IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

A CNN-based RNA N6-methyladenosine site predictor for multiple species using heterogeneous features representation

WALEED ALAM[1], SYED DANISH ALI[1,2], HILAL TAYARA[1,3], AND KIL TO CHONG.[1,3]

[1]Department of Electronics and Information Engineering, Jeonbuk National University, Jeonju 54896, South Korea
[2]Department of Electrical Engineering, The University of Azad Jammu and Kashmir, Muzaffarabad, 13100, Pakistan
[3]Advanced Electronics and Information Research Center, Jeonbuk National University, Jeonju 54896, South Korea

Equal Contribution: Waleed Alam and Syed Danish Ali.
Corresponding authors: Kil To Chong (kitchong@jbnu.ac.kr) and Hilal Tayara (hilaltayara@jbnu.ac.kr)

This research was supported by the Brain Research Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. NRF-2017M3C7A1044816).

ABSTRACT Post-transcriptional modification such as N6-methyladenosine (m6A) has a crucial role in the stability and regulation of gene expression. Therefore, the identification of m6A is highly required for understanding the functional mechanisms of biological processes. Several machine learning techniques based on handy craft feature extraction methods have been proposed to facilitate the laborious work. However, due to the inefficient feature extraction, these techniques increase the computational complexity and thereby affect the identification accuracy of m6A.This paper proposes a fast and reliable predictive model for the identification of m6A sites. The proposed model is based on the convolutional neural network (CNN) which extracts the most significant features from the RNA sequences encoded by concatenating one-hot and nucleotide chemical properties. The proposed model is applied and tested on multiple species benchmark datasets and evaluated against the state-of-art predictive models. The results indicate that the proposed model achieves high accuracy of 93.6 %, 93.8 %, 85.0 % and 92.5 % on the datasets of *Homo sapiens (H.sapien)*, *Mus musculus (M.musculs)*, *Saccharomyces cerevisiae (S.cerevisiae)*, and *Arabidopsis thaliana (A.thaliana)*, respectively.The proposed model could be used to facilitate the researcher's community in m6A identification. In addition, an easy to use web server is made available at https://home.jbnu.ac.kr/NSCL/pm6acnn.htm.

INDEX TERMS Post-transcription modification, RNA methylation , Sequence analysis, convolutional neural network, deep learning

I. INTRODUCTION

Methyladenosine (m6A) is the most frequently occurring RNA modification among more than 160 RNA modifications. It exists in eukaryotes including yeast, insects and mammals [1]–[5]. The m6A is the adenosine base methylated at the sixth position of the nitrogen. The m6A is associated with various biological processes including RNA structural dynamics [6], cell differentiation, and reprogramming [5], localization and degradation of RNA [7], alternative splicing [8], circadian clock regulation [9] and primary microRNA processing [10], Therefore, the understanding of the functional mechanism of this biological process is vital. In recent past, to identify m6A sites, high-throughput experimental methods were utilized such as m6A-seq [11] and MeRIP-Seq [12] . The m6A-seq provides mouse and human m6A modification landscape in a transcriptome-wide manner depending on massively parallel sequencing and antibody-mediated capture. While MeRIP-Seq identifies the transcripts which are adenosine methylation substrates and provides an understanding of mammalian transcriptome epigenetic regulation. The experimental methods are inefficient in cost and time as well as incompetent locating the position of m6A site. We aim to overcome the shortfall in the efficient identification of the m6A sites accurately and rapidly. Therefore, The development of computational models is very crucial.

Most of the previous works relied on machine learning and hand-crafted features. Chen et al. [13], [14] proposed two SVM models for m6A identification in *S.cerevisiae*

using nucleotide chemical properties (NCP) with nucleotide frequency and pseudo nucleotide composition. Liu et al. [15] used physical-chemical properties with the SVM model for m6A identification in *S.cerevisiae*. MethyRNA [16] was proposed to find m6A sites in *H.sapiens* and *M.musculus* using SVM and NCP with nucleotide frequency. iMethyl-STTNC [17] utilized split-tetra-nucleotide composition with SVM for m6A sites identification in *H.sapiens* and *S.cerevisiae*. M6AMRFS [18] used the dinucleotide binary encoding with the eXtreme Gradient Boosting algorithm for m6A identification in *H.sapiens*, *M.musculus*, *S.cerevisiae*, and *A.thalinia*. iMRM [19] used the NCP with the eXtreme Gradient Boosting algorithm for m6A identification in *H.sapiens*, *M.musculus*, and *S.cerevisiae*. The aforementioned machine learning-based algorithms were mainly oriented upon the handy crafted features that require the domain knowledge for accurate prediction of the proposed predictor. These features are fabricated in such a way that the information regarding the pattern in the sequence must be maintained. Whereas the deep learning-based computational architectures are capable of extracting the most important features from the sequences without any human intervention which leads to a significantly accurate and robust computational model. Currently, the deep learning-based models achieve remarkable results in the field of natural language processing [20], [21], image recognition [22], [23], speech recognition [24] and also in the field of computational biology [25]–[33].

