scientific reports

Check for updates

OPEN

A CNN based m5c RNA methylation predictor

Irum Aslam[1], Sajid Shah[2], Saima Jabeen[3]✉, Mohammed ELAffendi[2], Asmaa A. Abdel Latif[4], Nuhman Ul Haq[1] & Gauhar Ali[2]

Post-transcriptional modifications of RNA play a key role in performing a variety of biological processes, such as stability and immune tolerance, RNA splicing, protein translation and RNA degradation. One of these RNA modifications is m5c which participates in various cellular functions like RNA structural stability and translation efficiency, got popularity among biologists. By applying biological experiments to detect RNA m5c methylation sites would require much more efforts, time and money. Most of the researchers are using pre-processed RNA sequences of 41 nucleotides where the methylated cytosine is in the center. Therefore, it is possible that some of the information around these motif may have lost. The conventional methods are unable to process the RNA sequence directly due to high dimensionality and thus need optimized techniques for better features extraction. To handle the above challenges the goal of this study is to employ an end-to-end, 1D CNN based model to classify and interpret m5c methylated data sites. Moreover, our aim is to analyze the sequence in its full length where the methylated cytosine may not be in the center. The evaluation of the proposed architecture showed a promising results by outperforming state-of-the-art techniques in terms of sensitivity and accuracy. Our model achieve 96.70% sensitivity and 96.21% accuracy for 41 nucleotides sequences while 96.10% accuracy for full length sequences.

In the current era we are swimming in an extending sea of information. Data with big volume, high velocity, and variety is obtained from various fields of sciences and engineering[1–3]. Life researchers are also going to grapple with massive data because of high-throughput genomics. They are facing vast range of problems related to handling, processing, storing and interpreting biological data. The techniques used to generate biological data, spit out various types of information, such as interactions of proteins, genomic sequences or findings in medical records etc. Since, the biological data come from wide range of methods, that is why when compared to other domains of science it is highly heterogeneous in nature.

Learning from big sets of data (massive data) is highly challenging but undoubtedly it is the essential part of numerous fields in the current time. It needs new ways of thinking to acknowledge the challenges of learning with massive data and the related convenient solutions. The novel techniques of learning were required which possess the ability to fully making sense of big data. In other words we need the algorithms which are inherently efficient and powerful to tackle the data which poses the challenges due to its high dimensions, imbalance, heterogeneous and uncertain nature[4,5].

The amount of biological sequential data has also increased in the last decade with the advent of high-throughput sequencing projects. A biological sequence data is a continuous and single string or molecule of protein or nucleic acid which are made from amino acids or nucleotides respectively. The amino acids, nucleotides and ribonucleotides are the basic structural and functional blocks or units of the three fundamental and informative life's polymers that is proteins, DNA and RNA respectively.

The genetic information present in DNA molecule is interpreted and copied by different types of RNA polymerases. A specific well defined sites of DNA is decoded and transcribed into a variety of single-stranded transcripts (RNAs). Furthermore, there are four standard ribonucleotides i.e., U, A, G and C involved in the production of RNA molecules. Ribonucleic acid (RNA) is the main polymeric molecule that transfer instructions from genes to ribosomes in order to synthesize specific proteins. Each triplet codon within transcript e.g mRNA has been translated into an appropriate and relevant amino acid of protein chains. There are 21 standard amino acids which are categorized as essential and non-essential entailed in the synthesis of protein molecules.

[1]Department of Computer Science, COMSATS University Islamabad, Abbottabad Campus, Abbottabad 22060, KPK, Pakistan. [2]EIAS Data Science Lab, College of Computer and Information Sciences, Prince Sultan University, Rafha, Riyadh 12435, Saudi Arabia. [3]College of Engineering, AI Research Center, Alfaisal University, Riyadh 50927, Saudi Arabia. [4]Public Health and Community Medicine Department (Industrial medicine and occupational health specialty, Faculty of Medicine, Menoufia University, Shibîn el Kôm, Egypt. ✉email: sjabeen@alfaisal.edu

However, in every biopolymer type, the limited amount of basic building blocks seems to be necessary for the natural flow of biological information from DNA to RNA to protein. Although it appears that they are not sufficient to attain all expected and anticipated functions of these polymers in organisms. DNA, RNA and proteins are composed only from few basic units, so in order to perform such a variety of functions some targeted enzymes have altered the units at specific locations to produce new characteristics of molecules. The modifications in polymer molecules are termed as pre or post replicational , transcriptional or translational changes. Due to these modifications, the biopolymer got new functional or structural features thus allow them to perform more complicated functions in a well-organized manner[6–10].

Post-transcriptional modifications of RNA play a key role in performing a variety of biological processes, such as stability and immune tolerance, RNA splicing, protein translation and RNA degradation. One of these RNA modifications is m5c which participates in various cellular functions like RNA structural stability and translation efficiency, got popularity among biologists. More specifically, m5C modification also known as methylation occurs at 5th position of cytosine when methyl group (CH_3) is added to it as shown in Fig. 1. A comprehensive study is presented in[11] about the implication of RNA m5c modification in cancer.

By applying biological experiments to detect RNA m5c methylation sites would require much more efforts, time and money[12]. To know the logic of life it is essential to interpret the full spectrum of m5c methylation and its position in the RNA molecule. The m5c modified RNA molecules can be enunciated or presented as genetic or physical map, an actual sequence of amino acids or nucleic acids, or some more complicated data representation. To get insight into methylated molecule's function, there is a need to analyze hidden features with in these modified molecules.

In order to explore the hidden features in data, feature extraction techniques are widely used in data analysis field. Feature extraction refers to identifying an interpretable and discriminating representation of data for machine learning models that can enhance the prediction power of classifier and its performances. The performances of underlying classifier depends on the quality of extracted features.

