Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2020) 28: 1932 – 1948

doi:10.3906/elk-1909-119

A novel genome analysis method with the entropy-based numerical technique using pretrained convolutional neural networks

Bihter DAŞ[1,*], **Suat TORAMAN**[2], **İbrahim TÜRKOĞLU**[1]
[1]Department of Software Engineering, Faculty of Technology, Fırat University, Elazığ, Turkey
[2]Department of Informatics, Fırat University, Elazığ, Turkey

Received: 22.09.2019 • **Accepted/Published Online:** 07.02.2020 • **Final Version:** 29.07.2020

Abstract: The identification of DNA sequences as exon and intron is a common problem in genome analysis. The methods used for feature extraction and mapping techniques for the digitization of sequences affect directly the solution of this problem. The existing mapping techniques are not enough to detect coding and noncoding regions in some genomes because the digital representation of each base in a DNA sequence with an integer does not fully reflect the structure of an original DNA sequence. In the entropy-based mapping technique, we could overcome this problem because the technique deepens distinction rates of exon regions, and better reflects the complexity of DNA sequences. Moreover, in the literature, features are extracted by using various statistical techniques. The statistical features to be extracted are chosen by a system designer's experience. The other proposed approach in this study is to carry out the feature extraction using the transfer learning method. Transfer learning and feature extraction are performed automatically by convolutional neural network models as independent of the data set. In this study, we propose a new method to classify DNA sequences as exon and intron using two approaches. In the first approach, the entropy-based numerical technique was used for the numerical representation of DNA sequences. In the second approach, transfer learning was used to extract features. Then, the obtained features were classified by support vector machine and k-nearest neighbors algorithm. As a result of the classification, accurate performance with 97.8% was achieved. The performance of the current method was compared with the other numerical mapping techniques and feature extraction methods. The results showed that the developed method was much more successful than other methods.

Key words: DNA, genome analysis, convolutional neural network, classification, entropy-based mapping technique

1. Introduction

Genetic features are carried by chromosomes in a cell core. The chromosomes are composed of a combination of DNA and special proteins. Nucleotides are DNA's structure units. They consist of sugar, phosphate, and organic bases. These bases are adenine (A), guanine (G), thymine (T), and cytosine (C). If a nucleotide contains a base, it is characterized by the name of the base [1–3]. Figure 1 shows the structure of DNA.

During the production of proteins and enzymes, RNA copy sequences, which are corresponding to these bases, are extracted as an example of base sequences in DNA [4, 5]. These RNAs are named as mRNA. While mRNA is extracted, the noncoding parts of a DNA sequence are called introns, the coding parts of DNA are called exons [6–9]. Classification of the DNA sequence, which belong to a gene, as exons and introns is quite important. The investigation of a protein where, how, when, and how much is coded is very important. With the