Nazari et al. proposed iN6-Methyl (5-step) [34] in which they utilized the deep learning-based CNN architecture for the prediction of m6A sites in the benchmark datasets of *H.sapiens*, *M.musculus*, and *S.cerevisiae*. They extracted the feature using a word2vec model, which is a natural language processing model in which each sequence is manually divided into words having a length of k using a k-mer technique. The k was set to 3 and each word was mapped into its corresponding feature representation. The word2vec has to be trained on the whole-genome to produce a vector space, therefore, it is a computationally costly and time-consuming process.

In this regard, to fill the gaps in terms of performance and computational cost in existing computational models, we propose a simple and efficient CNN-based architecture for the identification of m6A sites in RNA sequences. We call it pm6A-CNN. The input RNA sequences are represented by the combination of one-hot encoding and nucleotide chemical properties (NCP). The nucleotide chemical properties are the most basic representation of nucleotides with respect to the functional groups, hydrogen bond and ring structure. The architecture of CNN is able to extract the most important features from RNA sequences representation automatically which enables the pm6A-CNN to identify the m6A sites more accurately and robustly. The grid search algorithm is utilized to select the optimum hyper-parameters of the pm6A-CNN. The performance of pm6A-CNN is evaluated using subsampling (k-fold cross-validation) method by setting the value of k to 10 for keeping the consistency with state of the art models. As an achievement, the pm6A-CNN outperformed the existing computational models. Finally, a user-friendly web server is constructed and made available at https://home.jbnu.ac.kr/NSCL/pm6acnn.htm.

II. MATERIALS AND METHODS

This section includes the benchmark dataset, the proposed model and the performance evaluation.

A. BENCHMARK DATASET

In this study, we used four different species benchmark datasets which are namely *H.sapien*, *M.musculs*, *S.cerevisiae*, and *A.thaliana*. All sequences of these four benchmark datasets have Adenines at the center. The positive sequences have experimentally validated methyladenosine (m6A) sites whereas the negative ones are not methyladenosine (m6A) sites. The *H.sapiens* benchmark dataset was prepared by Chen et al. [16] in 2017, consisting of 1130 positive sequences and 1130 negative sequences with the length of 41nt for each sequence. Dominissini et al. [11] in 2012 prepared the *M.musculus* benchmark dataset where the length of each sequence is 41nt. The benchmark dataset of *M.musculus* contains 725 positive sequences and 725 negative sequences. Chen et al. [13] created the benchmark dataset of the *S.cerevisiae* in 2015. It includes 1307 positive sequences and 1307 negative sequences where the length of each sequence is 51nt. The benchmark dataset of *A.thaliana* was prepared by Wang and Yan [35] in 2018. It includes 2100 positive sequences and 2100 negative sequences and the length of each sequence is 101nt. The summary of the benchmark datasets is listed in Table 1. In this paper, we used k-fold cross-validation in order to evaluate the performance of the proposed model. According to the recent literature for computational models, evaluation of the model using k-fold cross-validation or jackknife test, having a testing dataset is not mandatory as the outcome of k-fold combinations can be considered as different independent test datasets.

TABLE 1: The summary of multiple species benchmark datasets.

Species	Sequences		Length
H.sapiens	Positive	1130	41nt
	Negative	1130	
M.musculus	Positive	725	41nt
	Negative	725	
S.cerevisiae	Positive	1307	51nt
	Negative	1307	
A.thaliana	Positive	2100	101nt
	Negative	2100	

B. THE PROPOSED MODEL

We propose a deep learning-based CNN architecture that takes RNA sequence as an input as shown in Figure 1.

The optimum hyper-parameters are found using a grid search method. The ranges of hyper-parameters are enlisted in Table 2.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3002995, IEEE Access