Feature extraction can either be handcrafted or automated, depending on the nature of the problem, the amount of data and available resources. The manual feature extraction known as hand crafted features is not only difficult and time consuming process but also some time these features may not effectively represent the underlying objects/entities (sequences in our case). Apart from the above challenges, fully domain expertise is also required to carry out this task. Now a days large amount of data is available due to which both biologists and computer scientists are confronted with many difficulties to speedily perform data analysis tasks in life sciences. In addition to scalable and efficient methods, high performance computing (HPC) platforms and automatic feature extraction techniques are entailed to gain a keen insight into the biological functions from big data. The key feature of deep learning techniques is representation learning which extracts a diverse range of meaningful descriptors/ features that enhance the prediction capabilities of the underlying model. By using this approach, feature extraction and classification can be done in an end-to-end manner, enabling us to obtain the significant high level features automatically, resulting in improved performances[13–17].

Since further improvements have been enabled by the use of greater computational resources, especially graphics processing units (GPU), allowing training of deep networks containing various parameters in an appropriate time. It allows us to efficiently train specialized deep networks such as convolutional neural networks (CNN) and recurrent neural networks (RNN) with long short-term memory cells (LSTM). These networks have been successfully applied to many problems including image recognition and natural language processing tasks like language translation and speech recognition[18–20].

Prediction of m5c poses some of the challenges. For example, the nucleotide preference around m5c cite is not known. So lack of clear sequence context information of m5c cite would led to intricacy in the prediction method. It may possible that the motif of m5c is obscured so it is difficult to find local sequence context of m5c cites. Most of the researchers are using pre-processed RNA sequences of 41 nucleotides, therefore, it is possible that some of the information around these motif may have lost. The conventional methods are unable to process the RNA sequence directly and thus need optimized techniques for better features extraction. Being high dimensional, the methylated datasets usually posed a challenge to conventional analysis techniques. It is also termed as curse of dimensionality.

To handle the above challenges the goal of this study is to employ an end-to-end deep learning model as a powerful toolbox to classify and interpret m5c methylated data sites. The power of deep learning models for high dimensional data is proven in literature. Moreover, our aim is to analyze the sequence in its full length where the methylated cytosine may not be in the center. Thus, the contribution of this work to use analyze the sequences in a more natural way (closed to reality) using an end-to-end deep learning model to automatically extract the features.

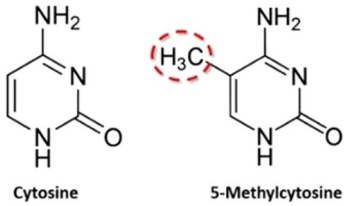


Figure 1. Cytosine with and without methylation.

We have obtained state-of-the-art results for both 41 nucleotides and full length datasets (see Table 3. This paper is organized as follows: Related work is discussed in section "Related work". Our proposed model is presented in section "Materials and methods". The obtained results are discussed in section "Results and discussion" and finally the paper is concluded in in section "Conclusions".

Related work

Methyl group attachment to RNA is a type of post-transcriptional modification controlling the mechanism of RNA interaction with other components of the cell. Recent research have been linked the RNA modifications to different processes ranges from alternative splicing to various diseases, including cancer. Understanding RNA modifications will let a new level of fine tuning of gene expression. It will have a significant impact on various field like fundamental biology, biotechnology, medicine and crop production etc.

In this section we mainly discussed only experimental or machine learning techniques which have been done previously on RNA methylation detection and classification.

The selection of experimental methods have shown in Table 1, depends on the type of modification, its abundance and pre-existing knowledge of context in the modified sequence[21,22]. Furthermore, these techniques are expensive in terms of time and money.

Many computational tools have been built due to the rapid developments of bioinformatics and machine learning techniques[24,25]. Considering the importance of RNA methylation specifically the m5c, there have been many computational tools designed till date that are used to detect or identify m5c RNA methylation. The developed tools or proposed approaches mainly worked on the primary sequence of RNA.

Three vital steps are used in the development of m5c methylation predictor: (1) data collection, (2) feature extraction, and (3) classification or prediction. The taxonomy of features which are used as input to machine learning models are given in Fig. 2.

A number of computational methods as shown in Table 3 had been developed to predict RNA m5c methylation. The description of these methods is discussed in this section.

Pengmian Feng et al.[26], used a support vector machine based-method to predict m5c sites in homo sapiens transcriptome. In the proposed method, RNA sequences were encoded applying the pseudo dinucleotide composition in which three RNA physiochemical features were incorporated. It was observed that the overall success rate that is gained by the developed model is 90.42%.

Sr.	Experimental methods
1	Radioisotope incorporation[21]
2	Thin-layer chromatography[23]
3	Mass spectrometry[21]
4	Differential enzyme or Chemical-RNA interactions[21]
5	Bisulphite RNA sequencing[21]
6	Antibody-based sequencing[21]

Table 1. Strategies for the detection of RNA methylation.

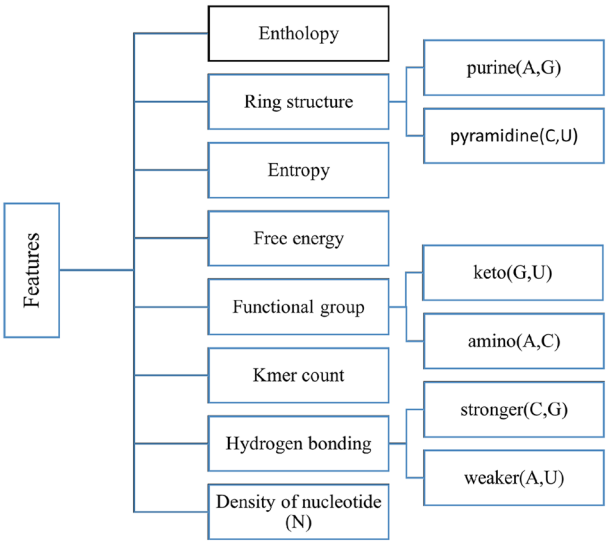


Figure 2. Taxonomy of features.

Pengmian Feng et al.[27] proposed a classification method that was applied for the classification of three kinds of RNA modification m1a, m6a and m5c. Local features (ring structure, functional group and hydrogen bonding) and density information of nucleotides have been employed to encode a RNA sequence. The encoded sequence is converted into general pseudo K-tuple nucleotide composition (PseKNC) vector which is used to train SVM based predictor called iRNA-PseColl. The classifier achieved 77.50% classification accuracy on a human transcriptome m5c methylated dataset.

