A hybrid multi-scale CNN-LSTM deep learning model for the identification of protein-coding regions in DNA sequences

A. Darvish, S. Shamekhi*

Faculty of Biomedical Engineering, Sahand University of Technology, Tabriz, Iran
E-mail addresses: a_darvish98@sut.ac.ir, shamekhi@sut.ac.ir
*Corresponding author: Tel: +984133458455

Received day month year, Revised day month year, Accepted day month year.

Abstract
Identification of the exact location of an exon in a DNA sequence is an important research area of bioinformatics. The main issues of the previous signal processing techniques are accuracy and robustness for the exact locating of exons. To address the mentioned issues, in this study, a method has been proposed based on deep learning. The proposed method includes a new preprocessing, a new mapping method, and a multi-scale modified and hybrid deep neural network. The proposed preprocessing method enriches the network to accept and encode genes at any length in a new mapping method. The proposed multi-scale deep neural network uses a combination of an embedding layer, a modified CNN, and an LSTM network. In this study, HMR195, BG570, and F56F11.4 datasets have been used to compare this work with previous studies. The accuracies of the proposed method have been 0.982, 0.966, and 0.965 on HMR195, BG570, and F56F11.4 databases, respectively. The results reveal the superiority and effectiveness of the proposed hybrid multi-scale CNN-LSTM network.

Keywords
Deep learning, DNA sequences, CNN, LSTM, Multi-scale, Protein coding region.

1. Introduction

Main Deoxyribonucleic Acid (DNA) is considered the main ingredient for the growth and transmission of heritability in living species [1]. Deoxyribonucleic Acid (DNA) is considered the main ingredient for the growth and transmission of heritability in living species [1]. A sequence of DNA is a long molecule based on biopolymer origin, which carries genetic information. DNA consists of two strands of linear polymer sequences. Each strand has a backbone consisting of alternating sugar (deoxyribose) and phosphate groups. One of four bases, adenine (A), cytosine (C), guanine (G), and thymine (T), is attached to each sugar [2, 3], and the coded information inside DNA consists of these four bases. Discovering a relationship between a protein structure and its function based on this information is an important research field [4]. For understanding this relationship, it is necessary to determine the exact locations of two regions of DNA sequences, protein-coding regions (exons) and non-coding regions (introns) [5-7]. The first step in introns and exons' separation is encoding the DNA sequence as a digital signal. For this purpose, different schemes, such as binary mapping [8], real numbers [9, 10] EIIP [11] QPSK-base [12, 13], fuzzy semantic similarity measure (FSSM) [14], and high-level structural information of physical properties of DNA molecule [15] have been proposed by researchers. A common drawback of these mapping methods, except for the FSSM, is the use of fixed numbers that cannot present all features of DNA sequences and interactions between their Constituent bases. It demonstrates the importance of a new generalized mapping method.

The next step in most of the recent exon detection algorithms is feature extraction of the mapped digital signal. In the last years of the previous century, Herzel et al. showed that there are hidden 3-, 10.5-, 200-, and 400-base periodicities in DNA sequences [16, 17]. Therefore, in most subsequent studies, the separation of introns and exons has been performed based on the period-3 property as a feature. These studies have extracted this feature using time-frequency methods such as Short-Time Fourier Transform (STFT) [18-20]. Time-frequency methods face with trade-off problem between the frequency and the spatial resolution. Also, the mapping method and the window size of STFT affect the shape of peaks and signal-to-noise ratio (SNR). As shown in Fig. 1, we have processed an example of these signals. The amplitude of the noise of this signal is high, which leads to unclear boundaries of regions.

In classic machine learning, such as the method described above, rules/features are designed/selected by humans. These rules/features are not complete and comprehensive. Surly, better rules could make more accurate results. Recently, deep learning, which is a subset of machine learning, has been used as the primary approach to overcome feature extraction and generalization problems [21]. Deep learning can automatically learn complex features of data by a mixture of simple features without

involving hand-coded rules. On the other hand, the accuracy of deep learning is usually more than the other machine learning methods [21]. To the best of our knowledge, most research in bioinformatics with deep learning methods has been done in genomic medicine and the medical imaging field [22-27], and we cannot find any published outstanding deep learning work on protein-coding region detection in DNA sequences.

In this study, we have proposed a new algorithm for identifying the location of exons based on a novel embedded encoding method with a hybrid multi-scale deep neural network. The proposed hybrid model consists of two deep networks with new architectures, a modified convolutional neural network (CNN), and a long short-term memory (LSTM) network. The main contributions of this work are:

- Using the neural network approach to detecting exons is impossible because of the different lengths of the genes. Our insight key is a preprocessing method that converts the input gene with any length to a constant length.
- We have proposed a new hybrid architecture to classify exons-introns. The proposed method can encounter the generalization challenge by using a considerable number of data. In contrast with the previous studies that usually extract just one feature (period-3 property), the proposed architecture consists of a multi-scale feature extractor perspective.

Also, we have proposed and employed a specific embedding layer to map DNA sequences and extract semantic relationships between nucleotides.