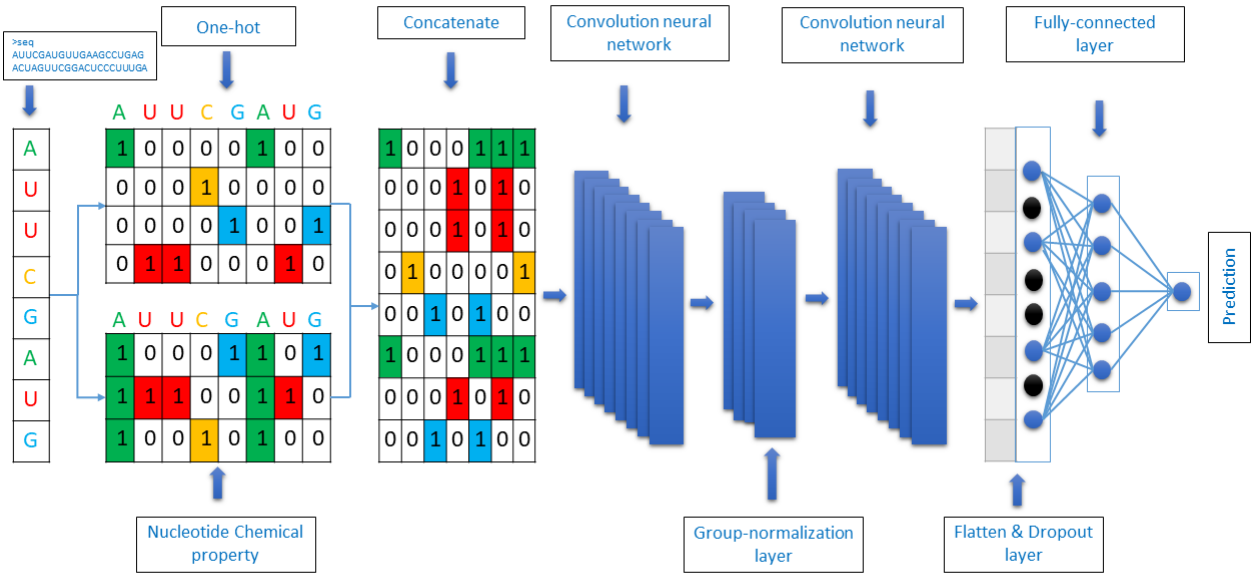

FIGURE 1: Illustration of the proposed model pm6A-CNN.

TABLE 2: The ranges of the tuned hyper-parameters.

The Hyper-Parameters	Range
Filters of convolution layers	[5,8,10,16]
Filter size	[3,5,7]
Stride lenght	[2,3,4]
Dropout	[0.2,0.3,0.4,0.5]
Dense layer neurons	[5,8,10,16]

The RNA sequence is represented by the combination of two commonly used encoding techniques one-hot encoding and nucleotide chemical properties (NCP). In one-hot encoding A is represented as (1,0,0,0), C is represented as (0,1,0,0), G is represented as (0,0,1,0) and U is represented as (0,0,0,1). While NCP is the representation of each nucleotide in the RNA sequence based on their three chemical groups in the three-dimensional Cartesian coordinate system. As each of the four nucleotides in the RNA sequences namely, Adenine (A), Cytosine (C), Guanine (G) and Uracil (U) consists of different chemical properties. Considering the ring structures, A and G are purines consisting of two rings. While C and U are pyrimidines consisting of one ring. In terms of secondary structure formation, the bonds between A and U are weak hydrogen bonds whereas between C and G are strong hydrogen bonds. Also, regarding the chemical functionality, A and C can be grouped into an amino group, while G and U into a group called the keto group. According to these three chemical properties, the four nucleotides can be categorized into three distinct groups which are represented in the three-dimensional Cartesian coordinate system by assigning the value of 1 or 0. Therefore, A is denoted by coordinates (1, 1, 1), C is represented by coordinates (0, 1, 0), G is represented by coordinates (1, 0, 0) and U is denoted by coordinates (0, 0, 1). The visual representation of NCP is shown in Figure 2.

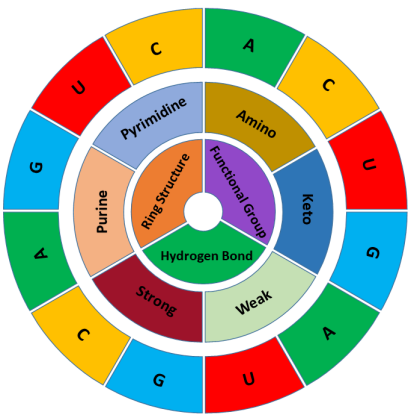

FIGURE 2: Diagrammatic representation of nucleotide chemical properties.

The vectors produced by utilizing one hot-encoding and NCP are concatenated together resulting in representing the RNA sequence by a seven-channel vector. The resulted vector is passed to a CNN model with two Conv1D layers and two fully connected layers. Each Conv1D layer is followed by a ReLU nonlinear activation function. In addition, the first Conv1D is followed by a group normalization [36] by setting the size of the group to four. The learned features using the two convolution layers are passed to the dropout

layer with a dropout rate of 0.5 and then to two fully connected layers. The first fully connected layer is followed by a ReLU activation function. The second fully connected layer is a one-node layer with a sigmoid activation function. L2 regularization method is used for weights and bias of the filters to avoid overfitting. Adam optimizer with the learning rate of 0.001 is utilized for model optimization. The binary cross-entropy is utilized as a loss function. The Batch size of 32 and early stopping based on validation loss is utilized for the maximum number of training iterations. Keras [37] framework is utilized for the implementation of the proposed model pm6A-CNN.

