JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 04 Issue 02
2023
Research Article
https://doi.org/10.56979/402/2023

Deep Learning Algorithms to Predict m7G from Human Genome

Hassan Kaleem[1 *], Malik Tahir Hassan[1], Sajid Mahmood[1], and Muhammad Noman Khalid[2]

[1] School of Systems and Technologies, University of Management and Technology, Lahore, Pakistan.
[2]Medicine and Surgery, Allama Iqbal Medical College, Lahore, Pakistan.
[*]Corresponding Author: Hassan Kaleem. Email: hassanrao.hr@gmail.com

Received: December 02, 2022 **Accepted:** February 20, 2023 **Published:** March 29, 2023.

Abstract: N7-methyl guanosine (m7G) is a common post-transcriptional RNA alteration that plays a role in various biological processes such as gene expression, protein synthesis and cell viability. It is also linked to several illnesses, thus a thorough understanding of the mechanism and biological activities of m7G sites is required. Several machine learning models have been developed to predict m7G from the human genome, but machine learning models require feature extraction from the dataset and model training, which is a complex and time taking process for biologists and biochemists. For the first time, deep learning based algorithm is used to predict m7G. The main benefit of using a deep learning model is it does not require any features extraction from the dataset before passing it to the model, instead it generates features by itself. The LSTM model has outperformed all the other machine learning algorithms and achieved 0.7977 MCC on the independent dataset and after parameter optimization through KerasTuner, the model achieved 0.9934 MCC on independent dataset.

Keywords: m7G Genome; Human RNA; LSTM; KerasTuner.

1. Introduction

It has been determined that N7- methyl guanosine (m7G) is a common post-transcriptional RNA alteration. Messenger RNA (mRNA) has a methyl (-CH3) group added at position N7 during the transcription initiation phase that specifies its function in a variety of biological processes, such as gene expression, RNA processing and metabolism, transcript integrity, protein synthesis, and cell viability [1] [2] [3] [4]. The m7G modification controls and regulates each stage of the mRNA life events, including transcription elongation, mRNA slicing, polyadenylation, and nuclear export [5] [6]. The m7G alteration plays a variety of biological roles, but it is also linked to a number of illnesses, including growth retardation [7], microcephalic primate dwarfism, brain maldevelopment, and the emergence of specific autoimmune disorders [8] [9]. Because of its value in modulating a small number of biological events and because of its value in modulating a few biological events and having a connection to several disorders. A thorough understanding of the mechanism and biological activities of m7G sites is required. Alkalanine-seq, MeRIP-seq, chemical-assisted m7G seq, and miCLIP-seq are a few of the experimental methods that have been developed to find m7G sites [10] [11] [12]. These experimental methods appear to be laborious, expensive, and unable to accurately discover m7G sites for transcriptome-wide detections. Therefore, it is crucial to design an efficient competition proposition for the accurate identification of m7G sites.

Artificial Intelligence (AI) is when machine performs tasks without the human involvement. Deep learning is a branch of machine learning based on artificial neural networks (ANN). Neural network is a kind of network which is inspired by human brain [13]. Deep Learning uses multiple layers of processing to extract better features from the dataset and learn things. Sound, text, sequences, images, videos could be used as the dataset. The results that deep learning has achieved, were never achieved before because it

uses multiple layered neural network architecture and large labels dataset [14]. Deep learning architecture has shown in the Figure 1.

