

Residual Neural Network for Predicting Super-enhancers on Genome Scale

Sara Sabba · Amina Hamrelaine ·
Maroua Smara · Mehdi Benhacine

Received: March 30th, 2021/ Accepted: date

Abstract Residual neural network (ResNet) is a Deep Learning model introduced by He et al. [13] in 2015 to enhance traditional Convolutional neural networks for computer vision problems. It uses skip connections over some layer blocks to avoid vanishing gradient problem. Currently, many researches are focused to test and prove the efficiency of the ResNet on different domains such as genomics.

In this paper, we propose a new ResNet model for predicting super-enhancers on genome scale. In fact, the prediction of super-enhancers (SEs) has prominent roles in biological and pathological processes; especially that related to the detection and progression of tumors. The obtained results are very promising and they proved the performance of our proposal compared to the CNN results.

Keywords Deep Learning · Residual Neural Network · Convolutional Neural Network · Bioinformatics · Transcriptional dysregulation · Super-Enhancers · Oncogene · Cancer

1 Introduction

Transcription factors are proteins that bind DNA regulatory elements of genes called enhancers. They play critical roles in the control of cell type-specific gene expression programs [8,16,30]. Super-enhancers (SEs) are clusters of enhancers. They are formed by binding of high levels of enhancer-associated chromatin features that drive high level expression of genes encoding key regulators of cell identity [18,28].

The identification of SEs is based on the differences in their ability to bind markers of promoter transcriptional activity [30], including cofactors such as

Address(es) of author(s) should be given

mediators (MED1, MED12) and cohesions (Nipbl, Smc1), histone modification markers (H3K27ac, H3K4me1, H3K4me3, H3K9me3), chromatin regulators (Brg1, Brd4, Chd7), chromatin molecules (p300, CBP) and many additional transcription factors (Nr5a2, Prdm14, Tcfcp2l1, Smad3, Stat3 and Tcf3)[26,27].

Recently, many studies [24,29,30] proved that gene transcriptional dysregulation is one of the core tenets of cancer development that involves in noncoding regulatory elements, such as TFs, promoters, enhancers, SEs, and RNA polymerase II (Pol II). In particular, SEs play core roles in promoting oncogenic transcription to accelerate cancer development [4,30]. Recent research showed that cancer cells acquire super-enhancers at oncogene and cancerous phenotype relies on these abnormal transcription propelled by SEs [14,25]. Accordingly, it is important to understand super-enhancers and their components since they control much disease-associated sequence variation occurs in these regulatory elements [12,16,21] large amounts of data in order to better understand biological processes.

In fact, the massive evolution of biological data implies the need and necessity to develop new techniques and tools to classify and benefit from them. In this context, Deep Learning is actually an extremely active research area in Machine Learning and bioinformatics [7,22]. Its algorithms proved their efficiency in many critical life situations. They allow predicting many diseases, treatments and biological phenomena from the analysis and interpretation of various types of data [1,9,10,23].

Many bioinformatics frameworks based on Machine Learning were developed in the literature to solve genomics problems. [33] proposed DeepEnhancer framework for predicting enhancers using convolutional neural networks (CNN). [34] developed a Deep Learningbased algorithmic framework, called DeepSEA to predict the noncoding-variant effects de novo from sequence. [3] used also deep convolutional neural networks to develop DeepBind approach for predicting the sequence specificities of DNA- and RNA-binding proteins. Likewise, SpliceFinder [32] and Splice2Deep [2] were designed to predict splice sites of human genomic using CNN model. The both works are trained and validated on some genomic sequences such as *Homo sapiens*, *Oryza sativa japonica*, *Mus musculus*, *Drosophila melanogaster*, and *Daniorerio*. In fact, there are so many critical frameworks worthy of our interest that we cannot cite them all.