The authors in[28] proposed a new predictor called iRNAm5C-PseDNC, which has been developed by embodying ten different types of physical-chemical features into pseudo dinucleotide composition through the auto/cross covariance technique. Rigorous jackknife tests has been used which shown that the anticipated accuracy of predictor is 92.37%.

In[29] the researchers built a model for the prediction of m5c cites. The proposed model based on composite features in which three features extraction techniques were combined. After feature extraction, MRMR (Minimum Redundancy Maximum Relevance) was applied as a feature selection method and SVM was used as classifier. The dataset used in this study acquired from RM-base data base in which each sequence is 41 nucleotide long and methylated cytosine is positioned in the center. The predictor have an accuracy up to 93.33%.

In[30] a new m5c site predictor called M5C-HPCR is proposed in which multiple base classifiers are properly integrated in order to improve the accuracy of classification tasks. This combination of multiple classifiers in this manner is called ensemble classification. In order to get discriminative features and encoding method has been done by applying a heuristic nucleotide physicochemical property reduction algorithm(HPCR). The predefined algorithm extracts multiple redacts of physical-chemical properties which were used as input for ensemble classifier. They have demonstrated results for two benchmark dataset using jackknife test and got MCC = 0.850 and AUC = 96.2%.

In[31] RNAm5Cfinder, a web-server developed based on random forest algorithm. It is an efficient tool uses RNA sequence features to identify RNA m5c sites in eight different cell types from mouse and human. The results show that the cell-specific predictors could perform better. For the tissue-specific m5c sites prediction in human the obtained, area under curve (AUC) is 77%.

In[12] a transfer learning based deep model DeepMRMP was built to predict different types of RNA modification sites. They have predicted multiple RNA site modifications for three species i.e., H. sapiens, M. musculus and S. cerevisiae. It is one of the reliable tool for predicting N1-methyladenosine (m1A), pseudouridine (Ψ) and 5-methylcytosine (m5c) modification sites. The designed predictor had achieved an accuracy up to 66% for m5c data set which is very less.

Dou et al.[32] have worked on multiple sequences using SVM and other machine learning techniques for arabidopsis thaliana. Chai et al.[33] have proposed a computational method called staem5 for m5c prediction of mus musculus and arabidopsis thaliana. A deep learning based ensembler classifier is used to predict N5-methylation in[34]. A CNN (Convolutional Neural Network) based model is proposed for prediting different kinds of RNA modifications in[35].

Materials and methods

Dataset

We have used the dataset of Squires et al.[36] for training and evaluation of proposed model. The used data denoted the widespread occurrence of modified cytosine throughout Human transcriptome[36]. The ensemble transcript IDs were available for each methylated and nonmethylated transcript's sequences.

The sequences for these IDs were obtained by applying the biomart tool. The tool is supported by ensemble genome browser. The tool provided many functions for obtaining sequence related information. The cDNA (Complementary DNA) for all transcripts have been downloaded from ensemble genome browser[37–39]. cDNA is similar to RNA, the presence of thymine (T) in place of Uracil (U). By simple T→U transition, the cDNA sequence is transformed into corresponding RNA sequence[40].

Apart from the whole transcript length we also trained the model on the dataset of corresponding data acquired from RM-base data base in which each RNA sequence is 41 nucleotide long and methylated cytosine is positioned in the center. This dataset is used by almost all research works mentioned in section "Related work". Redundant sequences were removed using CH-HIT[41].

Encoding data

All the sequences were first converted into k-mers by using overlapping sliding window. The selection of value for k in kmer is a strenuous process. Its value varies according to research domain. Basically, the sequence length, equals to L yields (L−k+1) total k-mers, and generate total n^k possible unique k-mers. Here, "n" is number of monomers which is four "U,A,C and G" in case of RNA[42,43]. On the basis of preceded work the proposed research has been selected the value of 3 for k. All the sequences were converted into 3-mers, so according to the formula we have got total 64 unique 3-mers. After obtaining k-mers the next step is to transform the k-mers into a digital vector. It is one of the fundamental phases in the process of feature learning and data representation, because machine learning models require numeric data.

In sequence analysis one-hot encoding is one of the common and effective encoding method, which map each sequence to a digital vector. One-hot digital vector designated every word as a $|V|$ dimensional vector with single "1" and the remaining "0s". Here, $|V|$ indicates the size of predefined vocabulary. For example each and every mono nucleotide in RNA can be encoded into a four-dimensional matrix or vector such as A = [1,0,0,0], C = [0,1,0,0], G = [0,0,1,0], U = [0,0,0,1]. So, in our proposed method each 3-mer was converted into a 64 dimentional one-hot digital vector as done by[44–47]. One-hot encoding is the simplest of all encoding techniques. There are advanced encoding techniques like whistle used in[48], Gene2vec[49], Geo2vec[50], Genomics features[48] etc, but we

have selected one-hot encoding because our main focus is to propose a powerfull deep learning classifier which rely less on the underlying encoding technique.

Data preparation for variable length input sequences

The input sequences are of variable length, so zeros are added (Zero padding) to each sequence up to a required longest common length. In this case the model will automatically learn that zeros carried no information, and they are added to generate same length vector[51].

Imbalance data

The biological data are usually imbalanced in which the negative class out numbered the positive one. It might yield awful results to train a model with such an imbalanced data. In order to overcome this issue, the proposed work relied on weighted cross entropy instead of simple loss function. The weighted loss penalizes the classifier if its performance is not well on the minor class. There are a lot of methods that can be applied to alleviate the issue of imbalance[52–58]. To handle the issue of imbalance data, Yang et al.[52] has devised a technique called sample subset optimization. The authors in[53] have proposed a level wise strategy to handle this issue. In[54] the researchers have used a technique called maximum-AUC to handle imbalanced data. A detailed survey is presented about techniques handling imbalance data in[55]. The authors in[56] have used NCR (neighborhood cleaning rule) to tackle the issue of imbalance data. The work in[57] presents a python package while[58] discusses a hybrid data level sampling technique to handle imbalance datasets. The data is splited into training, validation and testing sets in order to validate the generalization of our model(s). The over all methodology is given in Fig. 3.

