# DeepDSSR: Deep learning structure for human donor splice sites recognition

Tanvir ALAM[a,1], Mohammad Tariqul ISLAM[b], Mowafa Househ[a], Abdesselam BOUZERDOUM[a,c], Ferdaus Ahmed KAWSAR[d]
[a]*Information and Computing Technology Division, College of Science and Engineering, Hamad Bin Khalifa University (HBKU), Doha, Qatar;*
[b]*Computer Science Department, Southern Connecticut State University, USA;*
[c]*School of Electrical, Computer and Telecommunications Engineering University of Wollongong, Wollongong, NSW, Australia;*
[d]*Department of Computing, East Tennessee State University, USA*

**Abstract.** Human genes often, through alternative splicing of pre-messenger RNAs, produce multiple mRNAs and protein isoforms that may have similar or completely different functions. Identification of splice sites is, therefore, crucial to understand the gene structure and variants of mRNA and protein isoforms produced by the primary RNA transcripts. Although many computational methods have been developed to detect the splice sites in humans, this is still substantially a challenging problem and further improvement of the computational model is still foreseeable. Accordingly, we developed DeepDSSR (deep donor splice site recognizer), a novel deep learning based architecture, for predicting human donor splice sites. The proposed method, built upon publicly available and highly imbalanced benchmark dataset, is comparable with the leading deep learning based methods for detecting human donor splice sites. Performance evaluation metrics show that DeepDSSR outperformed the existing deep learning based methods. Future work will improve the predictive capabilities of our model, and we will build a model for the prediction of acceptor splice sites.

**Keywords:** deep learning; convolution neural network; bidirectional long short-term memory; donor splice sites

## 1. Introduction

More than 90% of mammalian genes are believed to be processed through an alternative splicing mechanism, which is crucial to understand the gene structure and transcript variants [1]. The exon-intron/intron-exon boundaries, where the splicing occurs, are called splice sites (SS), and introns are cut out from the pre-mRNA in the SS region [1]. The SS at the exon-intron boundary is called donor SS (DSS), and a highly conserved dinucleotide of GT is observed at the intron start side. The SS at the intron-exon boundary is called acceptor SS, and a highly conserved dinucleotide of AG is observed at the intron end side. However, SS identified by read aligner is not always reliable as there is a high chance of false mapping of short reads over a large reference genome [2]. Therefore absolutely precise computational model for identifying SS is necessary to identify the accurate gene structure and their transcript variants.

[1] Corresponding Author: Tanvir Alam, Information and Computing Technology Division, College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar. Email: talam@hbku.edu.qa

There are many existing methods that use traditional machine learning methods to predict human DSS [3-6]; they perform reasonably well. Recently deep convolutional neural networks (CNNs) have shown to achieve ground-breaking results in the classification of images [7]. The examples of using CNN for biological problems have increased recently, and in many cases, deep learning (DL) methods have been shown to generate more accurate results than traditional hand-curated feature-based machine learning methods [8]. Based on our literature search, we found three DL based architectures to predict the DSS. Naito developed DSSP, which used CNN and long short-term memory (LSTM) [9] based architecture to predict the DSS [10]. Zhang et al. developed DeepSplice, a CNN based model, to predict the human DSS [11]. Du et al. developed DeepSS, a two-layer CNN based architecture to predict DSS from humans and other organisms [12]. Naito and Zhang et al., both groups have used 1:10 ratio of true:false DSS and Due et al. used a ratio of 1:5 of true:false DSS sequence to report the model performance.

In this study, we have developed a novel deep learning architecture, DeepDSSR (deep donor splice site recognizer), using CNN [7] and bidirectional LSTM (BLSTM) [9] to predict the donor splice sites in human. Considering several metrics of model evaluation, DeepDSSR, outperformed the existing DL based models in predicting human DSS.

## 2. Methods

We collected publicly available human DSS dataset from HS3D [13]. This dataset contains information on 2,796 true DSS and 90,924 false DSS. The length of each sequence is 140 nucleotides and the conserved GT dinucleotide resides at the 71st and 72nd position of each sequence. Since this is an imbalanced dataset, and previously published DL based methods [10-12] used true/false DSS ratios of 1:1, 1:5, or 1:10 for evaluating their models, we also used these three ratios in our evaluation to compare the performance of our model to the existing DL based models.