Otherwise, there are few bioinformatics works based on Machine Learning proposed to predict super-enhancers of the genomes. [19] implemented and compared six different Machine Learning models to identify key features of SEs and to investigate their relative contribution in the prediction. The six models include: Random Forest, Support Vector Machine, k-Nearest Neighbor, Adaptive Boosting, Naive Bayes, and Decision Tree. To validate their idea, they used 10-fold stratified cross-validation, independent datasets in four human cell-types and a set of publicly available data. [5] proposed a new computational method called DEEPSEN for predicting super-enhancers based on convolutional neural network. The proposed method is trained and tested on

36 SEs features, where 32 ones are used by [19] and 4 others are selected from ChIP-seq and DNase-seq datasets.

In this paper, we propose a new solution for predicting super-enhancers on genome scale, based on supervised Deep Learning technique. Our proposal, called ResSEN, predicts super-enhancers using residual neural networks (ResNet) model. The work aims to improve the results obtained by DEEPSEN method. In fact, there are three reasons behind this motivation: (i) first, ResNet was created to optimize the performance of CNN for avoiding vanishing gradient problem, (ii) second, to the best of our knowledge, there is no approach that uses ResNet for predicting super-enhancers has been proposed in the literature [13,15,31], and (iii) third, the obtained results proved the performance of our proposal compared to the DEEPSEN results.

2 ResSEN: Residual Neural Network for predicting oncogenic super-enhancers

2.1 Data sets

The public database used to train and test our approach is used in previous works of [19] and [5]. In fact, there are 36 features (see Table 1) incorporate publicly available ChIP-seq and DNase-seq datasets of mouse embryonic stem cells (mESC) taken from Gene Expression Omnibus (GEO).

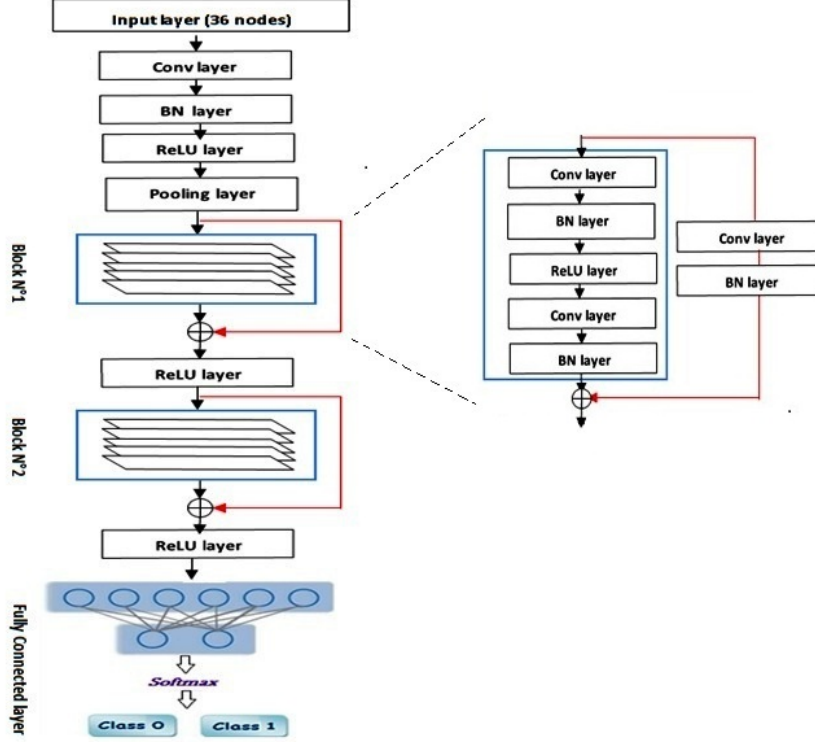
Table 1 Features of datasets used by [5] and our approach.

