iRhm5BiRNN: Identification of RNA 5-Hydroxymethylcytosine Modifications Using Bidirectional-Recurrent Neural Network

Ashfaq Ahmad[1], Ali Raza[1], Muhammad Fahid Waqas[2], Muneeba Shamas[3], Sanya Chaudhary[1], Kifayat ullah khan[4]

[1] Department of Computer Science, MY University Islamabad, Pakistan

[2] Department of Computer Science, University of Mianwali, Pakistan

[3] Department of Computer Science, Lahore College for Women University, Pakistan

[4] Department of Electrical Engineering, Sarhad University of science & information technology, Pakistan

Abstract

One of the essential areas of study in RNA epigenetics is the role of RNA 5-hydroxymethylcytosine (5hmC), which has been implicated in numerous biological processes. The identification of 5hmC can be achieved using multiple sequencing-based technologies; however, these existing approaches are time-consuming, expensive, and labor-intensive. Therefore, there is a significant need to develop more reliable, efficient, and feasible computational methods to replace or, ideally, complement high-throughput technologies. Despite developing different deep learning and machine learning models, their performance is currently insufficient and limited. In this study, we proposed a new identification scheme for deep learning, specifically a bidirectional recurrent neural network (BiRNN), called iRhm5BiRNN which overcomes the restriction of using only input information up to a present future frame for training. The BiRNN is trained simultaneously in both forward and backward directions, enabling it to accurately identify RNA 5hmC sites in genome-wide DNA sequences efficiently and reliably. Our approach utilizes a Bidirectional Recurrent Neural Network (BiRNN) to derive the most dependable features from the constantly changing RNA sequences. We achieved an accuracy of 85.51% on the benchmark dataset using our proposed architecture, surpassing the performance of existing methods in all evaluation metrics. These findings demonstrate the superiority of our approach and its potential to advance the field of RNA epigenetics.

***Keywords*:** *Deep learning, RNA 5-hydroxymethylcytosine, bidirectional recurrent neural networks, one-hot encoding*

1. INTRODUCTION

Over the past few years, the identification of RNA has presented numerous complex challenges. Among these challenges, RNA modification is one of the most essential and demanding scientific endeavors [1]. The discovery of chemically modified nucleosides and pseudouridines dates back to 1950 [2]. Since, more than 160 distinct RNA modifications have been identified in various RNA types, including mRNA, tRNA, rRNA, and snRNA [3]. These modifications also affect critical RNA processes such as pre-RNA splicing, RNA export, and microRNA translation. Furthermore, RNA alterations have been linked to various human disorders, including cardiovascular disease, cancer, obesity, and diabetes [4]. Characterizing the distribution of RNA modifications in the transcriptome is crucial for understanding their biological properties. For instance, the hydroxymethylcytosine (hMeC) modification is likely to be oxidized by the Tet family of enzymes, suggesting that 5hmC is predominantly found within exons and introns of coding regions [5-7]. Studies have reported a high concentration of 5hmC alterations in the Drosophila brain and significant levels of 5hmC modification in the brain stem, cerebellum, and hippocampus, with evidence that 5hmC identification and modification reduce the MPTP-induced Parkinson's model in mice [8]. These findings indicate that RNA 5hmC modification affects microRNA expression and proteins in brain tissue. Additionally, 5hmC is involved in the epigenetic regulation of gene expression through altered interactions between RNA and proteins [9]. To further understand the effects of 5hmC in different organisms, it is necessary to examine its presence in the transcriptomes of other species. Previous studies have employed the hMeRIP-seq method to study Drosophila 5hmC's transcriptome [10]. However, there are drawbacks to using hMeRIP-seq and wet lab investigations to detect 5hmC sites throughout the genome. These drawbacks include the high cost of experimental reagents and the time and labor-intensive nature of the procedures. To address these limitations, developing a computational model becomes crucial, as it can identify 5hmC modification sites with greater accuracy, efficiency, and cost-effectiveness compared to traditional methods. This is especially significant given the increasing number of genome samples that need analysis. The research indicates a high prevalence of 5hmC deficiencies in the Drosophila brain. Similarly, previous studies have shown significant enrichment of 5hmC modifications in the mouse brain stem, hippocampus, and cerebellum. These findings suggest that 5hmC modification may be critical in brain tissue development [11]. Understanding the biological functions of 5hmC in the transcriptomes of different species is essential; however, the distribution of 5hmC in most

animal species still needs to be more adequately examined. The iRNA5hmC model is currently being developed as the first machine-learning model to predict RNA 5hmC modifications solely based on RNA sequence information [12]. It utilizes the k-mer spectrum and positional nucleotide binary vector as feature representations, offering robust alternatives to the standard methods. Although the performance of iRNA5hmC is commendable, there is still room for further improvement. Another computational model called iRhm5CNN has been used recently, employing a CNN architecture to extract significant aspects of primary RNA sequence representations for accurate identification of RNA [13]. While these existing approaches have shown significant progress in predicting RNA 5hmC, their accuracy still needs enhancement [12]. In this study, we proposed a simple yet effective architecture based on Bidirectional Recurrent Neural Networks (BiRNNs) for identifying RNA 5hmC sites solely based on the RNA sequences. We aimed to develop a sophisticated system to accurately identify RNA 5hmC sites without relying on pre-selected features or categorization. This approach offers the advantages of speed and high accuracy in predicting RNA 5mC sites based on primary RNA sequences. Deep learning-based computational models have proven highly efficient and effective in various applications, including sequencing, sentiment analysis, and natural language processing [14-16]. RNA sequences are represented using one-hot encoding, and we provide an overview of the chemical components of nucleotides, including functional groups, hydrogen bonds, ring configurations, and functions. By extracting key features from primary RNA sequence representations, the BiRNN architecture reliably identifies 5hmC RNA sequences. To evaluate the effectiveness of our approach, we conducted subsampling with a five-fold cross-validation. The iRhm5BiRNN model outperforms current state-of-the-art techniques such as iRNA5hmC [12] and iRhm5CNN [13] by a large margin, as demonstrated by experimental results. The iRhm5BiRNN achieves superior performance in terms of accuracy, AU ROC, AU PR, sensitivity, specificity, and MCC, reaching values of 85.51%, 93.22%, 93.58%, 82.38%, 88.63%, and 71.16%, respectively. These findings surpass the results reported in previous studies by a significant margin. Considering these factors, our method has the potential to become an accurate and efficient tool for detecting 5hmC.

