ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 9, September 2021

DNA Sequence Classification using Machine Learning

K.V.Veeresh, kvveeresh591@gmail.com
PG Student,Department of Mathematics,Ramaiah University of Applied Science, Bengaluru,India

Abstract: Recently deeply neural network models have provided the state-of-the-art prediction of transcription factor binder (TFBS) on site classification (DNN). Nevertheless, how these approaches identify relevant indicators of the DNA sequences and why TFs are still unclear at specific locations. This paper proposes the Deep Motif Dashboard toolkit, which offers a number of viewing strategies for extracting motifs or sequence patterns from deep TFBS network templates. We demonstrate how you can imagine three main DNN models: convolutionary and recurring networks. The first method for visualisation is to find the saliency map of a test sequence using a first-order derivative that describes the importance for the prediction of every nucleotide. Secondly, we insert time output values to indicate the prediction score for a model for a sequential input over time, because of repeat models temporarily predicting (from one end to the other of the TFBS sequence. Finally, a class visualisation strategy finds the ideal input sequence for the given TFBS positive class by means of stochastic gradient optimisation.Our experimental results show that the best of the three architectures is the convolutionary and recurring architecture. The techniques of viewing suggest that both motifs as well as their dependence are predicted by CNN-RNN.

Key words: algorithms, consensus sequence, DNA sequence recognition, pattern matching, tools for computational biology

I. INTRODUCTION

DNA study is a major understanding factor for organisms in life science. DNA carries most of the genetic instructions for all organisms to develop, work and reproduce. We are now able to read a DNA sequence easily with the development of sequencing technologies. According to NHGRI Genome Sequencing Program data in October 2015, the costs of reading a million pairs of base-pairs quickly dropped from more than $5,000 in September 2001 to $0.014. There are also increasing exponentially the number of DNA sequence data. GenBank, a popular database of DNA sequences, grew by over 2 billion base pairs in December 2015, for example. It is very good to understand DNA by the power of our contemporary computer using this huge amount of information.
We have offered a profound understanding model using a single hot vector to preserve each nucleotide's sequence positions. Inspired by a profound model for learning text classification. In several validation datasets, this model has led to improvements in performance.

In recent years, several computer techniques have shown great results on various research problems[6-10]. Same methods for calculating a DNA sequence as a promoter or non promoter

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 9, September 2021

have been developed, along with some techniques for the identification of a certain sigma class of a promoter. For example, reference [11] introduced an identification sequence to predict a promoter presence. The method proposed, PseZNC, develops the sequence for nucleotide-based DNA. In addition, a Z-curve variable window method was presented to identify promoters [12].The LibSVM library-based software package BacSVM + has been reported to forecast Bacillus subtilis promoters[13].
The prediction method »28 and˙50 promoting in E. coli was also developed by De Avila e Silva et al.[14] based on the duplex stability of the neural network.

A deep function selection approach proposed by the [15] method evaluates the nonlinearity of the deep structure and selects the profound function subset to predict promoters within the DNA sequence at the input level. A hybrid technology was presented at et al[16] to predict promoters and their strengths, combining in depth learning and FastText N grammes. Finally, Rahman et al. [17] introduced a prediction method based on the classification function of a subspace-based ensemble.

II. RELATED WORKS

Classification techniques were developed in order to classify the protein sequence into a particular class, sub-class or family. All these methods tend to extract certain characteristics of protein and then compare them and then classify the protein sequence.

Jason T. L. Wang, Qic Heng Ma, Dennis Shasha, and Cathy H Wu[6] used a model in a neural network in their work to classify the existing protein sequences. Some characteristics from protein sequences for use as an input have been removed for this purpose.The high-level characteristics are extracted from the global and local sequence similarity using certain coding techniques. The methods used to discover global similarity are a method for encoding 2 grammes and an exchange group of six letters. To achieve local similarity, some user-defined variables were used. Also used to calculate the meaning of the motif was the MDL principle. Some features already defined have been used for intermediate layers of the neural network or hidden layers. This model's accuracy is between 90 and 92% [6].