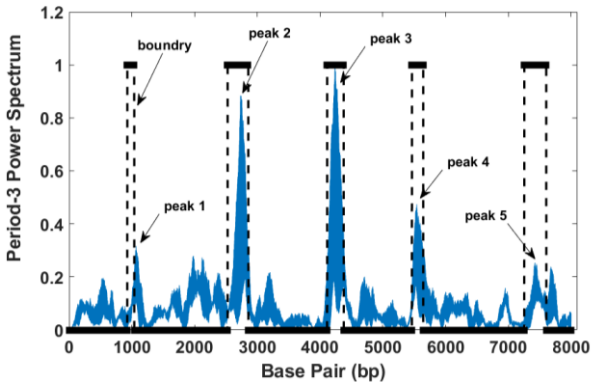

Fig.1. Period-3 extracted from the DNA signal using the STFT

2. Materials and methods

2.1. Dataset

In this study, we have used the HMR195 [28, 29], F56F11.4 [30], and BG570 [31] datasets for training and validating purposes. HMR195 contained 195 genes with single and multi-exon. In this dataset, the number of human, rat, and mouse sequences are 103, 10, and 82, respectively. The mean length of sequences is 7096 base pairs (BP), and the number of multi-exon and single-exon genes are 152 and 43, respectively. In this dataset, the relative proportions of coding, intronic, and intergenic regions of DNA sequences are 14%, 46%, and 40%, respectively. To balance the number of data in each class, we removed some non-coding areas. However, original imbalanced data have been used to test the performance of the proposed method.F56F11.4 is a protein-coding gene with a GenBank access number of AF099922 and a total length of 8000 BP. BG570 is a test dataset consisting of 570 single gene sequences prepared by Moisés Burset and Roderic Guigo.

2.2. Proposed Model

The proposed algorithm has two main steps. First, we have applied a set of pre-processing methods to the input dataset. Second, a new deep neural architecture has been proposed for classifying the regions of sequences. The proposed neural network consists of a modified CNN layer followed by an LSTM network. The computer code package of the proposed method is available on request from the corresponding author.

2.2.1 Proposed mapping method

Before processing and classifying the DNA sequences, we perform some proposed mapping operations to convert a gene with any length to a numerical matrix. This algorithm consists of five stages:

1. Reading the gene's information, and removing white spaces.

There is a lot of information for genes in the used datasets. One of them is the coding sequence of DNA (CSD) that is specified the coding regions of genes. In this study, we used CSD as a label and, the information that comes after Origin as a sequence.

2. Zero padding with a size of 400.

This Stage is to equalize the length of all characters.

3. Windowing with the length of 801, and slicing each sequence to subsequences.
4. Sliding a window with stride one, as illustrated in Fig. 2.

The output matrix of this step is fixed in one dimension and changeable in another dimension. Because of the fixed dimension, the proposed deep neural network can get genes of any size as an input. For example, an input sequence with 8000 characters, after windowing and slicing steps, is converted to a matrix with a size of 8000×801.

5. Using IUPAC nucleotide code to primary encode the 8000×801 matrix.

This preprocessing method generates a data matrix in which rows contain a specific part of a DNA sequence. Two rows of this matrix have been shown in Fig. 2. In each row, the central character needs to be determined whether it belongs to the exon region.

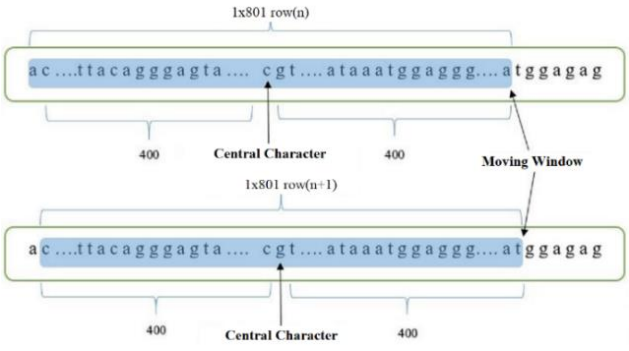

Fig.2. Moving window on a sequence in the proposed preprocessing.

2.2.2 Proposed CNN-LSTM Architecture

The proposed CNN-LSTM network has been illustrated in Fig. 3. As shown, it consists of some layers such as a word embedding layer, CNN layers, and an LSTM (a special kind of recurrent neural network (RNN)) layer, specialized in dealing with DNA sequences. The main idea of the proposed sequence encoding of this work has been inspired by natural language processing (NLP). We have implemented this idea as the embedding layer in our proposed method. One of the main challenges of classic machine learning in DNA sequence processing is extracting good features. In the proposed algorithm, the features have been extracted through the proposed CNN architecture. The next layer, LSTM, introduces memory units instead of conventional simple units to tackle exploding and vanishing problems of gradients and the long-term dependencies [32]. This ability helps the network identify the complex relationships between the characters in a DNA sequence. The detailed steps of the proposed algorithm are explained as follows:

1. The first proposed layer is a word embedding layer with a 1×801 input vector. Due to the use of 16 IUPAC codes, the output of this layer is a matrix with a size of 801×16. In previously published works, researchers usually have used four nucleobases (A, C, G, and T) instead of 16 polymorphism structures of DNA characters. The network can convert all 16 characters to numeric outputs by employing the proposed embedding layer. This layer learns all conversion details at the nucleotide level. Due to the similarity of the problem of this work and NLP problems, we have inspired the word embedding perspectives and Word2Vec to overcome the problem of the DNA sequence to number conversion. Word2vec is a framework for learning word embedding to convert the one-hot encoded categorical variables to vectors of floating-point numbers of smaller dimensionality than the input vectors. For more details, please refer to [21, 33].
2. We connected the word embedding layer to 6 parallel convolutional layers. Each convolutional layer has been designed with 16 filters, the ReLU activation function, and the padding option of "same." The size of the filter in the convolutional neural networks is a hyper-parameter. Also, we found that the larger filter size leads to a higher peak in the frequency domain and helps better classification between coding and non-coding regions. Therefore, we have studied the proposed deep neural network model on the HMR195 database with different convolutional layer numbers and filter sizes. We have used filter sizes of 420, 351, 120, 81, 51, and 33 for each parallel layer of the proposed multi-scale algorithm to trade-off between spatial and frequency resolution. Also, these 6 parallel filters with different sizes provide the concept of a multi-scale perspective of this work. A max-pooling and a dropout layer follow each filter in the convolution layer.
3. The outputs of the convolutional layers, the features of the network, have been concatenated.
4. The concatenated features have been considered inputs for two parallel LSTM layers. In these layers, the dropout factor is 0.5, and the recurrent dropout factor is 0.1. Moreover, the return sequences parameter for LSTM layers was set to TRUE.
5. The outputs of LSTM layers have been concatenated and followed by max-pooling and dropout layers.
6. The output of step 5 has been flattened and considered an input for a dense layer with 32 neurons, and a ReLU activation function has been followed by a dropout layer.