Our first step was to encode each input sequence using one-hot encoding, where each nucleotide of DNA was represented by a vector of length four, A:(1,0,0,0), C:(0,1,0,0), G:(0,0,1,0) and T:(0,0,0,1). These two-dimensional encoded sequences were then used as an input to the DeepDSSR architecture (Figure 1) for training and validation.

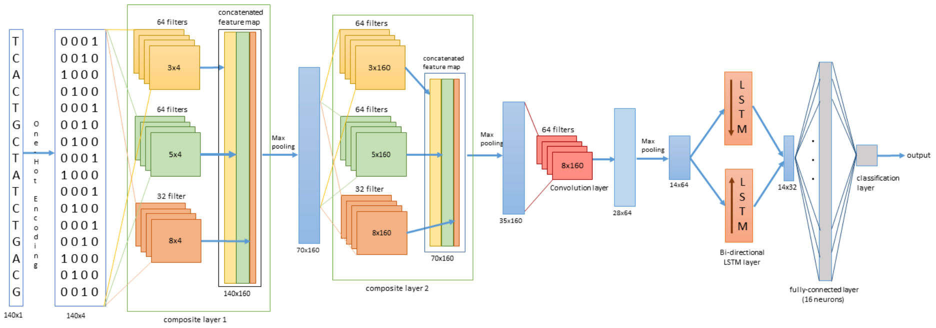

**Figure 1:** Proposed architecture for DeepDSSR

Our proposed network architecture consists of four layers, listed here in the order they are applied to the input: a pair of Inception-like [14] layers (hereafter referred to as composite layers), one conventional convolutional layer, a BLSTM, and two fully connected layers which include the output layer. Each composite layer consists of three, parallel one-dimensional convolution layers with a filter size of 3, 5, and 8, respectively. We used 64 kernels of size 3 and 5 and 32 of size 8.

We used two composite layers one after another with dropout and max-pooling layers for dimensionality reduction and regularization, respectively. After the two composite layers, a regular one-dimensional convolution layer with 64 kernels of size 8 was added into the network. We then add a dropout and a max-pooling layer. The next module, BLSTM, analyzes the output of this convolutional layer from two directions to discover left-to-right and right-to-left patterns that influence the label of this DNA sequence. The last two layers of the network are fully connected layers of size 16 and 1, respectively. The final layer is for classification: true DSS (1) and false DSS (0) predictions. The first fully-connected layer also has a dropout layer after it for regularization. We used ReLU as the activation function for the inner layers and sigmoid for the classification layer. The dropout rates for the two composite layers, the convolution layer, BLSTM layer, and the first fully-connected layer, are 0.4, 0.5, 0.6, 0.25, and 0.70, respectively. The max-pooling layers we used were one-dimensional and had a kernel size of 2.

We trained our model using the Adam [15] optimization algorithm and binary cross-entropy loss. We used the default values for the β1 (0.9) and β2 (0.999) parameters for the optimizer and used stochastic gradient descent with warm restart [16] as the learning rate scheduler with a minimum and maximum learning rate of 3x10$^{-5}$ and 10$^{-3}$, respectively, and a cycle length of 5 epochs. We used a batch size of 512.

We used Keras for implementing the DeepDSSR. The model was trained for 300 epochs on a single GPU machine having a GeForce GTX TitanX (Pascal). Each epoch took between 2 to 20 seconds to complete, depending on the version of the dataset (1:1, 1:5, or 1:10).

## 3. Results & Discussions

In the training process of DeepDSSR, we considered the true DSS and false DSS as the positive and negative class, respectively. To evaluate the performance of the model we used the following three evaluation metrics that were considered for most of the existing DL based models, Sensitivity (Sn) = (TP)/(TP+FN), Specificity (Sp)= (TN)/(FP+TN), and Matthew's Correlation coefficient (MCC):

$$MCC = (TP * TN - FP * FN) / \sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)},$$

where TP, FN, FP, and TN stand for true positive, false negative, false positive, and true negative, respectively. The performance of the model was evaluated using 10-fold cross-validation (see Table 1).