Super-enhancers data type	Features
Histone modifications	H3K27ac, H3K4me1, H3K4me3, H3K9me3
DNA hypersensitive site	DNaseI
RNA polymeraseII	Pol II
Transcriptional co-activating proteins	p300, CBP
P-TFEB subunit	Cdk9
Sub-units of Mediator complex	Med12, Cdk8
Chromatin regulators	Brg1, Brd4 and Chd7
Cohesin	Smc1, Nipbl
Subunits of Lsd1-NuRD complex	Lsd1, Mi2b
Histone deacetylase	H-DAC2, HDAC
Transcription factors	Oct4, Sox2, Nanog, Esrrb, Klf4, Tcfcp2l1, Prdm14, Nr5a2, Smad3, Stat3, Tcf3
Sequence signatures	AT content, GC content, phastCons, phastConsP, repeat fraction

The datasets contain 11100 samples. Among them, 1119 are positive and 9981 are negative. To train, test and compare our ResSEN approach, we divided those samples into training datasets and test datasets. Where 90% (i.e. 9990) are used for training and 10% (i.e. 1110) are used for performance testing (see Table 2).

Table 2 Division of samples.

Data sets	Samples size	Positive samples	Negative samples
Training data sets	9990	1006	8984
Test data sets	1110	113	997

**Fig. 1** Global architecture of ResSEN

2.2 ResSEN architecture

ResSEN architecture is composed of an input layer, a convolution layer, a pooling layer, two residual blocks and a fully connected layer.

2.2.1 Input data

Thirty six (36) characteristics are used to predict the super-enhancers (see Table 1). So, there are 36 nodes in the input layer. The values of these nodes are normalized and standardized before they are transmitted to the next network layers.

2.2.2 Convolutional layers

ResSEN is composed of six convolutional layers: i) a convolutional layer before the first residual block, ii) two convolution layers for each residual block (2×4), and iii) the final fully connected layer (In deep-learning the FC layer is considered as Conv layer).

In the first convolutional layer we applied 64 filters of size 17, followed by Max-pooling with pool-size 13 and stride 1. The first residual block has two convolutional layers, we applied 128 filters of size 13 in the first one, and 256 filters of the same size 13 in the second one. The second residual block has also two convolutional layers. In the first layer, we applied 256 filters of size 13, while in the second layer we applied 512 filters of the same size 13.

Fig. 2 illustrates the filters parameters of the five convolution layers. The last convolutional layer (FC layer) is detailed in section 2.2.6.

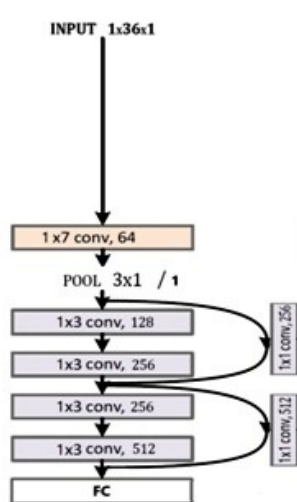


Fig. 2 Convolutional layers parameters of ResSEN

2.2.3 Batch normalization layer

Each convolutional layer is followed by a Batch Normalization layer (see Fig.1). Batch Normalization (BN) is a technique that was introduced to improve the speed, performance, and stability of deep neural networks [11]. It is used to automatically normalize the input layer. Each input xi is normalized using the mean and the variance of the input sample.

2.2.4 Activation layer

In ResNet, a non-linear activation function is generally used after each BN layer, to ensure the non-linearity of the model [11]. In ResSEN, we used ReLU (rectified linear unit) activation function:

$$ReLU(x) = \max(0, x) \quad (1)$$

In fact, ReLU function has become the default activation function for ResNet allowing model to train easier and faster and perform better [8].

2.2.5 Add identity

For each residual block, ResSEN uses convolution block strategy to add the blocks input to the blocks output. This type of design requires that the block's output and its input have the same shape (size), so they can be added together. The output of the first block will be the input of the second and the output of the second block will be the input of the fully connected layer. To transform the blocks input into the desired shape, we introduced 256 convolutions (256 filters) of size 13 for the first residual block and 512 convolutions (512 filters) of size 13 for the second residual block (see Fig.2).