Finally, we have used a dense layer with one output neuron and a sigmoid activation function.

2.3. Training

We have trained the proposed network using the Tensorflow framework and Keras. We have employed the Adadelta solver [34] with a learning rate of 0.5 and a binary cross-entropy loss function for optimization.

2.4. Evaluation and Statistical Analysis

To compare and evaluate the performance of methods, we have used criteria such as true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) at the base pair. Also, sensitivity (SN), specificity (SP), accuracy (ACC), and Mathews correlation coefficient (CC) of the methods [35] have been computed as follows:

$$SN = \frac{TP}{TP + FN} \tag{1}$$

$$SP = \frac{TP}{TP + FP} \tag{2}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$CC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4}$$

A receiver operating characteristic (ROC) is a technique for performance measurement in classification problems at various threshold settings, and the area under the ROC curve (AUC) and the minimum distance (MD) of the ROC curve and (0, 1) have been used as other criteria. MD is calculated as follows [36, 37]:

$$MD = min(\sqrt{(TP - 1)^2 + FP^2}) \tag{5}$$

The threshold corresponding to the MD is selected as the best value.

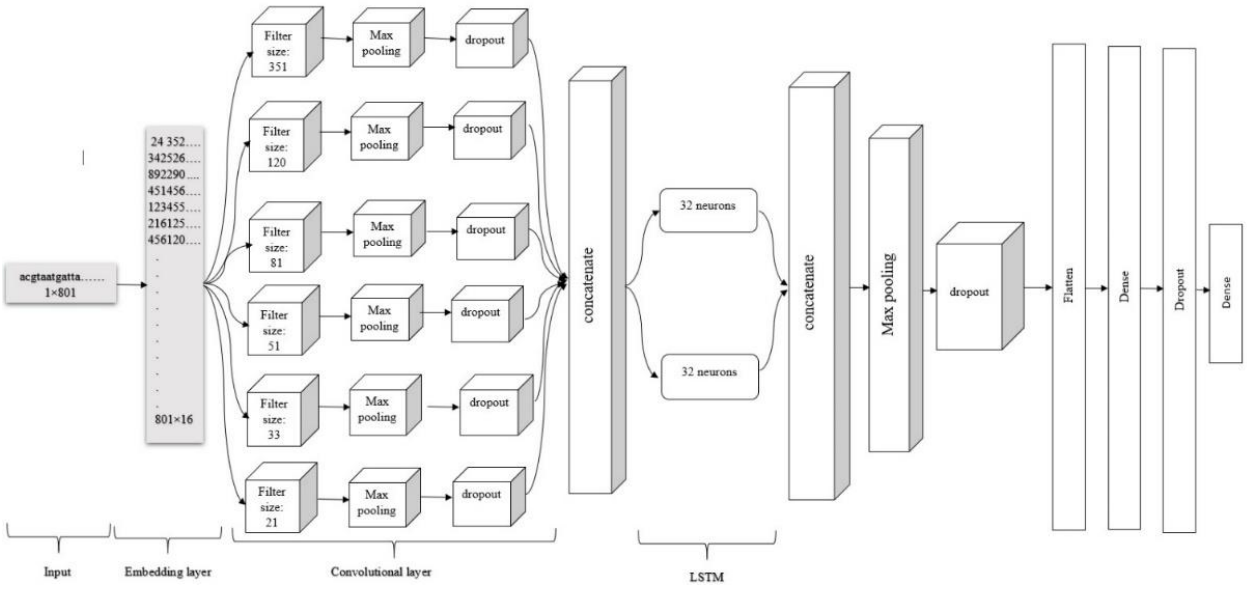

Fig.3. The proposed deep learning architecture

3. Results

Main In this section, we have presented the results in three parts: the results of the filter size study, the training results, and the comparative results.

3.1. Filter size effects

As mentioned before, we have studied the performance of the proposed model on the HMR195 database with a different number of convolutional layers and filter sizes. The plots of the period-3-PSD of F56F11.4 with two different window lengths are illustrated in Fig. 4. As shown, filters with smaller window sizes make more variations in the period-3-PSD that lead to increasing prediction error while larger window size misses small regions [38]. Small windows increase the predicting location resolution, and large windows increase the frequency resolution, so previous works usually selected and used a fixed window size [39].

We have designed layer configurations, including filter size and the number of convolutional layers, based on empirical results after one epoch. The comparison results between six configurations of the convolutional layers are presented in Table 1. The last configuration of the layers with the best accuracy and loss values has been selected for the rest of the work.

3.2. Training results of the proposed CNN-LSTM

In this work, we have used 20% and 80% of the total balanced HMR195 dataset as the validation and training data, respectively. The accuracy and the loss of the training and validation data have been shown in Fig. 5. As shown, the best accuracies and losses of the training and the validation data are 0.9745, 0.9893, 0.0743, and 0.0309, respectively. These values have been achieved after five epochs for the training data and four epochs for the validation data, respectively. The values for the validation phase are better than the training phase because of using the dropout in the training phase. We have used the BG570 database for fine-tuning. The result of this fine-tuning has been depicted in Fig. 5. As shown, the accuracy and loss for training data are 0.9658 and 0.0992, respectively. However, we found that with the increasing number of epochs from 4 to 5, the validation accuracy degraded from 0.9888 to 0.9833. Therefore, the weights related to epoch four have been saved.