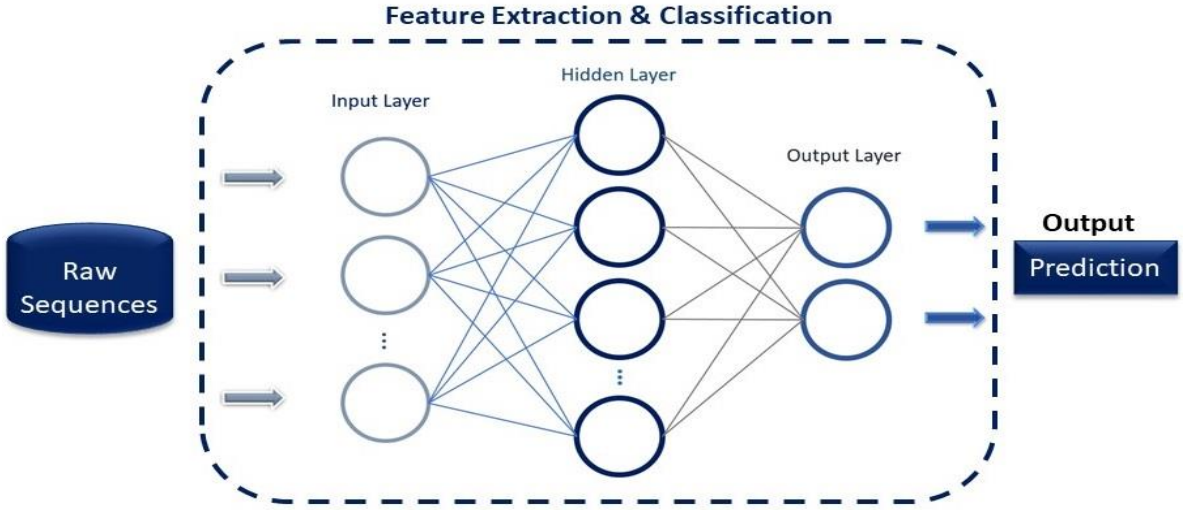

Figure 1. Architecture of Deep Learning Model

Recently, multiple machine learning models have been developed for the forecast of N7-methylguanosine from human genome, Chen et al [15] evolved the primary computational forecaster known as iRNA-N7-methylguanosine to recognize N7-methylguanoisine sites in the human metabolome. The characteristics synthesis approach was used to calculate mutual classification as well as characteristics grounded on structure.iRNA-N7-methylguanoisine attained an exactness of 89.88% in the jackknife test. The advantage of iRNA-N7-methylguanoisine for classifying N7-methylguanoisine sites was correspondingly established by associating with further approaches. They believe that iRNA-N7-methylguanoisine can be developed as a valuable means to classify N7-methylguanoisine sites. Song et al [16] conducted a research on current development in N7-methylguanosine Ribonucleic acid trainings has concentrated on the aforementioned inner ("instead of capped") occurrence inside mRNAs. Hundreds of thousands of inner mRNA N7-methylguanoisine sites have been recognized inside mammalian transcriptomes, and only one source to greatest portion, interpret and examine the enormous data produced freshly are deeply desirable. They report here m7GHub, a complete online podium for interpreting the site, guideline and pathogenesis of inner mRNA m7G. The m7GHub contains of four foremost mechanisms, counting: the first database recording 1218 disease-associated genetic mutations that may regulate m7G methylation. The primary inner mRNA m7G database includes 44 058 experimentally authenticated inside mRNA m7G sites, a high accuracy forecaster based on sequences, the primary web server for evaluating the influence of mutations on m7G status, and more. When used as a whole, m7GHub will be a helpful tool for study on internal mRNA m7G alteration. Yang et al [17] proposed an approach that is based on sequences to recognize the changing site in RNA sequences. In the beginning, numerous types of sequence structures were taken out to code N-7 methyl guanosine (m7G) and non-m7G examples. Consequently, they used mRMR and F-score to acquire the prime subsection of characteristics that could manufacture the maximum prediction accuracy. In 10-fold cross-validation, outcomes presented that the uppermost accuracy is 94.67% and attained by support vector machine for classifying m7G sites in genome of humans. They also found out that the outcomes of algorithms other than SVM and discovered that the SVM-based approach outperformed over all other approaches. Liu et al [18] proposed a AI based approach came about for the forecast of the methyl guanosine sites, and numerous five character withdrawal techniques (Pseudo dinucleotide composition, Pseudo k-tuple composition, K monomeric units, Ksnpf frequency and Nucleotide chemical property) remained used for the purpose of extracting features. To treasure trove the optimized feature subgroup, they used Random Forest based algorithm as well as a predictor based on SVM that attained the finest accuracy by gaining the topmost 240 characteristics for the training of model. With dissimilar calculation approaches, 10-fold cross validation, author test and self-regulating sample, N7-methylguanoisine Forecaster attained modest approaches matched with the contemporary interpreter iRNA-N7-methylguanoisine. The forecaster established in this

