IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 14, 2020, accepted December 26, 2020, date of publication January 5, 2021, date of current version January 14, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3049146

Prediction of RNA 5-Hydroxymethylcytosine Modifications Using Deep Learning

SYED DANISH ALI[1,2], JEE HONG KIM[3], HILAL TAYARA[4], AND KIL TO CHONG[1,5]

[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]Department of New and Renewable Energy, Vision College of Jeonju, Jeonju 55069, South Korea
[4]School of International Engineering and Science, Jeonbuk National University, Jeonju 54896, South Korea
[5]Advanced Electronics and Information Research Center, Jeonbuk National University, Jeonju 54896, South Korea

Corresponding authors: Kil To Chong (kitchong@jbnu.ac.kr) and Hilal Tayara (hilaltayara@jbnu.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) funded by the Korean Government (MSIT) under Grant 2020R1A2C2005612, in part by the National Research Foundation (NRF) funded by the Korean Government (MSIT) through the Brain Research Program under Grant NRF-2017M3C7A1044816.

ABSTRACT It is becoming increasingly clear that RNA 5-hydroxymethylcytosine (5hmC), which plays an important role in several biological processes, is one of the most important objects of study in the field of RNA epigenetics. Biochemical experiments using various sequencing-based technologies are capable of achieving high-throughput identification of 5hmC, but current methods are labor-intensive, costly, and time-consuming. There is an imperative need to develop more efficient and robust computational methods to replace, or at least complement, such high-throughput methods. Although one such machine learning-based model to achieve this has already been developed, its performance is limited. In this study, we developed iRhm5CNN, an efficient and reliable computational predictive model for the identification of RNA 5hmC sites. Our model is based on a convolution neural network (CNN) that extracts the most reliable feature from the RNA sequence inevitably. The results of our experiments show significant outperformance across all evaluation metrics of our proposed architecture when compared to the only existing state of the art computational model in all the evaluation metrics. The proposed model can be accessed for free at http://nsclbio.jbnu.ac.kr/tools/iRhm5CNN/.

INDEX TERMS Post-transcriptional modification, RNA 5-hydroxymethylcytosine, sequence analysis, convolutional neural network, deep learning.

I. INTRODUCTION

RNA epigenetics and epitranscriptomics are attracting increasing attention among researchers examining post-transcriptional modifications [1], [2]. These RNA modifications play a decisive role in the maturation and translation of mRNA, and regulation of RNA splicing [3]. 5-hydroxymethylcytosine is one of more than 170 distinct RNA modifications that may be found across all three domains of life, including Archae, Bacteria, and Eukarya [3], [4]. Racz *et al.* [5] initially detected 5hmC in wheat seedlings, though it has subsequently been identified in human and mammalian tissues [6], [7]. The 5hmC modification is formed by the oxidization of m5C, which is oxidizable by Tet-family enzymes into 5hmC [7]. Additionally, hydroxymethylcytosine RNA immunoprecipitation sequencing (hMeRIP-seq) reveals that Tet-family enzymes are most likely to oxidize m5C modifications in coding regions, suggesting that 5hmC is almost certainly situated in the introns and exons of coding transcripts [8]. Delatte *et al.* [9] observed an abundance of 5hmC modifications in the brain of Drosophila. Similarly, Miao *et al.* [10] used a dot blot analysis to determine that the brainstem, cerebellum, and hippocampus encompassed high levels of 5hmC modification and observed that 5hmC modification declined in MPTP-induced Parkinson's disease model in mice. Collectively, these outcomes intimate that the RNA 5hmC modification plays an important regulatory role in microRNA or protein expression in brain tissue, and that 5hmC contributes to the epigenetic regulation of gene expression by modulating RNA-protein interactions [11]. Determining the distribution of 5hmC in the transcriptome of various species is a critical

The associate editor coordinating the review of this manuscript and approving it for publication was Nuno Garcia.

step towards further understanding its biological functions in multiple species. Gaining insight into the functions of 5hmC and its presence in mammals would be appealing because of potential derivatives and precursors of this modification in RNA.