*Correspondence: bihterdas@gmail.com

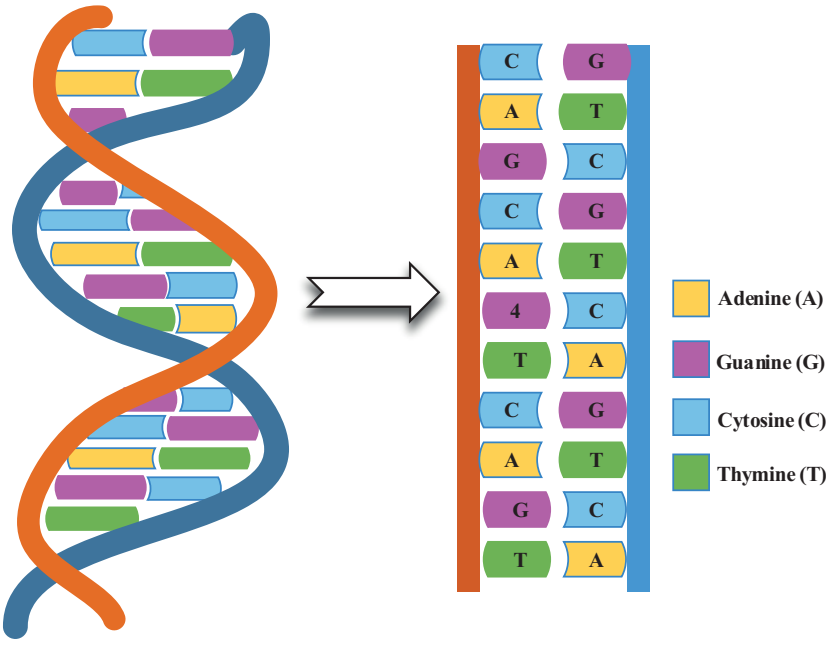

Figure 1. The double-helical structure of DNA.

classification of DNA sequences as exon and intron, it can be known where the stem cells will be transformed into organs, tissues, and cells, under which conditions the cells will be reproduced or killed. In traditional machine learning techniques, features are extracted using various statistical features to classify DNA sequences as exon and intron. Statistical features to be extracted are chosen by the system designer's experience. The model to be developed becomes specific for the dataset. Deep learning is a new machine learning method that automatically encodes the properties of datasets, regardless of data type. Thus, features of datasets are automatically extracted by a deep neural network without specifying any feature. In this work, we proposed a new approach to classify DNA sequences as exons and introns. In this approach, we used the entropy-based numerical mapping technique for numerical representation of the DNA sequences, and deep learning for the feature extraction. The obtained features of deep learning models were classified by two different classifiers.

This paper is organized as follows. In Section 2, the related work about classification of DNA sequences as exon and intron, and the usage of deep learning methods for genomic studies have been given. In Section 3, we mentioned the numerical representation of DNA datasets and described the proposed classification method based on deep learning. In Section 4, we showed the results of our approach and the classification accuracy. Finally, the conclusion is given in Section 5.

1.1. Contributions of the paper

When our study is compared with other methods in the literature, the achievement and originality of the study are given below.

- In order to recognize exon regions, the features were extracted using VGG16, VGG19, and ResNet from the convolutional neural network (CNN) models. The DNA sequences were classified as exons and introns using the obtained features.

- While various statistical features were used for the classification of exon and intron regions in previous studies, the features were automatically extracted from pretrained CNN models in the current study.
- For numerical representation of DNA sequences, the entropy-based mapping technique that we developed instead of existing mapping techniques was used.
- The digitized exon and intron sequences were examined by converting to spectrogram display for the first time.
- We achieved an average success with 97.8%

2. Related work

In this section, the studies about the classification of exon and intron regions in DNA and the applications of deep learning in this field are examined. Abo-Zahhad et al. [6] explained numerical mapping techniques in two types. These are fixed mapping techniques and physico-chemical property-based mapping techniques. They compared the sensibility and the accuracy of these mapping techniques in exon and intron classification. The human genome (GRch37hg19) with 8000 bases length was downloaded from the USCS database for the dataset. Digital signal processing was used for the classification application. Crosby and Gabbert [10] performed scalable parallelizable induction of decision trees algorithm to classify DNA sequences as introns or exons. For training and testing, the genomes of *Drosophila melanogaster* and the *Caenorhabditis elegans* were used. The test sample error rate with 15% was shown for the *Drosophila melanogaster*, and it was 1.6% for *Caenorhabditis elegans*. The authors in [11] presented a novel approach that is called wavelet-based time series approach for DNA analysis. This approach contained a variance information of amino/keto, purine/pyrimidine, and weak or strong hydrogen bond distribution of a DNA sequence. The feature vector was obtained using the variance information. A support vector machine (SVM) classifier is constructed to accurately classify DNA sequences as exon and intron. The dataset of DNA sequences of *Homo sapiens* was used, and a 10-fold cross-validation accuracy with 87.5% was achieved. Sahu and Panda [12] introduced a novel model that is independent of time-frequency filtering technique based on S-transform to accurately identify exon regions. For the dataset, a $F56F11.4a$ gene on chromosome III of *Caenorhabditis elegans* in NCBI Genbank was used. The electron–ion interaction potential (EIIP) mapping technique was used for numerical representation of the DNA sequence. For classification of exon and intron regions, the performance of the proposed method was compared with the other digital signal processing methods. Zhang and Yan [13] proposed an improved exon prediction method based on empirical mode decomposition and Fourier transform. They used the binary mapping technique to digitize the gene sequence of Z20656. The structural profile of the DNA sequence was used for feature selection. This method helped to improve the prediction accuracy of short exons. Hung and Tang [14] reviewed the GPU-based deep learning-based algorithms on biological data. The convolution neural network and recurrent neural network have been adopted in gene expression analysis, enhancer and regulatory region prediction, and methylation prediction. Sree et al. [15] proposed a novel deep learning component for quicker DNA processing. They took CA (Supervised) for preprocessing the initial DNA sequences. In this study the results showed that the deep learning component has increased the prediction accuracy with 12.3%. The authors in [15] developed a novel unsupervised classifier that used hybrid cellular automata with deep learning to address some problems in bioinformatics. This classifier predicts the gene with 98.7% accuracy.