RamadeviYellasiri, C.R.Rao[12], proposed a new classification model called the Rough Set Classifier to categorise voluminous protein information based on structural and functional protein properties. This model is fast and accurate, and can be used as an efficient classification tool. The accuracy of this classification is 97.7%. It is a hybridised tool which contains arithmetic sequences, concept grid and rough sets theory.The search area of the domain may be reduced to 9% without losing the protein classification potentials. The family information has special arithmetical identification and is used to reduce the search space of the domain. The rules generated are stored in the database of arithmetic.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 9, September 2021

SuprativSaha and RituparnaChaki[13] suggested a model in three phases for classifying unknown proteins in known families. In the first stage, the data set is reduced so that the noisy sequences are eliminated. It increases the accuracy and reduces the time required for computation.In the second stage a ranking algorithm will be applied which will provide the required characteristics such as molecular weight and isometric point. In this classification, the sequences are classified. In the final phase, the input sequence is classified into that particular class or family by neighbourhood analysis. This rule has the power, between the protein sequence and class, sub-class and family, to extract specific linkages. In addition to the data analysis technique, this type of classification produces information based on knowledge.

III. METHODOLOGY

DNA sequence representation

DNA sequences are transformed into pictures through three steps. First, we're a sequence as a list of k-mers. Then, every k-mer is changed into a single hot vector, which makes the sequence a one-hot vector list. Finally, with Hilbert curves each of the items in the list, we create a picture tensor. The following details are given for each of the steps.

In the light of molecular biology, nucleotides represent a DNA sequence does not mean much in one way. In protein synthesis, nucleotide motifs instead play a crucial role.The kmers of a sequence, defined in alphabet {A,C,G,T} k-letter words makeup the sequence, is common in bioinformatics. The term q-gram is most commonly used in computer science and in text mining [Tomovic et al., 2006]. For example, it is possible to convert the TGACGAC sequence to TGA, GAC, ACG, GAC list of 3 metres (note that these are overlapping). The first step in our approach is therefore to establish a km list of the DNA sequence. Prior studies have shown that the epigenetic status is useful at 3 mers and 4 mers [Pahm et al. 2005, Higashihara et al. 2008, 2005]. In preliminary experiments we have found that k = 4 produces the best performance: the lower K values reduce their accuracy, while the greater the risks of superposition.We used k = 1 only for splice data, to prevent overcasting (see experiments)

In the natural language processing, Word embedding like Glo Ve or words2vec or a single-hot [Goldberg, 2017] is common. Our approach is most appropriate for the latter. Every element in that vector is a word, and a N vector can therefore be used to represent N different words. N different words. - N different words. In all other positions a single-hot vector has one position that corresponds to the word position. A vector length of 4 k is necessary in a DNA sequence to represent all kmers, since this is the length of a series of words (k) from the alphabet {A,C,G,T}.One hot vector, length 4 for all 1 mers can be used, for example. A is the same as [1 0 0 0], C is the same [0 1 0 0], G is the same as [0 0 1 0], and T is the same as [0 0 0 1], for

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 9, September 2021

example. For us, the DNA sequence is a 4-mer list that can be converted into the single-hot vector lists of 44 = 256 each.

Model Architecture

First, as shown in figure 2, we present the picture diagram of DeepD2V. The convolution module is used to extract sequence features, then two completely connected layers and a drop-off layer for preventative purposes. The bi-LSTM module captures high order features.

