition (SNC), dinucleotide composition (DNC), trinucleotide composition (TNC), quad nucleotide composition (QNC), and penta nucleotide composition (PNC) to establish robust and intricate patterns from the RNA sequence. ese features are further enhanced by using structural and global sequence-order information and converting an RNA sequence into a feature vector. ese vectors are then followed by the integration process to develop a new feature set, which is a blended one, and it helps the model to represent the biological data more e ciently. e SHAP is set to improve this feature set, which enforces removing noisy and irrelevant features from the set of features to keep only the most discriminative ones for the classification. Furthermore, classifier training uses a DNN after selecting the relevant features under consideration. e model was validated through a tenfold cross-validation process by the benchmark data of Mature mRNA and Full Transcript data sets. e results were outstanding, with the identified accuracies of 91.47% and 95.86%, respectively. Also, it showed that the accuracy of the proposed model, Deep-m5U, is higher than that of existing predictors with enhanced performance in all the evaluation criteria. erefore, it can be used to predict RNA modification sites e ectively. Figure 1 illustrates the structure of the suggested model. e main outcomes of this study are as follows:

1. A novel deep computational model for predicting 5-methyluridine (m5U) modifications.
2. Integrating an optimized feature extraction process using nonlinear activation functions and multi-layer architectures to address dataset complexities.
3. A thorough assessment of model e cacy using both validation and testing datasets.
4. Superior performance in contrast to existing m5U predictors, as measured by various assessment parameters.

Materials and methods

Benchmark dataset

Datasets are essential for developing accurate and e ective machine-learning models for predicting and detecting RNA 5-methyluridine (m5U) modification sites. ese types of datasets, also known as Benchmark datasets, are essential to test the e ciency and capability of the ML algorithms. In this research, we used benchmark datasets that were used by [28, 30, 31]. We created a new sequence of forty-one nucleotides for the positive

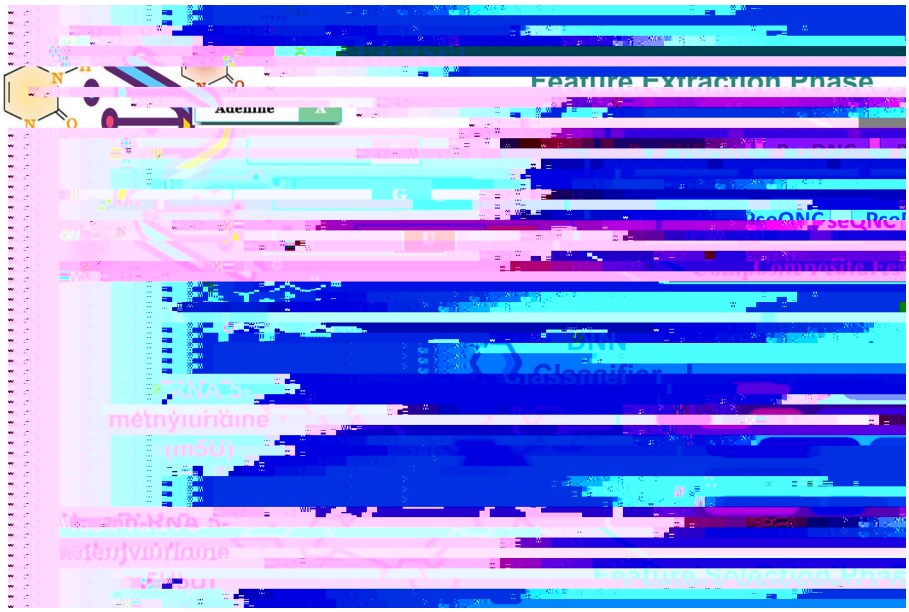

Fig. 1 Proposed model framework

Table 1 Benchmark dataset samples of RNA 5-methyluridine

Dataset	Positive sequences	Negative sequences
Full transcript	1,534	2,862
Full transcript-independent	500	731
Mature mRNA	983	985
Mature mRNA-independent	245	247

samples based on the sequence containing the experimentally defined modified uridine sites. e negative instances were randomly sourced from the unaltered uridine sites of the identical transcripts used for the positive samples. ese two databases for Full Transcript were the set of 3,696 positive and 3,696 negative samples in Mature mRNA databases, including 1,232 positive and 1,232 negative samples. At first, we did not exclude homologous sequences for the corresponding genomic regions, which could be problematic if sequence identity exceeded a certain threshold. We used the CD-HIT method to eliminate homologous sequences from the complete transcript m5U modification site data to tackle this issue, and an 80% homology similarity was used to enhance the data quality. Table 1 shows the details of the final m5U modification site dataset.

Feature formulation

Several approaches are designed to encode DNA, proteins, and RNA samples into mathematical structures with structural features of nucleotides maintained [6, 32–35]. ese approaches enable bioinformaticians to transform RNA sequences into di erent statistical features to retain uniqueness and patterns within the sequence [36]. e Pseudo Amino Acid Composition (PseAAC) technique fundamentally represents protein samples with a

discrete approach. Because of its e ciency, the PseAAC method was further developed to predict the RNA and DNA molecules, called the Pseudo k-tuple Nucleotide Composition (PseKNC) approach [37]. rough the series of transformations possible with this feature formulation technique, PseKNC has emerged as a prominent method for modeling RNA and DNA sequences in computational biology [38, 39]. Using the PseKNC method, RNA sequences are converted to the corresponding feature vectors, allowing for a representation of the RNA secondary structure without compromising the sequence order as much as possible. It makes the data after transformation similar within the RNA samples, therefore appropriate for use in di erent machine learning procedures [40]. To this end, all RNA sequences were encoded using the PseKNC method. RNA primary sequences were transformed into numerical feature vectors where the order of the nucleotides is maintained. For example, consider an RNA sequence N with L nucleotides, represented as:

$$N = N_1 N_2 N_3 ..... N_i .... N_L \tag{1}$$

$$N_i \in \{A, U, C, G\} \quad (i = 1, 2, 3, ..., L) \tag{2}$$

where L stands for the total number of nucleotides in an RNA sequence, which is the length of the sequence, in the above Eq. 2, the symbol A represents the Adenine nucleotide base, U is a Uracil nucleotide base, symbol G represents Guanine nucleotide base and the symbol C stands for Cytosine nucleotide base.N_i represents the nucleotide that occurs in the ith position concerning the sequence. Let's expand the general form of PseKNC representation discussed in [41] for the context of the sample presented in Eq. 1 below:

$$N = [\varphi_1 \varphi_2 \varphi_3 ..... \varphi_u ..... \varphi_z]^T \tag{3}$$

e transposed vector T, the numeric value z, and φ_u representing the actual value of the RNA sequence's function vector may be computed using Eq. 4.