Performance metrics or evaluation indicators and hyper-parameters settings

To evaluate the efficiency of classifier, many metrics were used including specificity (S_p), sensitivity (S_n), overall accuracy (ACC) and Area Under the Receiver Operating Characteristic Curve (AUC)[59]. The sequences containing methylated sites have been considered as the positive samples and the non methylated ones are negative samples, and all the metrics have computed according to the Formulas shown in Fig. 4. Different hyper-parameters for example learning rate, batch size, activation function, and dropout rate have been encountered. The proposed method has been maintained the default settings of almost all hyper-parameters.The value for the batch size has been selected up to 100. In order to overcome overfitting early stopping was adopted[60,61].

Our proposed models

Kunihiko Fukushima first proposed CNN in 1988[62]. There are three variants of CNN i.e., 1D, 2D and 3D CNN. In our work we have used 1D convolutional neural network, because it is best suited (among the three variants)

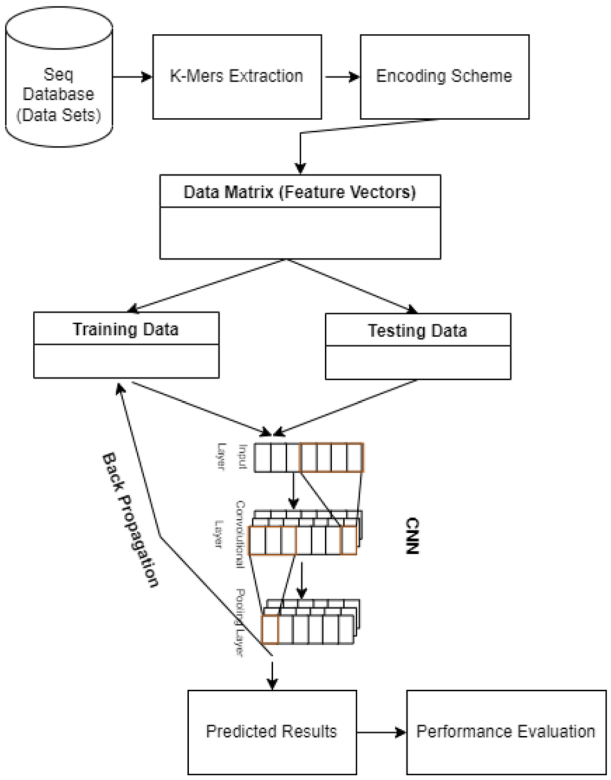


Figure 3. The over all methodology.

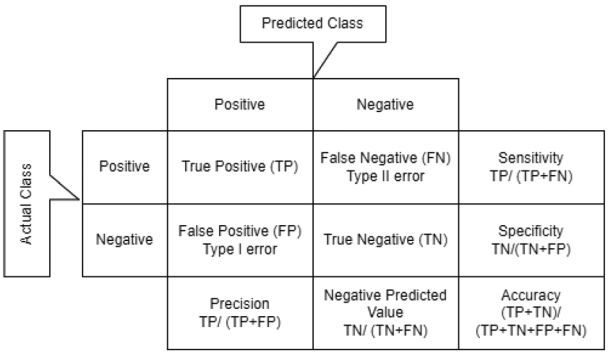


Figure 4. Evaluation metrics.

for sequential data analysis. A detailed and comprehensive review article about the applications of 1D CNN is presented in[63]. Kiranyaz et al.[64] used 1D CNN for the first time in 2015, on patient-specific ECG signals.

The basic idea to design classifier or network is to used variable size multichannel convolution layers[65,66]. Different sizes of convolution kernels are used to convolve the input data simultaneously at different resolutions or different n-grams (groups of words). Different sizes of convolution kernels increase diversity in the extracted features. The outputs of these layers are concatenated. Global max-pooling[67], have been applied after convolution. The purpose of max-pooling layer is to focus on the most active or important features in each feature map.

In the proposed model three convolutional layers followed by a global max-pooling, followed by a dropout layer were added. At the end a dense layer with one unit as output layer was used. The purpose of dropout layer is to overcome the issue of overfitting. All the layers parameters are shown in Table 2.The proposed model architecture is shown in Fig. 5.

Results and discussion

Using a CNN architecture as the one described earlier we analyzed methylated and non methylated sequences for human. We run the model both on the whole transcript as well as on 41 nucleotide lengthy sequences. The ratio of training and testing dataset is set to 90:10.

The accuracy, loss function's graph, for training while ROC and confusion matrix for testing are show in Fig. 6 respectively. These experiments were performed on 41 nucleotides lengthy sequences. We compare our model's results to the state-of-the-art classifiers in Table 3. These works also have used the same dataset of Human. The comparison is based on four different measures accuracy, sensitivity, specificity, and AUROC (see Table 3). For the human dataset taking 41 nucleotide our model shows prediction accuracy up to 96% for testing and 98% for training, outperforming all the methods.

These results demonstrate the ability of deep learning to extract the most significant patterns that characterize the different sequences. Although, our model has low specificity than some of the state of the arts techniques, but it has outperformed all in terms of sensitivity. It is worth mentioning that sensitivity is more critical than specificity. Furthermore, our model has outperformed some of the sate of the art techniques in terms of AUROC as well (see Table 3).

The second experiment consists of using the full sequence as input to the model after applying one-hot encoding. Our aim is to perform the experiment in real and most natural way and it is the main focus of this work. In real word the length of the underlying RNA sequences is not always 41 nucleotides and also the methylated cytosine may not be in the center. The training accuracy and loss and testing ROC and confusion matrix are shown in Fig. 7 respectively. It shows the model's ability to accurately classify the sequence without having a methylated cytosine in the center and considering 41 nucleotide length. Our model achieved accuracy up to 96.10% on test data.