Unfortunately, the limits of hMeRIP-seq and wet lab experimentation, including the high-price of experimental materials, as well as the time and labor-intensive nature of prolonging experiments, have made it difficult to identify 5hmC sites across the genome efficiently. Given the escalating availability of genomics samples generated in the post-genomics period, however, computational models can fill the gap and provide accurate, efficient, and cost-effective identification of 5hmC modification sites.

Recently, Liu *et al.* proposed iRNA5hmC, a computational model with a machine learning identifier based on a support vector machine (SVM) as a classifier to predict primary RNA sequences [12]. This model utilized the k-mer spectrum and positional nucleotide binary vector as a feature representation technique. The task of improving the low predictive performance of their model, however, was left to future research. The performance of iRNA5hmC can be further enhanced by proposing alternative robust computational methods. The existing method, which is based on domain knowledge, relies on drawn-out hand-designed input features. To overcome this constraint deep learning techniques could be effective alternative computational methods that are consequentially capable of learning the features by utilizing multiple levels of abstraction [13]–[15]. Computational models based on deep learning have proved to be very efficient and effective at image recognition [16], [17], information retrieval [18], natural language processing [19], speech recognition [20], [21], and computational biology [22]–[37]. Considering the effectiveness of deep learning methods in the field of computational biology; CNN implementation is the most popular implementation of deep learning.

In this study, we propose a simple and effective CNN based architecture for the identification of RNA 5hmC sites that utilizes only the primary RNA sequences shown in Figure 1. The primary RNA sequences are represented as the one-hot encoding. Our experiment also includes another most basic representation of the chemical components of nucleotides, concerning their chemical properties including functional groups, hydrogen bonds, and ring structures. The CNN architecture extracts the most important features from the primary RNA sequence representations, resulting in consistently accurate identification of the RNA 5hmC sites. The optimum hyperparameters were selected based on the grid search algorithm. The performance of our proposed method was evaluated using a subsampling (k-fold cross-validation) method where the value of k was set to five. Finally, a user-friendly publicly available web server is accessible at http://nsclbio.jbnu.ac.kr/tools/iRhm5CNN/.

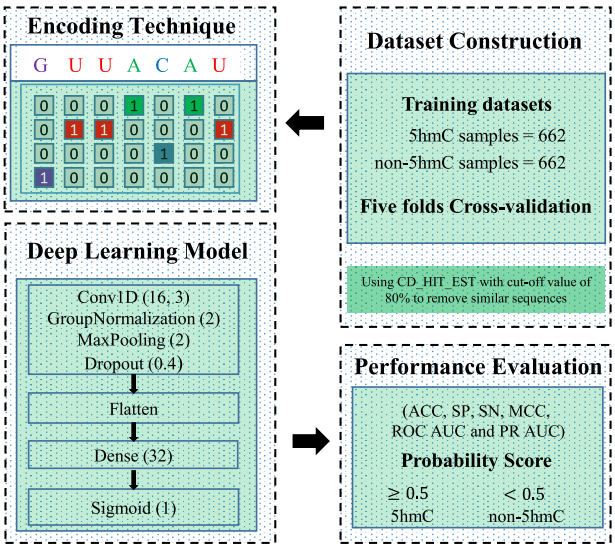

FIGURE 1. The detailed architecture of the iRhm5CNN model.

II. MATERIALS AND METHODS

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

A. BENCHMARK DATASET

The dataset used in the study was prepared and utilized by Liu *et al.* [12] and is available at http://server.malab.cn/iRNA5hmC/Download.html. The balance dataset consisted of 1324 samples. The 662 sequences having 5hmC in the center were regarded as positive samples collected from Delatte *et al.* [9]. Where the sequence similarity is less than 80%. The remaining randomly selected 662 sequences identified as having the intermediate cytosine (utilizing the method of hMeRIP-seq) were not identified as 5hmC and were regarded as negative samples. The length of each sample was 41 nucleotides.

