AIMS Bioengineering, 12(1): 145–161.
DOI: 10.3934/bioeng.2025006
Received: 06 January 2025
Revised: 06 February 2025
Accepted: 12 February 2025
Published: 04 March 2025

https://www.aimspress.com/journal/Bioengineering

Research article

Deep-m6Am: a deep learning model for identifying N6, 2'-O-Dimethyladenosine (m6Am) sites using hybrid features

Islam Uddin[1], Salman A. AlQahtani[2], Sumaiya Noor[3], Salman Khan[1,*]

[1] Department of Computer Science, Abdul Wali Khan University Mardan, KPK, Pakistan
[2] 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
[3] Business and Management Sciences Department, Purdue University, West Lafayette, IN, USA

* **Correspondence:** Email: salman@awkum.edu.pk.

Abstract: N6,2'-O-dimethyladenosine (m6Am) is a crucial RNA modification that plays a pivotal role in regulating gene expression and maintaining RNA stability. Given its dynamic involvement in various biological processes and diseases, accurately identifying m6Am is essential for understanding cellular mechanisms and pathogenesis. Furthermore, detecting m6Am modifications is key to deciphering regulatory pathways and elucidating disease mechanisms. In this study, we propose Deep-m6Am, a deep learning–based model for precisely identifying m6Am sites in RNA sequences. The proposed framework employs a comprehensive feature extraction process, i.e., integrating pseudo single nucleotide composition (PseSNC), pseudo dinucleotide composition (PseDNC), and pseudo trinucleotide composition (PseTNC) to capture complex sequence patterns. To enhance computational efficiency and eliminate noisy or redundant features, a supervised SHAP (SHapley Additive exPlanations) algorithm is utilized, ensuring the selection of the most informative features. Finally, a multilayer deep neural network (DNN) is used as a classification algorithm for identifying m6Am sites. The performance of the proposed model was evaluated in comparison with traditional machine learning (ML) algorithms and existing models. Experimental results demonstrate that Deep-m6Am outperforms previous approaches by 6.67% and traditional ML algorithms by 7.39%.

These findings highlight Deep-m6Am as a promising tool for advancing drug discovery and improving the diagnosis of diseases associated with m6Am modifications.

Keywords: machine learning; deep neural network; hybrid sequential model; deep-m6Am model; N6, 2'O-dimethyl adenosine; RNA modification

1. Introduction

N6,2'-O-dimethyladenosine (m6Am) is a significant RNA modification that plays a vital role in regulating various cellular processes, including gene expression, RNA stability, and the general integrity of RNA metabolism. This modification occurs at the five untranslated regions (UTRs) of messenger RNA (mRNA), influencing key RNA functions such as capping, translation initiation, and RNA decay [1]. m6Am has been shown to affect the interaction of RNA molecules with RNA-binding proteins, modulating critical processes like RNA splicing, transport, and stability. These modifications help regulate gene expression in response to cellular conditions and environmental cues, making them essential for maintaining cellular homeostasis. The dynamic and reversible nature of m6Am modifications in RNA is crucial for regulating mRNA's fate and ensuring the translation machinery's proper functioning [2]. The m6Am modification has gained attention due to its potential implications in disease pathogenesis and cellular dysfunction. The m6Am is linked to various biological processes, such as cell growth, differentiation, stress responses, and RNA surveillance mechanisms. The m6Am role in regulating mRNA stability suggests that it could regulate gene expression in response to stress or environmental changes, making it an essential factor in cellular adaptation and survival [3,4]. Similarly, alterations in m6Am modification patterns have been associated with several diseases, including cancer, neurological disorders, and metabolic conditions, highlighting its significance in both health and disease. Its importance and accurate identification of m6Am sites within RNA sequences is essential for advancing the understanding of gene regulation and the molecular mechanisms that govern disease progression [5]. The ability to detect m6Am modifications opens new avenues for therapeutic interventions, enabling the development of targeted strategies for diseases that involve aberrant RNA modifications. As a result, computational methods that allow efficient and precise detection of m6Am sites are critical for advancing research in RNA biology and molecular medicine.

Advancements in computational biology have led to several learning tools for predicting RNA modifications, particularly m6Am. For example, Song et al. [6] introduced MultiRM, an attention-based multi-label neural network capable of predicting 12 RNA modifications simultaneously. Using an attention mechanism, MultiRM identifies modification sites and interprets key sequence contexts, revealing strong associations between different RNA modifications. The model achieves 71.13% accuracy with an MCC of 0.427 and an AUC of 0.805 on sequence-based RNA modification mechanisms. Jiang et al. [7] proposed m6AmPred using the eXtreme gradient boosting with dart (XGBDart) algorithm and EIIP-PseEIIP encoding for feature representation. m6AmPred achieved 73.10% accuracy with an MCC of 0.462 and an AUC of 0.820 on cross-validation. Similarly, Luo et al. [8] developed another model named DLm6Am, i.e., an ensemble deep-learning framework combining one-hot encoding, nucleotide chemical property (NCP), and nucleotide density (ND) for

feature extraction. DLm6Am integrates CNN, BiLSTM, and multi-head attention modules, outperforming tools like m6AmPred and MultiRM with 79.55% accuracy, 81.71% sensitivity, 77.40% specificity, MCC of 0.591, and AUC of 0.863 on independent testing data. Recently, Jia et al. [9] proposed EMDL_m6Am, a stacking ensemble model employing one-hot encoding and integrating DenseNet, inflated convolutional network (DCNN), and deep multiscale residual network (MSRN) for feature extraction. EMDL_m6Am achieved 80.98% accuracy, 82.25% sensitivity, 79.72% specificity, MCC of 0.619, and AUC of 0.823 on training data, with independent testing (80.98% accuracy, AUC of 0.8211). Despite advancements, existing methods struggle with limited encoding schemes, inefficient feature selection, and reliance on single deep learning frameworks, leading to suboptimal performance and high computational costs. The lack of explainability in current models significantly hinders the interpretation and improvement of accuracy, robustness, and interpretability in m6Am site prediction techniques.