Layers	Parameters
Input	Sequence length = 5000 and 41, dimension = 64;
Convolution layer1	Filters = 128; Filter-length = 21; Activation = relu
Convolution layer2	Filters = 128; Filter-length = 41; Activation = relu
Convolution layer3	Filters = 128; Filter-length = 51; Activation = relu
Concatenate	[Convolution layer1, Convolution layer2, Convolution layer3]
Global max pooling	–
Dropout	0.5
Output	Activation = sigmoid

Table 2. Classifier parameters.

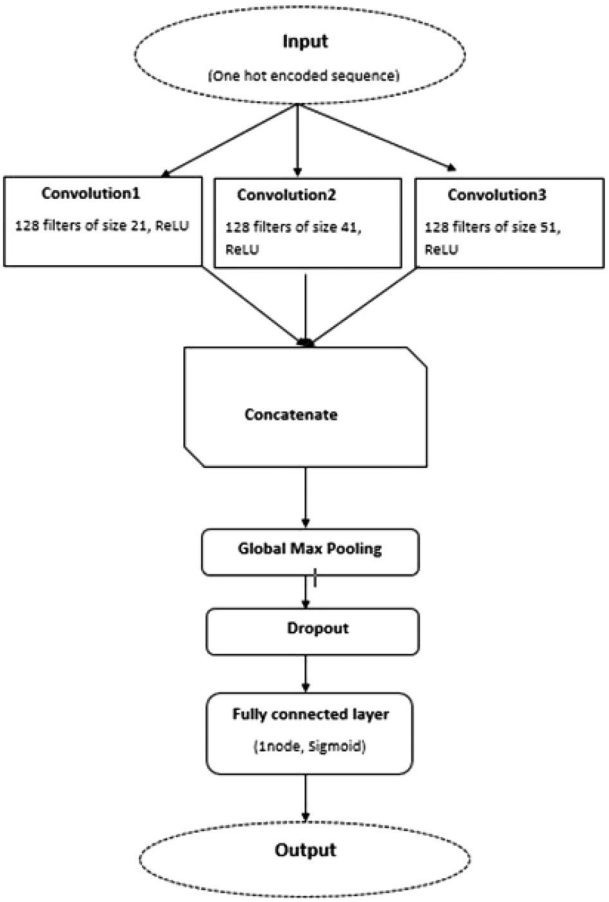
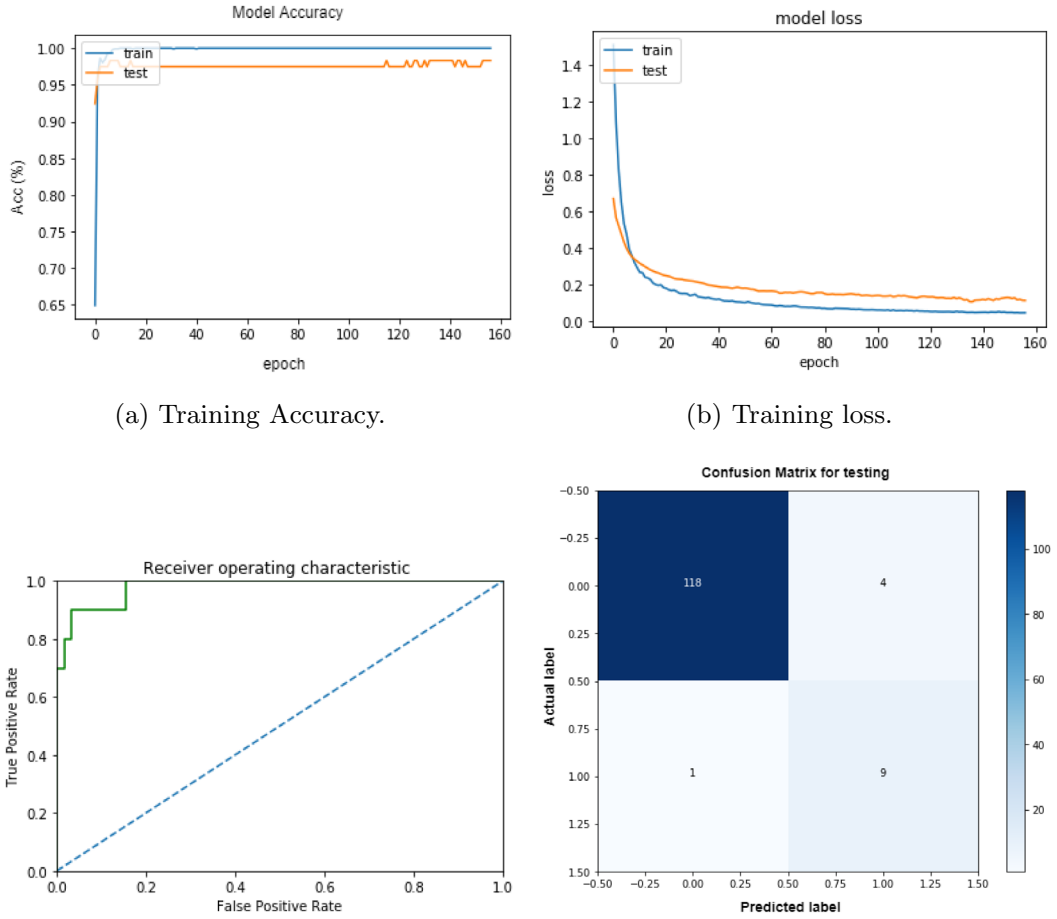


Figure 5. Model Architecture.

Conclusions

Accurate prediction of RNA methylated sequences is necessary for understanding the underlying mechanism of the regulation of genes. A convolutional neural network based model was introduced in this work to distinguish methylated sequences from non methylated sequences of human genome. The basic purpose of this research was to build a sequence based deep learning classifier that can m5c RNA methylation using full length sequences. The evaluation of the proposed architecture showed a promising results when compare to the stat-of-the-art techniques.In future we aim to focus on providing a web server for the current work. Furthermore, we want to extend this work to aberrant methylation classification and prediction.