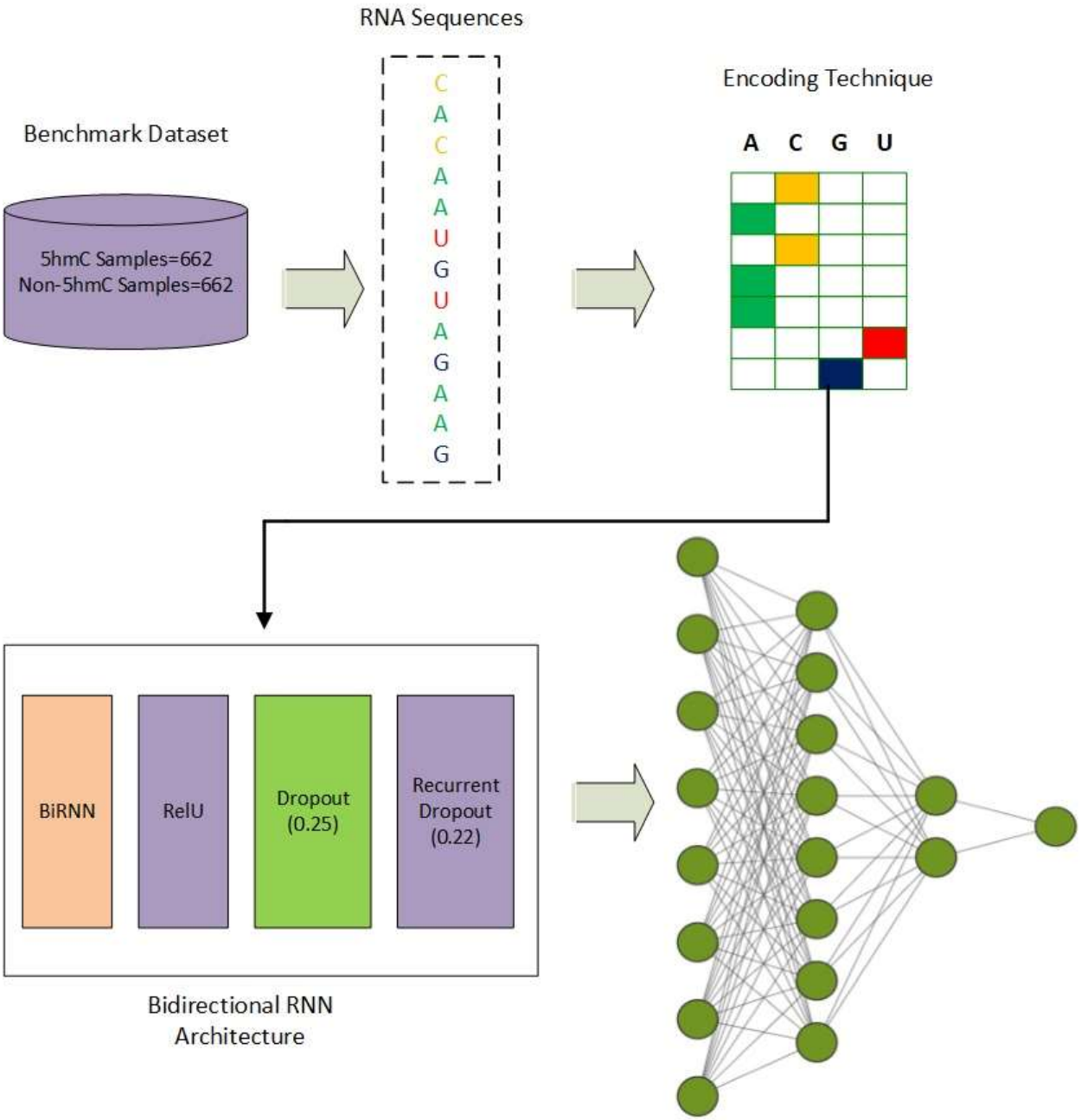

Figure 1. The proposed model framework.

2. MATERIALS AND METHODS DATASET

We present the benchmark dataset of deep learning techniques for predicting the RNA5hmC site, and the performance evaluation is all discussed in this section.