research can propose beneficial info for clarifying the procedure of the N7-methylguanoisine sites as well as associated investigational authentications. Bi et al [19] develop an innovative interpretable approach based on Machine learning model XG-N7-methylguanoisine for the contradictions of N7-methylguanoisine sites by way of the XG-Boost approach and six various kinds of sequence programming patterns. Either of 10-fold also jackknife cross-validation tests specify that XG- N7-methylguanoisine leave behind iRNA-N7-methylguanoisine. Furthermore, through powerful SHAP approach, this novel structure as well arrange for necessary analyses of the approach accuracy and highlight the maximum significant characteristics for recognizing N7-methylguanoisine sites. ExtremeGradient-N7-methylguanoisine is predicted to help as a valuable tool and direction for researchers in their study of mRNA variation sites. Dai et al [20] came up with a machine learning based approach known as m7G-IFL to recognize m7G sites via iterative feature demonstration procedure. M7G-IFL was appraised and matched with current forecasters to establish its supremacy. The consequences validate that their analyst leave behind current forecasters in respect of accuracy for recognizing m7G sites. By studying and matching the characteristics used in the interpreters, they set up that the optimistic and destructive trials in their feature space were extra disconnected than in current feature space. This consequence proves that their extracted features are much discriminative info through the iterative feature learning procedure, and consequently added to the analytical performance development. Ning et al[21] conducted an experiment, his study was motivated by present "deep learning and natural language processing" expertise, a technique termed "N7-methylguanoisine–Double Long Short Term Memory" is considered using "long short-term memory" system joined with completely linked net. After calculation of numerous characteristics, lone nucleotide (0, 1) code and nucleotide compound attribute are applied as feature training system. Through evaluation, "Double-LSTM" approach grounded on binary receptors torn apart via middle guanosine sites give superior outcome than Single Long Short Term Memory algorithm created on entire syndrome. Between numerous instructions, the way from boundary toward midpoint became the maximum accuracy in Double Long Short Term Memory algorithms. In conclusion, the outcome of N7-methylguanoisine-Double Long Short Term Memory is regulated with a "specificity of 94.37%, a sensibility of 92.96% and an accuracy of 93.66%", which pointed out that "N7-methylguanoisine–Double Long Short Term Memory" can be a beneficial tool for forecast of human RNA N'-meth guanosine sites. Zhang et al [22] offered a novel technique through integrating BERT-based polyglot approach in genetic genomics to characterize the statistics of RNA orders. Initially, they treat RNA sequences as normal sentences then work with BERT algorithm to alter them into static measurement using numerical grounds. Furthermore, a scheme based of feature selection on the flexible net technique is built to get rid of unnecessary features and keep significant features. Lastly, the designated feature subsection is contribution into an assembling collaborative algorithm to forecast N7-methylguanoisine sites, and the hyper parameters of the transformer based approach are altered with "tree-structured Parzen estimator (TPE)" method. Through "10-fold cross-validation", the execution of BERT- N7-methylguanoisine is calculate with an "accuracy of 95.48% and an MCC of 0.9100". The investigational outcomes specify that the projected technique considerably overtakes advanced forecast approaches in the documentation of m7G alterations. Shoombuatong et al [23] conducted an experiment using ensemble learning framework. In their experiment, four various computational structures were analyzed and the finest one was selected grounded on their performance. In total, 60 baseline models were developed, 54 on the first layer and 6 on the second. In the last random forest classifier was used for the final prediction. THRONE achieved 0.810 MCC on tenfold cross-validation testing. However, to check the model's robustness they created an independent dataset for testing. 334 samples were created for m7G and 3340 samples were created for non-m7G. THRONE achieved 0.568 MCC on independent testing. Kaleem et al used KerasTuner to predict anticancer peptides. KerasTuner is hyperparameter optimization framework it finds the best hyperparameter values for the model, in KerasTuner multiple units are passed to the model and then model finds out which unit is the most suitable for the dataset.