Based on the aforementioned considerations, in this study, we propose Deep-m6Am, a novel deep learning (DL) model designed to accurately identify m6Am sites in RNA sequences. The model integrates multiple feature extraction techniques, including pseudo single nucleotide composition (PseSNC), pseudo dinucleotide composition (PseDNC), and pseudo trinucleotide composition (PseTNC), to capture complex sequence patterns essential for precise prediction. A SHAP (SHapley Additive exPlanations)-based feature selection mechanism is incorporated to enhance computational efficiency and eliminate irrelevant or redundant features, ensuring that only the most informative features contribute to the model's predictions. The Deep-m6Am framework addresses the limitations of single-model approaches by leveraging a multilayer deep neural network (DNN) classifier, improving robustness and generalizability. The model's performance was rigorously evaluated using 5-fold cross-validation and independent testing. The Deep-m6Am demonstrates state-of-the-art results across multiple evaluation metrics, including accuracy, sensitivity, specificity, AUC, and MCC, outperforming existing models and traditional ML algorithms. Integrating cutting-edge feature extraction, selection, and deep learning methodologies, Deep-m6Am provides a powerful and interpretable tool for predicting RNA modifications. This advancement significantly contributes to RNA biology by offering more profound insights into RNA modifications and their roles in disease mechanisms, opening promising avenues for further research into RNA modification patterns. Therefore, Deep-m6Am is a robust computational framework for addressing key challenges in RNA modification analysis, as illustrated in Figure 1.

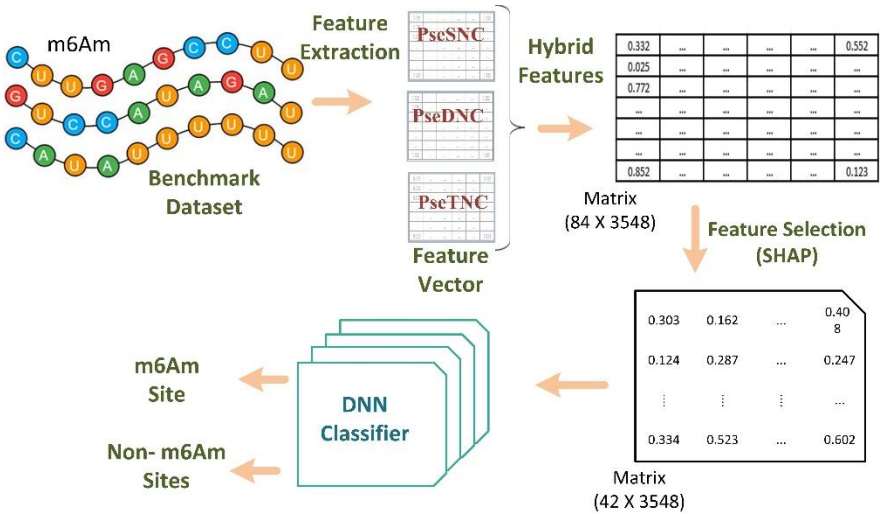

Figure 1. Architecture of the proposed model.

The rest of the paper is organized as follows: Section 2 presents material and methods, Section 3 illustrates performance metrics and evaluation, Section 4 provides experimental results and analysis, and the work is concluded in Section 5.

2. Materials and methods

2.1. Benchmark dataset

A valid and reliable benchmark dataset is essential for designing a powerful and robust computational model. In this study, we utilized the same benchmark datasets employed by Jia et al. [9]. These sites were regarded as highly confident, providing a solid foundation for accurate and reliable model development. Initially, sample sequences were extracted for the training dataset, as depicted in Eq. 1.

$$T_1 = (T_1^+ \cup T_1^-) \tag{1}$$

Where T_1 represents the total RNA sequences, T_1^+ represents the positive m6Am sequences, and T_1^- represents the non-m6Am sequences. U is a mathematical operator representing the union of the two subsamples. Moreover, a CD-HIT tool was employed to eliminate pairwise sequences with a similarity greater than 20%. Finally, we achieved a benchmark dataset comprising 3548 sequences with 1774 m6Am samples and 1774 non-m6Am samples. In addition, we randomly separated 15% of the samples with label stratification from the original dataset and generated an independent set. The remaining 85% of the samples were used as training sets. The independent benchmark dataset was mathematically formulated using the following Eq. 2.

$$T_2 = (T_2^+ \cup T_2^-) \tag{2}$$

Where T_2 represents the total RNA sequences, T_2^+ represents the positive m6Am sequences, and T_2^- represents the non-m6Am sequences. After separation, the training sets contained 2838 (i.e., 1419 Pos+ and 1419 Neg-) training instances and 710 (i.e., 355 Pos+ and 355 Neg-) independent instances. It is important to note that the independent test set was carefully saved separately as invisible data and was not used in learning and parameter tuning processes. The statistical distribution of the benchmark dataset, detailed in Table 1, ensures an equitable representation of positive and negative samples across training and independent testing, thereby enabling a robust and reliable model evaluation.

Table 1. Statistical distribution of the benchmark dataset.