$$\mathrm{Conv}(S)_{ij} = \mathrm{ReLU}\left(\sum_{s=0}^{Z-1}\sum_{n=0}^{I-1} W_{sn}^{k} S_{j+s,n}\right) \tag{1}$$

Equation 1 represents Conv1D where S represents the input of the RNA sample, k denotes the index of the filter, and j denotes the index of the output position. Each W^k is convolution filter having $Z \times I$ weight matrix, where Z represents the size of the filter while I denotes the number of input channels.

$$\mathrm{f} = w_{d+1}\sum_{k=1}^{d} m_k w_k z_k \tag{2}$$

Equation 2 represents the dense layer where w_{d+1} is an additive bias term, m_k shows the dropout operator which is appraised from Bernoulli distribution , z_k is the 1×d dimensional feature vector, and w_k denotes the weight of z_k from the previous layer.

$$\mathrm{ReLU}(x) = \begin{cases} x & if x > 0 \\ 0 & if x \leq 0 \end{cases} \tag{3}$$

Equation 3 shows the ReLU activation function where x is the input.

$$\mathrm{Sigmoid}(x) = \frac{1}{1+e^{-x}} \tag{4}$$

Equation 4 is the representation of the sigmoid activation function.

C. PERFORMANCE EVALUATION

To evaluate the performance of our proposed model we utilized the 10-fold cross-validation technique. The benchmark datasets were divided into ten mutually exclusive folds. Where one fold is reserved for testing of the proposed model, one fold for validation of the model, while the remaining folds were used for training of the proposed model. This is a recursive process which occurs ten times. The final approximation of the performance was taken into account by taking the average outcome of ten folds. The common performance metrics used in this study and several existing computational models [38]–[41] are Accuracy (ACC), Sensitivity (SN), Specificity (SP), and Mathew's correlation coefficient (MCC). These performance metrics are formulated as:

$$\mathrm{ACC} = 1 - \left(\frac{N_{-}^{+} + N_{+}^{-}}{N^{+} + N^{-}}\right) \tag{5}$$

$$\mathrm{SN} = 1 - \left(\frac{N_{-}^{+}}{N^{+}}\right) \tag{6}$$

$$\mathrm{SP} = 1 - \left(\frac{N_{+}^{-}}{N^{-}}\right) \tag{7}$$

$$\mathrm{MCC} = \frac{1 - \left(\frac{N_{-}^{+}+N_{+}^{-}}{N^{+}+N^{-}}\right)}{\sqrt{\left(1 + \frac{N_{+}^{-}-N_{-}^{+}}{N^{+}}\right)\left(1 + \frac{N_{-}^{+}-N_{+}^{-}}{N^{-}}\right)}} \tag{8}$$

where N^+ is the representation of methyladenosine sties while non-methyladenosine sites are represented by N^-. N_+^- represents the methyladenosine sites incorrectly identified as non- methyladenosine. N_-^+ states the number of non-methyladenosine site that was predicted as methyladenosine sties.

III. RESULTS AND DISCUSSION

The proposed model was evaluated using the 10-folds cross-validation on four species benchmark datasets. In order to study the effectiveness of incorporating nucleotide chemical properties, we conducted three experiments. The first experiment used only a one-hot encoding for RNA sequence representation. The second experiment used only nucleotide chemical properties encoding for RNA sequence representation. The third experiment integrated the representation of both encoding methods (one-hot and nucleotide chemical properties). The results of these experiments are shown in Figure 3 and Table 3.

TABLE 3: The performance of the proposed model using different sequence representation methods.

Species	Encoding method	ACC	SN	SP	MCC
H.sapiens	One-hot	0.928	0.878	0.978	0.862
	NCP	0.911	0.846	0.975	0.83
	One-hot + NCP	0.936	0.886	0.986	0.878
M.musculus	One-hot	0.935	0.893	0.976	0.874
	NCP	0.928	0.895	0.961	0.859
	One-hot + NCP	0.938	0.904	0.972	0.88
S.cerevisiae	One-hot	0.829	0.824	0.834	0.662
	NCP	0.833	0.85	0.816	0.668
	One-hot + NCP	0.85	0.846	0.855	0.703
A.thaliana	One-hot	0.918	0.926	0.91	0.837
	NCP	0.907	0.921	0.892	0.815
	One-hot + NCP	0.925	0.923	0.926	0.85

It can be seen that the integrated representation produced better results in the identification of m6A sites. Furthermore, Figure 4 shows the AUC of the proposed model along with standard deviation errors in 10 folds using the benchmark datasets of *H.sapiens*, *M.musculus*, *S.cerevisiae*, and *A.thaliana*. The confusion matrix of the proposed model is also visualized in Figure 5. Moreover, to assess the dominance of the pm6A-CNN on the base of performance. We