B. THE PROPOSED MODEL

The proposed architecture iRhm5CNN is a simplified CNN-based deep learning model as shown in Figure 1. CNN is a popular deep learning technique, with a well-deserved reputation for exceptional results and generalization. CNN extracts the most important features from an RNA sequence representation without any intervention of hand-designed features. The primary RNA sequence is represented as a one-hot vector input to the CNN architecture, with each RNA sequence, comprised of four nucleotides bases including Adenine (A), Cytosine (C), Guanine (G) and Uracil (U), being represented by vectors (1, 0, 0, 0), (0, 0, 1, 0), (0, 0, 0, 1), and (0, 1, 0, 0), respectively.

In general, CNNs are comprised of layers, including a convolution layer, a normalization layer, a pooling layer, and a fully connected layer. The most optimal hyperparameters are selected using the grid search algorithm based on a

TABLE 1. The ranges of the optimum hyper-parameters.

The Hyper-Parameters	Range
Convolution layers	[1,2]
Filters of convolution layers	[8,16,24]
Filter size	[3,5,7]
L2 regularizer	[1e-2,1e-3,1e-4]
Max pooling (Pool size)	[2,4]
Group size	[2,4]
Dropout	[0.25,0.3,0.4]
Neurons	[8,16,32]
Optimizer	[SGD, Adam]
Learning rate	[0.001, 0.0001]

manually defined range of hyperparameters. The local features of the input are extracted by the convolution layer, which has numerous convolutional units and parameters optimized using backpropagation [38]. The Rectified linear unit (ReLU) is used as an activation function for the convolution layer, followed by the group normalization layer, which is an effective alternative to batch normalization for dealing with a small batch size [39]. Normalization, which acts as a regularization technique, is performed in groups, and the group size selected for this study is two. The one-dimensional max-pooling layer with a pool size of two is utilized after the normalization layer. The max-pooling layer reduces the dimensionality and redundancy of the features from the previous layer. The max-pooling layer is followed by a dropout layer that has a dropout probability of 0.4. During training, the dropout layer randomly switches off the effect of some neurons to avoid overfitting [40]. The dropout layer is followed by a fully connected layer that uses ReLU as an activation function and L2 regularization with a value of 1×10^{-2}. The L2 regularization on bias and weight is the most sophisticated and effective techniques for mitigating overfitting, as L2 regularization penalizes the model with larger weights [41]. The last layer is the sigmoid activation function, which assigns probabilities to the outputs for the results to be mapped as 5hmC site or non-5hmC site. The range of optimum hyper-parameters for the grid search method are enlisted in Table 1.

Mathematically the architecture is expressed in the following Equations.

$$Conv(X)_{jk} = ReLU\left(\sum_{s=0}^{Z-1}\sum_{n=0}^{I-1} W_{sn}^{k} Xj + s, n\right) \tag{1}$$

The one-dimensional convolution layer is expressed in Equation 1 where input as an RNA sequence is X, index of the filter is k, 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.

$$ReLU(x) = \begin{cases} x & if\, x > 0 \\ 0 & if\, x \leq 0 \end{cases} \tag{2}$$

The ReLU activation function used in the architecture is given in Equation 2.

$$d = ReLU\left(w_{d+1}\sum_{k=1}^{d} m_k w_k z_k\right) \tag{3}$$

The fully connected layer along with dropout operation m_k having the probability of p which is sampled from Bernoulli distribution is mathematically expressed in Equation 3, where z_k is a 1×d dimensional feature vector, w_k is the weights of the z_k from the previous layer, and w_{d+1} is the additive bias term.

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

The sigmoid activation function is shown in Equation 4 where x is the input of function.

The Keras framework (available at https://keras.io/) was utilized for the construction of iRhm5CNN. Adam, which has a learning rate of 0.001, was used as an optimizer for the proposed model. Binary cross-entropy [42] was used as a loss function which measures the discrepancy between the probability distributions of principal class [41]. The maximum numbers of epochs for training were 81 with a batch size of 32. The early stopping on validation loss with the patience level of 15 was utilized to avoid overfitting which stops the training updates when there is no improvement in the loss for 15 epochs. Also, the Plateau learning rate reducer with the reduction factor of 0.01 was utilized. The patience level of the reducer was 10 epochs to reduce the learning rate with the factor of 0.01 if there was no improvement in validation loss after 10 epochs.