3. The proposed method and material

In this section, the evolution of the classification process including data, feature extraction, and model development is explained.

3.1. Dataset

The data usedin this application was obtained from Ensembl database (Homo_sapiens_IL18R1_sequence, Gene: IL18R1 ENSG00000115604)[1]. It consists of 86 exons and 86 introns which are 297 bases length. First of all, we used the Entropy–based Numerical Mapping Technique to digitize DNA sequences [16, 17]. The performance of this technique was compared with other two existing mapping techniques (e.g., EIIP, Integer) in the literature that are mostly used in classification applications.

3.2. Digitization of the data

In order to use DNA sequences in machine learning methods, they should be converted to numerical signals. In the literature there are some numerical mapping techniques. In this study, we used the entropy mapping technique in deep learning models for the first time. In this technique, fractional Shannon entropy is used to convert DNA sequences into numerical values. The entropy of codons' distribution in a DNA sequence is computed [16, 18]. There are 64 possible codon states in a DNA sequence. The proposed technique is defined by Equation 1.

$$Sf = -\sum \left[(-p(x_i))^{\alpha} p(x_i) \log(p(x_i)\right]. \tag{1}$$

Here $p(x_i)$ represents the frequency of each codon in the given DNA sequence. In Equation 1, there is an α value which is generally calculated by trial and error approach, but in the entropy-based mapping technique, this alpha value was extracted from the DNA data, and it was defined by a new approach in Equation 2.

$$\alpha = \frac{1}{\log(p(x_i))}. \tag{2}$$

The logarithm of the $p(x_i)$ value is divided by 1 to get the alpha value. DNA sequence is digitized using Equations 1 and 2. The performance of entropy-based mapping technique in deep learning models was compared to other two current mapping techniques. One of them is integer numerical technique. It is a widely used technique. Bases are represented by the following digital values: T=0, C=1, A=2, G=3 [19, 20]. Another one is EIIP technique. In this technique bases are represented by the following values: A=0.1260, G=0.0806, C=0.1340, T=0.1335 [19, 21]. Figure 2 shows an example of DNA sequence digitized by the three numerical mapping techniques.

Figure 3 shows signal representations of DNA sequence examples digitized by three numerical mapping techniques.

3.3. Spectrum images composing

Firstly, feature extraction was carried out to classify the digitized 86 exons and intron data. Spectrum images were obtained to extract features from each of 297 bases length exon and intron sequences. While obtaining these spectrum images, Hamming window width as 16 ms, overlap value as 8 ms and number of Fourier transforms

[1]Ensembl Genbank Database (2019) [online]. Website http://www.ensembl.org/index.html [accessed 3 September 2019]

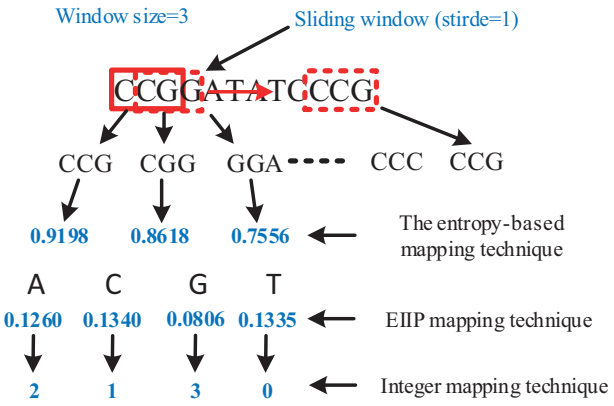

Figure 2. Vector representation of the three numerical mapping techniques for a DNA sequence