2.1. Dataset

A crucial aspect of developing a high-caliber bioinformatics tool is carefully selecting a reliable training dataset for the predictive model [17]. In this regard, Liu et al. [12] designed and implemented a dataset encompassing positive and negative samples. The balanced dataset consists of 1324 samples, with an equal distribution of 662 positive and 662 negative samples. Positive samples were obtained from Delatte et al. [10] and were characterized by 5hmC at the center. These positive samples were selected based on a less than 80% sequence similarity criterion. To obtain negative samples (i.e., non-5hmC sequences), the remaining intermediate cytosines that could not be identified as 5hmC using hMeRIP-seq were included. The length of each sample in the dataset was set to 41 nucleotides (nt). This comprehensive dataset forms the foundation for training the predictive model and enables the accurate identification of 5hmC sites.

2.2. Prediction Assuming Independent Outputs

2.2.1. Convolutional Neural Network

The Convolutional Neural Network (CNN) is a widely recognized discriminative deep learning model that eliminates the need for manual feature extraction by learning directly from the input data [14]. Figure 2 illustrates a CNN architecture consisting of multiple convolutional and pooling layers. The design of CNNs brings advantages to traditional Artificial Neural Networks (ANNs), such as regularized MLP networks. Each layer of a CNN considers the optimal parameters to generate useful outputs while simultaneously reducing model complexity. Additionally, CNNs employ a technique called dropout to mitigate over fitting issues in traditional networks. Due to its ability to handle a wide range of 2D shapes, CNNs find applications in various domains, including visual recognition, image analysis, image segmentation, and natural language processing [15]. Automatically detecting crucial properties without human intervention makes CNNs more powerful than regular networks. Numerous variations of CNNs exist in this field, including Visual Geometry Group (VGG) [17, 18], Xception [19], Inception [20] , ResNet [21], and many more.

Depending on their specific learning capabilities, these variations can be applied in diverse areas.

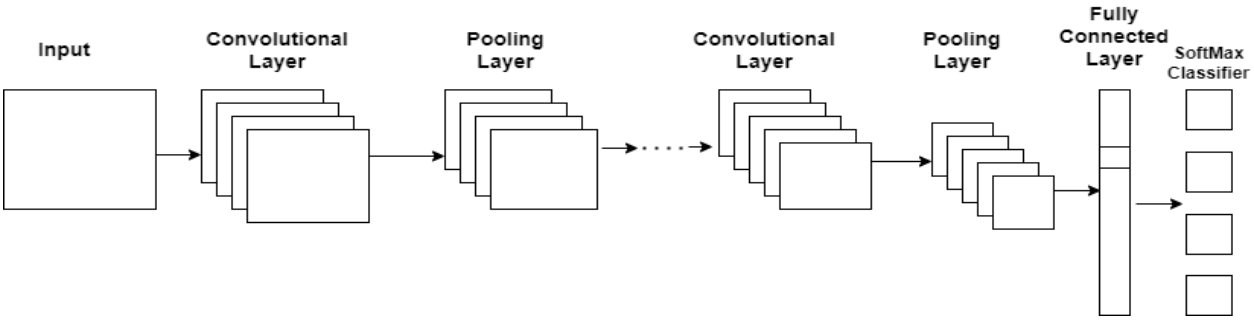

Figure 2: A Basic Structure of Convolutional neural network

2.2.2. Recurrent Neural Networks

The RNN model has gained significant popularity, especially in handling sequential data. Figure 3 depicts the unrolled structure of an RNN [22]. This architecture utilizes a feedback loop, where each node remains active by receiving input from the previous node at each time step. At each node, the input and the previous hidden state are combined to generate the current hidden state and output. The diagram below provides a visual representation of a simple recurrent neural network.

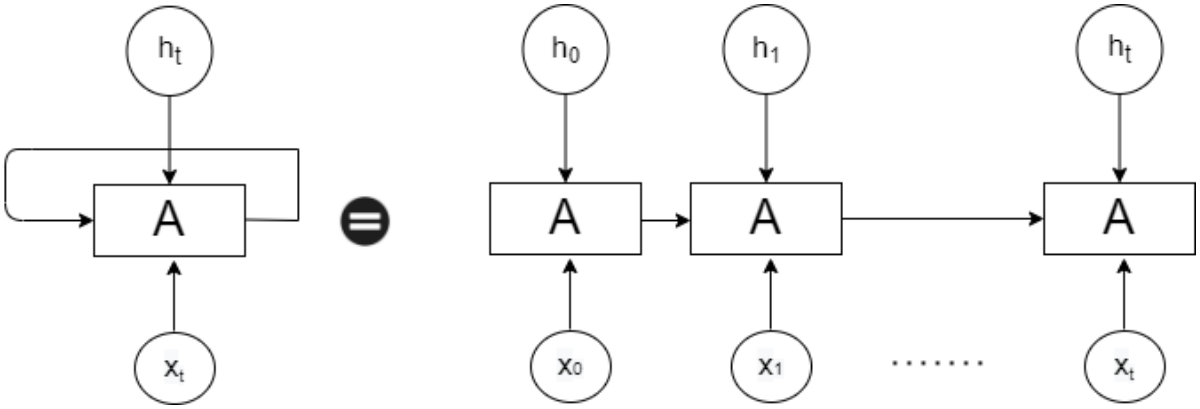

Figure 3: An unrolled recurrent neural network

2.3 The Proposed Predictor Model

The proposed model utilizes a Bidirectional Recurrent Neural Network (BiRNN), widely regarded as one of the most effective techniques for analyzing sequential data. This technique demonstrates a highly accurate prediction of RNA 5hmC sites compared to previous computational models. The BiRNN deep learning algorithm has gained significant recognition recently due to its exceptional performance and versatility. It efficiently captures the crucial features of RNA sequence

representations without the need for manually designed features. The following section will provide a detailed discussion of the proposed architecture.