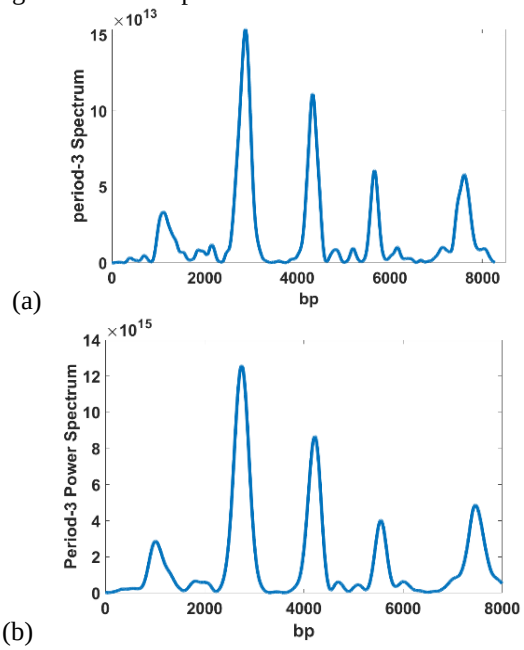

Fig.4 .The plots of the period-3-PSD of the F56F11.4 sequence with different window lengths: (a) 81, (b) 351.

3.3. Comparative Results

For a more accurate evaluation of the proposed method at the database level, the trained network has been used to label the complete set of the HMR195 database. In Table 2, the proposed method has been compared with methods MGWT [40], WTMM [41], DIT-FFT [42], and noise-assisted MEMD and wavelet transform (NA-MEMD-MGWT) [15]. In this work, we have implemented MGWT, and WTMM methods, so the CC values of these methods have been calculated. According to Table 2, the proposed method has achieved 0.98 in the CC benchmark on the whole HMR195, better than the

other implemented methods. Also, the proposed method is better in other criteria than the comparison methods, which reveal the robustness on not-seen data.

For the last performance assessment, the F56F11.4 sequence has been used as a test dataset and the corresponding results were compared with the results of eight previous methods: Singular Value Decomposition (SVD) [43], Linear Predictive Coding Model, and Goertzel Algorithm (LPCG or Goertzel) [44], MGWT [40], WRWW [45], Robust Singular Value Decomposition (RSVD) [46], Recursive Gauss-Newton tuned Adaptive Kaiser window (RGNAK) [47], sinusoidal-assisted variational mode decomposition (SAVMD) [7], and A-MEMD-MGWT [15]. Table 3 presents the results of these methods. In this Table, we have marked the implemented comparative methods as (Imp). The rest of the results have been extracted from the published works [7, 15, 47].

Table I. Comparison of 6 kinds of parallel convolutional layer configurations implemented in this work for one epoch.

	Num. of Conv. Layer	Size of Filters	Training Accuracy	Train Loss	Valid. Accuracy	Valid. Loss
1	4	3, 7, 9, 15	0.7302	0.5303	0.7707	0.4676
2	4	9, 15, 21, 33	0.7515	0.4976	0.7464	0.4968
3	5	9, 15, 21, 33, 51	0.7748	0.4671	0.8688	0.3167
4	5	9, 15, 21, 51, 81	0.7866	0.4473	0.8817	0.3077
5	6	9, 15, 21, 51, 81, 240	0.8118	0.3989	0.8824	0.2952
6	6	33, 51, 81, 120, 351, 420	0.8789	0.2919	0.9697	0.0992

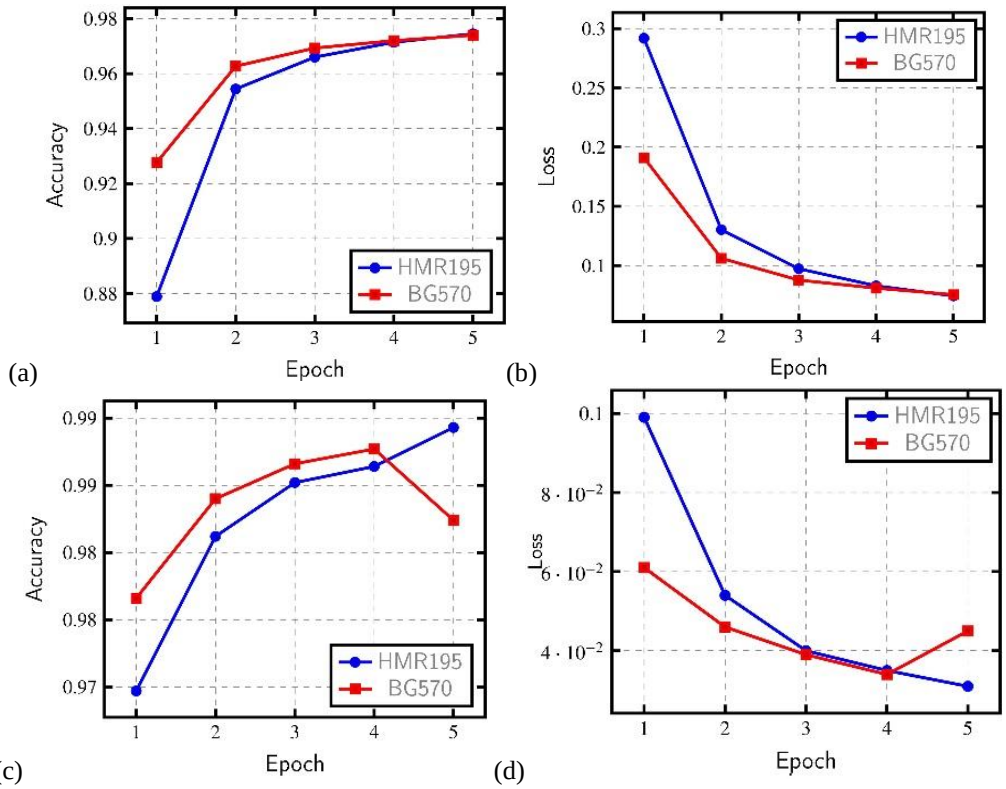

Fig.5. Training and validation graph results on HMR195 and BG570. (a)-(b) Train ACC and Loss, (c)-(d) Validation ACC and Loss

It is found that the sensitivity, specificity, accuracy, AUC, and CC values of the proposed algorithm are higher than all other methods, and the MD of the proposed method is less than the others.