**Table 1.** Performance of DeepDSSR and other existing DL based tools for human DSS prediction. *: Since the value of these metrics were not explicitly mentioned in the article [12], the values shown here are approximated through visual inspection from "Figure 4" of the corresponding article. NA: Not available in the literature

| Model (data ratio) | Sensitivity | Specificity | MCC |
|---|---|---|---|
| DeepSplice (1:1) | NA | NA | NA |
| DSSP (1:1) | 97.88 | 95.36 | 93.33 |
| DeepSS* (1:1) | 97.50 | 92.50 | 91.00 |
| **DeepDSSR (1:1)** | **97.50** | **96.42** | **93.93** |
| DeepSS* (1:5) | 95.50 | 97.50 | 90.50 |
| **DeepDSSR (1:5)** | 93.57 | **98.21** | **90.88** |
| DeepSplice (1:10) | 95.71 | 93.76 | NA |
| DSSP (1:10) | 90.31 | 98.75 | 87.99 |
| DeepSS* (1:10) | NA | NA | NA |
| **DeepDSSR (1:10)** | 91.43 | **98.85** | **89.15** |

From Table 1, we can observe that for the dataset with 1:1, 1:5 and 1:10 ratios, our model achieved 93.33, 90.88, 89.15 MCC, respectively, and it outperformed, considering MCC as an evaluation metric, all the existing DL based models [10-12] for all three data sets subsampled at the same ratio. Additionally, for 1:1 dataset, DeepDSSR outperformed all the existing methods in terms of all three evaluation metrics (MCC, Sn and Sp). For 1:5 and 1:10 ratio datasets, DeepDSSR achieved a Sp that surpasses all the existing methods' corresponding metric but at the cost of ~2% Sn for the 1:5 dataset. For the 1:10 dataset, our model outperformed DSSP in terms of Sn.

## 4. Conclusion

The paper introduced a new deep learning architecture, namely DeepDSSR, for the prediction of human donor splice sites. Experimental results were presented, which show that DeepDSSR outperforms existing DL models in terms of MCC and sensitivity. Future work will focus on improving the performance of DeepDSSR and building a new model for the prediction of acceptor splice sites.

## References

1.Wang, E.T., et al., *Alternative isoform regulation in human tissue transcriptomes.* Nature, 2008. **456**(7221): p. 470.

2.Nellore, A., et al., *Rail-RNA: scalable analysis of RNA-seq splicing and coverage.* Bioinformatics, 2016. **33**(24): p. 4033-4040.

3.Meher, P.K., T.K. Sahu, and A.R. Rao, *Prediction of donor splice sites using random forest with a new sequence encoding approach.* BioData mining, 2016. **9**(1): p. 4.

4.Wei, D., et al., *A novel splice site prediction method using support vector machine.* Journal of Computational Information Systems, 2013. **9**(20): p. 8053-8060.

5.Meher, P.K., et al., *A statistical approach for 5' splice site prediction using short sequence motifs and without encoding sequence data.* BMC bioinformatics, 2014. **15**(1): p. 362.

6.Xu, Z.-C., et al., *iSS-PC: Identifying Splicing Sites via Physical-Chemical Properties Using Deep Sparse Auto-Encoder.* Scientific reports, 2017. **7**(1): p. 8222.

7.Krizhevsky, A., I. Sutskever, and G.E. Hinton. *Imagenet classification with deep convolutional neural networks*. in *Advances in neural information processing systems*. 2012.

8.Ching, T., et al., *Opportunities and obstacles for deep learning in biology and medicine.* J R Soc Interface, 2018. **15**(141).

9.Hochreiter, S. and J. Schmidhuber, *Long short-term memory.* Neural computation, 1997. **9**(8): p. 1735-1780.

10.Naito, T., *Human Splice-Site Prediction with Deep Neural Networks.* Journal of Computational Biology, 2018. **25**(8): p. 954-961.

11.Zhang, Y., et al., *Discerning novel splice junctions derived from RNA-seq alignment: a deep learning approach.* BMC genomics, 2018. **19**(1): p. 971.

12.Du, X., et al., *DeepSS: Exploring Splice Site Motif Through Convolutional Neural Network Directly From DNA Sequence.* IEEE Access, 2018. **6**: p. 32958-32978.

13.Rampone, S., *HS3D: Homo Sapiens Splice Site Data Set.* Nucleic Acids Research, 2003.

14.Szegedy, C., et al. *Going deeper with convolutions*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

15.Kingma, D.P. and J.L. Ba, *ADAM: A method for stochastic optimization.*

16.Loshchilov, I. and F. Hutter, *SGDR: Stochastic gradient descent with warm restarts.* Learning. **10**: p. 3.