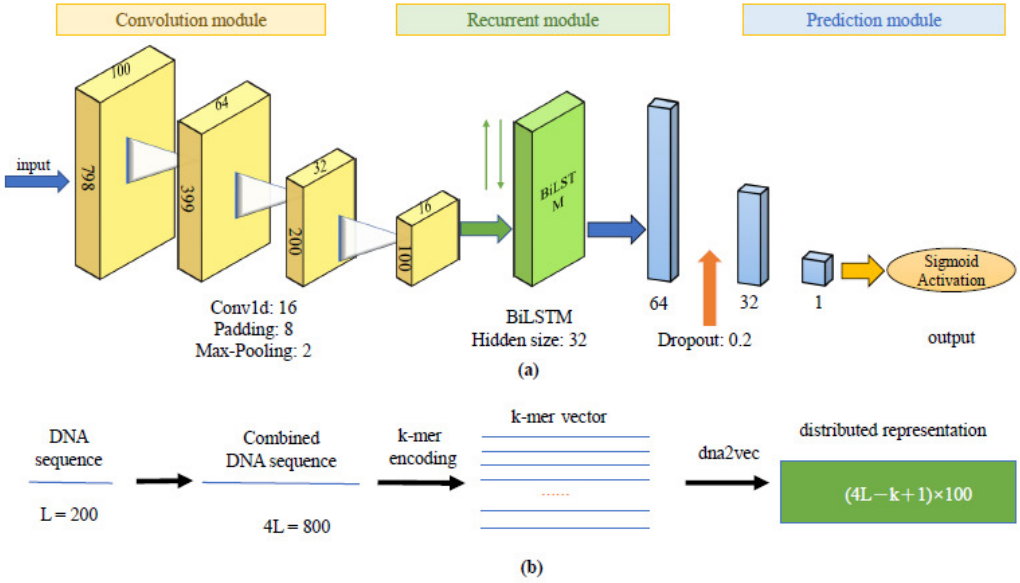


Figure 1. The illustrative diagram of our DeepD2V method for predicting protein–DNA binding. (a) The DeepD2V model includes three modules: convolution modules, recurrent module, and prediction module. Each module consists of four different laies, including convolution layer, rectified linear unit, normalisation and pooling of a series of batches; (b) DeepD2Vtakes a sequence of 200 bp each input and generates three variant types of sequences combined with the encoding and distribution k-mers by a dna2vec algorithm, forming an 800bp sequence.

The function of each module is described in detail as follows. The entire workflowcan be formulated as Equation (1):

$$Y = f_{pred}(f_{rnn}(f_{cnn}(x))) \qquad (1)$$

Compared with an RNN, a one-dimensional convolutional neural network can shortenthe sequence and extract the high-order feature of motifs, but requires a small computationalcost when dealing with long biological sequences. Therefore, we use a onedimensionalconvolution module before RNN. Each convolution module consists of aone-dimensional convolutional layer, a rectified linear layer (ReLU), and a max poolinglayer. A convolution layer is responsible

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 9, September 2021

for capturing motif features with a specifiednumber of kernels of filter. Relu is used as an activation function because it can reducethe gradient descent and back propagation efficiently, avoiding gradient explosion andgradient disappearance. Finally a max bundling layer is used to reduce the redundancy of the convolutionary layer output and to choose the maximum response Filters.

Long-short-term memory (LSTM) is RNN variants that resolve the problems with long-term dependency that traditional RNN can not manage. In consideration of thedouble-stranded structure of the DNA sequence, we selected bi-LSTM to extract long-termfeatures of the sequence. The specific calculation formula of bi-LSTM unit at position t ispresented as follows:

$$
\begin{aligned}
f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f), \\
i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i), \\
c_t &= f_t \odot c_{t-1} + i_t \odot tanh(W_c x_t + U_c h_{t-1} + b_c), \\
o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o), \\
h_t &= o_t \odot tanh(c_t)
\end{aligned}
\qquad (2)
$$

where _ is element-wise multiplication, bo, bc, bi, and b f are the biases, and Wi, Wf , Wo,Uo, Uf , and Ui are the weights.Two fully connected layers and one dropout layer constitute the final predictionmodule to integrate the feature learned from CNN and RNN. Dropout is widely used for regularizationto avoid overfitting by reducing the complex co-adaptation relationship between
neurons in a deep neural network. The dropout ratio is tuned via cross-validation. Finally, an active asigmoid function is used to calculate protein-DNA binding probability.

The initial value of learning rate is set to 0.001 and adjusted with the training epochsdynamically through an Adam optimizer. A DeepD2V model was trained on theworkstation with a Tesla M10 GPU, and training process on 50 data sets finished withinseven hours. The number of sequences simultaneously considered in gradient calculationwas determined by batch size. A large batch size can usually speed up the training processeffectively, but an oversized batch size easily results in a local minimum. Therefore, thebatch size in our study was set to the commonly used value 64.Table 1 summarizes recommendations and starting points for the most common hyper-parameters.