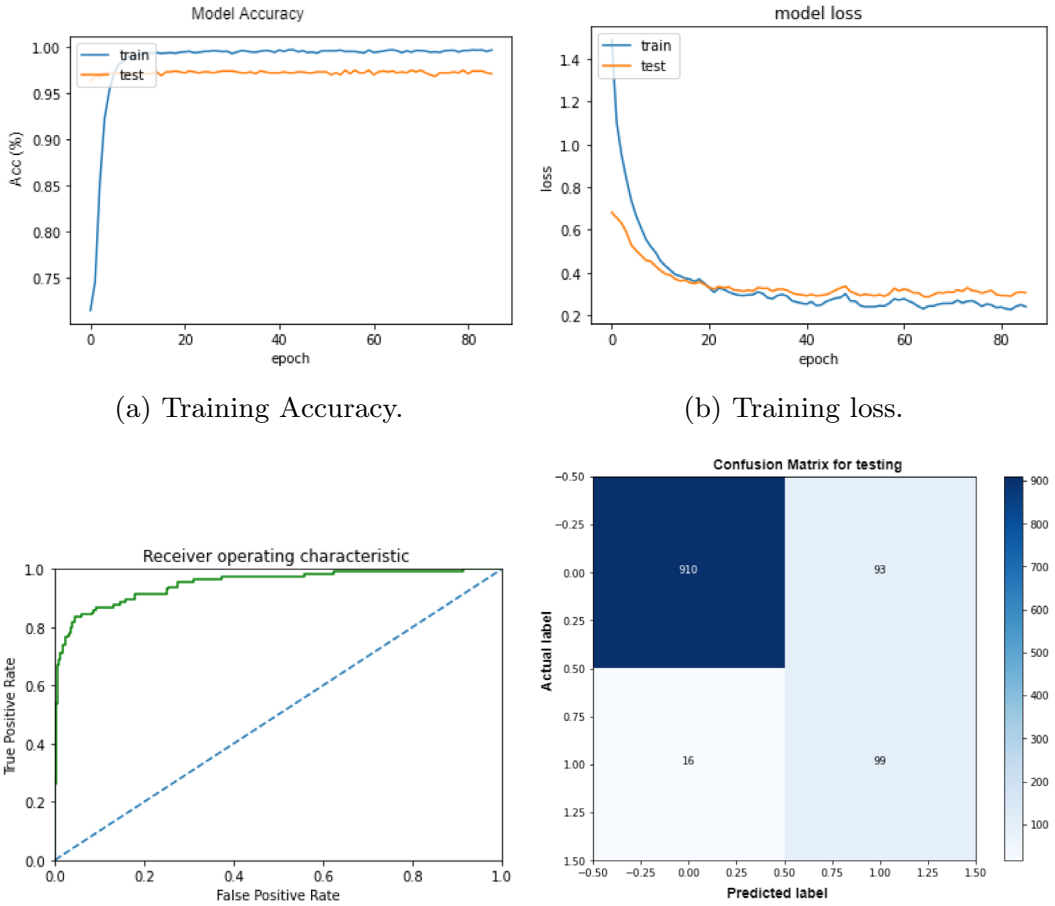
(a) Training Accuracy. (b) Training loss.

(c) Testing ROC. (d) Testing Confusion Matrix.

Figure 6. Performance of our model for training/testing data for 41 nucleotide sequences.

Predictor	Dataset(m5c)	Performance			
		S_n	S_p	Acc	AUROC
Identifying RNA 5-methylcytosine					–
sites[26] via pseudo					–
nucleotide compositions	H. sapiens	85.00	95.83	90.42	–
iRNA-PseColl[27]	H. sapiens	75.83	79.17	77.50	–
iRNA-PseColl[27]	RMbase database	69.89	99.86	92.37	–
Identifying 5-methylcytosine sites in					–
RNA sequence using composite encoding					–
feature into Chou's PseKNC[29]	H. sapiens	90.00	96.66	93.33	–
M5C-HPCR[30]	H. sapiens and Met1320	90.83	95.00	92.92	–
DeepMRMP[12]	H. sapiens, M. musculus	47.95	84.69	66.32	–
XGBoost Framework...[68]	H. sapiens	89	82.0	85.50	0.935
Attention based multi label...[69]	H. sapiens	92	78.0	85.00	0.910
Our model	H. sapiens	96.70	90.00	96.21	0.979

Table 3. Performance comparison of our model for 41 nucleotide sequences.



(a) Training Accuracy. (b) Training loss.

(c) Testing ROC. (d) Testing Confusion Matrix.

Figure 7. Performance of our model for training/testing data for full length sequences.

Data availability

The dataset used in this study is available in the NAR (Neuclic Acid Researcher) Online [https://academic.oup.com/nar] repository and it is discussed in the section "Dataset".

Received: 11 June 2023; Accepted: 29 November 2023
Published online: 11 December 2023

References

1. Hammad, M. *et al.* A novel end-to-end deep learning approach for cancer detection based on microscopic medical images. *Biocybern. Biomed. Eng.* **42**(3), 737–748 (2022).
2. Hammad, M. *et al.* Efficient multimodal deep-learning-based covid-19 diagnostic system for noisy and corrupted images. *J. King Saud Univ.-Sci.* **34**(3), 101898 (2022).
3. Abd El-Latif, A. A., Chelloug, S. A., Alabdulhafith, M. & Hammad, M. Tawalbeh: Accurate detection of alzheimer's disease using lightweight deep learning model on mri data. *Diagnostics* **10**, 2023 (2023).
4. Qiu, J., Wu, Q., Ding, G., Xu, Y. & Feng, S. A survey of machine learning for big data processing. *EURASIP J. Adv. Signal Process.* **2016**(1), 1–16 (2016).
5. Hammad, M. *et al.* Deep learning models for arrhythmia detection in iot healthcare applications. *Comput. Electr. Eng.* **100**, 108011 (2022).
6. Teperino, R., Lempradl, A. & Pospisilik, J. A. Bridging epigenomics and complex disease: The basics. *Cell. Mol. Life Sci.* **70**(9), 1609–1621 (2013).
7. Kumar, S., Chinnusamy, V. & Mohapatra, T. Epigenetics of modified dna bases: 5-methylcytosine and beyond. *Front. Genet.* **9**, 640 (2018).
8. Moore, P. B. & Steitz, T. A. The roles of rna in the synthesis of protein. *Cold Spring Harbor Perspect. Biol.* **3**(11), 003780 (2011).
9. Wang, Y.-C., Peterson, S. E. & Loring, J. F. Protein post-translational modifications and regulation of pluripotency in human stem cells. *Cell Res.* **24**(2), 143–160 (2014).
10. Helm, M. & Motorin, Y. Detecting rna modifications in the epitranscriptome: Predict and validate. *Nat. Rev. Genet.* **18**(5), 275–291 (2017).
11. Song, H. *et al.* Biological roles of rna m5c modification and its implications in cancer immunotherapy. *Biomark. Res.* **10**(1), 1–15 (2022).