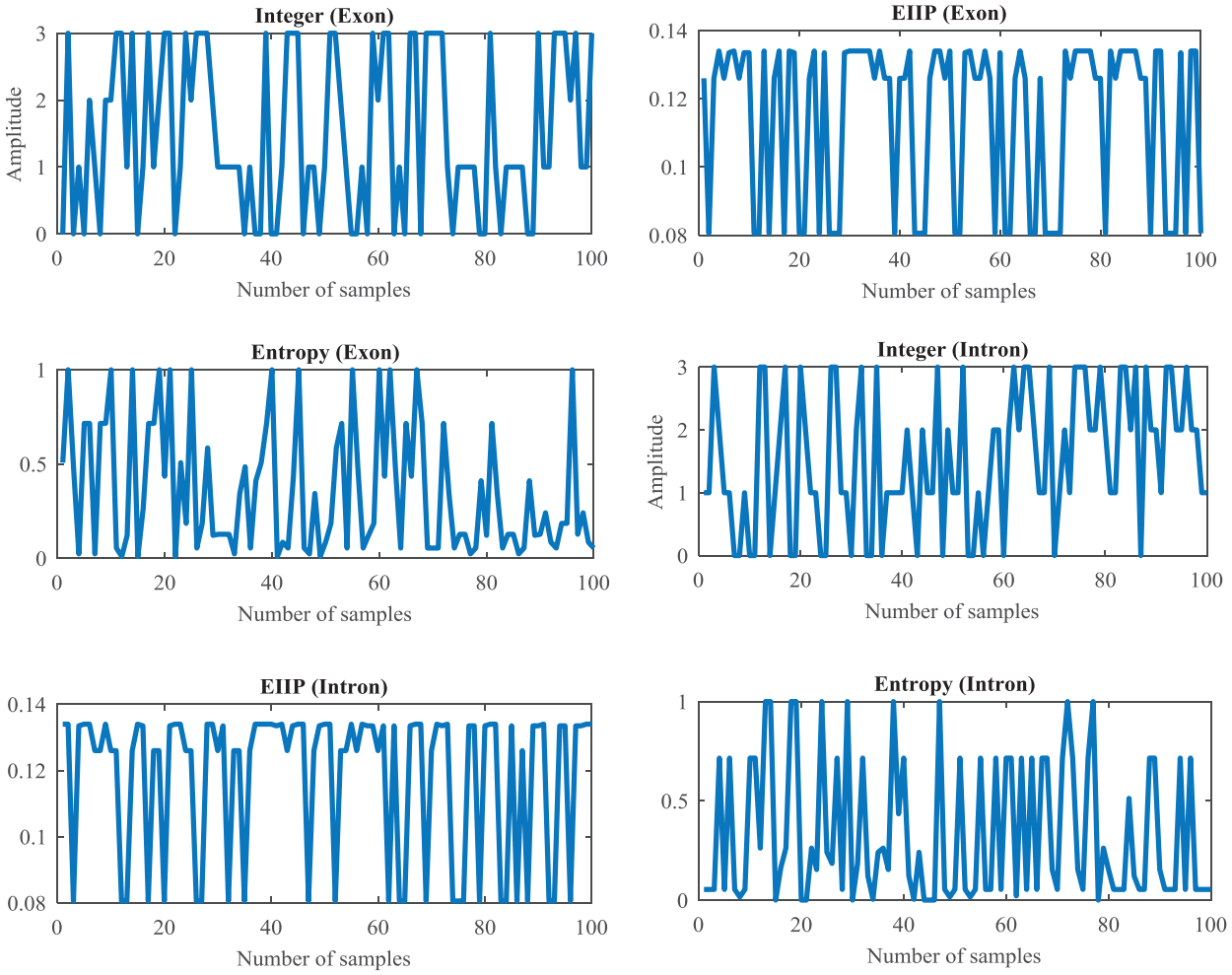

Figure 3. Signal representation of exon and intron data by integer, EIIP, and entropy-based techniques.