Also, the plots of the implemented methods have been illustrated in Fig. 6. The y-axis of the plot of the proposed method indicates the probability of membership, and the y-axes of the other charts are PSD values. The dashed lines in these plots are the coding regions plotted from the dataset. As shown, the output of the proposed method has a lower noise level in the non-coding regions and sharp edges in coding regions, which leads to better coding region identification based on a simple thresholding. In Fig. 7, the ROC curves of the proposed and implemented methods have been plotted. The analysis of the ROC curves reveals that the proposed method outperforms better than the others.

For the exon level investigation, the F56F11.4 with the length of 8000 has been segmented into five single exon levels as follows: 1-1783, 1784-3485, 3486-5020, 5021-

6449, 6450-8000 [27]. The result of the comparative analysis has been presented in Table 4. This signal segmentation leads to locally threshold single exon regions instead of conventional global thresholding, so the results of the comparison methods have been improved. But as a superiority, the proposed method has only one simple global thresholding at 0.5.

It should be noted, that we have trained the network on one Core i7 7700HQ CPU, Nvidia GeForce GTX 1060 with Max-Q Design GPU, and 24GB RAM for about 6 hours. Test time for a gene with 8000 nucleotides was about 30 seconds but the previous approaches are generally real-time.

3.4. Heat map representation

In Fig. 8, we have plotted the attention heat maps to explain more the performance of the proposed network. These figures help to view the genetic code through the network's eyes. In these maps, the intensity of the red color of the characters indicates the predicted probability of an exon. The green box is the true start codon, and the yellow box is the true stop codon.

Table II. Comparison between the proposed method with MGWT and WTMM methods on HMR195.

Method	SN	ACC	AUC	CC
MGWT(Imp)* (2008) [40]	-	-	0.8396	0.525
WRWW (2015) [45]	-	-	0.8317	-
WTMM(Imp) (2016) [41]	-	-	-	0.600
DIT-FFT (2019) [42]	0.8	0.88	-	-
NA-MEMD-MGWT (2021) [15]	-	-	0.7383	-
Proposed Method	**0.97**	**0.991**	**0.9995**	**0.9823**

*(Imp): Implemented in this work, -: not presented

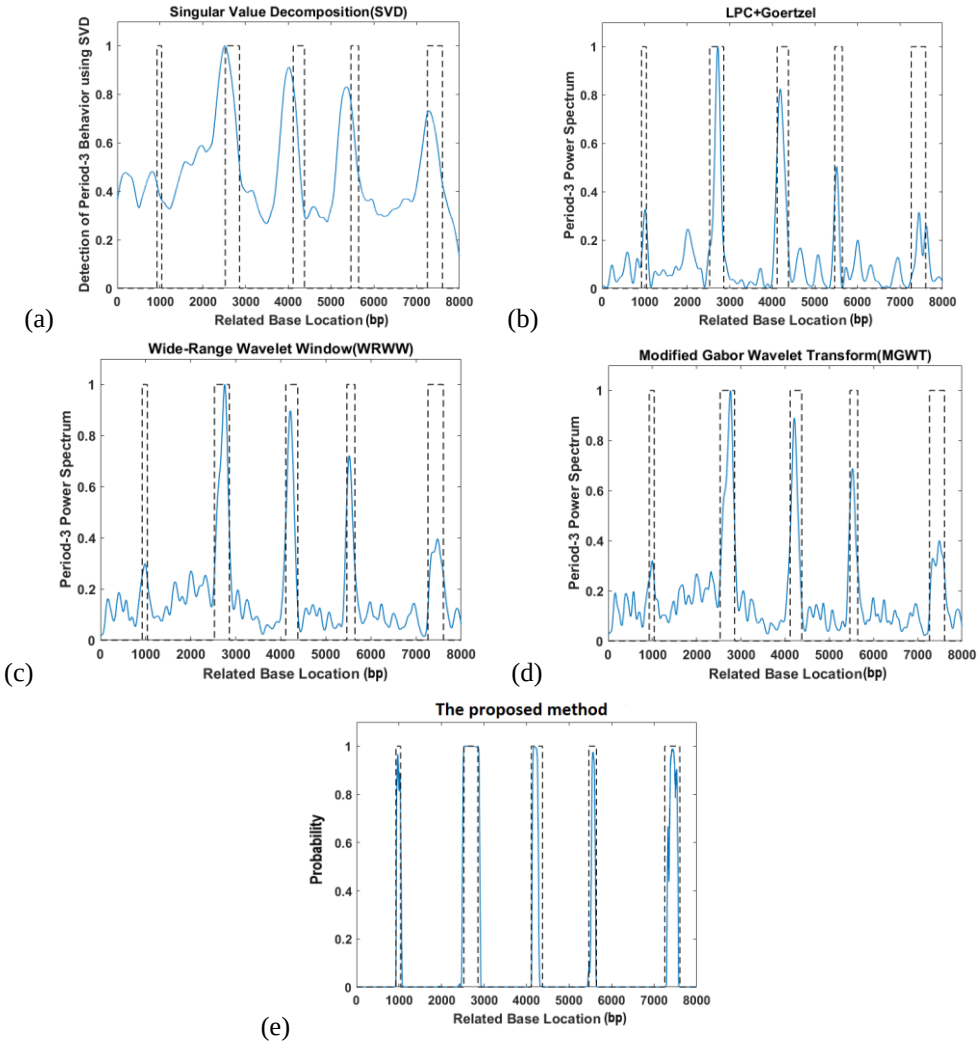

Fig 6. The plots of protein-coding region identification of the sequence F56F11.4 by some different methods based on period-3-PSD (a-d) and the proposed deep learning method (e). (a) SVD, (b) Goertzel, (c) WRWW (d) MGWT (e) The proposed Method.