Dataset	Number of samples	Positive samples	Negative samples
Cross validation	3548	1774	1774
Training dataset	2838	1419	1419
Independent dataset	710	355	355

2.2. Feature extraction techniques

Several techniques have been developed to convert DNA, protein, and RNA sequences into discrete mathematical models, maintaining the nucleotides' outstanding features and structural integrity. These methods ensure that the biological sequences are accurately described in numerical formats, enabling computational analysis without losing critical sequence-specific information. Accordingly, several bioinformatics approaches have been developed that can transform RNA sequences into various statistical equations with the preservation of the uniqueness and inherent patterns of the measures [10–13]. Following the second rule of Chou's 5-step guidelines, several feature extraction techniques have been implemented in this paper to improve the representation of RNA sequences. These techniques include pseudo K-tuple nucleotide composition (PseKNC), comprising methods like PseSNC (K = 1), PseDNC (K = 2), and PseTNC (K = 3). Feature extraction methods are explained in detail in the next section. The PseKNC approach represents RNA sequences as functional vectors by encoding their composition and sequence patterns. This method suppresses detailed order data, focusing on capturing essential features that suggest similarities between RNA samples. By transforming the sequences into structured mathematical representations, PseKNC facilitates efficient computational analysis while preserving key biological characteristics of the RNA [14]. Let us consider an RNA sequence R with N number of nucleotides, represented in Eq. 3.

$$R = R_1 R_2 R_3 \ldots R_i \ldots R_N \tag{3}$$

Where N represents the number of nucleotides in a RNA sequence (i.e. the length of a RNA sequence) and $R_i \in \{A, C, G, U\}$ $(i = 1,2,3,\ldots,L)$. Where R_i represents a nucleotides at the i^{th} sequence location and A,C,G, U represents Adenine, Cytosine, Guanine and Urine respectively [16,17].

The Eq. 3, can be expressed in the general form of the PseKNC as

$$R = \left[\phi_1 \quad \phi_2 \quad \phi_3 \quad \ldots \quad \phi_y \quad \ldots. \quad \varphi_z \right]^T \tag{4}$$

In RNA sequence representation, T is the transposed vector representing a mathematical transformation, z represents a numeric value typically corresponding to an output or dependent variable in the analysis, and ϕ_y represents the actual value of the RNA sequence's feature vector and can be computed using Eq. 5.

$$\phi_u = \begin{cases} \dfrac{f_u^{K-tuple}}{\sum_{i=1}^{4^k} f_u^{K-tuple} + w\sum_{j=1}^{\lambda} \theta_j} (1 \le u \le 4^k, u = 1,2,3,\ldots) \\ , \dfrac{w\theta_{u-4^k}}{\sum_{i=1}^{4^k} f_u^{K-tuple} + w\sum_{j=1}^{\lambda} \theta_j} (4^k + 1 \le u \le 4^k + \lambda) \end{cases} \tag{5}$$

$$\theta_j = \left\{ \frac{1}{L-K-(\lambda-1)} \sum_{i=1}^{L-K-(\lambda-1)} C_{i,i+j} \qquad j \to 1,2,\ldots..,\lambda, \lambda < L-K \right. \tag{6}$$

$$C_{i,i+j} = \frac{1}{u} \sum_{\xi=1}^{\lambda} \left[H_\xi (N_i N_{i+1} \ldots N_{i+K-1}) - H_\xi (N_{i+j} N_{i+j+1} \ldots N_{i+j+K-1}) \right]^2 \tag{7}$$

Where θ_j represent the j^{th} tier correlation factor or j^{th} rank correlation factor that reflects the sequence order correlation in most contiguous K-tuple nucleotides. λ represents the total number correlation rank and w represents the weight. This paper uses the PseKNC technique to convert the provided sequences into discrete feature vectors while maintaining the sequence order data. By designating different values to K (i.e., K = 1, 2, 3) in Eq. 4, three distinct modes of PseKNC were obtained, i.e., PseSNC (K = 1), PseDNC (K = 2), and PseTNC (K = 3), defined as follows:

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

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

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

2.3. Hybrid feature

This study used three distinct feature extraction methods to encode RNA sequences into discrete feature vectors, as summarized in Table 2. These features include PseSNC, PseDNC, and PseTNC, which integrate pseudo, composition, and transitional probability features to improve the

differentiation and interpretation of nucleotide sequences [17–19]. All individual features were incorporated to construct a comprehensive hybrid feature vector by capturing diverse sequence-derived attributes. Machine learning models leveraging hybrid features benefit from combining multiple extraction techniques, enhancing predictive performance by effectively capturing complex data patterns. This approach remains a widely adopted strategy in bioinformatics and genomics for improving model interpretability and accuracy.

Table 2. Dimension of feature vector with different values of K.

Feature extraction methods	Features
Pseudo single nucleotide composition (PseSNC)	4
pseudo dinucleotide composition (PseDNC)	16
Pseudo trinucleotide composition (PseTNC)	64
Hybrid features	84

2.4. Feature selection

Feature selection is critical in developing models to improve overall performance and computational efficiency. Feature selection involves identifying and retaining the most informative features while eliminating irrelevant or redundant ones, which can introduce noise and reduce prediction accuracy. This study employs SHAP (SHapley Additive exPlanations) as a robust feature selection technique. SHAP leverages cooperative game theory to quantify the contribution of each feature to the model's predictions, ensuring that only the most significant features are retained [20]. This approach reduces the dataset's dimensionality and enhances the model's interpretability by providing insights into the importance of individual features. By integrating SHAP into the Deep-m6Am framework, the model achieves optimized computational efficiency and improved generalization, enabling more accurate and reliable identification of m6Am sites in RNA sequences. This feature selection strategy is pivotal in addressing the challenges of high-dimensional data and ensuring the model's robustness and scalability. This approach enhances model interpretability and supports robust data analysis; it can be expressed as in Eq. 11.

$$SHAP_i(x) = \emptyset_i = \sum_{s \subseteq N\{i\}} \frac{|S|(|N|-|S|-1)}{|N|} [f(S \cup \{i\}) - f(S)] \tag{11}$$

Where ϕ_i represents the SHAP value for the feature i, N is the set of all features, and S is a subset of features excluding i. Then, $f(S)$ is the model's prediction given features in S, and $f(S \cup \{i\})$ is the model's prediction given features in S plus feature i. This equation captures the incremental effect of adding the feature i to different subsets of features.

2.5. Deep neural network architecture

The network topology of a deep neural network, an algorithm based on machine learning or artificial intelligence inspired by the human brain, includes input and output layers and multiple hidden

layers. The mechanism of neuron transmission and activation function in DNN is shown in Figure 2. Unlike traditional processing techniques, DNNs can self-learn and automatically acquire pertinent features from unstructured or raw data. Domains in which DNN has been successfully implemented include speech recognition, NLP (Natural Language Processing) and bioengineering, and imaging [21].