as 512 have been identified. The spectrogram images were formed using Viridis color map. Figure 4 shows the spectrum images of exon and intron sequences digitized by three numerical mapping techniques. After digitizing the DNA sequences using entropy and EIIP-based mapping techniques, we plotted the signals with MATLAB. Our aim is to create a visual representation of the digitized data. Then, the numerical data were converted into spectrogram image with MATLAB (spectrogram command). In Section 3.3, the parameters used to obtain the spectrogram image are given. The obtained spectrogram images are of 875×656 pixels. These images are resized to be 224×224 pixels for VGG16, VGG19, and ResNet model.

3.4. Convolutional neural network

Convolutional neural network (CNN) is one of the most used deep learning models today because of its high performance. CNN models are used in many areas such as object recognition, and signal and image processing [24]. In this study, CNN was used to classify exon and intron spectrogram data digitized by the entropy-based numerical mapping technique. The CNN models used in the study are VGG16, VGG19, and ResNet. It is necessary to have a large amount of data and also a good hardware infrastructure for training of a CNN model. It could be necessary to wait for hours or even days to process the data according to the available hardware features. If a large amount of data is not available to train a model, pretrained models can be used. This is called transfer learning. Transfer learning could be used to extract features from a dataset in a different field or a model that was pretrained using large datasets [22–24]. In this study, pretrained VGG16, VGG19, and ResNet models were used. In Figure 5, the feature extraction method used in this study is shown. It is not possible to distinguish the digitized exon and intron data by looking. At this point, deep learning models play an important role. The difference between deep learning models and traditional machine learning methods is that deep learning models do not require manual feature extraction. Therefore, in this study, CNN models capable of extracting features from raw data were used. The CNN models extract distinctive information from exon and intron data for training, and then find out which class the data for testing belongs to. As shown in Figure 5, a feature vector is obtained for each image at the output of all three CNN models. These feature vectors are then classified by SVM and k-nearest neighbors (k-NN) classifiers. VGG16 and VGG19 represent each image with a vector of 4096 dimensions, while ResNet represents 2048 dimensional. These vectors are the feature vectors that CNN models derive from each image. Therefore, in this study, manual feature extraction method was not used.

3.4.1. VGG16 and VGG19

VGG16 is a popular model that was developed after AlexNet. It was developed by Oxford University Visual Geometry Group (VGG). In the convolution layers of VGG16, smaller filters (3×3) were used differently from AlexNet. VGG16 consists of 13 convolution layers and 3 fully connected layers. There are 5 pieces of a max pooling layer with 2×2 size. The final layer is the Softmax layer that classifies incoming input data [25]. ReLu was used as activation function. The VGG19 model consists of 16 convolution layers and 3 fully connected layers. While VGG16 contains 138 million parameters, VGG19 contains approximately 144 million parameters [26, 27].

3.4.2. ResNet

The developments in computer hardware make deep CNN architectures very important. Classical deep learning models, which are formed by the successive addition of a number of layers, started with LeNet, then their