C. PERFORMANCE EVALUATION

The performance of the proposed model was quantitatively evaluated using the four metrics named as accuracy (ACC), sensitivity (SN), specificity (SP), and Matthews correlation coefficient (MCC), which are mathematically represented as:

$$\text{ACC} = \frac{TP + TN}{TP + TN + FN + FP} \tag{5}$$

$$\text{SN} = \frac{TP}{TP + FN} \tag{6}$$

$$\text{SP} = \frac{TN}{TN + FP} \tag{7}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \tag{8}$$

where TP (true positive) is the number of positive samples correctly identified; TN (true negative) is the number of negative samples correctly identified; FN (false negative) is the number of positive samples incorrectly identified as negative samples, and FP (false positive) is the number of negative samples incorrectly identified as positive samples. The range of values for accuracy, sensitivity, and specificity

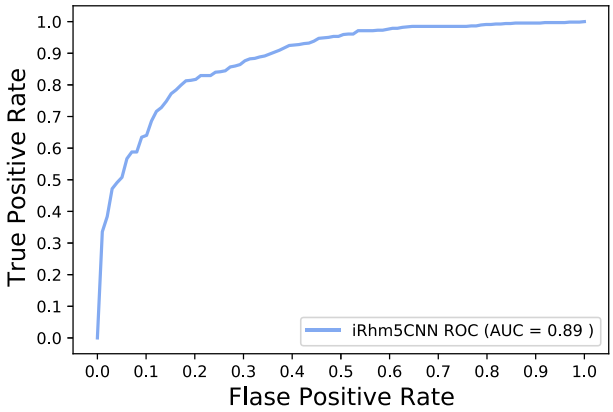

FIGURE 2. The illustration of the ROC along with area under the curve (AUC).

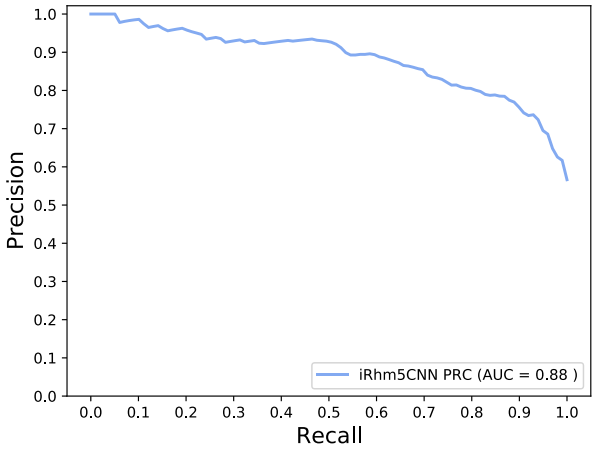

FIGURE 3. The illustration of the PR curve along with AUC.

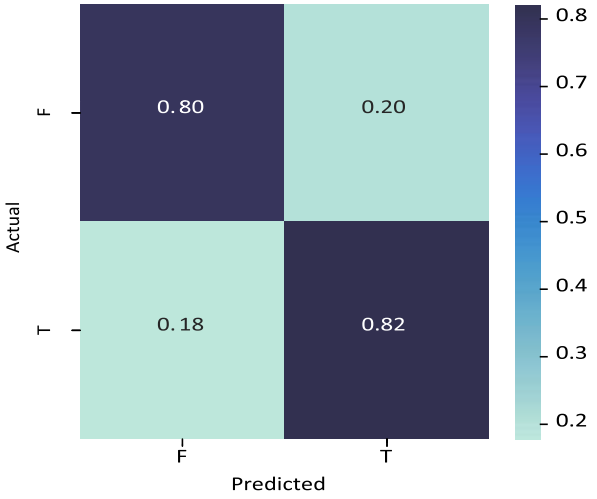

FIGURE 4. Illustration of confusion matrix of proposed model.

varies between [0, 1], whereas the range for the Matthews correlation coefficient is [−1, 1]. Higher values indicate superior performance.