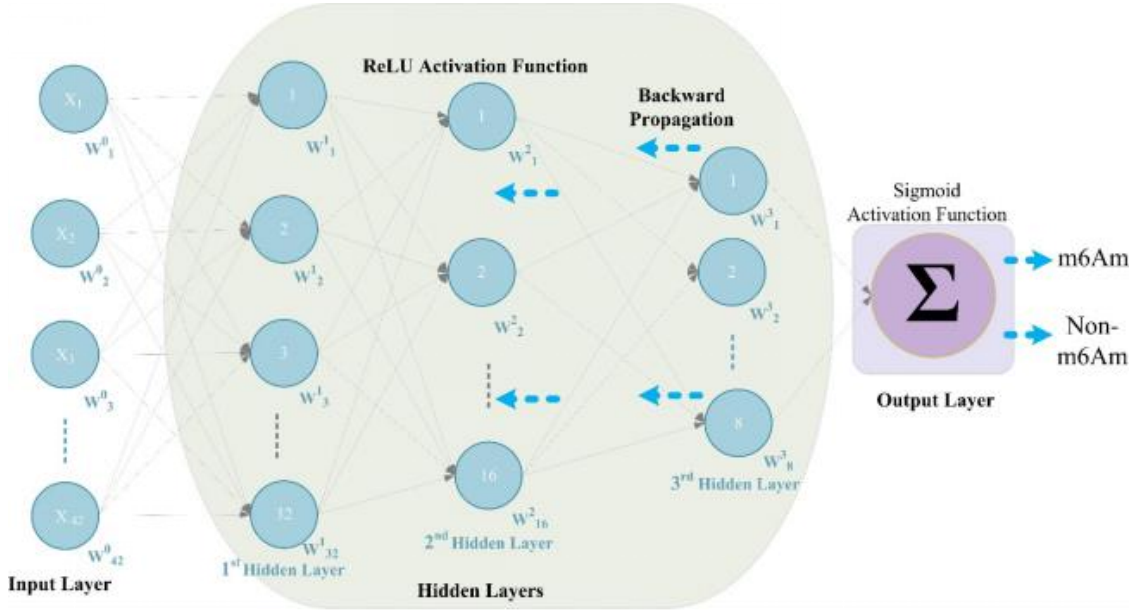

Figure 2. The architecture of the proposed deep neural network.

The proposed architecture utilizes fully connected layers to locate m6Am sites in RNA sequences. The input layer comprises 42 nodes linked to a first hidden layer of 32 nodes through weighted connections. A second hidden layer with 16 nodes processes outputs from the first layer, followed by a third and last layer with 8 nodes. Each layer employs the rectified linear unit (ReLU) activation function, enabling the model to detect nonlinear relationships and complex patterns [22]. The output layer uses the sigmoid activation function for normalized binary classification, distinguishing m6Am and non-m6Am sites in RNA sequences.

3. Performance evaluation criteria

The Deep-m6Am performance is rigorously evaluated using key metrics, including accuracy (ACC), sensitivity (SN), specificity (SP), Matthews correlation coefficient (MCC), and area under the curve (AUC) [23]. SN measures the model's ability to accurately identify true m6Am sites, while SP evaluates its capacity to predict negative cases correctly. ACC reflects the overall correctness of predictions, MCC provides a balanced classification performance assessment, and AUC highlights the model's ability to distinguish between positive and negative instances. These metrics comprehensively evaluate the model's predictive power, ensuring its reliability and effectiveness in identifying m6Am sites.

$$ACC = \frac{T^+ + T^-}{T^+ + F^+ + T^- + F^-} \tag{12}$$

$$SN = \frac{T^+}{T^+ + F^-} \tag{13}$$

$$SP = \frac{T^-}{T^- + F^+} \tag{14}$$

$$MCC = \frac{(T^- * T^+) - (F^- * F^+)}{\sqrt{(f^+ + T^+)(T^+ + F^-)(F^+ + T^-)(T^- + F^-)}} \tag{15}$$

Where T^+ symbolizes true positives, F^+ symbolizes false positives, T^- Symbolizes true negatives, and F^- false negatives, respectively.

4. Experimental results and analysis

4.1. Hyperparameters optimization

In this section, we analyze the hyperparameters of the Deep-m6Am model to optimize its performance. The key hyperparameters considered include learning rate (LR), batch size, number of layers, neurons per layer, and dropout rate. A dropout rate of 0.5 and L2 regularization (0.001) is applied to prevent overfitting, while Xavier initialization ensures stable weight distribution. The model is trained using the Adam optimizer with a learning rate of 0.01 and a momentum of 0.9 to accelerate convergence. Training is conducted for 100 epochs, utilizing ReLU activation functions in the hidden layers and Softmax activation in the output layer for effective learning and classification. A grid search technique was employed to assess the proposed model performance under various hyperparameters, exploring different combinations of parameters. Specifically, the analysis focused on the hyperparameters that significantly influence the performance of the DNN model, including the activation function, learning rate, and number of iterations. Table 3 presents the optimal hyperparameters for the Deep-m6Am.

Table 3. Optimal hyperparameters for the DNN model.

Parameter	Optimal value
Dropout rate	0.5
Weight initialization function	Xavier
Seed	12345L
Dropout	0.001
Number of hidden layers	3
Optimizer	Adam, SGD
L2 regularization	0.001
Epochs	100
Learning rate	0.01
Batch size	16
Activation functions	ReLU, Softmax
Momentum	0.9

4.2. *Performance analysis of DNN*

In this section, we conduct the performance analysis of the proposed Deep-m6Am model. We conducted experiments to examine the effects of LR. Table 4 presents a detailed comparison of performance metrics across different learning rates and shows how the chosen learning rate significantly impacts the model's effectiveness and reliability.

Table 4. Performance metrics across various learning rates (LR).