2. Materials and Methods

Benchmark dataset is used in the experiment for fair compression, the dataset has been proposed and used in previous work [15]. The dataset consists of 741 m7G samples and 741 non-m7G samples. To check the model's robustness independent testing dataset is used to evaluate the model. Shoombuatong et al [23]

created the independent dataset for their model THRONE. The independent dataset consists of 334 m7G samples and 3340 non-m7G samples.

Long short term memory (LSTM) is an artificial neural network (ANN) that was introduced by Hochreiter and Schmidhuber [24]. They work very well on a large variety of problems, especially on textual data, it is refined and popularized by many people in the past decade. They are designed to remember the information for a long period of time. The Figure 2 shows the structure of LSTM.

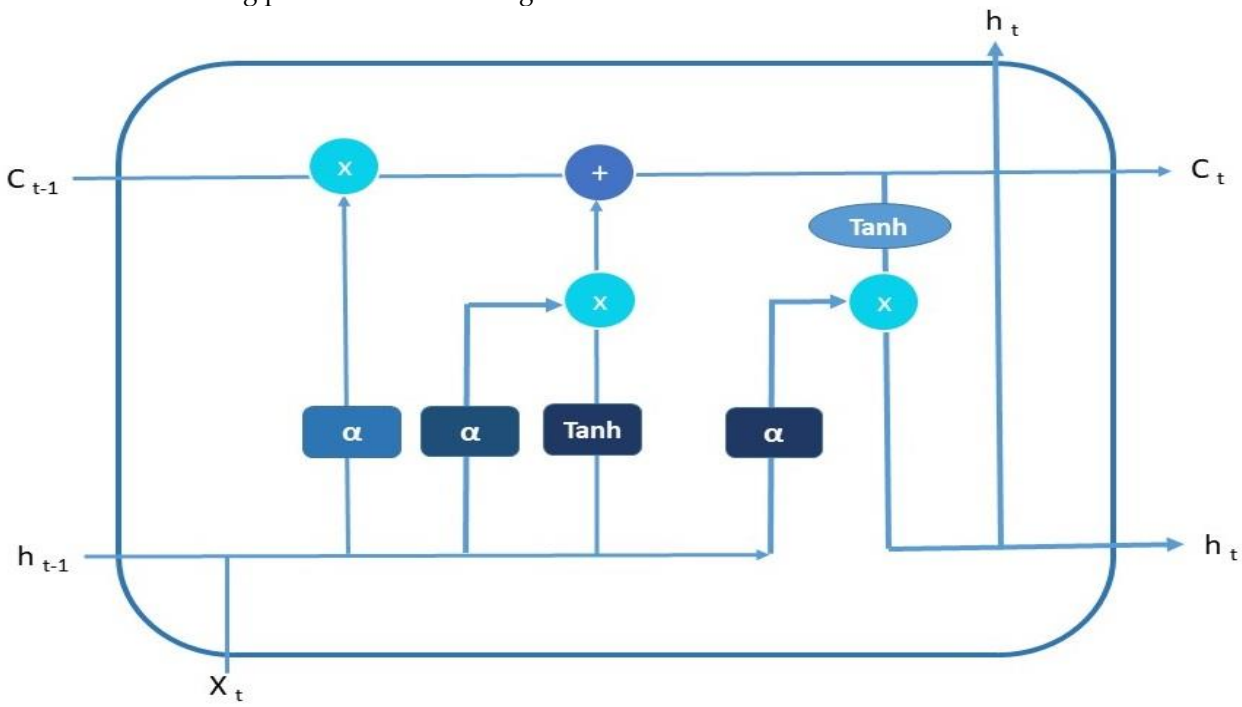

Figure 2. Architecture of LSTM Model

The horizontal line that runs through the structure's schematic and represents the central concept of LSTM is cell state. Information moves along the cell state without changing, much like a conveyer belt in an LSTM. The cell state, which is controlled by structures known as gates, can have information added to it or removed from it by the LSTM. Gates are the mechanisms used to allow information to pass voluntarily. A sigmoid neural network layer plus a pointwise multiplication operation make up these gates. The number between 0 and 1 is output by the sigmoid layer. Zero indicates that everything is permitted to pass through the gate. To regulate and safeguard the cell state, the LSTM has three of these gates.