Moreover, the receiver operating characteristics (ROC) curve showing the trade-off between the true positive and false positive rate is utilized. The precision-recall (PR) curve is used to demonstrate the trade-off between the true positive rate and the positive predictive values of a classifier through different probability thresholds. The area under the ROC curves and the PR curves is a significant indicator of the predictive efficiency of the binary classifiers. Finally, the confusion matrix is shown which is the visual representation of performance.

III. RESULTS AND DISCUSSION

The k-fold cross-validation technique is utilized to evaluate the robustness and sensitivity of the proposed computational model. The outcomes of k-fold cross-validation are fairly unbiased [43]. K-fold cross-validation is quite effective at proving the effectiveness of certain computational models using a benchmark dataset, as the results obtained are evaluated using k number of different training and validation sets [44]. The value of k is set to be 5. Among 5 folds three folds were utilized for training, one-fold for validation, and the remaining one fold for testing of the proposed model. The proposed model predicted RNA 5hmC sites with 81.20% accuracy, 82.03% sensitivity, 80.37% specificity, and an MCC of 0.62. These results are summarized in Table 3. A graphical representation of performance showing the area under the ROC curve is 0.89 and under the PR curve is 0.88 is presented in Figures 2 and 3 respectively. The confusion matrix is provided in Figure 4.

A. SEQUENCE ENCODING USING NUCLEOTIDE CHEMICAL PROPERTIES

Each of the four RNA nucleotides A, C, G, U has different chemical properties. The nucleotides are grouped into three groups depending on their chemical properties; base type, hydrogen bond strength, and functional (amino or keto) group. Concerning the base type, A and G are purines with two rings, while C and U are pyrimidines with one ring. The hydrogen bond between A and U is weak but strong between C and G. As to functional groups, A and C belong to the amino group while G and U are in the keto group. Based on these three chemical criteria, each nucleotide in the RNA sequence can be encoded in a three-dimensional cartesian coordinate system. The coordinates are assigned a value of 0 or 1. The first dimension reflects the base type, with 1 indicating purines and 0 indicating pyrimidines. Hydrogen bond strength is represented in the second dimension; 1 for a weak bond and 0 for a strong bond. The third dimension represents the functional group, with 1 indicating an amino group, and 0 indicating a keto group. Therefore, A, C, G, and U are represented as (1, 1, 1), (0, 0, 1), (1, 0, 0), and (0, 1, 0), respectively. The results of our proposed method using feature-based sequence encoding for the RNA sequences are summarized in Table 2. The ACC is 78.79%, the SN is 81.44%, the SP is 76.13%, the MCC is 0.58, the area under the ROC is 0.87, and under the PR is 0.86.

B. COMPARISON

We compared the outcomes of the proposed model with the only existing machine learning-based computational model,

TABLE 2. The performance of the proposed model iRhm5CNN using different sequence representation methods.

Encoding method	ACC	SN	SP	MCC
One-hot	0.81	0.82	0.80	0.62
NCP	0.79	0.81	0.76	0.58

TABLE 3. Performance comparison of proposed model iRhm5CNN comparison with the existing computational model.

Method	ACC	SN	SP	MCC
iRhm5CNN	0.81	0.82	0.80	0.62
iRNA5hmC	0.65	0.68	0.63	0.31

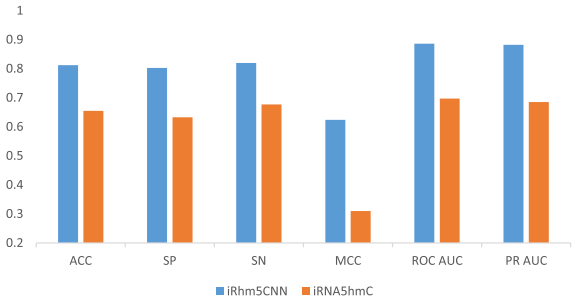

FIGURE 5. Illustration of comparison of existing models for predicting RNA 5hmC sites.

iRNA5hmC [12] which utilizes the SVM algorithm, where the features representation techniques are k-mer spectrum and positional nucleotide binary vectors. iRNA5hmC was evaluated using 5 folds cross-validation. The results of comparison are summarized in Table 3 and are illustrated in Figure 5. The outcomes of our proposed method showed significant improvement over iRNA5hmC. The empirical success rate in Table 3 reflects the significant improvement of our proposed model across all performance metrics using the benchmark dataset; 15.72% improvement in the accuracy, 14.36% improvement in the sensitivity, 17.08% improvement in the specificity, and an MCC of 31%. The significant improvement of the proposed model signifies that automatic feature selection of CNN outperforms the only existing machine learning-based method in all the performance metrics.