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3002995, IEEE Access

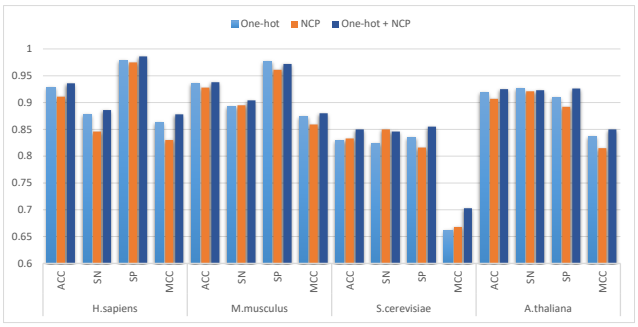

FIGURE 3: The performance of the proposed model using different feature representation

compared our model with the best known existing computational models including iMRM [19], iN6-Methyl (5-step) [34], M6AMRFS [18], and RFAthM6A [35]. The iMRM [19] and iN6-Methyl (5-step) [34] used only three benchmark datasets including *H.sapiens*, *M.musculus*, and *S.cerevisiae* and outperformed the M6AMRFS [18]. Whereas the RFAthM6A [35] showed a higher performance in comparison to M6AMRFS [18] by utilizing only the fourth benchmark dataset i.e. *A.thaliana*. The comparison between the computational models is illustrated in Table 4 and Figure 6. The '-' sign in Table 4 means that it was not possible to obtain these metrics for the corresponding tools.

TABLE 4: Summary outcomes of proposed model pm6A-CNN comparison with the existing computational models.

Species	Method	ACC	SN	SP	MCC
H.sapiens	M6AMRFS	0.91	0.82	1	0.833
	iN6-Methy	0.911	0.821	1	0.835
	iMRM	0.91	0.824	0.995	0.82
	pm6A-CNN	0.936	0.886	0.986	0.878
M.musculus	M6AMRFS	0.793	0.898	0.828	0.758
	iN6-Methy	0.895	0.788	1	0.807
	iMRM	0.889	0.783	0.995	0.779
	pm6A-CNN	0.938	0.904	0.972	0.88
S.cerevisiae	M6AMRFS	0.742	0.752	0.733	0.48
	iN6-Methy	0.753	0.761	0.746	0.507
	iMRM	0.777	0.77	0.78	0.555
	pm6A-CNN	0.85	0.846	0.855	0.703
A.thaliana	RFAthM6A	0.854	0.873	0.835	0.709
	M6AMRFS	0.81	0.806	0.814	0.621
	iMRM	-	-	-	-
	iN6-Methy	-	-	-	-
	pm6A-CNN	0.925	0.923	0.926	0.85

It is depicted that our proposed model outperformed all other competing methods. The performance of the proposed model for *H.sapiens*, and *M.musculus* benchmark datasets is higher than the state-of-the-art model iN6-Methyl (5-step) [34]. For the *H.sapiens* benchmark dataset, the improvements in terms of ACC, SN, MCC, and AUC are 2.5%, 6.5%, 4.3%, and 6.2%, respectively. In the benchmark dataset of *M.musculus*, the improvements in ACC, SN, MCC, and AUC are 4.3%, 11.9%, 7.3% and 5.8%, respectively. Whereas for the benchmark dataset of *S.cerevisiae*, the proposed model outperformed iMRM [19] in all the performance metrics ACC, SN, SP, MCC, and AUC by 7.3%, 7.6%, 7.5%, 15.6%, and 7.4%, respectively. Finally, for the benchmark dataset of *A.thaliana*, the proposed model improved ACC, SN, SP, and MCC by 7.1%,5.0%,9.1%,14.1%, respectively. The achieved results of the proposed model in terms of all performance metrics for all the benchmark datasets show the robustness of the proposed model in the identification of m6A site using the combination of two different encoding techniques for the representation of RNA sequences.

IV. WEB-SERVER

The proposed model has been made available for free access at https://home.jbnu.ac.kr/NSCL/pm6acnn.htm. This web server supports processing direct sequence inputs as well as uploading Fasta files. The webserver was built using the Flask library in Python.

CONCLUSION

In this study, we proposed an efficient deep learning-based CNN architecture for the identification of m6A sites in multiple species. The CNN based predictor extracts the most important features by utilizing the combination of one-hot encoding and nucleotide chemical properties for the representation of RNA sequences. This combination helped the predictor to accomplish more effective and efficient outcomes for the identification of m6A sites. Moreover, it is anticipated that the established predictor along with webserver would be effective facilitation for the researchers to explore the functional procedure of m6A sites.

REFERENCES

[1] P. Boccaletto, M. A. Machnicka, E. Purta, P. Piątkowski, B. Bagiński, T. K. Wirecki, V. de Crécy-Lagard, R. Ross, P. A. Limbach, A. Kotter et al., "Modomics: a database of rna modification pathways. 2017 update," Nucleic acids research, vol. 46, no. D1, pp. D303–D307, 2018.