2.3.1. Proposed BiRNN Architecture

In this section, we present and describe the architecture of the proposed BiRNN model, as illustrated in Figure 1. The model builds upon the foundation of RNN by incorporating more cell units in their hidden states, ranging from h_t to h_n. A bidirectional training approach is employed to enhance the model's performance. The architecture consists of four layers, each of which will be discussed in the following subsections.

Layer 1 (RNA Nucleotide's): The RNA sequences were fed into the Encoding Scheme before presented RNN models after being selected and preprocessed.

Layer 2 (One Hot Encoding):

The primary RNA sequences are encoded using the widely employed one-hot encoding method. One-hot encoding is considered the most significant, efficient, and frequently used approach for transforming categorical data, such as DNA and RNA sequences, into numerical form. In this encoding scheme, each nucleotide is represented by a binary vector. Specifically, Adenine (A) is represented as (1,0,0,0), Uracil (U) is represented as (0,0,0,1), Cytosine (C) is represented as (0,1,0,0), and Guanine (G) is represented as (0,0,1,0). Consequently, an RNA sequence of length n can be represented by a matrix of dimensions 4 by n, where each column corresponds to the one-hot encoded representation of a nucleotide. This encoding method enables the efficient numerical representation of RNA sequences, facilitating subsequent analysis and modeling.

Layer 3 (Bidirectional RNN): This section introduces the hidden blocks of the proposed model, namely BiRNN. These blocks consist of a forward track that processes the data section from left to right and a reverse track that analyzes the input from right to left. The following formulas can represent the forward recurrent sequence and the backward recurrent sequence:

$$\vec{h}_t = f(W_{z,\vec{h}} X_t + W_{\vec{h},\vec{h}} \vec{h}_{t-1} + b_{\vec{h}}) \quad (1)$$

$$\overleftarrow{h}_t = f(W_{z,\overleftarrow{h}} X_t + W_{\overleftarrow{h},\overleftarrow{h}} \overleftarrow{h}_{t+1} + b_{\overleftarrow{h}}) \quad (2)$$

$$y_t = (W_{\vec{h}} \vec{h}_t + W_{\overleftarrow{h},y} \overleftarrow{h}_t + b_y) \quad (3)$$

$$y_t = (W_{\vec{h},y}\vec{h}_t + W_{\overleftarrow{h},y}\overleftarrow{h}_t + b_y) \quad (4)$$

The formulas for the forward and backward recurrent sequences are defined as follows:

For the forward sequence: $\vec{h}_t = f(W_{p,q} * x_t + b_r)$ (5)

For the backward sequence: $\overleftarrow{h}_t = f(W_{p,q} * x_t + b_r)$ (6)

In these formulas, x represents the input feature vector, $\vec{h}$ denotes the activation vector on the forward (or backward) hidden layer. $W_{p,q}$ denotes the weight matrix, b_r represents the bias term, f represents the activation function applied to each node in the hidden layers, and y represents the posterior probability vector for the output label.

Layer 4 (Output layer): The predicted probability of characters for each step of t in the alphabet is calculated using a normal sigmoid function applied to the output layer. This performance can be shown in the following equation:

$$y_t = sigmoid(W_{\vec{h},y}\vec{h}_t + W_{\overleftarrow{h},y}\overleftarrow{h}_t + b_y) \quad (7)$$

2.4 Performance Evaluation

The proposed deep learning method named iRhm5-BiRNN was evaluated using several classification measures to predict RNA5hmC sites, including accuracy, sensitivity, specificity, MCC, and F1-score. These metrics are calculated below, and a corresponding confusion matrix was generated [13, 23-32]. Our proposed framework demonstrates comparable performance across five key metrics: accuracy, sensitivity, specificity, F1-score, and Matthew's correlation coefficient (MCC). These metrics are defined as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$Sen = \frac{TP}{TP + FN} \quad (9)$$

$$Spe = \frac{TN}{TN + FP} \quad (10)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (11)$$

$$F1 - Score = 2 * \frac{precision \times recall}{precision + recall} \quad (12)$$

The performance of the proposed model was evaluated based on the correct identification of positive samples (TP), correct identification of negative samples (TN), incorrect identification of positive samples (FP) as 5hmC sites, and incorrect identification of negative samples (FN) as non-5hmC sites. Four metrics were used to assess the performance: sensitivity, specificity, accuracy, and Matthew's correlation coefficient (MCC). Accuracy, sensitivity, and specificity range from 0 to 1, while Matthew's coefficient ranges from -1 to +1, with higher values indicating better performance. To visualize the diagnostic capacity of our predictor, a receiver operating characteristic (ROC) curve was plotted by comparing the true positive rate (1-SN) against the false positive rate (1-SP). The area under the ROC curve (AUC) was calculated as an effective performance parameter, as shown in Figure 4.