IV. WEB-SERVER

Experimental scientists may use a web server to get their desired results without engaging in complicated mathematics. A web-server can document computationally evaluated outcomes [45]. These factors are responsible for improving computational biology performance in the medical sciences [46]. Considering the effectiveness of web server a publicly available web-server can be accessible at http://nsclbio.jbnu.ac.kr/tools/iRhm5CNN/ which is built using python.

V. CONCLUSION

Accurate identification of RNA 5hmC is a necessary step to the continued exploration of vast and diverse biological functions. This study proposed an efficient and effective computational tool (iRhm5CNN) for the identification of 5hmC using a deep learning framework. Our proposed model uses a simple CNN architecture for appropriate feature extraction followed by a fully connected layer to discriminate between RNA 5hmC and non-5hmC sites. The proposed model outperformed the existing state-of-the-art model, and the outcomes confirm iRhm5CNN's efficacy. We anticipate that our developed computational model along with its publicly available web server will be utilized for drug development and academic research related to the investigation of the functional procedures of 5hmC sites, including but not limited to abnormalities in brain development.

ACKNOWLEDGMENT

(Syed Danish Ali and Jee Hong Kim equally contributed to this work.)

REFERENCES

[1] C. He, "Grand challenge commentary: RNA epigenetics?" *Nature Chem. Biol.*, vol. 6, no. 12, pp. 863–865, Dec. 2010.

[2] Y. Saletore, K. Meyer, J. Korlach, I. D. Vilfan, S. Jaffrey, and C. E. Mason, "The birth of the epitranscriptome: Deciphering the function of RNA modifications," *Genome Biol.*, vol. 13, no. 10, p. 175, 2012.

[3] S. M. Huber, P. van Delft, L. Mendil, M. Bachman, K. Smollett, F. Werner, E. A. Miska, and S. Balasubramanian, "Formation and abundance of 5-Hydroxymethylcytosine in RNA," *ChemBioChem*, vol. 16, no. 5, pp. 752–755, Mar. 2015.

[4] P. J. McCown, A. Ruszkowska, C. N. Kunkler, K. Breger, J. P. Hulewicz, M. C. Wang, N. A. Springer, and J. A. Brown, "Naturally occurring modified ribonucleosides," *WIREs RNA*, vol. 11, no. 5, p. e1595, Sep. 2020.

[5] I. Rácz, I. Király, and D. Lasztily, "Effect of light on the nucleotide composition of rRNA of wheat seedlings," *Planta*, vol. 142, no. 3, pp. 263–267, 1978.

[6] W. Li and M. Liu, "Distribution of 5-Hydroxymethylcytosine in different human tissues," *J. Nucleic Acids*, vol. 2011, pp. 1–5, Oct. 2011.

[7] L. Fu, C. R. Guerrero, N. Zhong, N. J. Amato, Y. Liu, S. Liu, Q. Cai, D. Ji, S.-G. Jin, L. J. Niedernhofer, G. P. Pfeifer, G.-L. Xu, and Y. Wang, "Tet-mediated formation of 5-Hydroxymethylcytosine in RNA," *J. Amer. Chem. Soc.*, vol. 136, no. 33, pp. 11582–11585, Aug. 2014.

[8] I. A. Roundtree, M. E. Evans, T. Pan, and C. He, "Dynamic RNA modifications in gene expression regulation," *Cell*, vol. 169, no. 7, pp. 1187–1200, Jun. 2017.