[2] Z. Bodi, J. D. Button, D. Grierson, and R. G. Fray, "Yeast targets for mrna methylation," Nucleic acids research, vol. 38, no. 16, pp. 5327–5335, 2010.

[3] R. Levis and S. Penman, "5′-terminal structures of poly (a)+ cytoplasmic messenger rna and of poly (a)+ and poly (a)- heterogeneous nuclear rna of cells of the dipteran drosophila melanogaster," Journal of molecular biology, vol. 120, no. 4, pp. 487–515, 1978.

[4] J. M. Adams and S. Cory, "Modified nucleosides and bizarre 5′-termini in mouse myeloma mrna," Nature, vol. 255, no. 5503, pp. 28–33, 1975.

[5] T. Chen, Y.-J. Hao, Y. Zhang, M.-M. Li, M. Wang, W. Han, Y. Wu, Y. Lv, J. Hao, L. Wang et al., "m6a rna methylation is regulated by micrornas and promotes reprogramming to pluripotency," Cell stem cell, vol. 16, no. 3, pp. 289–301, 2015.

[6] C. Roost, S. R. Lynch, P. J. Batista, K. Qu, H. Y. Chang, and E. T. Kool, "Structure and thermodynamics of n6-methyladenosine in rna: a spring-loaded base modification," Journal of the American Chemical Society, vol. 137, no. 5, pp. 2107–2115, 2015.

[7] X. Wang, Z. Lu, A. Gomez, G. C. Hon, Y. Yue, D. Han, Y. Fu, M. Parisien, Q. Dai, G. Jia et al., "N 6-methyladenosine-dependent regulation of messenger rna stability," Nature, vol. 505, no. 7481, pp. 117–120, 2014.

[8] N. Liu, Q. Dai, G. Zheng, C. He, M. Parisien, and T. Pan, "N 6-methyladenosine-dependent rna structural switches regulate rna–protein interactions," Nature, vol. 518, no. 7540, pp. 560–564, 2015.

[9] S. Geula, S. Moshitch-Moshkovitz, D. Dominissini, A. A. Mansour, N. Kol, M. Salmon-Divon, V. Hershkovitz, E. Peer, N. Mor, Y. S. Manor et al., "m6a mrna methylation facilitates resolution of naïve pluripotency toward differentiation," Science, vol. 347, no. 6225, pp. 1002–1006, 2015.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3002995, IEEE Access