$$\varphi_{\mathbf{u}} = \frac{\frac{\mathbf{f_u^{K-tuple}}}{\sum_{\mathbf{i}=1}^{4^{\mathbf{k}}} \mathbf{f_u^{K-tuple}} + \mathbf{w} \sum_{\mathbf{j}=1}^{\lambda} \theta_{\mathbf{j}}} (1 \le \mathbf{u} \le 4^{\mathbf{k}}, \mathbf{u} = 1, 2, 3, ...)}{, \frac{\mathbf{w}\theta_{\mathbf{u}-4^{\mathbf{k}}}}{\sum_{\mathbf{i}=1}^{4^{\mathbf{k}}} \mathbf{f_u^{K-tuple}} + \mathbf{w} \sum_{\mathbf{j}=1}^{\lambda} \theta_{\mathbf{j}}} (4^{\mathbf{k}} + 1 \le \mathbf{u} \le 4^{\mathbf{k}} + \lambda)} \tag{4}$$

where θ_j stands for the jth tier correlation factor or rank; this exhibits the correlation of the sequence order of the consecutive K-tuple nucleotides in RNA sequence. e parameter λ represents the overall degrees of correlation ranks to be accounted (or tiers) for through a weight factor w. e impact of the correlation factors is stabilized. eoretical and simulation work has shown that $\lambda = 1$ and $w = 0$ are the best values for the two parameters. 1, which, in turn, provides the best performance results. e correlation factor θ_j can be calculated as follows:

$$\theta_j = \left\{ \frac{1}{L - K - (\lambda - 1)} \sum_{i=1}^{L-K-(\lambda-1)} C_{i,i+j} \quad j \to 1, 2, ..., \lambda, \lambda < L - K \right. \tag{5}$$

where, $C_{i,i+j}$ is the correlation function and can be computed using Eq. 6.

$$C_{i,i+j} = \frac{1}{u}\sum_{\xi=1}^{\lambda}\left[H_{\xi}(N_iN_{i+1}...N_{i+K-1}) - H_{\xi}(N_{i+j}N_{i+j+1}...N_{i+j+K-1})\right]^2 \quad (6)$$

where N_i represents any nucleotide (Ref. Equation 1), $H_{\xi}(N_iN_{i+1}...N_{i+K-1})$ is the numerical value of the ξth physicochemical property of $N_iN_{i+1}...N_{i+K-1}$ (K-tuple nucleotide) in the RNA sequence, and $H_{\xi}(N_{i+j}N_{i+j+1}...N_{i+j+K-1})$ is the corresponding value for the K-tuple nucleotide.$N_{i+j}N_{i+j+1}...N_{i+j+K-1}$.

Hybrids features

In this study, every RNA sequence was written down as an isolated function vector using the PseKNC method. Analyzing the five feature extraction methods of the PseKNC, in which K varied from 1 to 5, we demonstrated that the PseKNC has five distinct functional modes. ese modes correspond to di erent compositions: PseSNC: the proportion for single nucleotide composition; PseDNC: the proportion for dinucleotide composition; PseTNC: the proportion for trinucleotide composition; PseQNC: the proportion for quad nucleotide composition; PsePNC: the proportion for penta nucleotide composition. e number of features generated for each mode is indicated in Table 2 below. To construct a comprehensive hybrid feature vector, we combined all five feature vectors as follows:

$$\mathbf{N_{m5U}} = \mathbf{N_{PseSNC}} \cup \mathbf{N_{PseDNC}} \cup \mathbf{N_{PseTNC}} \cup \mathbf{N_{PseQNC}} \cup \mathbf{N_{PsePNC}} \quad (7)$$

where N_{m5U} denote the hybrid feature vector, $\cup$ denotes the union, N_{PseSNC}, N_{PseDNC}, N_{PseTNC} N_{PseQNC} and N_{PsePNC} are the individual feature vectors and defined as follows:

$$N_{PseSNC} = \left|f_{j=1,...4D}^{1-Tuple} \xrightarrow{f} (A, C, G, U)\right. \quad (8)$$

$$N_{PseDNC} = \left|f_{j=1,...16D}^{2-Tuple} \xrightarrow{f} (AA, CC, GG, UU)\right. \quad (9)$$

$$N_{PseTNC} = \left|f_{j=1,...64D}^{3-Tuple} \xrightarrow{f} (AAA, CCC, GGG, UUU)\right. \quad (10)$$

$$N_{PseQNC} = \left|f_{j=1,...256D}^{4-Tuple} \xrightarrow{f} (AAAA, CCCC, GGGG, UUUU)\right. \quad (11)$$

Table 2 The total number of pseKNC features from k = 1 to k = 5

Methods	No. of features
PseSNC	4
PseDNC	16
PseTNC	64
PseQNC	256
PsePNC	1024
Hybrid feature	1364

$$N_{PsePNC} = \left| f_{j=1,\ldots 1024D}^{5-Tuple} \xrightarrow{f} (AAAAA, CCCCC, GGGGG, UUUUU) \right. \tag{12}$$

SHAP features selection

Decoding the biological import of selected features in machine learning models is sometimes problematic since these algorithms are known as black boxes, and their internal workings are complex to understand [42]. Another critical idea in machine learning is the data shape; it involves aspects like the organization, size, and arrangement of datasets utilized in a classification or regression function. Some behavioral patterns are exhibited by a machine learning algorithm based on the shape of the data sets it consists of. It is beneficial during data partition, such as dividing data into training, testing datasets, data normalization, and feature selection. Data cleaning is crucial because when data is well structured, it can perform optimally, hence the basis for decision-making. Through cooperative game theory, SHAP can provide a solution to explain the contributions of 'input features' present in a model [43]. SHAP scores each feature, and this numeric value encodes how informative that feature is to resulting decisions. The approach computes the prediction variation when a particular characteristic is included or excluded and quantifies its effect on the model. This incremental effect is mathematically formalized through Eq. 13, which points out how feature i impacts the result when interacting with different components of features.

$$SHAP_i(x) = \phi_i = \sum\nolimits_{s \subseteq N \setminus \{i\}} \frac{|S|(|N| - |S| - 1)}{|N|} [f(S \cup \{i\}) - f(S)] \tag{13}$$

where:

- ϕ_i, denotes the SHAP value for the feature i.
- N, represents the set of all features.
- S, is a subset of features excluding feature i.
- $f(S)$ is the model's prediction given the features in S.
- $f(S \cup \{i\})$ is the model's prediction given the features in S and feature i.

In this study, we use BorutaSHAP-based wrapper feature selection to identify the most influential features from the extracted vector, as it evaluates the contribution of each feature to model performance. BorutaSHAP enhances the training process by highlighting the global importance of features and facilitating the selection of the optimal feature set. For our model, we selected the top 125 features for the Full Transcript dataset and 80 features for the Mature mRNA from a hybrid feature vector with a total dimension of 1364 from both datasets. Figure 2 (a-b) presents the summarized BorutaSHAP plots for the top features, where each row represents a chosen feature. Red points indicate high-contributing features, while blue points signify those with lower contributions. The horizontal axis shows the SHAP values, where positive values push the prediction towards m5U, and negative values predict the non-m5U class.