Table 1. DeepD2V Hyper-parameters, search space, and recommendation.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 9, September 2021

Calibration Hyper-Parameters	Search Space	Recommendation
Convolutional layer number	{1, 3, 5, 7}	3
Learning rate	$\{1 \times 10^{-2}, 1 \times 10^{-3}, 1 \times 10^{-4}, 1 \times 10^{-5}\}$	1×10^{-3}
Batch Size	{1, 32, 64, 128, 256, 512}	64
Loss Function	/	Binary cross entropy
Optimizer	{Adam, AdaDelta}	Adam
Convolutional neurons number	{8, 16, 32, 64, 128}	16
Convolutional kernel size	{3, 9, 16, 24}	3
Max Pooling window size	{2, 4, 8}	2
Number of bi-LSTM neurons	{8, 16, 32, 64}	16
Dropout ratio	{0.1, 0.2, 0.5}	0.1

IV. RESULTS

We first assessed the robustness of DeepD2V with different strides and k-mervalues.To verify the performance improvement by combined DNA sequences, the performance ofDeepD2V model with different inputs was compared. Finally, DeepD2V was compared toother simplified models with only CNN, RNN or Bi-LSTM modules alone.

For systematic performance evaluation, we compare DeepD2V with other four stateof-the-art protein–DNA binding methods, including DanQ, DeepBind, WSCNN, andWSCNNLSTM. The performance comparison experiments were conducted on 50 ChIP-seqbenchmark data sets.

As shown in Figure 2, when the value of k increases, the performance of the modelgradually decreases. The model achieves better performance at k = 3. We suppose that thisresult may be attributed to the fact that three bases constitute an amino acid, and k = 3exactly captures the biological essentiality of the sequence. For the subsequent tuning ofstride step s, we set k to 3 and then use grid search for an optimal value of s.Figure 3 shows that the larger the stride, the worse the performance of the model. Thisresult may be due to the large stride leading to a loss of fine-grained sequence information.Previous studies encourage the use of larger values of k-mer length and suggeststhat small stride values (s = 1)may decrease the performance of the embedding algorithm.

However, in our empirical experiments, we find that the model with smaller values ofstride and k-mer length (k = 3 and s = 1) perform better than other combinations of k and s.Therefore, we set the k and s values to 3 and 1 in the following performance comparisonexperiments. Note that the experiment utilizes a raw DNA sequence as input instead ofthe combined sequence.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 9, September 2021

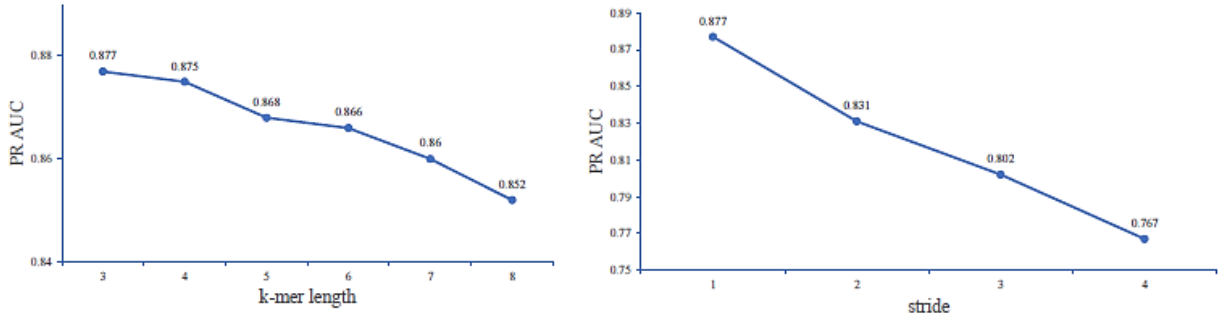
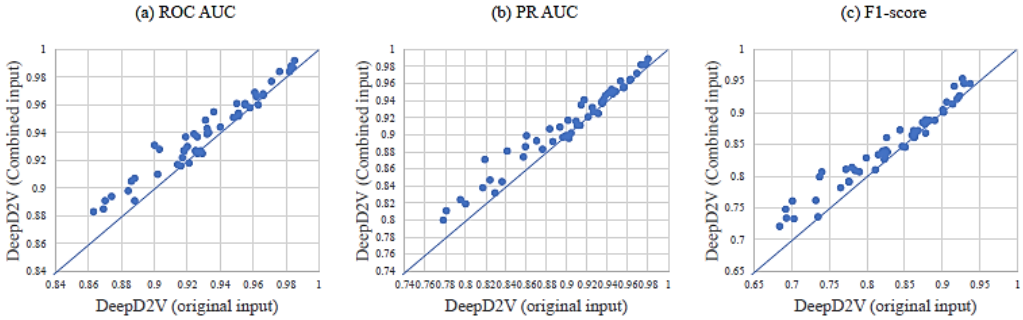