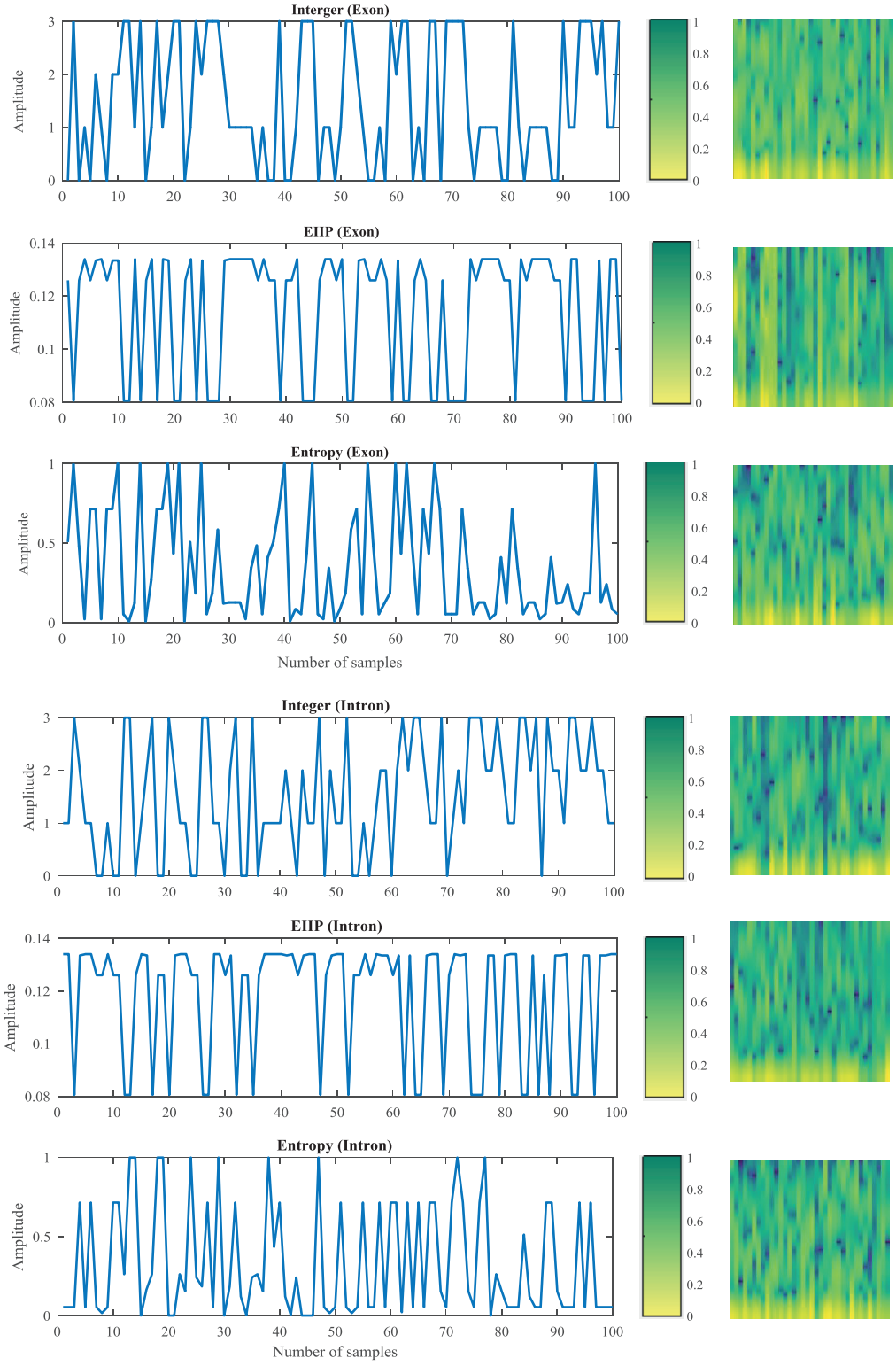

Figure 4. Spectrum images of the exon and intron sequences digitized by three mapping techniques.