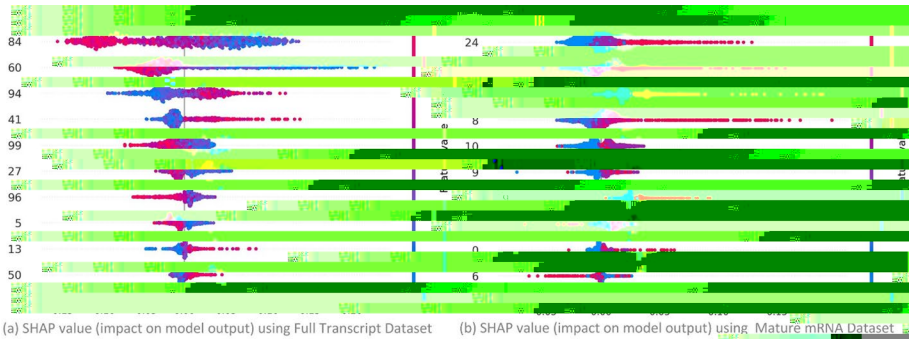

Fig. 2 Feature selection via SHAP analysis

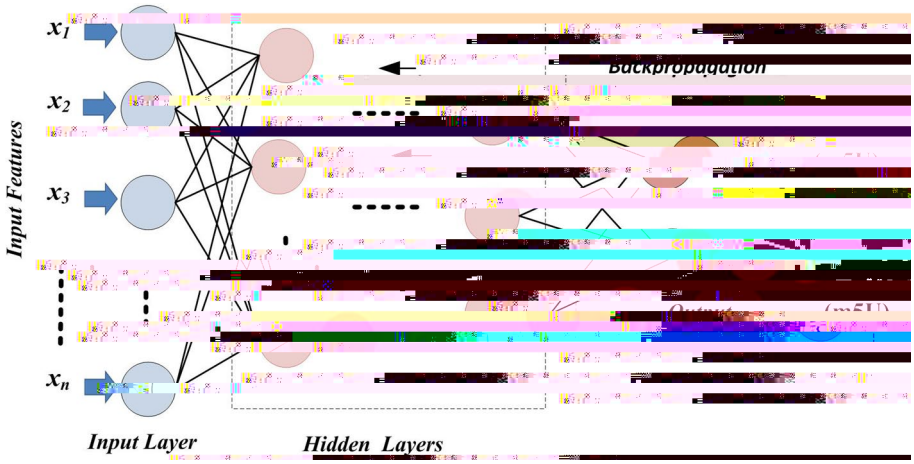

Fig. 3 Neuron representation in deep neural network configuration

Deep neural network architecture

Deep Neural Networks (DNNs) are a sub-classification of machine learning inspired by the structure and functionality of the human brain. DNN architecture involves an input layer, several hidden layers, and an output layer in between, as shown in Fig. 3.

e hidden layers are essential for the network to learn about features and patterns in data that it can't detect in the raw data. Whereas the number of hidden layers increases the predictive power to map complex patterns, it also increases the difficulty, computational costs, and over-fitting. Feature extraction is one of the most prominent advantages of DNNs. ey do not need any feature engineering of the data since they can learn the features from the data independently, even if the data is unlabeled or su ers from unstructured data. As pointed out in [44], this capability is realized through standard learning methods. In this work, we will employ regular learning approaches. Experts have proved that DNNs are more e ective in addressing complex classification problems than previous machine learning techniques because of their depth and flexibility. DNNs have been extensively used in several domains, including bioengineering [45], speech recognition, image recognition [35], and natural language processing [33]. DNNs show that they are equally e cient in such fields, pointing to their potential to solve numerous complex issues.

Model training

Based on a benchmark dataset, this study used the DNN model to identify m5U sites. From the complete sequences, 80% of the samples were used as the training dataset, while the remaining 20% were utilized as an independent test dataset. Consequently, the m5U modification site dataset consisted of 1,534/2,862 samples for training and 500/731 samples for an independent test set in the Full Transcript mode. 983/985 samples were for the training set, and 245/247 samples were for the independent test set in the Mature mRNA mode. e proposed multi-layer DNN model comprises an input–output layer and four hidden layers, as shown in Fig. 4 above. As with the previous novel architecture, each layer has multiple neurons, and the inputs and outputs correspond to the feature vectors shown in Eq. 14. e weights stored at each neuron are set by the Xavier initialization method [46], ensuring that the variance is well-conserved and that practical learning is promoted across the layers. In order to improve the learning technique of the model, a backpropagation algorithm was adopted to change the weights iteratively, enabling the reduction of errors between the output and target classes. e hyperbolic tanh (Tanh) activation function is used in both the input and hidden layers to incorporate nonlinearity into the developed model. is activation function enables the network to capture intricate patterns and the presence of relationships within data to decide whether a neuron should be activated because of the output generated. When measuring in the output layer, the activation function applied here is the softmax activation function. Since the probabilities of classifying the points or samples into an individual class, the values obtained are probabilities from 0 to 1. is approach ensures that the DNN model can develop ballistic meanings of the predictions, improving the accuracy of the number of m5U sites it can identify.

$$y_a = f(B_a + \sum_{b=1}^{m} x_b w_b^a) \tag{14}$$

where y_a denote output at a layer, B_a denote bias value, w_b^a represent weight used at a layer b by a neuron, x_b denote input feature, and f denote a nonlinear activation Tanh function, which can be calculated using Eq. 15.

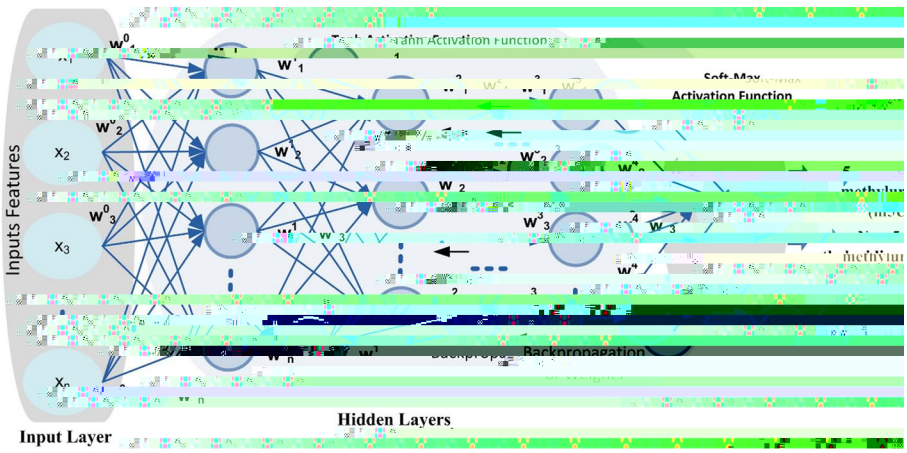

Fig. 4 RNA m5U sites proposed DNN model configuration

$$f(i) = \frac{e^i}{1 + e^i} \tag{15}$$

Evaluation of performance

Several performance indicators are usually employed during the assessment of the output by the machine learning algorithm to determine the performance and reliability of the model. (1) Accuracy (ACC) is formulated to estimate the classifier's precision of instances and is calculated by actual instances, including true positives and negatives, divided by total instances [47–49]. is o ers an overall picture of the model's e ectiveness in achieving each objective. (2) Sensitivity (SN) or actual positive rate or recall rate, which measures the capability of the classifier in identifying the correct attributes for the particular positive instances [50–54]. is one is for assessing the probability that, indeed, the test will correctly identify subjects with the condition it is looking for and is computed by dividing the number of test subjects that are correctly identified by the disease by the total number of test subjects that have the disease plus those that are erroneously identified as having the disease. is is important, especially when it is relevant to identify positive instances, as in medical diagnosis or fraud detection, where it would be highly undesirable to 'overshoot' a favorable instance. Altogether, the proposed metrics provide a holistic picture of the model's performance, its accuracy, and, at the same time, its ability to filter the right instances [55]. (3) SP stands for Specificity because a negative rate calculates the antagonistic classes the classifier has recognized. (4) Mathew's Correlation Coe cient measures binary classification consistency [13–15]. e performance metrics can be expressed mathematically as follows:

$$ACC = 1 - \frac{m^5 u_-^+ + m^5 u_+^-}{m^5 u^+ + m^5 u^-} \tag{16}$$

$$SN = 1 - \frac{m^5 u_+^-}{m^5 u^-} \tag{17}$$

$$SP = 1 - \frac{m^5 u_-^+}{m^5 u^+} \tag{18}$$

$$MCC = \frac{1 - (\frac{m^5 u_-^+ + m^5 u_+^-}{m^5 u^+ + m^5 u^-})}{\sqrt{(1 + \frac{m^5 u_-^+ + m^5 u_+^-}{m^5 u^+})(1 + \frac{m^5 u_-^+ + m^5 u_+^-}{m^5 u^-})}} \tag{19}$$