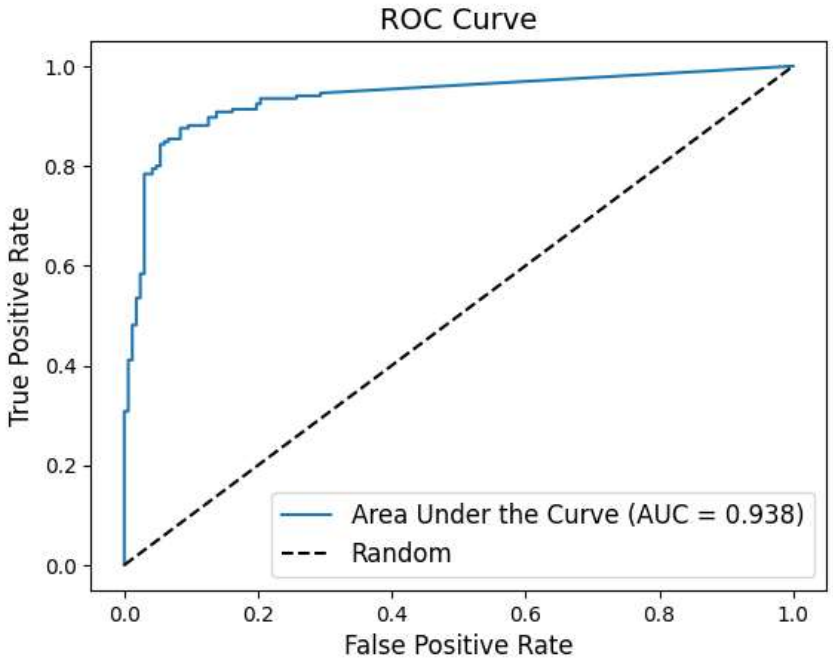

Figure 4. The ROC analysis with of the Proposed Model.

In addition, we employed the 5-fold cross-validation approach to evaluate the predictor's ability to forecast future outcomes [33]. To measure the prediction performance of the model, we utilized Precision-Recall (PR) curves and Receiver Operating Characteristic (ROC) curve. The ROC curve compares the true positive rate (TPR; 1-specificity) with the false positive rate (FPR; 1-specificity) at different thresholds, while the precision-recall curve plots precision (the proportion of real positives out of all predicted positives) against recall (sensitivity) at various thresholds. The PR curve is more sensitive and responsive to false positives compared to the ROC curve, which is particularly valuable when dealing with imbalanced sample sizes. Furthermore, the area under the curve (AUC) of the receiver operating characteristic is an objective measure of the prediction model's quality and is influenced by the number of observations. The AUC ranges from 0.5 to 1, with higher values indicating better prediction accuracy [34]. Finally, the confusion matrix provides a visual representation of the model's performance and is displayed to facilitate the evaluation.

3. RESULTS AND DISCUSSION

3.1. Performance Comparison using Different Deep Learning Classifier

In this experiment, we compared three different internal structures: CNN, RNN, and bidirectional RNN, to determine the best-performing model with higher metrics. The final architecture was selected after evaluating the performance of each model and adding a dropout layer. We utilized Benchmark datasets to assess the superiority of CNN, RNN, and bidirectional RNN. Table 1 presents the prediction accuracy achieved by each architecture. Among CNN, RNN, and BiRNN, accuracies of 84.94%, 82.67%, and 85.51% were obtained, respectively. BiRNN achieved the highest accuracy of 85.51% as shown in Table 1. Additionally, when considering the other five evaluation metrics, BiRNN demonstrated superior performance. BiRNN is considered more complex and faster than the other two architectures, enabling it to capture more features. Figure 6 and Table 1 indicate that the accuracies of CNN and RNN are significantly lower than that of BiRNN. Thus, based on our benchmark dataset, it is evident that BiRNN outperforms other classifiers in effectiveness. Figure 6 provides a graphical representation of various methods for predicting RNA 5hmC sites, and Table 1 summarizes the findings of RNA sequences using different models. To encode the sequences, we proposed a feature-based technique. The results

show an accuracy of 85.51%, sensitivity of 82.38%, specificity of 88.63%, MCC of 0.71, and F1-score of 85.04%.

3.2. Nucleotide Properties of Chemical for Sequence Encoding

RNA is composed of four nucleic acids: adenine (A), cytosine (C), guanine (G), and uracil (U), each having its distinct chemical properties [35, 36]. These chemical features of the nucleotides can be categorized into three classes: hydrogen bond strength, base type, and functional group (keto or amino) [37]. The basic structure of RNA nucleotides is illustrated in Figure 5.

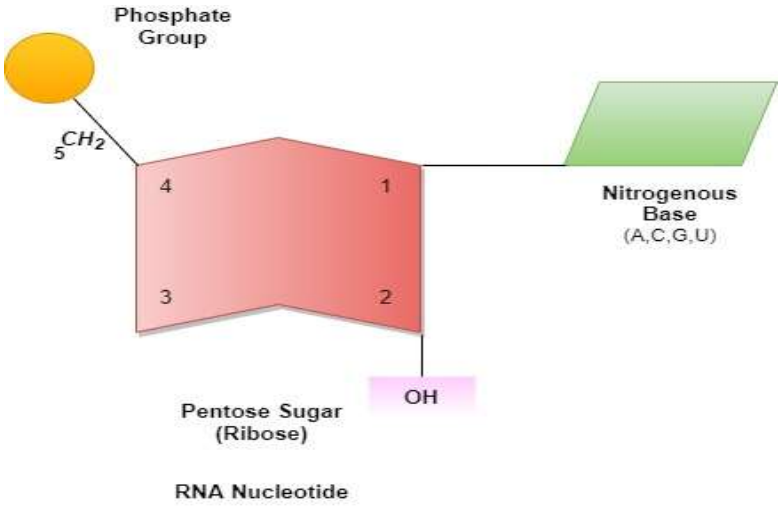

Figure 5: shows the RNA Nucleotide Structure

The RNA nucleotides consist of two purines, A and G, each containing two rings and C and U, which have a single-ring structure [37]. A weak association exists between A and U, while C and G have a strong relationship. A and C belong to the amino group, while G and U belong to the keto group. These three chemical conditions allow for the mapping of RNA nucleotides to three-dimensional Cartesian coordinates. Each coordinate is assigned a binary value of either 0 or 1. A value of 1 represents a purine or pyrimidine, while all other values are represented by 0 or 1. A weak hydrogen bond is represented by a value of 1, while a strong one is represented by 0. The amino group is denoted by a 1, whereas the keto group is denoted by a 0. Accordingly, A and G are represented as 1, C as 0, U as 1, and G as 1.