Figure 2. A comparison of DeepD2V with different k-mer length and stride step on the 23 ChIP-seq datasets in a Gm12878

Combined Sequence Outperforms Original DNA Sequence

For double-stranded DNA sequences, reverse or complement sequences are taken intoaccount in the training of DeepD2V because proteins may bind to the DNA complementaryor reverse the sequence strand. Hence, we explore the significance of a combined sequenceand the original DNA sequence again. Figure 6 shows the performance of DeepD2V takingas input the original DNA sequence and combined sequences on all 50 benchmark datasets. It can be found that the combined sequences greatly promote DeepD2V performancecompared to original DNA sequences over all of the three performance measures. We thinkthat the combined sequences potentially capture high-order dependencies among differentpatterns of DNA sequences, such as implicit DNA shape.



Performance Comparison between CNN, RNN, bi-LSTM, and DeepD2V

To verify the performance improvement by the combination of CNN and bi-LSTM,we compare DeepD2V to three simplified models: CNN-only, RNN-only, and the hybridstructure of CNNs and RNNs. Figure 3 displays DeepD2V's performance in all 50 public ChIP-seq data sets with those deeplearning architectures. It should be noted that the four models have been optimised with the specificichyper parameters so that objective performance assessment can be performed.We think CNN performs well when extracting high-order characteristics from

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 9, September 2021

sequences, while bi-LSTM is able to track long-term reliance on motifs, and avoid disappearing gradient and explosive gradient problems while training for DNAlong sequence. DeepD2V achieved better predictor performance with an average F1 score of 0.809 in comparison with CNN and bi-LSTM alone. DeepD2V takes full advantageof the CNN and bi-LSTM model and thus achieves the best performance. It extracts thehigh-order abstract features from the combined DNA sequences by CNN and capturesbidirectional semantic dependencies of sequences with bi-LSTM.

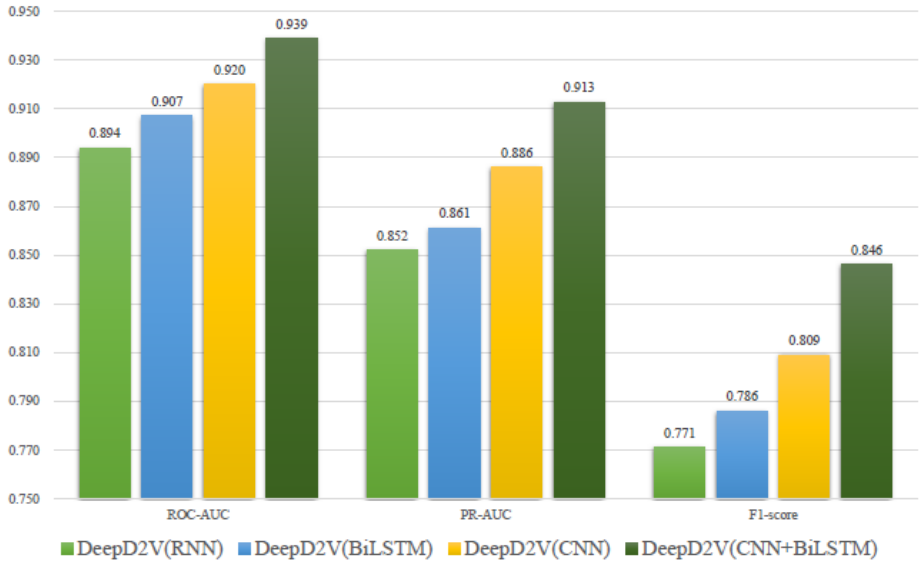


Figure 3. Comparison of DeepD2V with CNN-only, RNN-only, and hybrid structure of CNNs and RNNs on 50 public