According to the above Equation, the variables represent the positive and negative values in the above-given equations. Where $m^5 u^+$ represent True Positive, $m^5 u^-$ True Negative, and $m^5 u_-^+$ False Positive and $m^5 u_+^-$ represent False Negative accordingly.

Discussion and experimental analysis

is section evaluates and discusses the proposed model's e ectiveness in depth. Several validation tests, including the K-fold and independent tests, can be utilized to assess the overall performance of the machine learning training algorithm in bioinformatics. e K-fold cross-validation approach is a typical validation technique that uses evenly balanced findings [56]. Consequently, a tenfold cross-validation test employing such benchmarking datasets was used to examine the overall accuracy of the suggested prescription in this work.

System con guration

To experiment, we used the sixth-generation Intel Core i5 processor, an average desk work option that confidently performs its functions, such as data processing and basic computing tasks. SSD 256-GB, booting, reading, and writing speeds and application performance are much better than what HDD could provide. e system configuration also includes the 8 GB of RAM, which achieves a good level of multitasking. Typical Python 3 libraries such as Numpy and Scipy, common in data science workflows, were pre-installed onto the system for training and testing ML models. We also included Tensorflow and Keras [57] for building deep neural networks and Pandas and Matplotlib to do heavy work with data analysis, cleaning, and collating data for running machine learning models. is setup is well-suited for a data-centric individual or small members-focused team. Advanced tasks associated with larger datasets or resource-intensive tasks may require enhancement of the CPU and RAM. e overall system configuration, consisting of HP Core i5 6th generation, 256 GB SSD, and 8 GB RAM, is presented in Table 3.

Analysis of nucleotide composition

In this section, we utilized the Two-Sample Logo software [44] to compare sequences with and without m5U modifications and determine if nucleotides with m5U modification sites di er in composition. is approach generated two-sample logos highlighting areas where residues are significantly enriched or reduced in m5U-modified sequences.

e statistical analysis, performed using a t-test with a significance threshold of $p < 0.05$, revealed notable di erences, as illustrated in Fig. 5.

e di erences in complete transcript mode sequences with and without m5U modifications are illustrated in Fig. 5-A, and Fig. 5-B shows the Full Transcript and Mature

Table 3 The system configuration, including software and hardware

System	Dell Core i5 6th generation
SSD	256 GB
RAM	8 GB
Language	Python 3
Framework	Tensorflow, keras, pandas, matplotlib

mRNA sequences, respectively. Regarding the consensus motifs, both modes had a conservative UUC at the positions 0–2. We observed specific nucleotide enrichments toward the regions containing m5U modification sites. For instance, in both modes, G was mainly found enriched with a ratio of 35 at position -8 and 50 at position -3. C primarily was concentrated at positions 7 and 8 while U was at position 1. Furthermore, another example of the homopolymer and homopolymeric region is found in G (positions -10 to -8) and C (positions 6 to 9). e analysis also revealed compositional di erences between sequence types: C was enriched at position nineteen in the Full Transcript mode, while G was enriched in the Mature mRNA mode. In particular, when C Sch is at position -1, Full Transcript mode enriched A and Mature mRNA mode gave a higher concentration of G. us, nucleotide deviation analysis can predict RNA m5U modification sites.

Hyper parameters and optimizations

In this section, we intend to find the best values for the hyperparameters in the DNN model. We used a grid search algorithm [58] to assess DNN performance under di erent configurations [59, 60]. We noticed that the values of some parameters with the potential to improve DNN's performance were stochastic[61, 62]. We included the following parameters in the grid search algorithm: activation function, learning rate, and number of iterations. Based on the results, Table 4 identifies a set of the best-obtained hyperparameter values.

We ran experiments to evaluate how di erent activation functions and learning rates impact performance. e results, shown in Table 5, include tests using ReLU, Sigmoid, and Tanh as activation functions, with learning rates ranging from 0.1 to 0.3. According to the table, the DNN classifier achieved the highest accuracy, 91.47% for the Full Transcript dataset and 95.86% for the Mature mRNA dataset when using Tanh as the activation function and a learning rate of 0.1.

Table 5 shows that a reduction in the learning rate results in an equal enhancement of the accuracy of the DNN model. However, increasing the learning rate to less than 0. 1 did not produce much higher increases in accuracy . erefore, we can also state that regarding the value of the learning rate, the DNN model reached the maximum

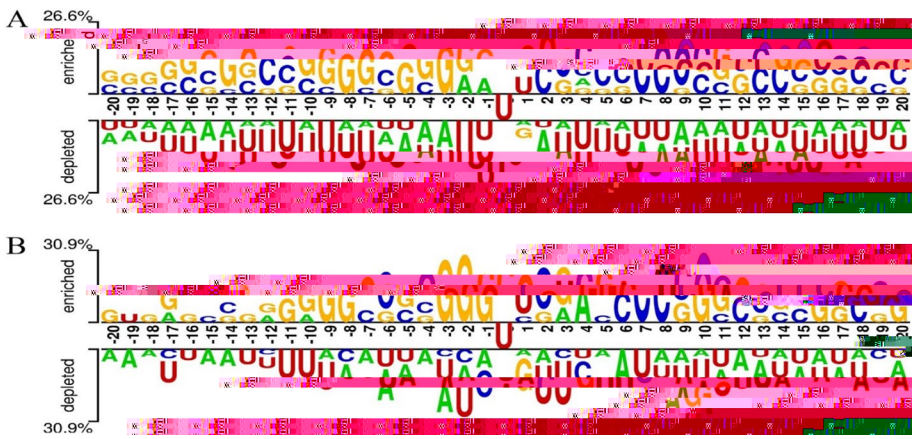

Fig. 5 m5U sites nucleotide composition analysis

Table 4 List of optimal hyper-parameters value of proposed DNN model

List of parameters	Optimal values
Activation functions	Tanh and softmax
Dropout	0.25
Number of hidden layers	4
Regularization l2	0.001
Learning rates	0.1
Number of neurons at hidden layers	70–45-21–6, 68,52,18,4,
Optimizer	SGD Method
Updater	ADAGRAD function
Weight initialization function	XAVIER function
Seed	12345L
Training epoch	50, 30
Momentum	0.9

Table 5 Impact of di□erent learning rates and activation functions on the accuracy of DNN model using tenfold

Dataset	LR	Tanh (%)	Sigmoid (%)	ReLU (%)
Full transcript	0.1	91.47	90.89	89.54
	0.2	91.22	90.31	89.01
	0.3	91.01	90.25	88.21
Mature mRNA	0.1	95.86	93.78	94.71
	0.2	95.21	93.08	94.31
	0.3	94.97	92.93	94.01

state of accuracy at the level of 0. 1 when using the Tanh activation function, which is expected. Details of the best hyperparameters for a few of the main ones are summarized in Table 4 below.