LR	ACC (%)	SN (%)	SP (%)	MCC
0.01	83.43	82.64	84.22	0.669
0.02	80.05	79.71	80.38	0.601
0.03	79.43	80.27	78.58	0.589
0.04	78.86	78.58	79.14	0.577
0.05	78.70	75.00	82.40	0.672

As shown in Table 4, the Deep-m6Am model achieves optimal performance with a learning rate of 0.01, attaining the highest accuracy (ACC) of 83.43%, sensitivity (SN) of 82.64%, specificity (SP) of 84.22%, and MCC of 0.669. However, as the learning rate increases, the model's performance declines, highlighting that excessively higher learning rates negatively influence overall metrics.

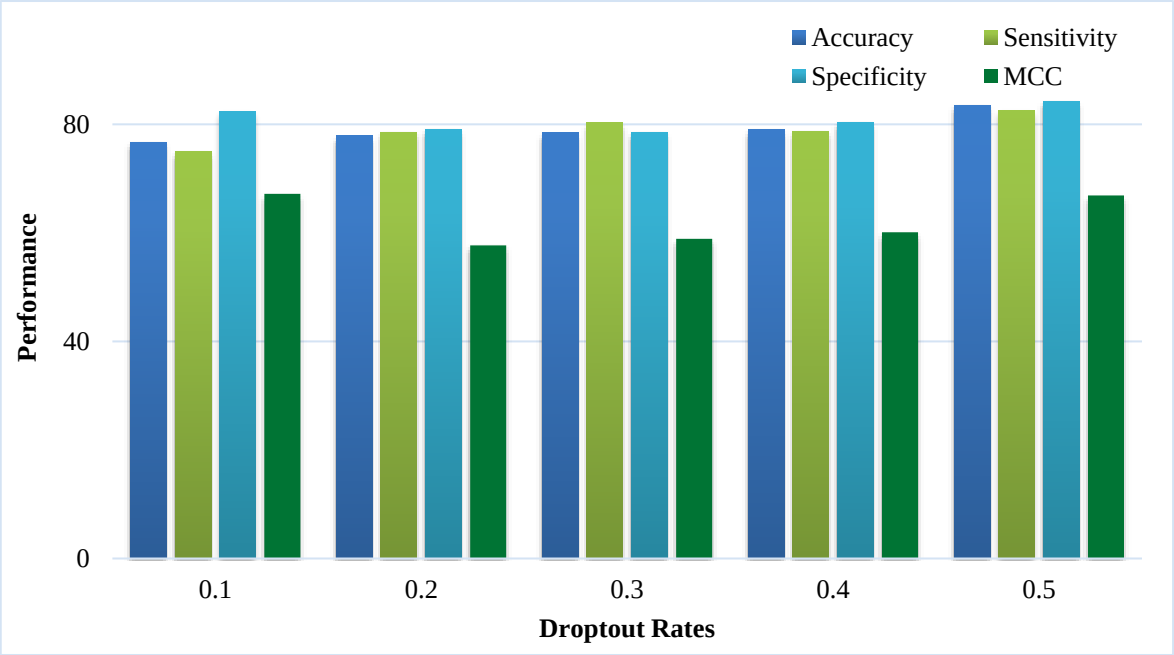

Figure 3. Performance comparison across various dropout rates.

Furthermore, in Figure 3, we analyze the model's fluctuation in performance with different dropout rates, offering valuable guidance for optimizing this hyperparameter. Proper optimization is crucial for balancing generalization and overfitting prevention, ensuring a robust and reliable model. Figure 3 shows that the model achieves optimal performance at a dropout rate of 0.5, with the highest ACC (83.43%) and MCC (66.90%). Performance improves as the dropout rate increases from 0.1 to 0.5, highlighting 0.5 as the most effective rate for balancing generalization and accuracy.

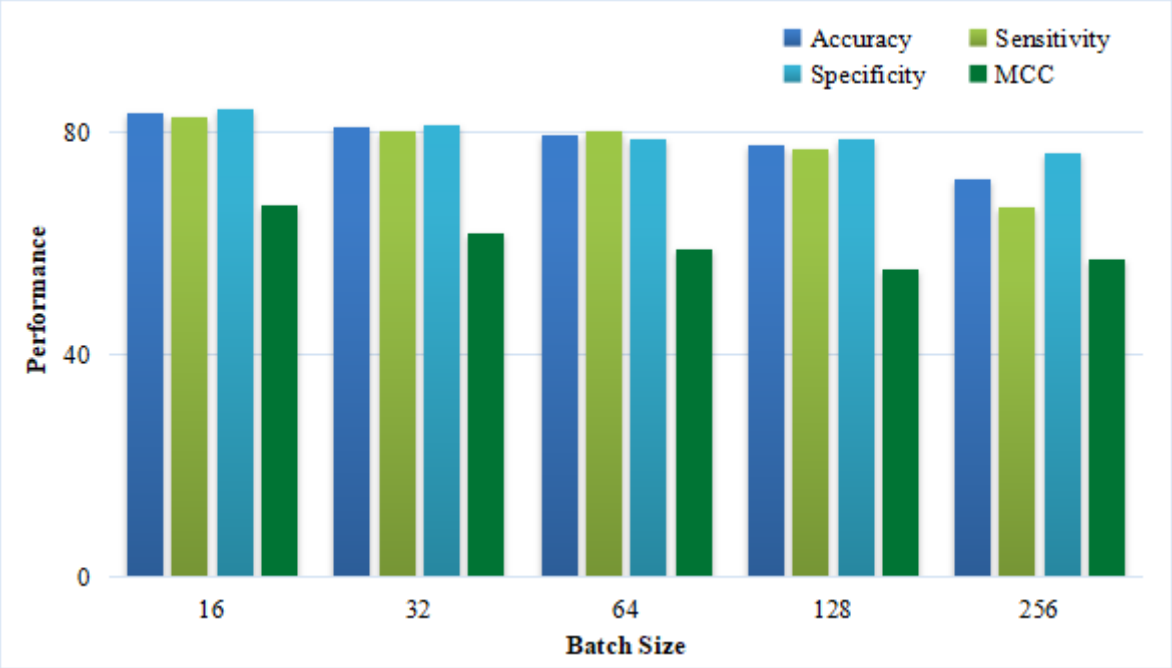

Figure 4. Performance comparison across various batch sizes.