[9] B. Delatte, F. Wang, L. V. Ngoc, E. Collignon, E. Bonvin, R. Deplus, E. Calonne, B. Hassabi, P. Putmans, S. Awe, and C. Wetzel, "Transcriptome-wide distribution and function of RNA hydroxymethylcytosine," *Science*, vol. 351, no. 6270, pp. 282–285, Jan. 2016.

[10] Z. Miao, N. Xin, B. Wei, X. Hua, G. Zhang, C. Leng, C. Zhao, D. Wu, J. Li, W. Ge, M. Sun, and X. Xu, "5-hydroxymethylcytosine is detected in RNA from mouse brain tissues," *Brain Res.*, vol. 1642, pp. 546–552, Jul. 2016.

[11] H.-Y. Zhang, J. Xiong, B.-L. Qi, Y.-Q. Feng, and B.-F. Yuan, "The existence of 5-hydroxymethylcytosine and 5-formylcytosine in both DNA and RNA in mammals," *Chem. Commun.*, vol. 52, no. 4, pp. 737–740, 2016.

[12] Y. Liu, D. Chen, R. Su, W. Chen, and L. Wei, "IRNA5hmC: The first predictor to identify RNA 5-Hydroxymethylcytosine modifications using machine learning," *Frontiers Bioeng. Biotechnol.*, vol. 8, p. 227, Mar. 2020.

[13] Z. Zhang, Y. Zhao, X. Liao, W. Shi, K. Li, Q. Zou, and S. Peng, "Deep learning in omics: A survey and guideline," *Briefings Funct. Genomics*, vol. 18, no. 1, pp. 41–57, Feb. 2019.

[14] H. Tayara and K. T. Chong, "Improving the quantification of DNA sequences using evolutionary information based on deep learning," *Cells*, vol. 8, no. 12, p. 1635, Dec. 2019.

[15] H. Tayara and K. Chong, "Improved predicting of the sequence specificities of RNA binding proteins by deep learning," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Mar. 18, 2020, doi: 10.1109/TCBB.2020.2981335.