Next, we conducted numerous experiments to evaluate the DNN model's performance by varying the number of training epochs. e findings are illustrated in Figs. 6 and 7.

e data shows that the error rate consistently decreases as training epochs increase. For instance, in Fig. 6, which represents the Full Transcript dataset, the DNN model started with an error loss of 1.287 at the initial epoch, steadily dropping to 0.004 by the 50^{th} epoch. Similarly, Fig. 7 shows the results for the Mature mRNA dataset, where the initial error loss was 1.241 and reduced to 0.003 after the 30^{th} epoch. From these results, we can conclude that 50 epochs for the Full Transcript dataset and 30 epochs for the Mature mRNA dataset are optimal, as the error rates stabilize at these points. e optimal configuration derived from this analysis is summarized in Table 4.

Performance analysis using sequence formulation techniques

In this section, we evaluate the proposed model using di erent sequence formulation methods, summarised in Table 6, using the Full Transcript dataset. As presented in Table 6, the findings show that the best performance was realized whenever feature combination or hybrid features were employed instead of the individual features methods.

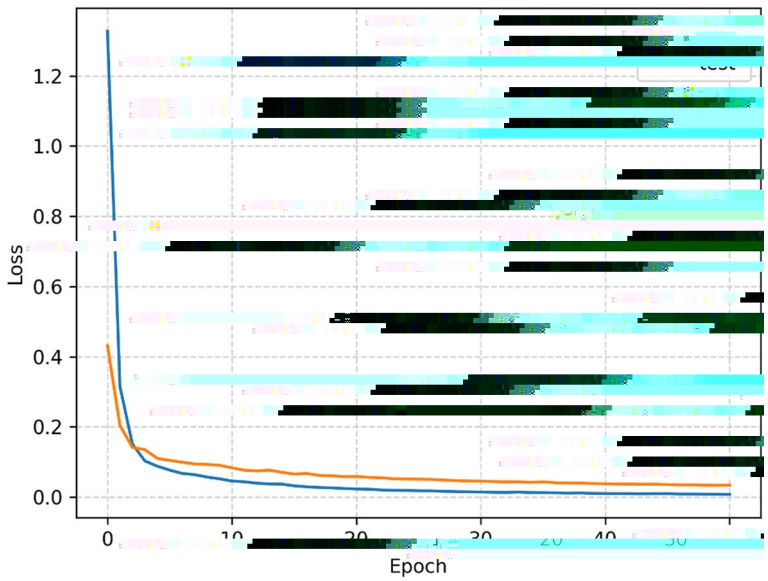

Fig. 6 Error loss on the full transcript dataset using the tanh activation functions

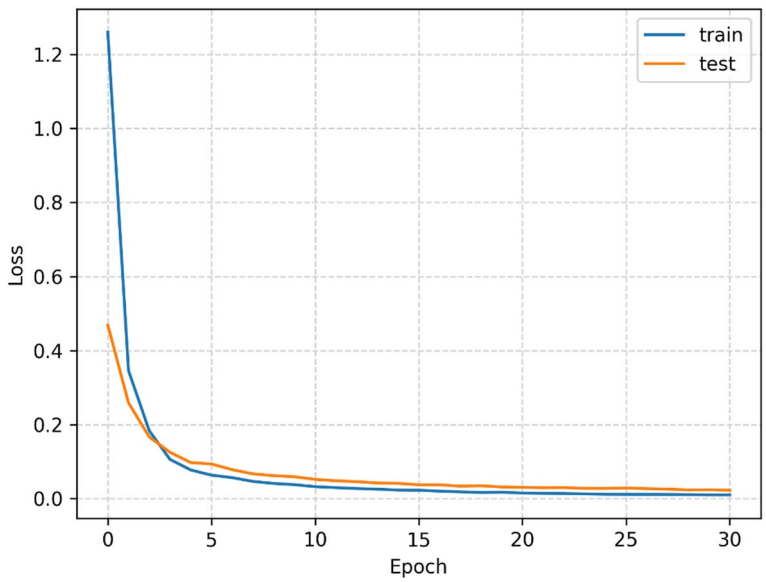

Fig. 7 Error loss on the mature mRNA dataset using the Tanh as activation functions

Initially, the model's performance metrics produced an accuracy of 89.03%, sensitivity of 89.12%, Specificity of 88.92%, and MCC of 0.798. All these metrics were captured before the implementation of any feature selection algorithms. Finally, to improve the model's overall performance, we used one more step in the feature space, feature selection, to bring down the dimension of the hybrid feature vector. is adjustment on the part of the assessment led to significant gains on the following ranges: Overall accuracy to the rate of 91.47%, sensitivity to the new rate of 92.94%, Specificity rising to the level of 90.01%, and lastly, MCC at 0.830.

Table 6 Performance comparison using sequence formulation techniques and hybrid feature vector using full transcript dataset

Methods	ACC (%)	SN (%)	PRE (%)	F1 (%)	SP (%)	MCC
PseSNC	86.34	87.34	86.37	86.33	86.34	0.727
PseDNC	87.37	88.35	87.47	87.36	86.39	0.748
PseTNC	88.01	89.59	88.03	87.53	86.41	0.760
PseQNC	87.04	88.46	88.46	88.85	85.71	0.741
PsePNC	88.68	90.35	86.89	86.85	86.99	0.772
Hybrid feature (without feature selection)	89.03	89.12	90.78	88.91	88.92	0.798
Hybrid Features (with feature selection)	91.47	92.94	91.51	91.47	90.01	0.830

Table 7 Performance comparison using sequence formulation techniques and hybrid feature vector using mature mRNA

Methods	ACC (%)	SN (%)	PRE (%)	F1 (%)	SP (%)	MCC
PseSNC	91.47	92.47	91.51	91.67	90.47	0.830
PseDNC	92.94	93.94	93.00	92.94	91.94	0.859
PseTNC	93.01	94.01	93.15	93.00	92.01	0.862
PseQNC	92.75	94.15	92.81	92.75	91.34	0.856
PsePNC	93.39	94.39	93.57	93.86	92.39	0.870
Hybrid feature (without feature selection)	94.57	95.04	95.01	95.04	94.12	0.891
Hybrid features (with feature selection)	95.86	96.15	95.59	95.52	94.58	0.917

Similarly, we evaluated the proposed model's performance on di erent sequence formulation methods using the Mature mRNA dataset, as presented in Table 7. e results show that the model performed best when applying hybrid features, outperforming the individual formulation methods. For example, the proposed model achieved a success rate of 94.57%, with sensitivity at 95.04%, Specificity at 94.12%, and an MCC of 0.891. To further enhance its performance, we applied a feature selection method to reduce the dimensionality of the hybrid features. is led to a noticeable improvement, with the success rate increasing to 95.86%, sensitivity to 95.15%, Specificity to 96.58%, and an MCC of 0.917 accordingly.