Moreover, we analyze the effect of varying batch sizes on model performance, comparing outcomes across different sizes to identify optimal configurations. Figure 4 illustrates the impact of batch size on model performance, showing a decline as the batch size increases from 16 to 256. The model achieves optimal performance at a batch size of 16, with the highest ACC (83.43%) and MCC (69.90%). As batch size increases, performance gradually decreases, emphasizing the importance of tuning this hyperparameter for optimal results.

4.3. Performance evaluation using cross-validation

Evaluating the robustness of statistical learning models is essential, and this is typically achieved through validation techniques such as jackknife, k-fold cross-validation, and subsampling. Among these methods, k-fold cross-validation is particularly effective for objectively assessing model performance by dividing the dataset into multiple test sets. This approach ensures a thorough evaluation of the model's generalizability and reliability. Table 5 presents a performance comparison of the proposed Deep-m6Am model using various feature extraction techniques, including individual, hybrid, and SHAP-based feature selection methods.

Table 5. 5-fold cross-validation performance evaluation.

Method	ACC (%)	SN (%)	SP (%)	MCC
PseSNC	71.70	66.55	76.40	0.570
PseDNC	77.73	76.89	78.58	0.555
PseTNC	79.43	80.27	78.58	0.589
Hybrid features	80.83	80.27	81.40	0.617
Hybrid features after SHAP	83.43	82.64	84.22	0.669

Table 5 highlights the varying predictive performance of individual features, with PseSNC, PseDNC, and PseTNC achieving ACCs of 71.70%, 77.73%, and 79.43%, respectively. The hybrid feature approach significantly improves classification, reaching an ACC of 80.83%. Further enhancement through SHAP-based feature selection optimizes feature importance, achieving the highest ACC (83.43%) and MCC (0.669). These results underscore the effectiveness of hybrid features in capturing complex patterns and the role of SHAP in refining feature selection for improved model performance.

4.4. Performance comparison with different ML algorithms

In this section, we provide an analysis of the DNN model in comparison to well-known machine learning algorithms such as K-nearest neighbor (KNN), random forest (RF), decision tree (DT), naive Bayes (NB), and support vector machine (SVM) [16,24–26]. Table 6 illustrates the importance of evaluating model performance across different classifiers. We employed a 5-fold cross-validation scheme to ensure a reliable and unbiased performance assessment.

Table 6. Performance comparison with ML algorithms on 5-fold cross-validation.

Classifiers	ACC (%)	SN (%)	SP (%)	MCC
RF	68.72	66.91	70.53	0.667
DT	71.80	70.22	73.38	0.706
KNN	77.15	75.62	78.68	0.741
NB	79.99	78.35	81.63	0.712
SVM	82.53	81.96	83.09	0.651
Deep-m6Am	83.43	82.64	84.22	0.669

Table 6 shows that Deep-m6Am outperforms other ML algorithms, achieving the highest ACC (83.43%) and MCC (0.669). SVM follows with an ACC of 82.53%, while NB and KNN achieve 79.99% and 77.15%, respectively. DT (71.80%) and RF (68.72%) perform lower. These results highlight Deep-m6Am as the most effective model for m6Am site identification. To analyze further, we evaluate the proposed model performance using the Area Under the ROC Curve (AUC), as shown in Figure 5. Figure 5 show that the proposed model achieved an AUC value of 0.853, indicating excellent performance compared with widely used ML algorithms.