2.1. Performance Evulation Stretagies

To assess the model, five performance measurements were used. Matthew's correlation coefficient (MCC), accuracy (Acc), sensitivity (Sn), specificity (Sp), and area under the curve are some examples of these metrics (AUC). The definition of the metrics is as follow:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (1)$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Sn = \frac{TP}{TP + FN} \quad (3)$$

$$Sp = \frac{TN}{TN + FP} \quad (4)$$

$$AUC = Sn\,(1 - Sp) \quad (5)$$

True positives, true negatives, false positives, and false negatives are denoted by the letters TP, TN, FP, and FN, respectively.

3. Results

Deep Learning algorithms take raw input and generate features by itself. The results are changed if we changes the maximum features for LSTM. 700 is the best maximum features value, it gives highest results on it for the Main dataset for 10 fold cross validation testing. For independent testing, 500 are the best values because it gives highest accuracy on that.

Table 1. Parameter used in LSTM

Maximum Features	**500-700**
Activation Function	Softmax
Optimizer	Adam
Epochs	100
Batch size	32

3.1. Comparative Analysis

Tenfold cross-validation testing have been performed on the Main dataset, in tenfold cross validation testing, dataset is divided into ten different parts and every part is tested to get the most accurate results. At the end mean is calculated on the basis of given accuracies. LSTM results are in Table 2.

Table 2. Comparison of LSTM with THRONE on tenfold cross validation testing.

Methods	**Main Dataset**				
	Acc	**Sn**	**Sp**	**AUC**	**MCC**
LSTM	0.97	1.0	0.9328	0.9665	0.9402
THRONE	0.950	0.950	0.951	0.966	0.900

The LSTM model achieved MCC, Acc, Sn, Sp, and AUC of 0.9402, 0.97, 1.0, 0.9328, and 0.9665 respectively. Independent dataset is used to evaluate the model's robustness. LSTM achieved MCC, Acc, Sn, Sp, and AUC of 0.7977, 0.96, 0.95, 1.0, and 0.97 respectively. LSTM model gives 0.2297 better MCC results then THRONE. Figure 3 shows the results comparison between LSTM, THRONE, and iRNA-m7G on independent testing.

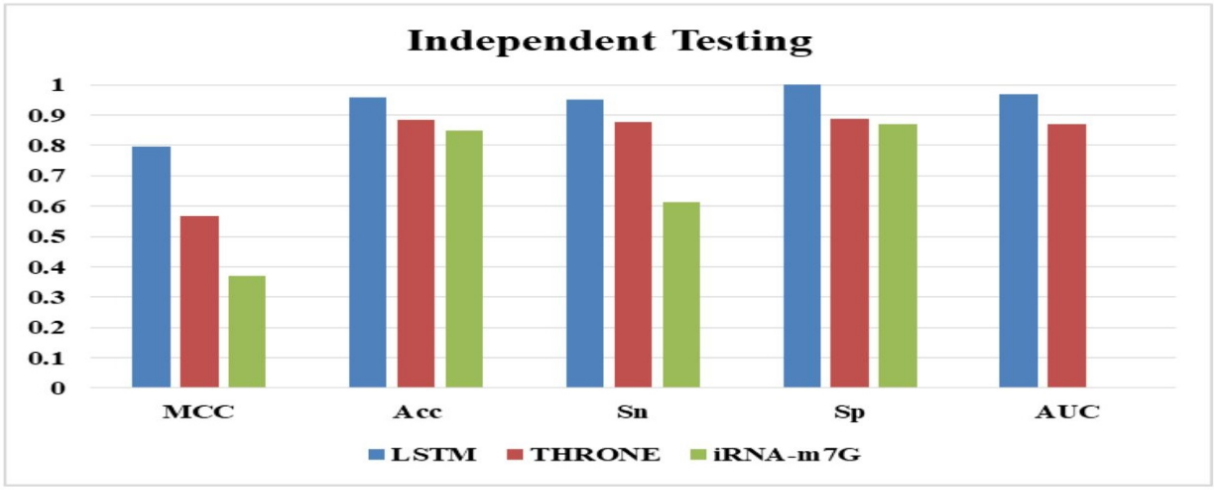

Figure 3. Performance comparison between LSTM, THRONE, and iRNA-m7G on independent dataset.