Performance comparison of di erent classi ers

In this section, the performance of the proposed model is examined by testing it with several well-known supervised machine learning algorithms using hybrid features. e nature of the algorithms under consideration for this comparison are Random Forest (RF) [63], Support Vector Machine (SVM), Logistic Regression (LR), Naive Bayes (NB), and K- Nearest Neighbor (KNN) [64]. Random Forests is another ensemble learning technique that builds several decision trees by utilizing di erent bootstrapping methods. e result from each tree decision is combined by applying voting to improve the

classification performances. It can be used in almost all classification and regression problems. K-Nearest Neighbor is a non-parameterized learning algorithm widely used in image processing. It divides instances into classes depending on the distance from the neighbors, and due to the straightforward approach, it fits most of the problems. Support Vector Machines are especially e ective when dealing with linear and nonlinearly separable data; this algorithm searches for the best hyperplane to classify di erent classes e ectively. is method is widely used, especially in bioinformatics, because of its e ectiveness in working with large data sets. Naive Bayes, which derives from Bayes' eorem, is a probabilistic classifier that analyzes features independently. It is particularly e ective for text categorization, having small data sets, and working in high-dimensional spaces. Additional information about the performance of each of the algorithms is provided in Table 8.

Table 8 compares di erent models applied to the Full Transcript dataset using various evaluation criteria. e presented models' performance can be analyzed based on accuracy: the Logistic Regression (LR) model reached 88.45% and an MCC of 0. 769. Respectively, the NB, RF, and KNN models had better performances with a mean accuracy of 89.09%, 89.73%, and 90.63%. e MCC values of the three sets were 0.782, 0.795, and 0.812. e SVM model recorded the highest accuracy amongst the traditional models at 91.19% with an MCC of 0.824. However, the proposed Deep-m5U model yielded the highest performance with an accuracy of up to 91.47% and an MCC of up to 0.830. erefore, it demonstrates the higher e ciency of the proposed Deep-m5U model compared to the other methods used in the analysis.

In addition to the mentioned performance measures, the performance of the proposed model was evaluated with the AUC (Area Under the ROC Curve) metric, widely used to assess binary classifiers. Figure 8 shows the AUC outcomes for the proposed type using Full Transcript collection. e AUC metric ranges between 0 and 1, and the higher value of AUC suggests a classifier's better performance [65]. ese two features combine in a plot having the FPR, false positive rate, on the x-axis and TPR, true positive rate, on the y-axis. e results of using the proposed model for the given dataset were quite encouraging, with the obtained AUC value of about 0.972 with the Full Transcript dataset higher than other machine learning algorithms, including NB, KNN, and SVM. is high AUC value gives more weight to the model of identifying the di erence between positive cases and negative ones, thus proving e ective.

Moreover, Table 9 presents the performance metrics for various models on the Mature mRNA dataset. e SVM model had the highest accuracy of 94.44% and an MCC of

Table 8 Performance comparison of di□erent classifiers using Full transcript dataset

Methods	ACC (%)	SN (%)	PRE (%)	F1 (%)	SP (%)	MCC
LR	88.45	88.89	88.22	88.02	88.03	0.769
NB	89.09	89.55	88.52	88.23	88.65	0.782
RF	89.73	90.41	89.11	89.41	89.04	0.795
KNN	90.63	90.91	90.31	90.61	90.36	0.812
SVM	91.19	92.50	90.91	90.95	89.91	0.824
Deep-m5U	91.47	92.94	91.51	91.47	90.01	0.830

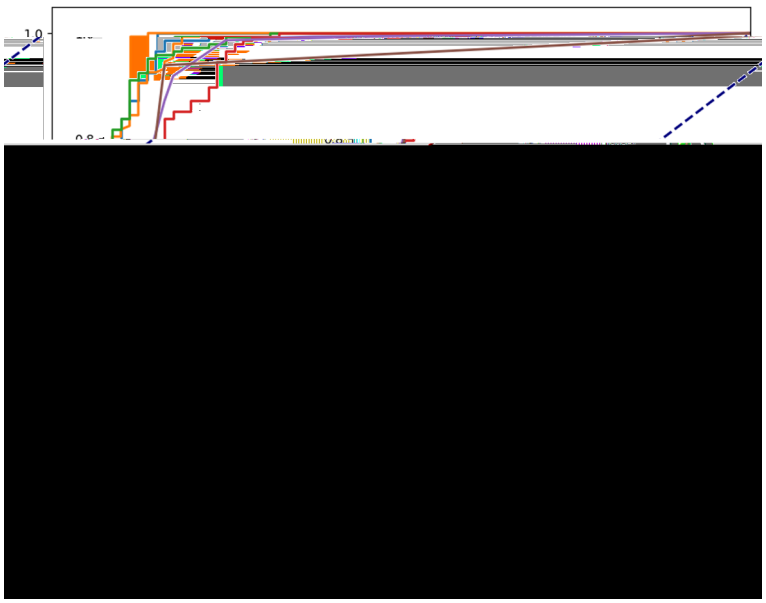

Fig. 8 The e□ciency of AUC using the Full Transcript dataset

Table 9 Performance comparison of di□erent classifiers using a mature mRNA dataset

Methods	ACC (%)	SN (%)	PRE (%)	F1 (%)	SP (%)	MCC
LR	92.01	92.28	91.75	91.73	91.76	0.840
NB	92.31	92.57	92.05	92.08	92.06	0.846
RF	93.68	93.96	93.40	93.47	93.41	0.874
KNN	94.03	94.32	93.71	93.72	93.75	0.881
SVM	94.44	94.78	94.17	94.16	94.12	0.889
Deep-m5U	95.86	96.15	95.59	95.52	94.58	0.917

0.889 compared to the traditional machine learning techniques. e proposed Deep-m5U outperformed all others, with the highest accuracy of 95.86% and an MCC of 0.917. Similarly, in the other parameters, the F1 score was 95.52%, sensitivity 96.15%, and Specificity 94.58%. e DNN model outperformed the SVM algorithm and other traditional ML because it uses a single processing layer. Traditional ML struggles with complex datasets that have high nonlinearity.

In contrast, the DNN model utilizes multiple processing layers, allowing it to handle complex and nonlinear data more e ectively. e AUC values for the proposed method using Mature mRNA datasets are illustrated in Fig. 9. e AUC graph visually represents the model's performance, with an increasing area under the curve indicating improved performance. Conversely, a decrease in this area suggests a reduction in model e ectiveness. e proposed model achieved AUC values of 0.981 using Mature mRNA datasets compared to the other machine-learning algorithms.