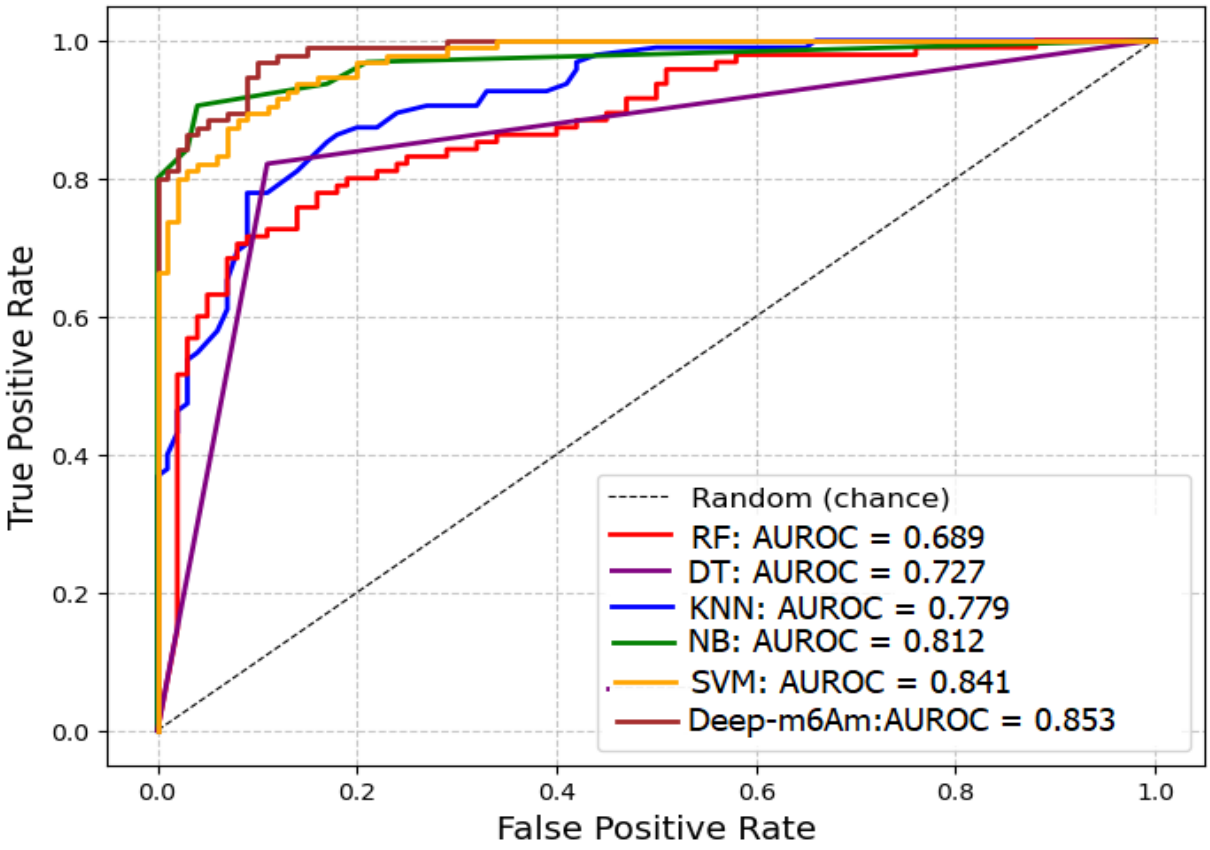

Figure 5. AUC performance comparison with commonly used classifiers using 5-fold cross-validation.

Furthermore, Table 7 evaluates various ML algorithms on an independent dataset to assess generalizability and robustness. From Table 7, the proposed Deep-m6Am model demonstrated superior performance among the ML classifiers, achieving an ACC of 82.86% with an MCC of 0.657%. The SVM classifier had an ACC of 81.64% with an MCC of 0.632. In contrast, KNN achieved an ACC of 75.45%, while NB, DT, and RF performed at 77.29%, 68.88%, and 66.22%, respectively. This analysis identifies Deep-m6Am as the top-performing model, showcasing its superiority in handling dataset complexities and ensuring reliable and accurate m6Am site prediction.

Table 7. Performance comparison with ML algorithms on the independent dataset.

Classifiers	ACC (%)	SN (%)	SP (%)	MCC
RF	66.22	64.50	68.00	0.325
DT	68.88	67.10	70.66	0.378
KNN	75.45	73.20	77.70	0.509
NB	77.29	75.20	79.38	0.545
SVM	81.64	79.50	83.78	0.632
Deep-m6Am	82.86	83.65	82.07	0.657

4.5. Performance comparison on the independent dataset

In this section, we conduct a detailed comparative analysis of the proposed Deep-m6Am predictor against several state-of-the-art predictors, including MultiRM [6], m6AmPred [7], DLm6Am [8], and EMDL_m6Am [9]. This comparison is shown in Table 8.

Table 8. Performance comparison with existing models on the independent dataset.

Predictor	ACC (%)	SN (%)	SP (%)	MCC
MultiRM [6]	71.13	78.59	63.66	0.427
m6AmPred [7]	73.10	72.11	74.08	0.462
DLm6Am [8]	79.55	81.71	77.40	0.591
EMDL_m6Am [9]	80.98	82.25	79.72	0.619
Deep-m6Am	82.86	83.65	82.07	0.657

From Table 8, the MultiRM achieved an ACC of 71.13% and MCC of 0.427, while m6AmPred had an ACC of 73.10% and MCC of 0.462. DLm6Am demonstrated an ACC of 79.55% and MCC of 0.591, and EMDL_m6Am obtained an ACC of 80.98% and MCC of 0.619. In comparison, the proposed Deep-m6Am outperformed all these models, achieving the highest ACC of 82.86% and MCC of 0.657. These results highlight the superior predictive accuracy and robustness of Deep-m6Am for m6Am site identification, making it the most effective model among the evaluated predictors.

5. Conclusions

The biological function of N6,2'-O-dimethyladenosine (m6Am) in RNA sequences underscores its critical role in regulating post-transcriptional processes, RNA stability, and translation. This study introduces the Deep-m6Am model, which employs a hybrid feature extraction approach, incorporating SHAP (SHapley Additive exPlanations) feature selection and DNN classifier to precisely identify m6Am sites within RNA sequences. Through 5-fold cross-validation, compared with popular ML methods, Deep-m6Am demonstrated unique advantages that resulted in more precise m6Am site

predictions. Furthermore, the proposed Deep-m6Am model showed superior performance metrics, achieving an average accuracy of 82.86% compared to the existing models. These results underscore the potential of Deep-m6Am as a reliable and efficient tool for advancing RNA modification analysis.

Future research could expand Deep-m6Am to analyze other RNA modifications and integrate multi-omics data for enhanced predictive accuracy. Exploring its role in disease-specific studies could advance precision medicine. Optimizing computational efficiency through transfer learning, hyperparameter optimization, and parallel programming will improve the model's scalability and applicability in RNA biology and medical research [27].

Acknowledgments

This work was supported by Research Supporting Project Number (RSPD2025R585), King Saud University, Riyadh, Saudi Arabia.

Data Availability

The datasets used and analyzed during the current study are available on the GitHub link: https://github.com/islamuddinw1/Deep-m6Am1.git.

Conflict of interest

The authors declare no conflict of interest.