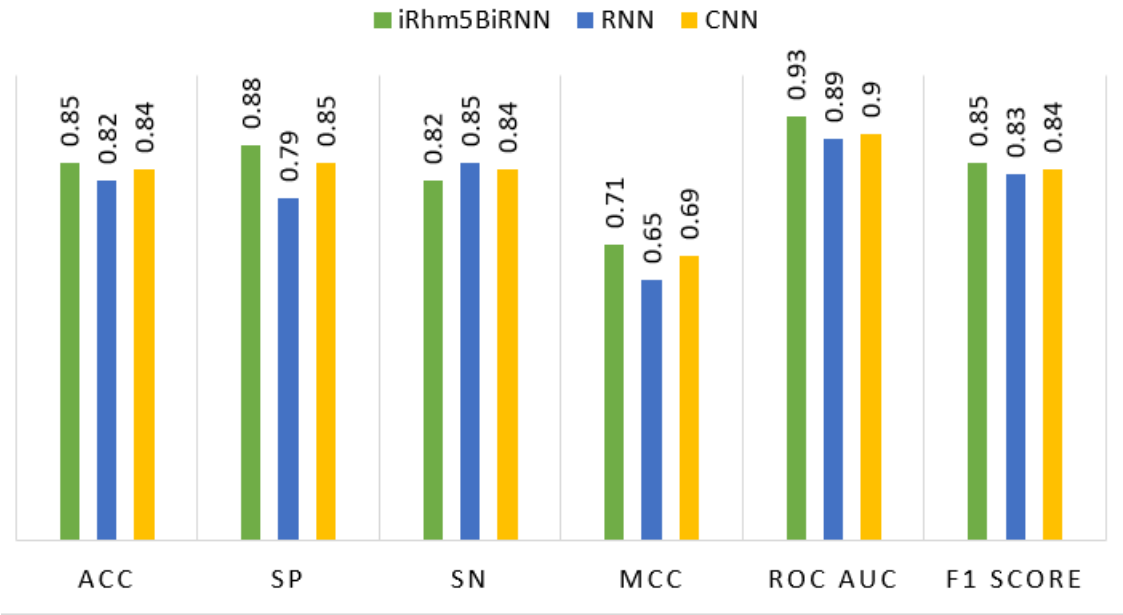

Figure 6. A graphical representation of various methods for predicting RNA 5hmC sites

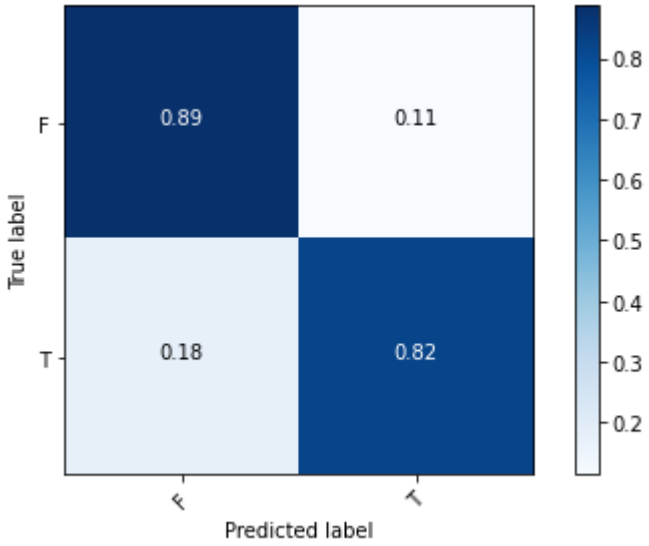

Figure 7. The proposed model confusion matrix.

TABLE 1. Result of different comparison of BiRNN with Simple RNN and CNN using different Model architectures

Methods	Acc (%)	Sn (%)	Sp (%)	MCC	F1-Score (%)
BiRNN	85.51	82.38	88.68	0.71	85.04
RNN	82.67	85.39	79.88	0.65	83.28
CNN	84.94	84.48	85.39	0.69	84.72