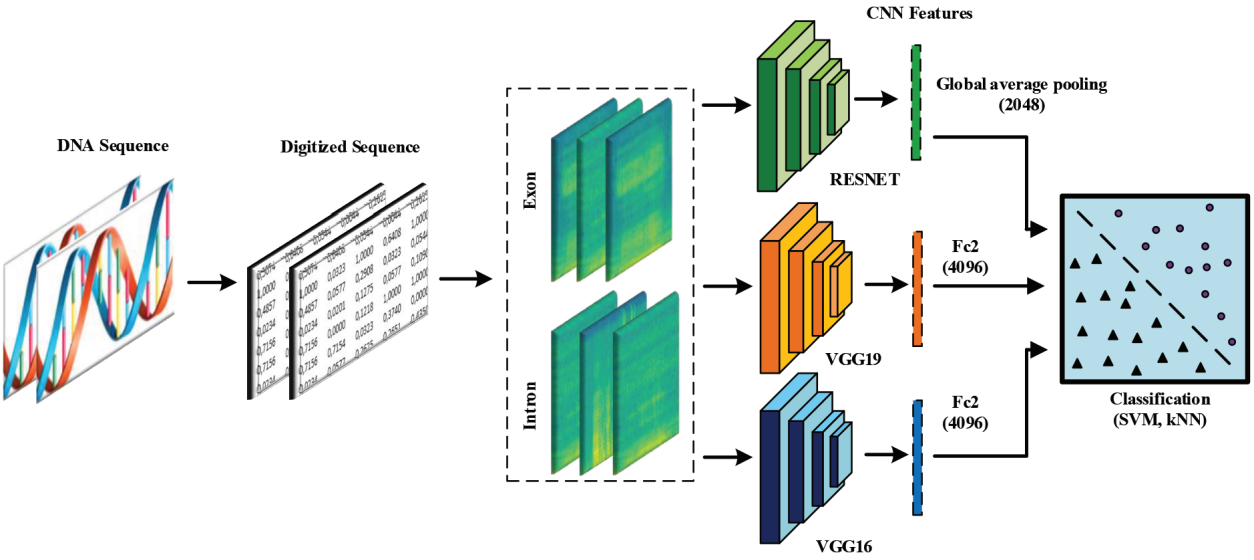

Figure 5. The feature extraction method.

popularity increased with AlexNet, and has continued with VGG16 and GoogleNet. The number of layers has also increased over time, but the increase in the number of layers or increasing of the depth does not mean that the learning will increase at the same ratio too. As the depth increases, training becomes difficult. Therefore, the input and gradient values (vanishing gradients) begin to disappear [28]. In order to solve this problem, ResNet makes a shortcut connection from the input (x) to the output instead of mapping using the nonlinear $F(x)$ function in the classic CNN models. Thus, by skipping certain layers, the input (x) is added to the function F(x) as arithmetic $(F(x)+x)$. Figure 6 shows a sample of ResNet connection. Consequently, it carries out more effective training process. Resnet50 has a 50-layer structure and contains approximately 25.6 million parameters [29]. As it compared to other previous CNN models, it has a fairly low parameter value. There are also ResNet101 and ResNet152 models of ResNet.

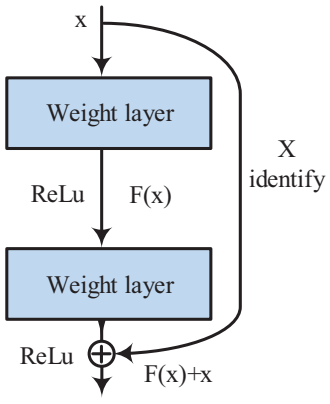

Figure 6. A sample of ResNet connection. [29]

Figure 6 shows a sample of ResNet connection. Consequently, it carries out more effective training process. Resnet50 has a 50-layer structure and contains approximately 25.6 million parameters. Compared to

other previous CNN models, it is a fairly low parameter value. There are also ResNet101 and ResNet152 models of ResNet. Deep features were extracted from each spectrogram image. While VGG16 and VGG19 obtain a 4096 dimensional feature vector from each image, ResNet obtains a 2048 dimensional vector [29, 30].

3.5. Classification

SVM and k-NN, which are frequently used in the literature, were used to classify DNA data. The following information for both classifiers is given in detail.