12. Sun, P. P. *et al.* Deepmrmp: A new predictor for multiple types of rna modification sites using deep learning. *Math. Biosci. Eng* **16**, 6231–6241 (2019).
13. Asgari, E. & Mofrad, M. R. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PloS one* **10**(11), 0141287 (2015).
14. Liang, H., Sun, X., Sun, Y. & Gao, Y. Text feature extraction based on deep learning: A review. *EURASIP J. Wirel. Commun. Netw.* **2017**(1), 1–12 (2017).
15. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**(7553), 436–444 (2015).
16. Wang, H. & Raj, B. On the origin of deep learning. arXiv:1702.07800 (2017).
17. Yin, Z. *et al.* Computing platforms for big biological data analytics: Perspectives and challenges. *Comput. Struct. Biotechnol. J.* **15**, 403–411 (2017).
18. Cireşan, D., Meier, U., Masci, J. & Schmidhuber, J. A committee of neural networks for traffic sign classification. *Int. Joint Conf. Neural Netw.* **2011**, 1918–1921 (2011).
19. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25**, 1097–1105 (2012).
20. Geiger, J. U. T., Zhang, Z., Weninger, F., Schuller, B. & Rigoll, G. Robust speech recognition using long short-term memory recurrent neural networks for hybrid acoustic modelling. In *Fifteenth Annual Conference of the International Speech Communication Association* (2014).
21. Mongan, N. P., Emes, R. D. & Archer, N. Detection and analysis of rna methylation. *F1000Research* **8**, 1 (2019).
22. Trixl, L. & Lusser, A. The dynamic rna modification 5-methylcytosine and its emerging role as an epitranscriptomic mark. *Wiley Interdiscipl. Rev.: RNA* **10**(1), 1510 (2019).
23. Stahl, E. *et al.* Thin-layer chromatography: A laboratory handbook. *Thin-layer chromatogr. Lab. Handb.* **1962**, 1 (1962).
24. Lv, H. *et al.* Evaluation of different computational methods on 5-methylcytosine sites identification. *Brief. Bioinf.* **21**(3), 982–995 (2020).
25. Wang, H., Wang, S., Zhang, Y., Bi, S. & Zhu, X. A brief review of machine learning methods for rna methylation sites prediction. *Methods* **1**, 141 (2022).
26. Feng, P., Ding, H., Chen, W. & Lin, H. Identifying rna 5-methylcytosine sites via pseudo nucleotide compositions. *Mol. BioSyst.* **12**(11), 3307–3311 (2016).
27. Feng, P. *et al.* irna-psecoll: Identifying the occurrence sites of different rna modifications by incorporating collective effects of nucleotides into pseknc. *Mol. Therapy-Nucleic Acids* **7**, 155–163 (2017).
28. Qiu, W.-R., Jiang, S.-Y., Xu, Z.-C., Xiao, X. & Chou, K.-C. irnam5c-psednc: Identifying rna 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition. *Oncotarget* **8**(25), 41178 (2017).
29. Sabooh, M. F., Iqbal, N., Khan, M., Khan, M. & Maqbool, H. Identifying 5-methylcytosine sites in rna sequence using composite encoding feature into Chou's Pseknc. *J. Theor. Biol.* **452**, 1–9 (2018).
30. Zhang, M. *et al.* Accurate rna 5-methylcytosine site prediction based on heuristic physical-chemical properties reduction and classifier ensemble. *Anal. Biochem.* **550**, 41–48 (2018).
31. Li, J., Huang, Y., Yang, X., Zhou, Y. & Zhou, Y. Rnam5cfinder: A web-server for predicting rna 5-methylcytosine (m5c) sites based on random forest. *Sci. Rep.* **8**(1), 1–5 (2018).
32. Dou, L., Li, X., Ding, H., Xu, L. & Xiang, H. Prediction of m5c modifications in rna sequences by combining multiple sequence features. *Mol. Therapy-Nucleic Acids* **21**, 332–342 (2020).
33. Chai, D., Jia, C., Zheng, J., Zou, Q. & Li, F. Staem5: A novel computational approach for accurate prediction of m5c site. *Mol. Therapy-Nucleic Acids* **26**, 1027–1034 (2021).
34. Hasan, M. M. *et al.* Deepm5c: A deep learning-based hybrid framework for identifying human rna n5-methylcytosine sites using a stacking strategy. *Mol. Therapy* **2022**, 1 (2022).
35. Tahir, M. M., Hayat, G. & Chong, K. T. A convolution neural network-based computational model to identify the occurrence sites of various rna modifications by fusing varied features. *Chemometr. Intell. Lab. Syst.* **211**, 104233 (2021).
36. Squires, J. E. *et al.* Widespread occurrence of 5-methylcytosine in human coding and non-coding rna. *Nucleic Acids Res.* **40**(11), 5023–5033 (2012).
37. Aken, B. L. *et al.* The ensembl gene annotation system. *Database* **2016**, 45 (2016).
38. Kinsella, R. J. *et al.* Ensembl biomarts: A hub for data retrieval across taxonomic space. *Database* **2011**, 4123 (2011).
39. Tahir, M., Tayara, H. & Chong, K. T. Convolutional neural networks for discrimination of rna pseudouridine sites. *IBRO Rep.* **6**, 552 (2019).
40. Zhou, Y., Zeng, P., Li, Y.-H., Zhang, Z. & Cui, Q. Sramp: Prediction of mammalian n6-methyladenosine (m6a) sites based on sequence-derived features. *Nucleic Acids Res.* **44**(10), 91–91 (2016).
41. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. Cd-hit: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**(23), 3150–3152 (2012).
42. Zhang, Q., Pell, J., Canino-Koning, R., Howe, A. C. & Brown, C. T. These are not the k-mers you are looking for: Efficient online k-mer counting using a probabilistic data structure. *PloS one* **9**(7), 101271 (2014).
43. Manekar, S. C. & Sathe, S. R. Estimating the k-mer coverage frequencies in genomic datasets: A comparative assessment of the state-of-the-art. *Curr. Genom.* **20**(1), 2–15 (2019).
44. Wu, C. H. Neural networks for molecular sequence classification. In *The Protein Folding Problem and Tertiary Structure Prediction* 279–305 (Springer, 1994).
45. Zhu, L., Zhang, H.-B. & Huang, D.-S. Direct auc optimization of regulatory motifs. *Bioinformatics* **33**(14), 243–251 (2017).
46. Zhang, H., Zhu, L. & Huang, D.-S. Wsmd: Weakly-supervised motif discovery in transcription factor chip-seq data. *Sci. Rep.* **7**(1), 1–12 (2017).
47. Chuai, G. *et al.* Deepcrispr: Optimized crispr guide rna design by deep learning. *Genome Biol.* **19**(1), 1–18 (2018).
48. Chen, K. *et al.* Whistle: A high-accuracy map of the human n 6-methyladenosine (m6a) epitranscriptome predicted using a machine learning approach. *Nucleic Acids Res.* **47**(7), 41–41 (2019).
49. Zou, Q., Xing, P., Wei, L. & Liu, B. Gene2vec: Gene subsequence embedding for prediction of mammalian n6-methyladenosine sites from mrna. *RNA* **25**(2), 205–218 (2019).
50. Huang, D. *et al.* Geographic encoding of transcripts enabled high-accuracy and isoform-aware deep learning of rna methylation. *Nucleic Acids Res.* **50**(18), 10290–10310 (2022).
51. Dwarampudi, M. & Reddy, N. Effects of padding on lstms and cnns. arXiv:1903.07288 (2019).
52. Yang, P., Zhang, Z., Zhou, B. B. & Zomaya, A. Y. Sample subset optimization for classifying imbalanced biological data. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* 333–344 (Springer, 2011).
53. Li, Y. *et al.* Deepre: Sequence-based enzyme ec number prediction by deep learning. *Bioinformatics* **34**(5), 760–769 (2018).
54. Wang, S., Sun, S. & Xu, J. Auc-maximized deep convolutional neural fields for sequence labeling. arXiv:1511.05265 (2015).
55. Buda, M., Maki, A. & Mazurowski, M. A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw.* **106**, 249–259 (2018).
56. Liu, Z., Xiao, X., Qiu, W.-R. & Chou, K.-C. idna-methyl: Identifying dna methylation sites via pseudo trinucleotide composition. *Anal. Biochem.* **474**, 69–77 (2015).