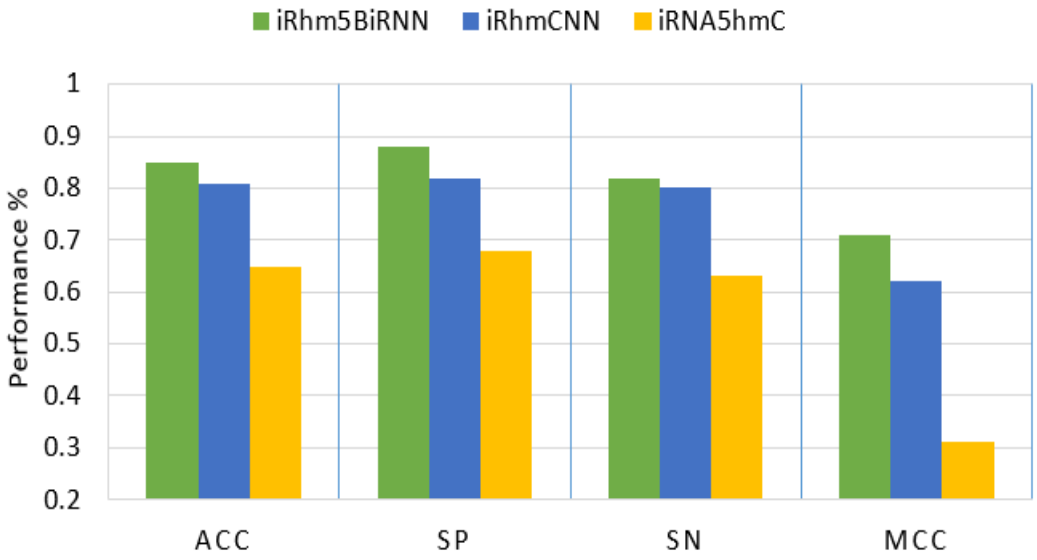

Figure 8. A graphical representation of Predicting RNA 5hmC sites comparison of the existing models

TABLE 2. A comparison of the proposed iRhm5BiRNN model's performance against the existing computational model.

Methods	Acc	Sn	Sp	MCC
iRhm5BiRNN	**0.85**	**0.82**	**0.88**	**0.71**
iRhm5CNN	0.81	0.82	0.80	0.62
iRNA5hmC	0.65	0.68	0.63	0.31

3.3. Performance comparison of the proposed model with existing methods

In this section, we compared the outcomes of the proposed model with those of existing deep learning and machine learning-based computational models that were available. We conducted a comparative analysis using 5-fold cross-validation to assess the performance of two models on the same dataset: iRNA5hmC [12], and iRhmCNN [13]. Each model was fine-tuned individually to achieve optimal performance, and the results are presented in Table 2. The iRhm5BiRNN model

outperforms the other classifiers in all four metrics, with an accuracy (ACC) of 85.51%, sensitivity (SN) of 82.38%, specificity (SP) of 88.63%, and Matthew's correlation coefficient (MCC) of 0.71. Furthermore, we employed ROC curve to compare the performance of different classifiers, as depicted in Figure 4, respectively. These results highlight the superior discriminative power of the iRhm5BiRNN model in distinguishing between 5hmC and non-5hmC sites. Our model demonstrates significantly improved performance when comparing our proposed technique to iRhm5CNN [12] and iRNA5hmC [13]. As shown in Table 2 and Figure 8, our proposed model outperforms the benchmark dataset in all performance parameters, with notable improvements of 4% accuracy, 8% specificity, and a 7% increase in MCC. These results indicate that iRhm5-BiRNN surpasses the existing methods and delivers superior performance across all four measures.

4. CONCLUSION

The identification of RNA 5hmC sites is of significant importance in various studies. In this study, we proposed a computational predictor, iRhm5BiRNN, which utilizes deep learning techniques to identify RNA 5hmC sites. Based on Bidirectional Recurrent Neural Networks (BRNN), our approach does not rely on prior knowledge or experimental information. Accurately identifying RNA 5hmC sites is crucial for exploring their diverse and yet unknown biological functions. Our proposed model, iRhm5BiRNN, leverages a straightforward BiRNN architecture to extract relevant features for distinguishing between RNA 5hmC and non-5hmC sites. We also conducted a study to determine the optimal hyper-parameters for BiRNN models, including the choice of the optimizer. The findings revealed that selecting the best hyper-parameters leads to superior results. Compared to previous models, our proposed model achieved a classification accuracy of 85.56%. The iRhm5-BiRNN model outperformed other state-of-the-art techniques across all evaluation metrics. Consequently, our proposed iRhm5BiRNN model demonstrates a more robust predictive capability for RNA 5hmC sites, making it a valuable tool for more accurate clinical decisions. In the future, we aim to evaluate the effectiveness of our approach in identifying other RNA modification sites. This ongoing investigation will contribute to further advancements in RNA modification detection.