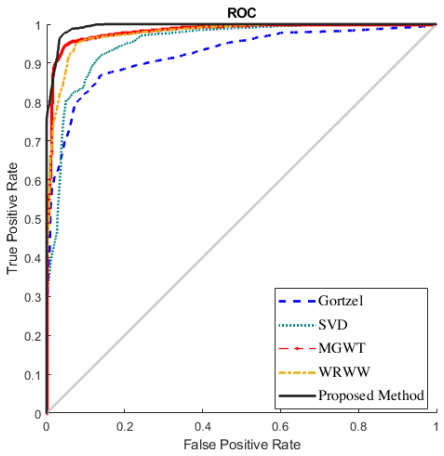

Fig 7. ROCs of implemented methods on F56F11.4.

4. Discussion

As observed in Fig. 4, according to the size of the filter, there is a trade-off between undesired variations and missing regions. Also, location resolution and frequency resolution are other influenced requirements by the size of the filter. Therefore, the proposed multi-scale approach can deal with these problems. As shown in Table 1 the best result of the proposed method after one epoch as the initial investigation was for the filter sizes of 33, 51, 81, 120, 351, and 420.

Based on the curves illustrated in Fig 5, to provide the generalization of the proposed method, and to prevent overfitting problems, the weights related to epoch four have been used for the rest of the work.

As illustrated in Fig. 6, the proposed method predicts the regions of exons with lower noise, sharper edges, and higher probability picks. This quality of prediction helps to use a simple fixed probability threshold of 0.5 to find the location of exon regions. In other words, the probability output plot of our proposed method has higher SNR that leads to clear boundaries of the regions with a simple fixed threshold of 0.5. While in previous methods, SVD [43], LPC+ Goertzel [44], WRWW [45], and MGWT [40], which are based on traditional period-3 property, a specific threshold value should be calculated for each database. Also, previous studies have worked on mapping methods and noise suppression filters, and the period-3 property has been the only extracted feature. While in the proposed approach, mapping and feature extraction have been calculated together in an automated multi-scale manner. For digitizing DNA nucleotides, several methods have existed that use a fixed mapping. Ahmad et. al [14] observed that mapping based on genetic context improves accuracy and minimizes the false-positive rate of identification. So, for considering the genetic context and the semantics of the nucleotide relationship, we have used the embedding layer for mapping. The noteworthy point is that the result of the model before using this layer was undesirable.

As seen in the heat map representation (Fig. 8), the proposed neural network does not learn to find the start and stop codons for detecting exons. According to the embedding layer performance, it can be inferred that the proposed method detects exons based on the nucleotide relationship.

According to the aforementioned, the proposed method in this study is accurate and general. But as mentioned in section 3.3, the computational load of the proposed method in comparison to the previous methods is high and needs more time to train.

5. Conclusions

This research used the deep machine learning paradigm instead of period-3 property-based classical machine learning methods, and a modified deep neural network model with a hybrid multi-scale CNN-LSTM network was introduced with a new preprocessing method. The main advantages of the proposed model are the embedded feature extraction and unsupervised mapping of nucleotides with various lengths. Also, in the proposed model, we used IUPAC which leads to covering all possible nucleotides. The comparative study has shown that the proposed method outperforms other methods used in the experiments. A comparison with the mentioned methods reveals the superiority and effectiveness of the proposed method. In future works, the proposed hybrid method can be used for splice site detection via a multi-class classifier instead of a binary classifier.

Table III. Comparative analysis of different methods on F56F11.4.

Method	MD	SP	SN	ACC	AUC	CC
SVD(Imp)* (2005) [43]	0.156	0.565	0.908	0.878	0.952	0.653
Goertzel(Imp) (2013) [44]	0.190	0.539	0.867	0.865	0.922	0.611
MGWT(Imp) (2008) [40]	0.072	0.791	0.934	0.863	0.972	0.838
WRWW(Imp) (2015) [45]	0.093	0.70	0.948	0.929	0.967	0.777
RSVD (2017) [46]	-	0.9	0.8	0.85	-	0.833
RGNAK (2019) [47]	-	0.92	0.96	0.94	-	0.813
SAVMD (2021) [7]	-	-	-	-	0.93	-
NA-MEMD-MGWT (2021) [15]	-	-	-	-	-	0.8354
Proposed Method	0.048	0.965	0.968	0.965	0.994	0.877