2.2.6 Fully connected layer

The fully connected (FC) layer of the ResSEN is structured as follow:

- The number of input neurons is 17408;
- The activation function is ReLU;
- The number of output layer is 2 neurons;
- The function used to calculate the probability of the output classes is: Softmax.

$$Softmax(x_j) = \max \frac{e^{x_i}}{\sum_j e^{x_i}}, j \in \{1, 2, \dots, k\} \quad (2)$$

Where, k is the number of classes. Moreover, to obtain the predicted class A , we applied the *argmax* function to the *Softmax* function output:

$$A = \text{argmax}(Softmax(x_j)) \quad (3)$$

- So, if $A = 1$: the predicted class is positive, that means, the presence of the super-enhancers in the genome;
- if $A = 0$: the predicted class is negative, that means, the absence of super-amplifiers in the genome.

2.2.7 ResSEN training

ResSEN training is based on supervised learning, which consists of calculating the optimal weights (weights) using the input matrix D (the data samples) and the output matrix A (the desired output or the class label) corresponding to D. D is a matrix of size N36 and A is a binary matrix of size N1, where N is the number of samples which is set to 11100. $A[i] = 1$ if the corresponding sample represents the super-enhancer class, otherwise, $A[i] = 0$.

During the training phase, ResSEN uses the cross entropy loss function to measure the difference between the calculated output and the desired output, and Backpropagation model and Adam method [20] to update network weight's.

3 Experiment results and comparison

In the context of binary classification, the evaluation of models performance is based on some performance measures that are computed from the confusion matrix. Thus, to evaluate and compare the ResSEN performance, we calculated four measures: accuracy, recall, precision and F1-score.

Table 3 Confusion matrix.

		Actuel class	
		-	+
Predicted class	+	True Negatives	False Negatives
	-	False Positives	True Positives

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN},$$

$$Recall = \frac{TP}{TP+FN}, Precision = \frac{TP}{TP+FP}$$

$$F1Score = 2 * Precision * Recall / (Precision + Recall)$$

The best results obtained by testing the best model of ResSEN on the test samples are presented in Table 4.

Table 4 ResSEN results.

Precision	Recall	F1-Score	Accuracy
94,79%	94,94%	94,85%	94,95%

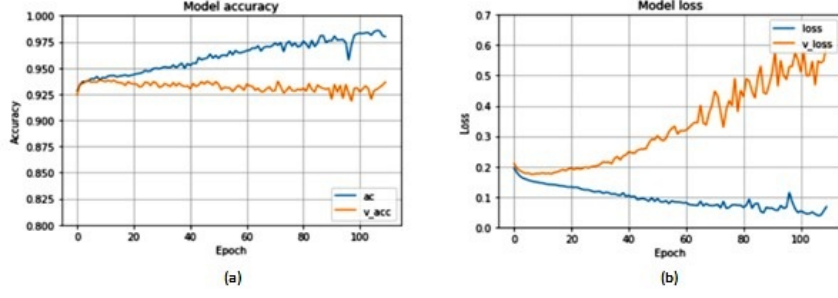


Fig. 3 (a) DeepSEN accuracy curve, (b) DeepSEN loss curve.

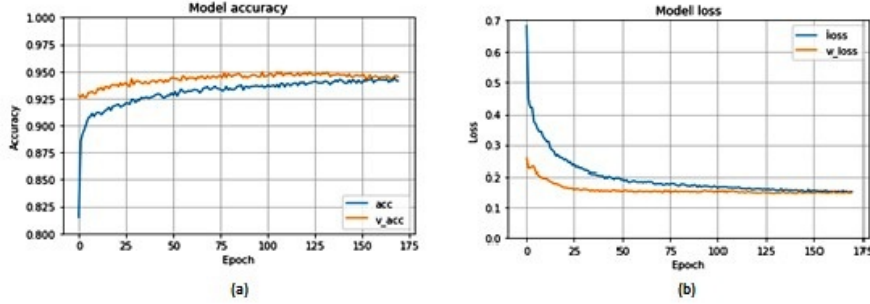


Fig. 4 (a) ResSEN accuracy curve, (b) ResSEN loss curve.