V. CONCLUSION

We developed a new framework in this paper based on the binding training in deep protein-DNA. First, the reverse complementary sequence has been taken intoaccount to train the prediction model. We made a hypothesis that proteins may bind tothe inverse sequence or complementary sequence rather than original sequence. Therefore,we consider various combinations of original, reverse, complementary, and reversecomplementary sequences. Our experimental results verified that the combined sequenceis remarkably beneficial for feature extraction. Second, the word embedding algorithmis employed in DeepD2V to train distributed representation of k-mers. In addition,we presented a new framework based on deep learning, which takes the combined DNA sequenceas input and employs both CNN and bi-LSTM to extract high-order and long-termfeatures. Compared with other deep learning methods based on sequence feature alone forprotein–DNA binding prediction, DeepD2V achieved better robustness and performance.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 9, September 2021

REFERENCES

[1] Min, X.; Zeng,W.; Chen, N.; Chen, T.; Jiang, R. Chromatin accessibility prediction via convolutional long short-term memorynetworks with k-mer embedding. Bioinformatics **2017**, 33, i92–i101

[2] Le, N.; Nguyen, Q.H.; Chen, X.; Rahardja, S.; Nguyen, B.P. Classification of adaptor proteins using recurrent neural networks andPSSM profiles. BMC Genom. **2019**, 20, 966.

[3] Le, N.; Nguyen, B.P. Prediction of FMN Binding Sites in Electron Transport Chains based on 2-D CNN and PSSM Profiles.IEEE/ACM Trans. Comput. Biol. Bioinform. **2019**.

[4] Nguyen, Q.H.; Nguyen-Vo, T.H.; Le, N.; Do, T.; Nguyen, B.P. iEnhancer-ECNN: Identifying enhancers and their strength usingensembles of convolutional neural networks. BMC Genom. **2019**, 20, 951.

[5] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need.2017;pp. 5998–6008.

[6] Alipanahi, B., Delong, A., Weirauch, M.T., et al. 2015. Predicting the sequence specificities of DNA-and RNA-bindingproteins by deep learning. Nat. Biotechnol. 33, 831.

[7] Britz, D. 2015. Recurrent neural networks tutorial, Part 1–Introduction to RNNs. Available at: www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns. Accessed January 20, 2018.

[8] Busia, A., Dahl, G.E., Fannjiang, C., et al. 2019. A deep learning approach to pattern recognition for short DNAsequences. bioRxiv 353474.

[9] Zhang S, Zheng D, Hu X, Yang M. Bidirectional long short-term memory networks for relation classification.InProceedings of the 29th Pacific Asia Conference on Language, Information and Computation2015 (pp. 73–78).

[10]Dobzhansky T. Nothing in biology makes sense except in the light of evolution. The americanbiologyteacher. 2013 Feb; 75(2):87–91.

[11] Chollet F. Keras: The python deep learning library[J]. Astrophysics Source Code Library, 2018.

[12] Guo Y, Yu L, Wen Z, Li M. Using support vector machine combined with auto covariance to predict

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 9, September 2021

[13] Goldberg, Y. & Levy, O. word2vec Explained: deriving Mikolov*et al.*'s negative-sampling word-embedding method. *arXiv preprintarXiv:1402.3722* (2014).

[14] Asgari, E. &Mofrad, M. R. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PloSone***10**, e0141287 (2015).

[15] Johnson, J. M. &Khoshgoftaar, T. M. Survey on deep learning with class imbalance. *J. Big Data.* **6**, 1–54. https:// doi. org/ 10. 1186/s40537- 019- 0192-5 (2019).

[16] Kotsiantis, S., Kanellopoulos, D., Pintelas, P. E., Kanellopoulos, D., &Pintelas, P. Handling imbalanced datasets: a review datapreprocessing view project machine learning and data mining view project handling imbalanced datasets: a review (n.d.). https://www. researchgate. net/ publi cation/ 22808 4509. Accessed June 8, 2021.

[17] Greff, K., Srivastava, R. K., Koutnik, J., Steunebrink, B. R. &Schmidhuber, J. LSTM: a search space odyssey. *IEEE Trans. NeuralNetw. Learn. Syst.* https:// doi. org/ 10. 1109/ TNNLS. 2016. 25829 24 (2017).

[18] He, H., & Ma, Y. Imbalanced learning: foundations, algorithms, and applications (2013).

[19]. Kingma, P. E. & Ba, J. L. Adam: a method for stochastic optimization. İn *3rd The International Conference on Learning RepresentationsCLR 2015—Conference on Tracking Proceedings, International Conference on Learning Representations, ICLR* (2015