Comparison of proposed predictor with existing predictors

We evaluate the performance of the proposed Deep-m5U model against the existing models on both Full Transcript and Mature mRNA datasets. Aggregate the

abovementioned comparison, which is provided in Table 10. erefore, moderate improvements in the m5UPred predictions were observed for the Full Transcript dataset with an accuracy of around 83.60%, sensitivity of 72.82%, and Specificity of 89.38% in MCC and MCC levels, respectively 0.634. As for the evaluation, the m5U-SVM model [29

Table 11 The performance of the proposed model compared to the existing models on the independent datasets

Mode	Method	ACC (%)	SN (%)	SP (%)	MCC
Full Transcript	Deep-m5U	92.94	91.72	94.14	0.831
	m5U-SVM	90.82	87.40	93.16	0.809
	m5UPred	87.17	80.60	91.66	0.732
Mature mRNA	Deep-m5U	95.17	93.48	96.87	0.916
	m5U-SVM	94.11	93.06	95.14	0.882
	m5UPred	89.70	87.440	91.95	0.795

Table 12 Performance comparison on the cross-cell-type and cross-technique validation

Testing method	Model	Cross-technique validation			Cross-cell-type validation		
		miCLIP-Seq ACC (%)	FICC-Seq ACC (%)	Average ACC (%)	HEK293 ACC (%)	HAP1 ACC (%)	Average ACC (%)
Cross-validation	m5UPred	86.76	90.58	88.67	86.72	80.15	83.44
	m5U-GEPred	96.14	96.42	96.28	96.79	90.62	93.71
	Deep-m5U	96.52	96.92	96.72	97.06	91.28	94.17
Independent	m5UPred	82.29	73.29	77.79	86.2	73.99	80.10
	m5U-GEPred	86.26	90.83	88.55	74.82	78.71	76.77
	Deep-m5U	88.93	91.27	90.10	75.62	78.98	77.30

To further examine the proposed model's generalization, the proposed model's performance comparison on independent datasets is shown in Table 11. From Table 11, the m5U-SVM predictor achieved an accuracy of 90.82% and an MCC of 0.809 using the Full Transcript dataset, while the proposed Deep-m5U model significantly outperformed it with an accuracy of 92.94% and an MCC of 0.831. Similarly, for the Mature mRNA dataset, m5U-SVM had lower performance metrics with an accuracy of 94.11%, whereas the proposed Deep-m5U model again showed superior results with an accuracy of 95.17%.

Performance evaluation by cross-technique and cross-cell-type validation

is section assesses cross-technique validation (tenfold) and cross-cell-type validation on the benchmark dataset used in [31]. Initially, we categorized the experimentally verified m5U sites based on their profiling methodologies, namely miCLIP and FICC-seq, and the cell lines, namely HEK293. e research outcomes are shown in Table 12, which provides a comparative performance study of the current models m5UPred [28] and m5U-GEPred [31] across two assessment scenarios: cross-validation and independent testing. In cross-validation, the proposed model Deep-m5U regularly attains the most outstanding average accuracy, i.e., 97.28% and 94.17%, surpassing m5UPred and m5U-GEPred. m5U-GEPred demonstrates robust performance; nevertheless, its precision is somewhat inferior to Deep-m5U. Similarly, Deep-m5U again exhibits greater average accuracy in the independent testing findings, i.e., 90.10% and 77.30%. m5U-GEPred exhibits superior accuracy relative to m5UPred; nevertheless, both models see a marginal decrease in performance when compared to the cross-validation scenario. e overall performance of the proposed

Deep-m5U model shows exceptional accuracy and generalization capabilities in both testing methodologies.

Conclusions

5-methyluridine (m5U) is a prominent RNA modification critical in various biological functions and disease pathogenesis. As a posttranscriptional modification involving methylation at the C5 position of uridine, understanding its biological significance requires precise computational tools. is study introduced a novel deep learning-based model, i.e., Deep-m5U, designed to accurately predict RNA m5U sites by leveraging hybrid features and utilizing tenfold cross-validation and independent datasets. e model e ectively addressed the over-fitting issue by optimizing hyper-parameters and demonstrated robust performance, achieving accuracies of 91.47% and 95.86% on the Full Transcript and Mature mRNA datasets, respectively. Additionally, the model attained 92.94% and 95.17% accuracy on independent test samples, outperforming traditional machine learning methods and existing state-of-the-art approaches. e promising results of Deep-m5U highlight its potential to significantly contribute to further research in RNA modifications and their implications in disease, particularly in areas such as stress response and breast cancer, as well as its utility in developing therapeutic strategies.

In future, our research will be extended in three key directions: (a) developing a publicly accessible web server for large-scale RNA 5-methyluridine (m5U) prediction, enabling more comprehensive community access and application [66, 67] (b) implementing a multi-view feature encoding strategy to capture better the properties of biological sequences [68–70] and (c) exploring additional deep learning models to improve the identification of m5U modification sites in RNA sequences [71].

Acknowledgements
This work was supported by Research Supporting Project Number (RSPD2024R585), King Saud University, Riyadh, Saudi Arabia.

Author contributions
All authors contributed equally to this work. SK and SN wrote the main manuscript. SAQ and NA were responsible for debugging the code and providing the datasets. AN, HHA, and WA conducted the independent testing and reviewed the manuscript for grammar.

Funding
This research is not funded.

Availability of data and materials
The datasets used and analyzed during the current study are available on the GitHub link: https://github.com/salman-khan-mrd/Deep-m5U.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that there is no conflict of interest regarding the publication of this paper.

Received: 13 September 2024 Accepted: 6 November 2024
Published online: 19 November 2024