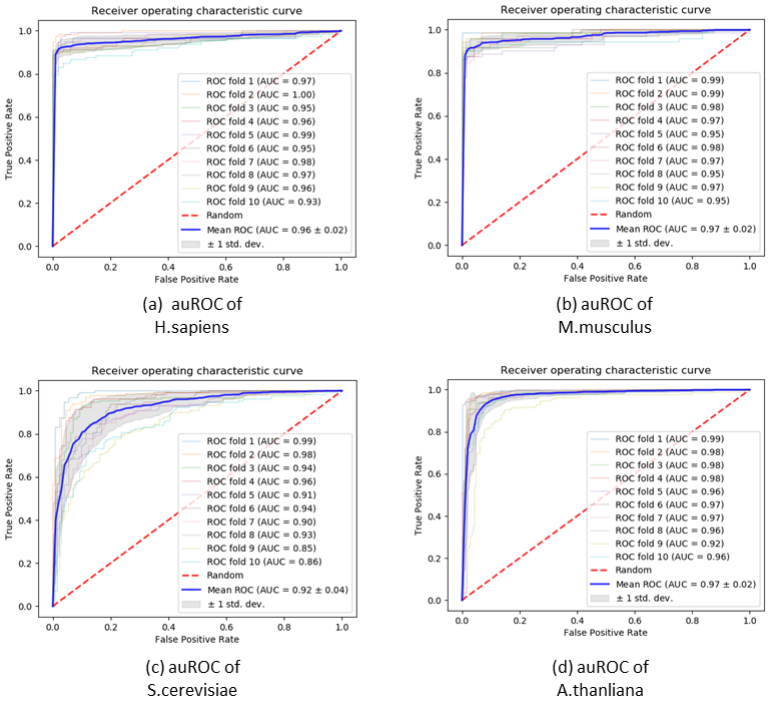

FIGURE 4: The auROC of the proposed model for multiple species

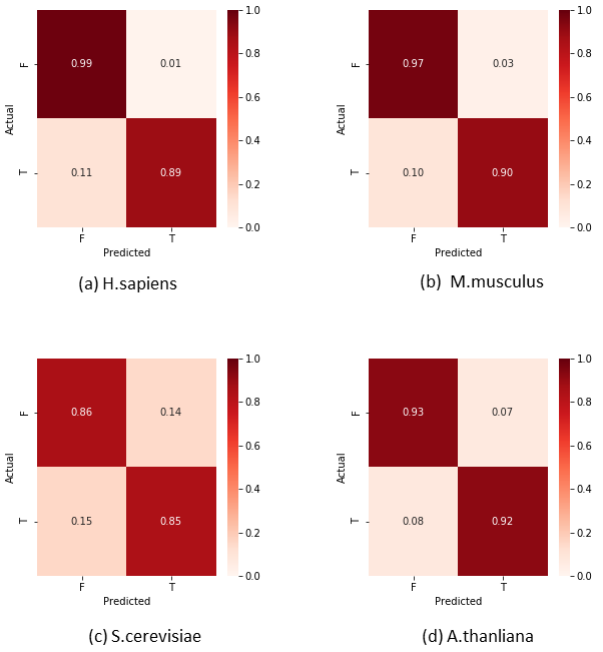

FIGURE 5: The confusion matrix of proposed model for multiple species

[10] C. R. Alarcón, H. Lee, H. Goodarzi, N. Halberg, and S. F. Tavazoie, "N 6-methyladenosine marks primary micrornas for processing," Nature, vol. 519, no. 7544, pp. 482–485, 2015.

[11] D. Dominissini, S. Moshitch-Moshkovitz, S. Schwartz, M. Salmon-Divon, L. Ungar, S. Osenberg, K. Cesarkas, J. Jacob-Hirsch, N. Amariglio, M. Kupiec et al., "Topology of the human and mouse m 6 a rna methylomes revealed by m 6 a-seq," Nature, vol. 485, no. 7397, pp. 201–206, 2012.

[12] K. D. Meyer, Y. Saletore, P. Zumbo, O. Elemento, C. E. Mason, and S. R. Jaffrey, "Comprehensive analysis of mrna methylation reveals enrichment in 3′ utrs and near stop codons," Cell, vol. 149, no. 7, pp. 1635–1646, 2012.

[13] W. Chen, H. Tran, Z. Liang, H. Lin, and L. Zhang, "Identification and analysis of the n 6-methyladenosine in the saccharomyces cerevisiae transcriptome," Scientific reports, vol. 5, p. 13859, 2015.

[14] W. Chen, P. Feng, H. Ding, H. Lin, and K.-C. Chou, "irna-methyl: Identifying n6-methyladenosine sites using pseudo nucleotide composition," Analytical biochemistry, vol. 490, pp. 26–33, 2015.

[15] Z. Liu, X. Xiao, D.-J. Yu, J. Jia, W.-R. Qiu, and K.-C. Chou, "prnam-

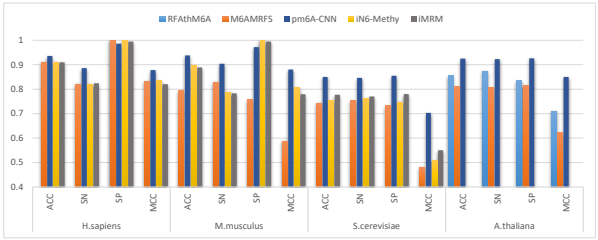

FIGURE 6: Illustration of comparison between the proposed model and existing models for the multiple species

pc: Predicting n6-methyladenosine sites in rna sequences via physical–chemical properties," Analytical biochemistry, vol. 497, pp. 60–67, 2016.

[16] W. Chen, H. Tang, and H. Lin, "Methyrna: a web server for identification of n6-methyladenosine sites," Journal of Biomolecular Structure and Dynamics, vol. 35, no. 3, pp. 683–687, 2017.

[17] S. Akbar 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, vol. 455, pp. 205–211, 2018.

[18] X. Qiang, H. Chen, X. Ye, R. Su, and L. Wei, "M6amrfs: robust prediction of n6-methyladenosine sites with sequence-based features in multiple species," Frontiers in genetics, vol. 9, p. 495, 2018.

[19] K. Liu and W. Chen, "imrm: a platform for simultaneously identifying multiple kinds of rna modifications," Bioinformatics, 2020.

[20] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," Journal of machine learning research, vol. 12, no. Aug, pp. 2493–2537, 2011.

[21] M. Sundermeyer, T. Alkhouli, J. Wuebker, and H. Ney, "Translation modeling with bidirectional recurrent neural networks," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 14–25.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097–1105.

[23] H. Tayara and K. T. Chong, "Object detection in very high-resolution aerial images using one-stage densely connected feature pyramid network," Sensors, vol. 18, no. 10, p. 3341, 2018.

[24] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," Neural networks, vol. 18, no. 5-6, pp. 602–610, 2005.

[25] L. Wei, R. Su, B. Wang, X. Li, Q. Zou, and X. Gao, "Integration of deep feature representations and handcrafted features to improve the prediction of n6-methyladenosine sites," Neurocomputing, vol. 324, pp. 3–9, 2019.

[26] Z. Lv, C. Ao, and Q. Zou, "Protein function prediction: from traditional classifier to deep learning," Proteomics, vol. 19, no. 14, p. 1900119, 2019.