KerasTuner is used to predict m7G from human genome. KerasTuner achieved MCC, Acc, Sn, Sp, and AUC of 0.9935, 1.00, 0.9988, 1.00, and 0.9994 respectively.

4. Conclusions

Multiple machine learning model were developed to predict Human RNA, machine learning model requires features extraction from the dataset, this is a time taking process and results were not satisfactory. On the other hand, deep learning models takes raw input, generates features by themselves and provide

better results. LSTM model was used to conduct the experiment. LSTM model outperformed all the other predictors and achieved 0.7977 MCC on independent testing. During the experiment, one thing that was identified that if the maximum features value is changed from 700 for Main dataset, the results starts to decline. And for Independent dataset, 500 is the optimal features value that the model should generates, it gives highest accuracy on it. Deep learning algorithms require large amount of dataset to train otherwise they do not perform very well, but if the model is over trained, it's results starts to decline because the model is overfitting. In order to get the best results from deep learning algorithms, their parameters should be selected carefully otherwise they will not perform very well. To solve the problem about large dataset for deep learning models, KerasTuner is used in the experiments which achieved better results than LSTM and it requires less dataset to train.

References

1 Potential regulatory role of epigenetic RNA methylation in cardiovascular diseases - ScienceDirect. https://www.sciencedirect.com/science/article/pii/S075333222100161X (accessed Oct. 03, 2022).

2 Regulation of mRNA cap methylation | Biochemical Journal | Portland Press. https://portlandpress.com/biochemj/article/425/2/295/44780/Regulation-of-mRNA-cap-methylation (accessed Oct. 03, 2022).

3 Discovery of m7G-cap in eukaryotic mRNAs." https://www.jstage.jst.go.jp/article/pjab/91/8/91_PJA9108B-01/_article/-char/ja/ (accessed Oct. 03, 2022).

4 Y. Wu, S. Zhan, Y. Xu, and X. Gao, "RNA modifications in cardiovascular diseases, the potential therapeutic targets," Life Sci., vol. 278, p. 119565, Aug. 2021, doi: 10.1016/j.lfs.2021.119565.

5 Specific regulation of mRNA cap methylation by the c-Myc and E2F1 transcription factors | Oncogene. https://www.nature.com/articles/onc2008463 (accessed Oct. 03, 2022).

6 L. Pandolfini et al., "METTL1 Promotes let-7 MicroRNA Processing via m7G Methylation," Mol. Cell, vol. 74, no. 6, pp. 1278 1290.e9, Jun. 2019, doi: 10.1016/j.molcel.2019.03.040.

7 tRNA m7G methyltransferase Trm8p/Trm82p: Evidence linking activity to a growth phenotype and implicating Trm82p in maintaining levels of active Trm8p. https://rnajournal.cshlp.org/content/11/5/821.short (accessed Oct. 03, 2022).

8 S. Lin, Q. Liu, V. S. Lelyveld, J. Choe, J. W. Szostak, and R. I. Gregory, Mettl1/Wdr4-Mediated m7G tRNA Methylome Is Required for Normal mRNA Translation and Embryonic Stem Cell Self-Renewal and Differentiation, Mol. Cell, vol. 71, no. 2, pp. 244-255.e5, Jul. 2018, doi: 10.1016/j.molcel.2018.06.001.

9 P. L. Pereira et al., A new mouse model for the trisomy of the Abcg1–U2af1 region reveals the complexity of the combinatorial genetic code of down syndrome, Hum. Mol. Genet., vol. 18, no. 24, pp. 4756–4769, Dec. 2009, doi: 10.1093/hmg/ddp438.