4 Comparison and discussion

In order to ensure a fair comparison with our ResSEN algorithm, we re-executed the DeepSEN [6] using 90% of samples for training and 10% of samples for testing. The obtained results (see Fig.5), show that the best model of DeepSEN achieves an accuracy of 93,64% and a precision of 90%. However, in both cases (validation with all the data sets or with 10% of samples) we noticed the presence of the overfitting problem. The latter is clearly modeled in the accuracy and loss curves that we generated after re-executed DeepSEN (see Fig.3). Knowing that, the blue and the orange curves represent the development of accuracy / loss in the training phase and in the testing phase respectively.

Fig. 4 shows the accuracy and loss curves of ResSEN model. In this case, there is no overfitting problem. We notice a harmonization between the curves generated in training and test phases.

Finally, the comparison graph (see Fig. 5) shows that our proposed model outperforms that of DeepSEN.

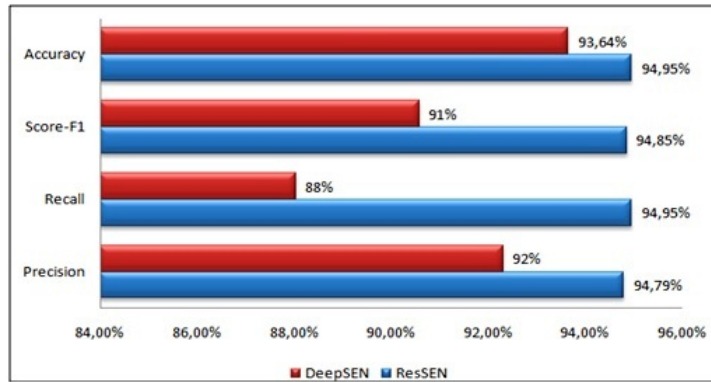


Fig. 5 Performance comparison graph of ResSEN and DeepSEN.

5 Conclusion

In this paper, we presented our prediction approach for detecting super-enhancers in human genomes, called ResSEN. This proposal is based on the ResNet Deep Learning technique aiming at improving the results of existing approaches.

The ResSEN was evaluated using 36 features of mESC datasets taken from Gene Expression Omnibus (GEO). The obtained results were compared with that obtained by DeepSEN approach which is based on CNN architecture. Comparison shows that our proposed ResSEN outperforms DeepSEN, and proves the effectiveness of ResNet architecture as a classifier of genomic problems.

References

- [1] Alazab, M. et al. (2020). COVID-19 Prediction and Detection Using Deep Learning. International Journal of Computer Information Systems and Industrial Management Applications, Vol. 12, pp. 168-181.
- [2] Albaradei, S. et al. (2020). Splice2Deep: An ensemble of deep convolutional neural networks for improved splice site prediction in genomic DNA. Gene X, Vol. 5.
- [3] Alipanahi, B., DeLong, A., Weirauch, M. and Frey, B.J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature Biotechnol, Vol. 33, pp. 831838.
- [4] Bradner, J. E., Hnisz, D. and Young, R. A. (2017). Transcriptional addiction in cancer. Cell, Vol. 168, pp. 629643.
- [5] Bu, H., Hao, J., Gan, Y. et al. (2019). DEEPSEN: a convolutional neural network based method for super-enhancer prediction. BMC Bioinformatics, Vol. 20.
- [6] Bu, H., Hao, J., Gan, Y. et al. (2019). DEEPSEN code: <https://github.com/1991Troy/DEEPSEN>.
- [7] Cao, Y., Geddes, T., Yang, J., and Yang, P. (2020). Ensemble deep learning in bioinformatics. Nature Machine Intelligence, pp.1-9.
- [8] Chen, S., Jia, Q., Tan, Y., Li, Y. and Tang, F. (2020). Oncogenic super-enhancer formation in tumorigenesis and its molecular mechanisms. Experimental qnd Molecular Medicine, Vol. 52, pp. 713723.
- [9] Ching, T. et al., (2018). Opportunities and obstacles for deep learning in biology and medicine. Journal of the Royal Society Interface, Vol. 15, No. 141.