References

1. Cohn, W.E. and E. Volkin, *Nucleoside-5'-phosphates from ribonucleic acid.* Nature, 1951. **167**: p. 483-484.
2. Qian, S.-B., *1, 2 1 Division of Nutritional Sciences, Cornell University, Ithaca, New York 14853, USA; email: 2 Graduate Field of Biomedical and Biological Sciences, Cornell University, Ithaca, New York 14853.* USA Annual Review of Nutrition. **40**: p. 51-75.
3. Roundtree, I.A., et al., *Dynamic RNA modifications in gene expression regulation.* Cell, 2017. **169**(7): p. 1187-1200.
4. Jonkhout, N., et al., *The RNA modification landscape in human disease.* Rna, 2017. **23**(12): p. 1754-1769.
5. Miao, Z., et al., *5-hydroxymethylcytosine is detected in RNA from mouse brain tissues.* Brain research, 2016. **1642**: p. 546-552.
6. Yuan, F., et al., *Bisulfite-free and base-resolution analysis of 5-methylcytidine and 5-hydroxymethylcytidine in RNA with peroxotungstate.* Chemical communications, 2019. **55**(16): p. 2328-2331.
7. Huber, S.M., et al., *Formation and abundance of 5-hydroxymethylcytosine in RNA.* Chembiochem, 2015. **16**(5): p. 752-755.
8. Rácz, I., I. Király, and D. Lásztily, *Effect of light on the nucleotide composition of rRNA of wheat seedlings.* Planta, 1978. **142**: p. 263-267.
9. Li, B., A.D. Ellington, and X. Chen, *Rational, modular adaptation of enzyme-free DNA circuits to multiple detection methods.* Nucleic acids research, 2011. **39**(16): p. e110-e110.
10. Delatte, B., et al., *Transcriptome-wide distribution and function of RNA hydroxymethylcytosine.* Science, 2016. **351**(6270): p. 282-285.
11. Zhang, H.-Y., et al., *The existence of 5-hydroxymethylcytosine and 5-formylcytosine in both DNA and RNA in mammals.* Chemical Communications, 2016. **52**(4): p. 737-740.
12. Liu, Y., et al., *iRNA5hmC: the first predictor to identify RNA 5-hydroxymethylcytosine modifications using machine learning.* Frontiers in bioengineering and biotechnology, 2020. **8**: p. 227.
13. Ali, S.D., et al., *Prediction of rna 5-hydroxymethylcytosine modifications using deep learning.* IEEE Access, 2021. **9**: p. 8491-8496.
14. LeCun, Y., et al., *Gradient-based learning applied to document recognition.* Proceedings of the IEEE, 1998. **86**(11): p. 2278-2324.
15. Géron, A., *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. 2022: " O'Reilly Media, Inc.".
16. Sarker, I.H., *Deep cybersecurity: a comprehensive overview from neural network and deep learning perspective.* SN Computer Science, 2021. **2**(3): p. 154.
17. Krizhevsky, A., I. Sutskever, and G.E. Hinton, *Imagenet classification with deep convolutional neural networks.* Communications of the ACM, 2017. **60**(6): p. 84-90.
18. He, K., et al., *Spatial pyramid pooling in deep convolutional networks for visual recognition.* IEEE transactions on pattern analysis and machine intelligence, 2015. **37**(9): p. 1904-1916.
19. Chollet, F. *Xception: Deep learning with depthwise separable convolutions*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
20. He, K., et al. *Deep residual learning for image recognition*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
21. Szegedy, C., et al. *Going deeper with convolutions*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

22. Hinton, G., et al., *Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups.* IEEE Signal processing magazine, 2012. **29**(6): p. 82-97.
23. Razaa, A., et al., *iAFP-ET: A robust approach for accurate identification of antifungal peptides using extra tree classifier and multi-view fusion.*
24. Akbar, S., et al., *Identifying Neuropeptides via Evolutionary and Sequential based Multi-perspective Descriptors by Incorporation with Ensemble Classification Strategy.* IEEE Access, 2023.
25. Akbar, S., et al., *Prediction of Amyloid Proteins using Embedded Evolutionary & Ensemble Feature Selection based Descriptors with eXtreme Gradient Boosting Model.* IEEE Access, 2023.
26. Akbar, S., et al., *Prediction of Antiviral peptides using transform evolutionary & SHAP analysis based descriptors by incorporation with ensemble learning strategy.* Chemometrics and Intelligent Laboratory Systems, 2022. **230**: p. 104682.
27. Ahmad, A., et al., *Identification of antioxidant proteins using a discriminative intelligent model of k-space amino acid pairs based descriptors incorporating with ensemble feature selection.* Biocybernetics and Biomedical Engineering, 2022. **42**(2): p. 727-735.
28. Akbar, S., et al., *iAtbP-Hyb-EnC: Prediction of antitubercular peptides via heterogeneous feature representation and genetic algorithm based ensemble learning model.* Computers in Biology and Medicine, 2021. **137**: p. 104778.
29. Akbar, S., et al., *iHBP-DeepPSSM: Identifying hormone binding proteins using PsePSSM based evolutionary features and deep learning approach.* Chemometrics and Intelligent Laboratory Systems, 2020. **204**: p. 104103.
30. Akbar, S., et al., *cACP: Classifying anticancer peptides using discriminative intelligent model via Chou's 5-step rules and general pseudo components.* Chemometrics and Intelligent Laboratory Systems, 2020. **196**: p. 103912.
31. Akbar, S., et al., *iAFP-gap-SMOTE: an efficient feature extraction scheme gapped dipeptide composition is coupled with an oversampling technique for identification of antifreeze proteins.* Letters in Organic Chemistry, 2019. **16**(4): p. 294-302.
32. Akbar, S. and M. Hayat, *iMethyl-STTNC: Identification of N6-methyladenosine sites by extending the idea of SAAC into Chou's PseAAC to formulate RNA sequences.* Journal of theoretical biology, 2018. **455**: p. 205-211.
33. Li, C.-C. and B. Liu, *MotifCNN-fold: protein fold recognition based on fold-specific features extracted by motif-based convolutional neural networks.* Briefings in Bioinformatics, 2020. **21**(6): p. 2133-2141.
34. Greiner, M., D. Sohr, and P. Göbel, *A modified ROC analysis for the selection of cut-off values and the definition of intermediate results of serodiagnostic tests.* Journal of immunological methods, 1995. **185**(1): p. 123-132.
35. Chen, W., et al., *PAI: Predicting adenosine to inosine editing sites by using pseudo nucleotide compositions.* Scientific reports, 2016. **6**(1): p. 1-7.
36. Chen, W., et al., *Identifying 2'-O-methylationation sites by integrating nucleotide chemical properties and nucleotide compositions.* Genomics, 2016. **107**(6): p. 255-258.
37. Chen, W., et al., *iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties.* Bioinformatics, 2017. **33**(22): p. 3518-3523.