References

1. Khanal J, Tayara H, Zou Q, Chong KT. Identifying DNA N4-methylcytosine sites in the rosaceae genome with a deep learning model relying on distributed feature representation. Comput Struct Biotechnol J. 2021;19:1612–9.
2. El Allali A, Elhamraoui Z, Daoud R. Machine learning applications in RNA modification sites prediction. Comput Struct Biotechnol J. 2021;19:5510–24.
3. Chou K-C. Progresses in predicting post-translational modification. Int J Pept Res Ther. 2020;26(2):873–88.
4. Jonkhout N, Tran J, Smith MA, Schonrock N, Mattick JS, Novoa EM. The RNA modification landscape in human disease. RNA. 2017;23(12):1754–69.
5. da Silva Oliveira JP, de Oliveira RT, Guedes AL, da Costa OM, Macedo AF. Metabolomic studies of anthocyanins in fruits by means of a liquid chromatography coupled to mass spectrometry workflow. Current Plant Biology. 2022;32:100260.
6. Khan S, Khan MA, Khan M, Iqbal N, AlQahtani SA, Al-Rakhami MS, Khan DM. Optimized feature learning for anti-inflammatory peptide prediction using parallel distributed computing. Appl Sci. 2023;13(12):7059.
7. Su D, Chan CT, Gu C, Lim KS, Chionh YH, McBee ME, Russell BS, Babu IR, Begley TJ, Dedon PC. Quantitative analysis of ribonucleoside modifications in tRNA by HPLC-coupled mass spectrometry. Nat Protoc. 2014;9(4):828–41.
8. Sarkar A, Gasperi W, Begley U, Nevins S, Huber SM, Dedon PC, Begley TJ. Detecting the epitranscriptome. Wiley Interdiscip Rev RNA. 2021;12(6):e1663.
9. Xiao S, Cao S, Huang Q, Xia L, Deng M, Yang M, Jia G, Liu X, Shi J, Wang W. The RNA N 6-methyladenosine modification landscape of human fetal tissues. Nat Cell Biol. 2019;21(5):651–61.
10. Laptev I, Shvetsova E, Levitskii S, Serebryakova M, Rubtsova M, Bogdanov A, Kamenski P, Sergiev P, Dontsova O. Mouse Trmt2B protein is a dual specific mitochondrial metyltransferase responsible for m5U formation in both tRNA and rRNA. RNA Biol. 2020;17(4):441–50.
11. Powell CA, Minczuk M. TRMT2B is responsible for both tRNA and rRNA m5U-methylation in human mitochondria. RNA Biol. 2020;17(4):451–62.
12. Carter J-M, Emmett W, Mozos IR, Kotter A, Helm M, Ule J, Hussain S. FICC-Seq: a method for enzyme-specified profiling of methyl-5-uridine in cellular RNA. Nucleic Acids Res. 2019;47(19):e113–e113.
13. Khan F, Khan M, Iqbal N, Khan S, Muhammad Khan D, Khan A, Wei D-Q. Prediction of recombination spots using novel hybrid feature extraction method via deep learning approach. Front Genet. 2020;11:539227.
14. Inayat N, Khan M, Iqbal N, Khan S, Raza M, Khan DM, Khan A, Wei DQ. iEnhancer-DHF: identification of enhancers and their strengths using optimize deep neural network with multiple features extraction methods. Ieee Access. 2021;9:40783–96.
15. Ahmad W, Ahmad A, Iqbal A, Hamayun M, Hussain A, Rehman G, Khan S, Khan UU, Khan D, Huang L. Intelligent hepatitis diagnosis using adaptive neuro-fuzzy inference system and information gain method. Soft Comput. 2019;23:10931–8.
16. Chen K, Wei Z, Zhang Q, Wu X, Rong R, Lu Z, Su J, De Magalhães JP, Rigden DJ, Meng J. Whistle: a high-accuracy map of the human n 6-methyladenosine (m6a) epitranscriptome predicted using a machine learning approach. Nucleic Acids Res. 2019;47(7):e41–e41.
17. Zhou Y, Zeng P, Li Y-H, Zhang Z, Cui Q. SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. Nucleic Acids Res. 2016;44(10):e91–e91.
18. Chen W, Feng P, Ding H, Lin H, Chou K-C. iRNA-Methyl: Identifying N6-methyladenosine sites using pseudo nucleotide composition. Anal Biochem. 2015;490:26–33.
19. Ma J, Zhang L, Chen S, Liu H. A brief review of RNA modification related database resources. Methods. 2022;203:342–53.
20. Wang X, Zhang Y, Chen K, Liang Z, Ma J, Xia R, de Magalhães JP, Rigden DJ, Meng J, Song B. m7GHub V2. 0: an updated database for decoding the N7-methylguanosine (m7G) epitranscriptome. Nucleic Acids Res. 2024;52:D203–12.
21. Zhang Y, Jiang J, Ma J, Wei Z, Wang Y, Song B, Meng J, Jia G, De Magalhães JP, Rigden DJ. DirectRMDB: a database of post-transcriptional RNA modifications unveiled from direct RNA sequencing technology. Nucleic Acids Res. 2023;51(D1):D106–16.
22. Cappannini A, Ray A, Purta E, Mukherjee S, Boccaletto P, Moafinejad SN, Lechner A, Barchet C, Klaholz BP, Stefaniak F. MODOMICS: a database of RNA modifications and related information. 2023 update. Nucleic Acids Res. 2024;52:D239–44.
23. Song B, Chen K, Tang Y, Wei Z, Su J, De Magalhães JP, Rigden DJ, Meng J. ConsRM: collection and large-scale prediction of the evolutionarily conserved RNA methylation sites, with implications for the functional epitranscriptome. Brief Bioinf. 2021;22:bbab088.
24. Feng P, Chen W. iRNA-m5U: a sequence based predictor for identifying 5-methyluridine modification sites in saccharomyces cerevisiae. Methods. 2022;203:28–31.
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. 2022;23(21):13493.
26. Li Y-H, Zhang G, Cui Q. PPUS: a web server to predict PUS-specific pseudouridine sites. Bioinformatics. 2015;31(20):3362–4.
27. Li J, Huang Y, Yang X, Zhou Y, Zhou Y. RNAm 5Cfinder: a web-server for predicting RNA 5-methylcytosine (m5C) sites based on random forest. Sci Rep. 2018;8(1):17299.
28. Jiang J, Song B, Tang Y, Chen K, Wei Z, Meng J. m5UPred: a web server for the prediction of RNA 5-methyluridine sites from sequences. Mol Therapy-Nucleic Acids. 2020;22:742–7.
29. 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. 2023;21(1):93.
30. Yu L, Zhang Y, Xue L, Liu F, Jing R, Luo J. Evaluation and development of deep neural networks for RNA 5-Methyluridine classifications using autoBioSeqpy. Front Microbiol. 2023;14:1175925.
31. 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. 2023;14:1277099.

32. Qiyas M, Naeem M, Khan N, Khan S, Khan F: Confidence levels bipolar complex fuzzy aggregation operators and their application in decision making problem. *IEEE Access* 2024.
33. Khan S, Khan M, Iqbal N, Dilshad N, Almufareh MF, Alsubaie N. Enhancing sumoylation site prediction: a deep neural

62. Ahmad A, Akbar S, Khan S, Hayat M, Ali F, Ahmed A, Tahir M. Deep-AntiFP: prediction of antifungal peptides using distanct multi-informative features incorporating with deep neural networks. Chemom Intell Lab Syst. 2021;208:104214.
63. Fawagreh K, Gaber MM, Elyan E. Random forests: from early developments to recent advancements. Syst Sci Control Eng An Open Access J. 2014;2(1):602–9.
64. Cheng D, Zhang S, Deng Z, Zhu Y, Zong M: k NN algorithm with data-driven k value. In: advanced data mining and applications: 10th international conference, ADMA 2014, Guilin, China, December 19–21, 2014 Proceedings *10: 2014*. Springer: 499–512.
65. Zhou G-P, Chen D, Liao S, Huang R-B. Recent progresses in studying helix-helix interactions in proteins by incorporating the Wenxiang diagram into the NMR spectroscopy. Curr Top Med Chem. 2016;16(6):581–90.
66. Arif M, Fang G, Fida H, Musleh S, Yu D-J, Alam T. iMRSAPred: improved prediction of Anti-MRSA peptides using physicochemical and pairwise contact-energy properties of amino acids. ACS Omega. 2024;9(2):2874–83.
67. Arif M, Fang G, Ghulam A, Musleh S, Alam T. DPI_CDF: druggable protein identifier using cascade deep forest. BMC Bioinf. 2024;25(1):145.
68. Ge F, Arif M, Yan Z, Alahmadi H, Worachartcheewan A, Yu D-J, Shoombuatong W. MMPatho: leveraging multilevel consensus and evolutionary information for enhanced missense mutation pathogenic prediction. J Chem Inf Model. 2023;63(22):7239–57.
69. Hu J, Zeng W-W, Jia N-X, Arif M, Yu D-J, Zhang G-J. Improving DNA-binding protein prediction using three-part sequence-order feature extraction and a deep neural network algorithm. J Chem Inf Model. 2023;63(3):1044–57.
70. Hu J, Chen K-X, Rao B, Ni J-Y, Thafar MA, Albaradei S, Arif M. Protein-peptide binding residue prediction based on protein language models and cross-attention mechanism. Anal Biochem. 2024;694:115637.
71. Sikander R, Arif M, Ghulam A, Worachartcheewan A, Thafar MA, Habib S. Identification of the ubiquitin–proteasome pathway domain by hyperparameter optimization based on a 2D convolutional neural network. Front Genet. 2022;13:851688.

Publisher's Note