*(Imp): Implemented in this work,
-: not presented

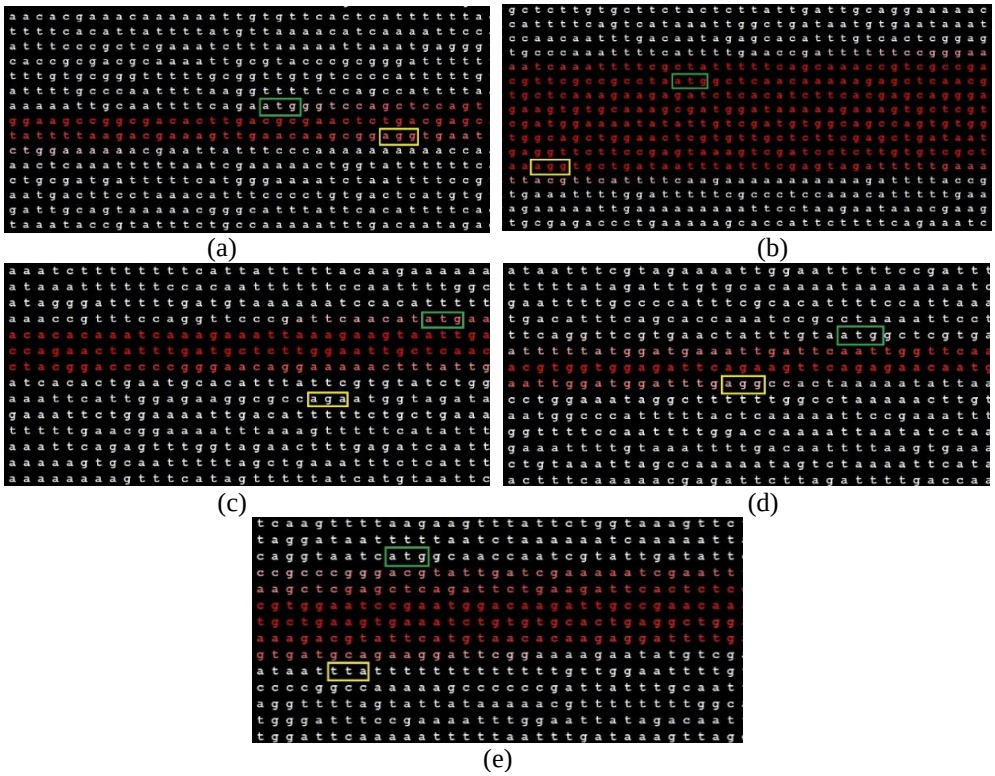

Fig 8. The attention heat maps of the exons in F56F11.4. (a)-(e) Exon 1-5.

Table IV. The result of the comparative in exon level.

	E1			E2			E3			E4			E5		
Method	**ACC**	**AUC**	**CC**	**ACC**	**AUC**	**CC**	**ACC**	**AUC**	**CC**	**ACC**	**AUC**	**CC**	**ACC**	**AUC**	**CC**
SVD	0.964	.756	0.657	0.68	0.777	0.477	0.867	0.844	0.61	0.905	0.946	0.713	0.985	0.976	0.956
Goertzel	0.934	0.939	0.65	0.817	0.879	0.631	0.863	0.913	0.681	0.823	0.78	0.439	0.877	0.81	0.639
MGWT	0.998	0.995	0.986	0.888	0.928	0.739	0.949	0.928	0.829	0.953	0.973	0.831	0.969	0.942	0.91
WRWW	0.943	0.969	0.70	0.848	0.901	0.675	0.945	0.915	0.813	0.948	0.97	0.817	0.968	0.951	0.908
Proposed	**0.978**	**0.988**	**0.85**	**0.954**	**0.971**	**0.872**	**0.986**	**0.972**	**0.952**	**0.92**	**0.954**	**0.746**	**0.982**	**0.962**	**0.948**

6. References

[1] D. P. Snustad and M. J. Simmons, *Principles of genetics*. John Wiley & Sons, 2015.

[2] E. R. Dougherty and I. Shmulevich, *Genomic signal processing and statistics*. Hindawi Publishing Corporation, 2005.

[3] H. JE, "Guyton and Hall textbook of medical physiology," *Philadelphia, PA: Saunders Elsevier*, vol. 107, p. 1146, 2011.

[4] A. M. Oudelaar and D. R. Higgs, "The relationship between genome structure and function," *Nature Reviews Genetics*, vol. 22, no. 3, pp. 154-168, 2021.

[5] P. Vaidyanathan, "Genomics and proteomics: A signal processor's tour," *IEEE Circuits and Systems Magazine*, vol. 4, no. 4, pp. 6-29, 2004.

[6] F. B. Nasr and A. E. Oueslati, "CNN for human exons and introns classification," in *2021 18th International Multi-Conference on Systems, Signals & Devices (SSD)*, 2021, pp. 249-254: IEEE.

[7] Q. Zheng, T. Chen, W. Zhou, S. A. Marhon, L. Xie, and H. Su, "SAVMD: An adaptive signal processing method for identifying protein coding regions," *Biomedical Signal Processing and Control*, vol. 70, p. 102998, 2021.

[8] R. F. Voss, "Evolution of long-range fractal correlations and 1/f noise in DNA base sequences," *Physical review letters*, vol. 68, no. 25, p. 3805, 1992.

[9] P. D. A. Cristea, "Genomic signals of chromosomes and of concatenated reoriented coding regions," in *Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues II*, 2004, vol. 5322, pp. 29-41: International Society for Optics and Photonics.

[10] G. L. Rosen, *Signal processing for biologically-inspired gradient source localization and DNA sequence analysis*. Georgia Institute of Technology, 2006.

[11] A. S. Nair and S. P. Sreenadhan, "A coding measure scheme employing electron-ion interaction pseudopotential (EIIP)," *Bioinformation*, vol. 1, no. 6, p. 197, 2006.

[12] D. Anastassiou, "Genomic signal processing," *IEEE signal processing magazine*, vol. 18, no. 4, pp. 8-20, 2001.

[13] P. D. Cristea, "Genetic signal representation and analysis," in *Functional Monitoring and Drug-*

Tissue Interaction, 2002, vol. 4623, pp. 77-84: International Society for Optics and Photonics.

[14] M. Ahmad, L. T. Jung, and M. A.-A. Bhuiyan, "On fuzzy semantic similarity measure for DNA coding," *Computers in biology and medicine*, vol. 69, pp. 144-151, 2016.

[15] Q. Zheng, T. Chen, W. Zhou, L. Xie, and H. Su, "Gene prediction by the noise-assisted MEMD and wavelet transform for identifying the protein coding regions," *Biocybernetics and Biomedical Engineering*, vol. 41, no. 1, pp. 196-210, 2021.

[16] H. Herzel, E. N. Trifonov, O. Weiss, and I. Grosse, "Interpreting correlations in biosequences," *Physica A: Statistical Mechanics and Its Applications*, vol. 249, no. 1-4, pp. 449-459, 1998.

[17] H. Herzel, O. Weiss, and E. N. Trifonov, "10-11 bp periodicities in complete genomes reflect protein structure and DNA folding," *Bioinformatics (Oxford, England)*, vol. 15, no. 3, pp. 187-193, 1999.