- [10] Esteva, A. et al. (2019). A guide to deep learning in healthcare. *Nature Medecine*. Vol. 25, No.1.
- [11] Furusho, Y., Ikeda, K. (2019). ResNet and Batch-normalization Improve Data Separability. *Proceedings of Machine Learning Research*, Vol. 101, pp. 94-108.
- [12] Grossman, S.R. et al. (2013). Identifying recent adaptations in large-scale genomic data. *Cell*, Vol. 152, pp. 703713.
- [13] He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, pp. 770-778.
- [14] He, Y., Long, W. and Liu, Q. (2019). Targeting Super-Enhancers as a Therapeutic Strategy for Cancer Treatment. *Frontiers in Pharmacology*, Vol. 10.
- [15] Alzantot, M., Wang, Z., Srivastava, M. (2019). Deep Residual Neural Networks for Audio Spoofing Detection. *arXiv:1907.00501v1*.
- [16] Hnisz, D. et al. (2013). Super-enhancers in the control of cell identity and disease. *Cell*, Vol. 155, No. 4, pp. 934947.
- [17] Huang, J., Li, K., Cai, W., Liu, X., Zhang, Y., Orkin, S.H, Xu, J. and Yuan, G-C. (2018). Dissecting super-enhancer hierarchy based on chromatin interactions. *Nature Communications*, Vol. 9, No. 943.
- [18] Khan, A., Zhang, X. (2019). Integrative modeling reveals key chromatin and sequence signatures predicting super-enhancers. *Scientific Reports*, Vol. 9.
- [19] Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.
- [20] Lee, T.I. and Young, R.A. (2013). Transcriptional regulation and its misregulation in disease. *Cell*, Vol. 152, pp. 12371251.
- [21] Li, Y., Huang, C., Ding, L., Li, Z., Pan, Y. and Gao, X. (2019). Deep learning in bioinformatics: Introduction, application, and perspective in the big data era. *Methods*, Vol. 166, pp. 4-21.
- [22] Litjens, G. et al. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, Vol. 42, pp. 6088.
- [23] Lu, J., Li, Y., Wu, Y., Zhou, S., Duan, C., Dong, Z., Kang, T. and Tang, F. (2018). MICAL2 mediates p53 ubiquitin degradation through oxidating p53 methionine 40 and 160 and promotes colorectal cancer malignance. *Theranostics*, Vol. 8, No. 19, pp. 52895306.
- [24] Mansour, M. R. et al. (2014). Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science (New York, N.Y.)*, Vol. 346, No. 6215, pp. 13731377.
- [25] Ng, H.H., Surani, M.A. (2011). The transcriptional and signalling networks of pluripotency. *Nature cell biology*, Vol. 13, pp.490496.
- [26] Orkin, S.H., Hochedlinger, K. (2011). Chromatin connections to pluripotency and cellular reprogramming. *Cell*, Vol. 145, pp. 835850.
- [27] Qu, J., Ouyang, Z., Wu, W., Li, G., Wang, J., Lu, Q. and Li, Z. (2020). Functions and Clinical Significance of Super-Enhancers in Bone-Related Diseases. *Frontiers in Cell and Developmental Biology*, Vol. 8.
- [28] Sengupta, S. and George, R. E. (2017). Super-enhancer-driven transcriptional dependencies in cancer. *Trends Cancer*, Vol. 3, pp. 269281.
- [29] Tang, F., Yang, Z., Tan, Y. and Li, Y. (2020). Super-enhancer function and its application in cancer targeted therapy. *NPJ Precision Oncology*, Vol. 4, No. 2.
- [30] Tang, R. and Lin, J. (2018). Deep Residual Learning for Small-Footprint Keyword Spotting. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, pp. 5484-5488.
- [31] Wang, R., Wang, Z., Wang, J. and Li, S. (2019). SpliceFinder: ab initio prediction of splice sites using convolutional neural network. *BMC Bioinformatics*, Vol. 20.
- [32] Xu, M., Ning, C., Ting, C. and Rui, J. (2016). DeepEnhancer: Predicting enhancers by convolutional neural networks. *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Shenzhen, pp. 637-644.
- [33] Zhou, J. and Troyanskaya, O. (2015). Predicting effects of noncoding variants with deep learningbased sequence model. *Nature Methods*, Vol. 12, pp. 931934.