[16] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[17] H. Tayara and K. Chong, "Object detection in very high-resolution aerial images using one-stage densely connected feature pyramid network," *Sensors*, vol. 18, no. 10, p. 3341, Oct. 2018.
[18] B. Mitra and N. Craswell, "Neural models for information retrieval," 2017, *arXiv:1705.01509*. [Online]. Available: http://arxiv.org/abs/1705.01509
[19] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12 pp. 2493–2537, Aug. 2011.
[20] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.
[21] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Netw.*, vol. 92, pp. 60–68, Aug. 2017.
[22] M. Tahir, H. Tayara, and K. T. Chong, "IPseU-CNN: Identifying RNA pseudouridine sites using convolutional neural networks," *Mol. Therapy Nucleic Acids*, vol. 16, pp. 463–470, Jun. 2019.
[23] H. Tayara, M. Tahir, and K. T. Chong, "ISS-CNN: Identifying splicing sites using convolution neural network," *Chemometric Intell. Lab. Syst.*, vol. 188, pp. 63–69, May 2019.
[24] M. Oubounyt, Z. Louadi, H. Tayara, and K. T. Chong, "DeePromoter: Robust promoter predictor using deep learning," *Frontiers Genet.*, vol. 10, p. 286, Apr. 2019.
[25] 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, Aug. 2019.
[26] H. Tayara, M. Tahir, and K. T. Chong, "Identification of prokaryotic promoters and their strength by integrating heterogeneous features," *Genomics*, vol. 112, no. 2, pp. 1396–1403, Mar. 2020.
[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, Feb. 2019.
[28] Z. Lv, C. Ao, and Q. Zou, "Protein function prediction: From traditional classifier to deep learning," *Proteomics*, vol. 19, no. 14, Jul. 2019, Art. no. 1900119.
[29] 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," *Chemometric Intell. Lab. Syst.*, vol. 189, pp. 96–101, Jun. 2019.
[30] J. Khanal, D. Y. Lim, H. Tayara, and K. T. Chong, "i6mA-stack: A stacking ensemble-based computational prediction of DNA N6-methyladenine (6mA) sites in the Rosaceae genome," *Genomics*, 2020, doi: 10.1016/j.ygeno.2020.09.054.
[31] 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. 178577–178583, 2019.
[32] 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," *Chemometric Intell. Lab. Syst.*, vol. 193, Oct. 2019, Art. no. 103811.
[33] M. U. Rehman and K. T. Chong, "DNA6mA-MINT: DNA-6mA modification identification neural tool," *Genes*, vol. 11, no. 8, p. 898, Aug. 2020.
[34] T. Chantsalnyam, D. Y. Lim, H. Tayara, and K. T. Chong, "NcRDeep: Non-coding RNA classification with convolutional neural network," *Comput. Biol. Chem.*, vol. 88, Oct. 2020, Art. no. 107364.
[35] A. Wahab, O. Mahmoudi, J. Kim, and K. T. Chong, "DNC4mC-deep: Identification and analysis of DNA N4-methylcytosine sites based on different encoding schemes by using deep learning," *Cells*, vol. 9, no. 8, p. 1756, Jul. 2020.
[36] W. Alam, S. D. Ali, H. Tayara, and K. T. Chong, "A CNN-based RNA N6-methyladenosine site predictor for multiple species using heterogeneous features representation," *IEEE Access*, vol. 8, pp. 138203–138209, 2020.
[37] O. Mahmoudi, A. Wahab, and K. T. Chong, "IMethyl-deep: N6 methyladenosine identification of yeast genome with automatic feature extraction technique by using deep learning algorithm," *Genes*, vol. 11, no. 5, p. 529, May 2020.
[38] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.
[39] Y. Wu and K. He, "Group normalization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[40] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[41] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*, vol. 1. Cambridge, MA, USA: MIT Press, 2016.
[42] P.-T. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Ann. Oper. Res.*, vol. 134, no. 1, pp. 19–67, Feb. 2005.
[43] Y. Bengio and Y. Grandvalet, "No unbiased estimator of the variance of K-fold cross-validation," *J. Mach. Learn. Res.*, vol. 5, pp. 1089–1105, Dec. 2004.
[44] K.-C. Chou and H.-B. Shen, "Recent progress in protein subcellular location prediction," *Anal. Biochemistry*, vol. 370, no. 1, pp. 1–16, Nov. 2007.
[45] K.-C. Chou and H.-B. Shen, "Recent advances in developing Web-servers for predicting protein attributes," *Natural Sci.*, vol. 1, no. 2, p. 63, Sep. 2009.
[46] K.-C. Chou, "Impacts of bioinformatics to medicinal chemistry," *Medicinal Chem.*, vol. 11, no. 3, pp. 218–234, Mar. 2015.

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 with Jeonbuk National University, Jeonju, South Korea. He is working with the Department of Electrical Engineering, The University of Azad Jammu and Kashmir, Pakistan. His research interests include bioinformatics and machine learning.

JEE HONG KIM received the Ph.D. degree in control & measurement engineering from Jeonbuk National University, in 2010. He is currently a Professor with the Department of New & Renewable Energy, Vision College of Jeonju, South Korea, and a member of the Advanced Information and Electronics Research Center, Jeonbuk National University. His research interests include the areas of bioinformatics, artificial intelligence, drug discovery, and big-data in the renewable energy sector.

HILAL TAYARA received the B.Sc. degree in computer engineering from Aleppo University, Aleppo, Syria, in 2008, the M.S. and Ph.D. degrees from Jeonbuk National University, Jeonju, South Korea, in 2015 and 2019, respectively, both in electronics and information engineering. He is currently an Assistant Professor with the Department of International Science and Engineering, Jeonbuk National University. His research interests include bioinformatics, machine learning, and image processing.

KIL TO CHONG received the Ph.D. degree in mechanical engineering from Texas A&M University, in 1993. He is currently a Professor with the School of Electronics and Information Engineering, Jeonbuk National University, Jeonju, South Korea, Head of the Advanced Information and Electronics Research Center, Jeonbuk National University, and the President of Korean Electronics Engineering Society, Systems and Control. His research interests include the areas of bioinformatics, artificial intelligence, brain disease, and new drug discovery.