[18] H. Saberkari, M. Shamsi, and M. H. Sedaaghi, "A punctual algorithm for small gene prediction in DNA sequences using a time-frequency approach based on the z-curve," *GSTF Journal of Engineering Technology (JET)*, vol. 2, no. 1, p. 1, 2013.

[19] M. Ahmad, L. T. Jung, and A.-A. Bhuiyan, "A biological inspired fuzzy adaptive window median filter (FAWMF) for enhancing DNA signal processing," *Computer methods and programs in biomedicine*, vol. 149, pp. 11-17, 2017.

[20] A. K. Singh and V. K. Srivastava, "Improved filtering approach for identification of protein-coding regions in eukaryotes by background noise reduction using S–G filter," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 10, no. 1, pp. 1-16, 2021.

[21] F. Chollet, *Deep learning with Python*. Simon and Schuster, 2021.

[22] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," *Briefings in bioinformatics*, vol. 18, no. 5, pp. 851-869, 2017.

[23] N. K. Vaegae, "Walsh code based numerical mapping method for the identification of protein coding regions in eukaryotes," *Biomedical Signal Processing and Control*, vol. 58, p. 101859, 2020.

[24] A. K. Singh and V. K. Srivastava, "The three base periodicity of protein coding sequences and its application in exon prediction," in *2020 7th international conference on signal processing and integrated networks (spin)*, 2020, pp. 1089-1094: IEEE.

[25] N. Naderi and B. Nasersharif, "Robust sub-band speech feature extraction using multiresolution convolutional neural networks," *TABRIZ JOURNAL OF ELECTRICAL ENGINEERING*, vol. 49, no. 3, pp. 1393-1404, 2019.

[26] M. Afrasiabi, H. Khotanlou, and M. Mansoorizadeh, "Deep neural network for interaction prediction in video using fuzzy relationship and optical flow," *TABRIZ JOURNAL OF ELECTRICAL ENGINEERING*, vol. 50, no. 3, pp. 1035-1046, 2020.

[27] A. Saeedi, M. Saeedi, A. Maghsoudi, and A. Shalbaf, "Major depressive disorder diagnosis based on effective connectivity in EEG signals: A convolutional neural network and long short-term memory approach," *Cognitive Neurodynamics*, vol. 15, no. 2, pp. 239-252, 2021.

[28] S. Rogic. [Online]. Available: http://srogic.wordpress.com/datasets/hmr195-dataset/

[29] S. Rogic, A. K. Mackworth, and F. B. Ouellette, "Evaluation of gene-finding programs on mammalian sequences," *Genome research*, vol. 11, no. 5, pp. 817-832, 2001.

[30] A. Saito, A. Tomita, R. Ando, K. Watanabe, and H. Akima, "Similarity of muscle synergies extracted from the lower limb including the deep muscles between level and uphill treadmill walking," *Gait & posture*, vol. 59, pp. 134-139, 2018.

[31] M. Burset and R. Guigo, "Evaluation of gene structure prediction programs," *genomics*, vol. 34, no. 3, pp. 353-367, 1996.

[32] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[33] Y. Goldberg and O. Levy, "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method," *arXiv preprint arXiv:1402.3722*, 2014.

[34] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[35] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442-451, 1975.

[36] R. Kumar and A. Indrayan, "Receiver operating characteristic (ROC) curve for medical researchers," *Indian pediatrics*, vol. 48, no. 4, pp. 277-287, 2011.

[37] S. Shamekhi, M. H. M. Baygi, B. Azarian, and A. Gooya, "A novel multi-scale Hessian based spot enhancement filter for two-dimensional gel electrophoresis images," *Computers in biology and medicine*, vol. 66, pp. 154-169, 2015.

[38] C. Yin and S. S.-T. Yau, "Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence," *Journal of theoretical biology*, vol. 247, no. 4, pp. 687-694, 2007.

[39] M. Ahmad, L. T. Jung, and A.-A. Bhuiyan, "From DNA to protein: Why genetic code context of nucleotides for DNA signal processing? A review," *Biomedical Signal Processing and Control*, vol. 34, pp. 44-63, 2017.

[40] J. Mena-Chalco, H. Carrer, Y. Zana, and R. M. Cesar Jr, "Identification of protein coding regions using the modified Gabor-wavelet transform," *IEEE/ACM Transactions on*

Computational Biology and Bioinformatics, vol. 5, no. 2, pp. 198-207, 2008.

[41] X. Zhang *et al.*, "Short exon detection via wavelet transform modulus maxima," *PloS one*, vol. 11, no. 9, p. e0163088, 2016.

[42] S. Kar, M. Ganguly, and S. Das, "Using DIT-FFT algorithm for identification of protein coding region in eukaryotic gene," *Biomedical Engineering: Applications, Basis, and Communications*, vol. 31, no. 01, p. 1950002, 2019.

[43] M. Akhtar, E. Ambikairajah, and J. Epps, "Detection of period-3 behavior in genomic sequences using singular value decomposition," in *Proceedings of the IEEE Symposium on Emerging Technologies, 2005.*, 2005, pp. 13-17: IEEE.

[44] H. Saberkari, M. Shamsi, H. Heravi, and M. H. Sedaaghi, "A novel fast algorithm for exon prediction in eukaryotic genes using linear predictive coding model and goertzel algorithm based on the Z-curve," *International Journal of Computer Applications*, vol. 67, no. 17, 2013.

[45] S. A. Marhon and S. C. Kremer, "Prediction of protein coding regions using a wide-range wavelet window method," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 13, no. 4, pp. 742-753, 2015.

[46] L. Das, J. Das, and S. Nanda, "Advanced protein coding region prediction applying robust SVD algorithm," in *2017 2nd International Conference on Man and Machine Interfacing (MAMI)*, 2017, pp. 1-6: IEEE.

[47] L. Das, S. Nanda, and J. Das, "An integrated approach for identification of exon locations using recursive Gauss Newton tuned adaptive Kaiser window," *Genomics*, vol. 111, no. 3, pp. 284-296, 2019.