10 AlkAniline-Seq: Profiling of m7G and m3C RNA Modifications at Single Nucleotide Resolution - Marchand - 2018 - Angewandte Chemie International Edition - Wiley Online Library." https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201810946 (accessed Oct. 03, 2022).

11 L.-S. Zhang et al., "Transcriptome-wide Mapping of Internal N7-Methylguanosine Methylome in Mammalian mRNA," Mol. Cell, vol. 74, no. 6, pp. 1304-1316.e8, Jun. 2019, doi: 10.1016/j.molcel.2019.03.036.

12 Dynamic methylome of internal mRNA N7-methylguanosine and its regulatory role in translation | Cell Research. https://www.nature.com/articles/s41422-019-0230-z (accessed Oct. 03, 2022).

13 J. Schmidhuber, "Deep learning in neural networks: An overview," Neural Netw., vol. 61, pp. 85–117, Jan. 2015, doi: 10.1016/j.neunet.2014.09.003.

14 A. Javaid, Q. Niyaz, W. Sun, and M. Alam, "A Deep Learning Approach for Network Intrusion Detection System," EAI Endorsed Trans. Secur. Saf., vol. 3, no. 9, pp. 21–26, May 2016.

15 Kaleem, H., Rukhsar, S., & Khalid, M. N. (2022). Anticancer Peptides Prediction: A Deep Learning Approach. Journal of Computing & Biomedical Informatics, 3(02), 144–151. https://doi.org/10.56979/302/2022/81

16 W. Chen, P. Feng, X. Song, H. Lv, and H. Lin, "iRNA-m7G: Identifying N7-methylguanosine Sites by Fusing Multiple Features," Mol. Ther. - Nucleic Acids, vol. 18, pp. 269–274, Dec. 2019, doi: 10.1016/j.omtn.2019.08.022.

17 m7GHub: deciphering the location, regulation and pathogenesis of internal mRNA N7-methylguanosine (m7G) sites in human | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/36/11/3528/5803644 (accessed Oct. 04, 2022).

18 Prediction of N7-methylguanosine sites in human RNA based on optimal sequence features - ScienceDirect. https://www.sciencedirect.com/science/article/pii/S0888754320309228 (accessed Oct. 04, 2022).

19 X. Liu, Z. Liu, X. Mao, and Q. Li, "m7GPredictor: An improved machine learning-based model for predicting internal m7G modifications using sequence properties," Anal. Biochem., vol. 609, p. 113905, Nov. 2020, doi: 10.1016/j.ab.2020.113905.

20 An Interpretable Prediction Model for Identifying N7-Methylguanosine Sites Based on XGBoost and SHAP - ScienceDirect. https://www.sciencedirect.com/science/article/pii/S2162253120302511 (accessed Oct. 04, 2022).

21 Iterative feature representation algorithm to improve the predictive performance of N7-methylguanosine sites | Briefings in Bioinformatics | Oxford Academic." https://academic.oup.com/bib/article-abstract/22/4/bbaa278/5964186 (accessed Oct. 05, 2022).

22 Q. Ning and M. Sheng, "m7G-DLSTM: Intergrating directional Double-LSTM and fully connected network for RNA N7 methlguanosine sites prediction in human, Chemom. Intell. Lab. Syst., vol. 217, p. 104398, Oct. 2021, doi: 10.1016/j.chemolab.2021.104398.

23 BERT-m7G: A Transformer Architecture Based on BERT and Stacking Ensemble to Identify RNA N7-Methylguanosine Sites from Sequence Information. https://www.hindawi.com/journals/cmmm/2021/7764764/ (accessed Oct. 04, 2022).

24 W. Shoombuatong, S. Basith, T. Pitti, G. Lee, and B. Manavalan, "THRONE: A New Approach for Accurate Prediction of Human RNA N7-Methylguanosine Sites," J. Mol. Biol., vol. 434, no. 11, p. 167549, Jun. 2022, doi: 10.1016/j.jmb.2022.167549.

25 S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory, Neural Comput., vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.