[27] Q. Zou, P. Xing, L. Wei, and B. Liu, "Gene2vec: gene subsequence embedding for prediction of mammalian n6-methyladenosine sites from mrna," Rna, vol. 25, no. 2, pp. 205–218, 2019.

[28] M. Oubounyt, Z. Louadi, H. Tayara, and K. T. Chong, "Deep learning models based on distributed feature representations for alternative splicing prediction," IEEE Access, vol. 6, pp. 58 826–58 834, 2018.

[29] I. Nazari, H. Tayara, and K. T. Chong, "Branch point selection in rna splicing using deep learning," IEEE Access, vol. 7, pp. 1800–1807, 2018.

[30] Z. Louadi, M. Oubounyt, H. Tayara, and K. T. Chong, "Deep splicing code: Classifying alternative splicing events using deep learning," Genes, vol. 10, no. 8, p. 587, 2019.

[31] M. Oubounyt, Z. Louadi, H. Tayara, and K. T. Chong, "Deepromoter: Robust promoter predictor using deep learning," Frontiers in genetics, vol. 10, 2019.

[32] H. Tayara, M. Tahir, and K. T. Chong, "iss-cnn: Identifying splicing sites using convolution neural network," Chemometrics and Intelligent Laboratory Systems, vol. 188, pp. 63–69, 2019.

[33] M. Tahir, H. Tayara, and K. T. Chong, "ipseu-cnn: Identifying rna pseudouridine sites using convolutional neural networks," Molecular Therapy-Nucleic Acids, vol. 16, pp. 463–470, 2019.

[34] I. Nazari, M. Tahir, H. Tayara, and K. T. Chong, "in6-methyl (5-step): Identifying rna n6-methyladenosine sites using deep learning mode via chou's 5-step rules and chou's general pseknc," Chemometrics and Intelligent Laboratory Systems, vol. 193, p. 103811, 2019.

[35] X. Wang and R. Yan, "Rfathm6a: a new tool for predicting m 6 a sites in arabidopsis thaliana," Plant molecular biology, vol. 96, no. 3, pp. 327–337, 2018.

[36] Y. Wu and K. He, "Group normalization," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.

[37] F. Chollet et al., "Keras," https://keras.io, 2015.

[38] M. Tahir, H. Tayara, and K. T. Chong, "idna6ma (5-step rule): Identification of dna n6-methyladenine sites in the rice genome by intelligent computational model via chou's 5-step rule," Chemometrics and Intelligent Laboratory Systems, vol. 189, pp. 96–101, 2019.

[39] J. Khanal, I. Nazari, H. Tayara, and K. T. Chong, "4mccnn: Identification of n4-methylcytosine sites in prokaryotes using convolutional neural network," IEEE Access, vol. 7, pp. 145 455–145 461, 2019.

[40] H. Tayara, M. Tahir, and K. T. Chong, "Identification of prokaryotic promoters and their strength by integrating heterogeneous features," Genomics, 2019.

[41] A. Wahab, S. D. Ali, H. Tayara, and K. T. Chong, "iim-cnn: Intelligent identifier of 6ma sites on different species by using convolution neural network," IEEE Access, vol. 7, pp. 178 577–178 583, 2019.

WALEED ALAM received his M.Sc degree in Information Technology from the Institute of Information and Technology Quaid-i-Azam University, Islamabad, Pakistan in 2018. Now he is the Master student of Electronics and information engineering at Chonbuk National University, Jeonju, South Korea since 2019. His research interests include artificial intelligent, machine learning, deep learning, image processing and currently focused on bioinformatics application using deep learning.

SYED DANISH ALI received the B.Sc. degree in electronics engineering from the Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Pakistan, in 2013, and the M.S. degree in electrical engineering from Abasyn University, Pakistan, in 2018. He is currently pursuing the Ph.D. degree in electronics and information engineering from Chonbuk National University, Jeonju, South Korea. He is currently working with the Department of Electrical Engineering, The University of Azad Jammu and Kashmir, Pakistan. His research interests include bioinformatics and machine learning.

HILAL TAYARA received his B.Sc. degree in Computer Engineering from Aleppo University in Aleppo, Syria, in 2008. In 2015 and 2019, he received his M.S. and PhD degree in Electronics and Information Engineering from Chonbuk National University in Jeonju, South Korea. He is currently researcher at Chonbuk National University. His research interests include bioinformatics, machine learning, and image processing.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3002995, IEEE Access

KIL TO CHONG received his Ph.D. in Mechanical Engineering from Texas A&M University in 1995. Currently, he is a professor at the School of Electronics and Information Engineering at Chonbuk National University in Jeonju, Korea, and a head of the Advanced Research Center of Electronics. His research interests are in the areas of machine learning, signal processing, motor fault detection, network system control, and time-delay systems.