References

1. Ye F, Wang T, Wu X, et al. (2021) N6-Methyladenosine RNA modification in cerebrospinal fluid as a novel potential diagnostic biomarker for progressive multiple sclerosis. *J Transl Med* 19: 1–14. https://doi.org/10.1186/S12967-021-02981-5
2. Janaki Ramaiah M, Divyapriya K, Kartik Kumar S, et al. (2020) Drug-induced modifications and modulations of microRNAs and long non-coding RNAs for future therapy against glioblastoma multiforme. *Gene* 723: 144126. https://doi.org/10.1016/J.GENE.2019.144126
3. Dieterich C, Völkers M (2021) RNA modifications in cardiovascular disease—an experimental and computational perspective. *Epigenetics Cardiovasc Dis* 24: 113–125. https://doi.org/10.1016/B978-0-12-822258-4.00003-1
4. Akbar S, Khan S, Ali F, et al. (2020) iHBP-DeepPSSM: identifying hormone binding proteins using PsePSSM based evolutionary features and deep learning approach. *Chemom Intell Lab Syst* 204: 104103. https://doi.org/10.1016/J.CHEMOLAB.2020.104103
5. Chen Y, Ouyang X, Yu X, et al. (2021) N6-adenosine methylation (m^6A) RNA modification: an emerging role in cardiovascular diseases. *J Cardiovasc Transl Res* 14: 857–872. https://doi.org/10.1007/S12265-021-10108-W

6. Song Z, Huang D, Song B, et al. (2021) Attention-based multi-label neural networks for integrated prediction and interpretation of twelve widely occurring RNA modifications. *Nat Commun* 12: 4011. https://doi.org/10.1038/s41467-021-24313-3
7. Jiang J, Song B, Chen K, et al. (2022) m6AmPred: identifying RNA N6, 2'-O-dimethyladenosine (m^6A_m) sites based on sequence-derived information. *Methods* 203: 328–334. https://doi.org/10.1016/J.YMETH.2021.01.007
8. Luo Z, Su W, Lou L, et al. (2022) DLm6Am: a deep-learning-based tool for identifying N6,2'-O-Dimethyladenosine sites in RNA sequences. *Int J Mol Sci* 23: 11026. https://doi.org/10.3390/ijms231911026
9. Jia J, Wei Z, Sun M (2023) EMDL_m6Am: identifying N6, 2'-O-dimethyladenosine sites based on stacking ensemble deep learning. *BMC Bioinf* 24: 397. https://doi.org/10.1186/s12859-023-05543-2
10. Khan S, Uddin I, Khan M, et al. (2024) Sequence based model using deep neural network and hybrid features for identification of 5-hydroxymethylcytosine modification. *Sci Rep* 14: 9116. https://doi.org/10.1038/s41598-024-59777-y
11. Khan S, Khan M, Iqbal N, et al. (2023) Enhancing sumoylation site prediction: a deep neural network with discriminative features. *Life* 13: 2153. https://doi.org/10.3390/life13112153
12. Khan S, AlQahtani SA, Noor S, et al. (2024) PSSM-Sumo: deep learning based intelligent model for prediction of sumoylation sites using discriminative features. *BMC Bioinf* 25: 284. https://doi.org/10.1186/s12859-024-05917-0
13. Uddin I, Awan HH, Khalid M, et al. (2024) A hybrid residue based sequential encoding mechanism with XGBoost improved ensemble model for identifying 5-hydroxymethylcytosine modifications. *Sci Rep* 14: 20819. https://doi.org/10.1038/s41598-024-71568-z
14. Liu B, Wu H, Chou KC (2017) Pse-in-One 2.0: an improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nat Sci* 9: 67–91. https://doi.org/10.4236/ns.2017.94007
15. Chen W, Lin H, Chou KC (2015) Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. *Mol BioSyst* 11: 2620–2634. https://doi.org/10.1039/c5mb00155b
16. Khan S, Naeem M, Qiyas M (2023) Deep intelligent predictive model for the identification of diabetes. *AIMS Math* 8: 16446–16462. https://doi.org/10.3934/math.2023840
17. Kaur G, Sharma A (2023) A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis. *J Big Data* 10: 1–23. https://doi.org/10.1186/S40537-022-00680-6
18. Gumaei A, Hassan MM, Hassan MR, et al. (2019) A hybrid feature extraction method with regularized extreme learning machine for brain tumor classification. *IEEE Access* 7: 36266–36273. https://doi.org/10.1109/ACCESS.2019.2904145
19. Noor S, Naseem A, Awan HH, et al. (2024) Deep-m5U: a deep learning-based approach for RNA 5-methyluridine modification prediction using optimized feature integration. *BMC Bioinf* 25: 1–23. https://doi.org/10.1186/S12859-024-05978-1
20. Demir S, Sahin EK (2023) An investigation of feature selection methods for soil liquefaction prediction based on tree-based ensemble algorithms using AdaBoost, Gradient boosting, and XGBoost. *Neural Comput Appl* 35: 3173–3190. https://doi.org/10.1007/S00521-022-07856-4

21. Al-Jumaili MHA, Siddique F, Abul Qais F, et al. (2023) Analysis and prediction pathways of natural products and their cytotoxicity against HeLa cell line protein using docking, molecular dynamics and ADMET. *J Biomol Struct Dyn* 41: 765–777. https://doi.org/10.1080/07391102.2021.2011785
22. Gütter J, Kruspe A, Zhu XX, et al. (2022) Impact of training set size on the ability of deep neural networks to deal with omission noise. *Front Remote Sens* 3: 932431. https://doi.org/10.3389/frsen.2022.932431
23. Chicco D, Jurman G (2023) The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Min* 16: 1–23. https://doi.org/10.1186/S13040-023-00322-4
24. Noor S, AlQahtani SA, Khan S (2025) Chronic liver disease detection using ranking and projection-based feature optimization with deep learning. *AIMS Bioeng* 12: 50–68. https://doi.org/10.3934/bioeng.2025003
25. Khan S, Khan M, Iqbal N, et al. (2022) Deep-piRNA: Bi-layered prediction model for PIWI-interacting RNA using discriminative features. *Comput Mater Contin* 72: 2243–2258. https://doi.org/10.32604/cmc.2022.022901
26. Bibi N, Khan M, Khan S, et al. (2024) Sequence-based intelligent model for identification of tumor t cell antigens using fusion features. *IEEE Access* 12: 155040–155051. https://doi.org/10.1109/ACCESS.2024.3481244
27. Khan S, Khan MA, Khan M, et al. (2023) Optimized feature learning for anti-inflammatory peptide prediction using parallel distributed computing. *Appl Sci* 13: 7059. https://doi.org/10.3390/app13127059

AIMS Press