57. Lematre, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* **18**(1), 559–563 (2017).
58. Kaur, P. & Gosain, A. Robust hybrid data-level sampling approach to handle imbalanced data during classification. *Soft Comput.* **24**(20), 15715–15732 (2020).
59. Flach, P.: Performance evaluation in machine learning: The good, the bad, the ugly, and the way forward. In *Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33* 9808–9814 (2019).
60. Angermueller, C., Parnamaa, T., Parts, L. & Stegle, O. Deep learning for computational biology. *Mol. Syst. Biol.* **12**(7), 878 (2016).
61. Ying, X.: An overview of overfitting and its solutions. In *Journal of Physics: Conference Series, vol. 1168* 022022 (IOP Publishing, 2019).
62. Fukushima, K. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Netw.* **1**(2), 119–130 (1988).
63. Kiranyaz, S. *et al.* 1d convolutional neural networks and applications: A survey. *Mech. Syst. Signal Process.* **151**, 107398 (2021).
64. Kiranyaz, S., Ince, T. & Gabbouj, M. Real-time patient-specific ecg classification by 1-d convolutional neural networks. *IEEE Trans. Biomed. Eng.* **63**(3), 664–675 (2015).
65. Rakhlin, A. *Convolutional Neural Networks for Sentence Classification* (GitHub, 2016).
66. Yin, W. & Schutze, H. Multichannel variable-size convolution for sentence classification. arXiv:1603.04513 (2016).
67. Li, W., Liu, K., Zhang, L. & Cheng, F. Object detection based on an adaptive attention mechanism. *Sci. Rep.* **10**(1), 1–13 (2020).
68. Abbas, Z. & ur-Rehman, M., Tayara, H., Zou, Q., & Chong, K.T. Xgboost framework with feature selection for the prediction of rna n5-methylcytosine sites. *Mol. Therapy* **2023**, 14 (2023).
69. Song, Z. *et al.* Attention-based multi-label neural networks for integrated prediction and interpretation of twelve widely occurring rna modifications. *Nat. Commun.* **12**(1), 4011 (2021).

Acknowledgements

This work is supported by EIAS (Emerging Intelligent Autonomous Systems) Data Science Lab, Prince Sultan University, KSA. Intelligent Autonomous Systems) Data Science Lab, Prince Sultan University, KSA. The authors would like to acknowledge the support of Prince Sultan University for paying the Article Processing Charges (APC) of this publication.

Author contributions

I.A. reviewed literature, wrote code and prepared the first draft. S.S. provided supervision, formulated the problem, helped in proposed solutionand obtaining results. Revised the manuscript. M.E. provided supervision, helped in obtaining results and reviewed the manuscript. S.J. provided supervision, helped in proposed solution and review the manuscript. A.A.A.L. provided supervision, helped in obtaining the results and reviewed the manuscript. N.U.H. provided supervision, helped in proposed solution and review the manuscript. G.A. provided supervision, helped in proposed solution and review the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.J.