3.5.1. Support vector machines

Support vector machines are machine learning algorithms used in classification problems, and designed according to the principle of structural risk minimization. The SVM aims to find the most appropriate hyperplane to differentiate between classes for the classification of a 2-class data (e.g., exon and intron). In the classification of a two-class dataset, $\{x_i, y_i\}, i = 1, 2, 3, \ldots n$ $x_i \in R^d$ d-dimensional space of the training data, and $y_i \in -1, +1$ represents class labels. The SVM tries to find a distinctive hyperplane between the training data [31–33]. For a linearly separable two-class dataset, the hyperplane can be defined in Equation 3.

$$w.x_i + b \geq +1, \quad y = +1 \qquad w.x_i + b \leq -1, \quad y = -1 \tag{3}$$

Equation 4 is used to find a hyperplane that divides data with the same class label.

$$y_i(w.x_i + b) \geq 1 \tag{4}$$

The points forming these hyperplanes are called support vectors and these planes are expressed as $w.x_i + b = \pm 1$. In order to obtain the optimal hyperplane, $\parallel w \parallel$ should be minimized. In this case, the limited optimization problem in Equation 5 should be solved [33, 34].

$$\arg min \frac{1}{2} = \parallel w \parallel^2 \tag{5}$$

In cases where the data cannot be separated linearly, the optimization problem can be solved by adding slack variable (ξ);

$$min \left[\frac{\parallel w \parallel^2}{2} + C \sum_{i=1}^{n} \xi_i \right] \tag{6}$$

In Equation 6, α is Lagrange multiplier and x_i is support vector. As a result, the decision function for the two-class problem that can be separated linearly is as Equation 7 [34, 35].

$$y_i(w.x_i + b) - 1 + \xi_i \geq 0 \tag{7}$$

$$\xi_i \geq 0 \qquad i = 1, 2, 3, ..., n \tag{8}$$

Here, C is the penalty parameter. When it is not possible to define the hyperplane with linear equations, the data (x) is mapped to a high dimensional property field via some nonlinear mapping function $(\phi(x))$. This transformation makes it easier to find linear hyperplanes. The calculated cost of $(\phi(x).\phi(x_i))$ is reduced by

using the kernel function (K) [34, 35]. As a result, the decision function for the problem of two classes that can be separated linearly is as Equation 9.

$$f(x) = sign\left(\sum_{i=1}^{n} y_i \alpha_i K(x.x_i) + b\right) \tag{9}$$

In this study, radial basis function (RBF), polynomial, and linear that have the most commonly used core functions in the SVM classifier are used.

3.5.2. k-nearest neighbors algorithm

Another method for the classification is the k-nearest neighbor algorithm (k-NN). k-NN is a nonparametric method, which is used for the classification [37]. k is the neighboring number of new data to be classified. According to the $k-$NN method, if new data is mostly close to a group, it is included in that group. Euclidean distance measurement is generally used for proximity measurement. Thus, new data is added to the nearest class [38].

3.6. Performance metrics

In order to evaluate the performance of the proposed method k-fold cross-validation is used. The value of k was chosen as 10. Thus, the dataset was divided into 10 parts. Nine parts were used for the training, and the remaining part was used for the testing. This process was applied to all parts. The average of obtained 10 values was calculated for the performance evaluation. The classification performance of the models was compared for the sensitivity, specificity, accuracy, precision, and f1 score [39]. The parameters used to calculate the sensitivity, specificity, accuracy, precision, and f1 score are defined as follows. TP, FN, FP, and TN are as follows:

- *TP (True Positive)*: The number of correctly defined exons.
- *FN (False Negative)*: The number of incorrectly defined exons.
- *FP (False Positive)*: The number of positive examples misclassified as negative.
- *TN (True Negative)*: The number of incorrectly defined introns.

In Figure 7, flowchart of the proposed method is shown.

4. Application results and discussion

The proposed and developed classification approaches have been successfully implemented. One hundred and seventy-two DNA sequences including 86 exon and 86 intron data were used. After the exon and intron sequences were digitized by the entropy-based numerical mapping technique, they could be processed in deep learning models. For this, all the exon and intron data were converted to spectrogram images, and feature vectors were obtained. The obtained spectrogram images were 875×656 pixels. These images were resized to be 224×224 pixels for VGG16, VGG19, and ResNet models. Then, deep features were extracted from each spectrogram image. While VGG16 and VGG19 obtained a 4096-dimensional feature vector from each image, ResNet obtained a 2048-dimensional vector. Subsequently, the obtained feature vectors from spectrogram images were classified by SVM and k-NN. In order to process all the features in the classification as both training and test data,

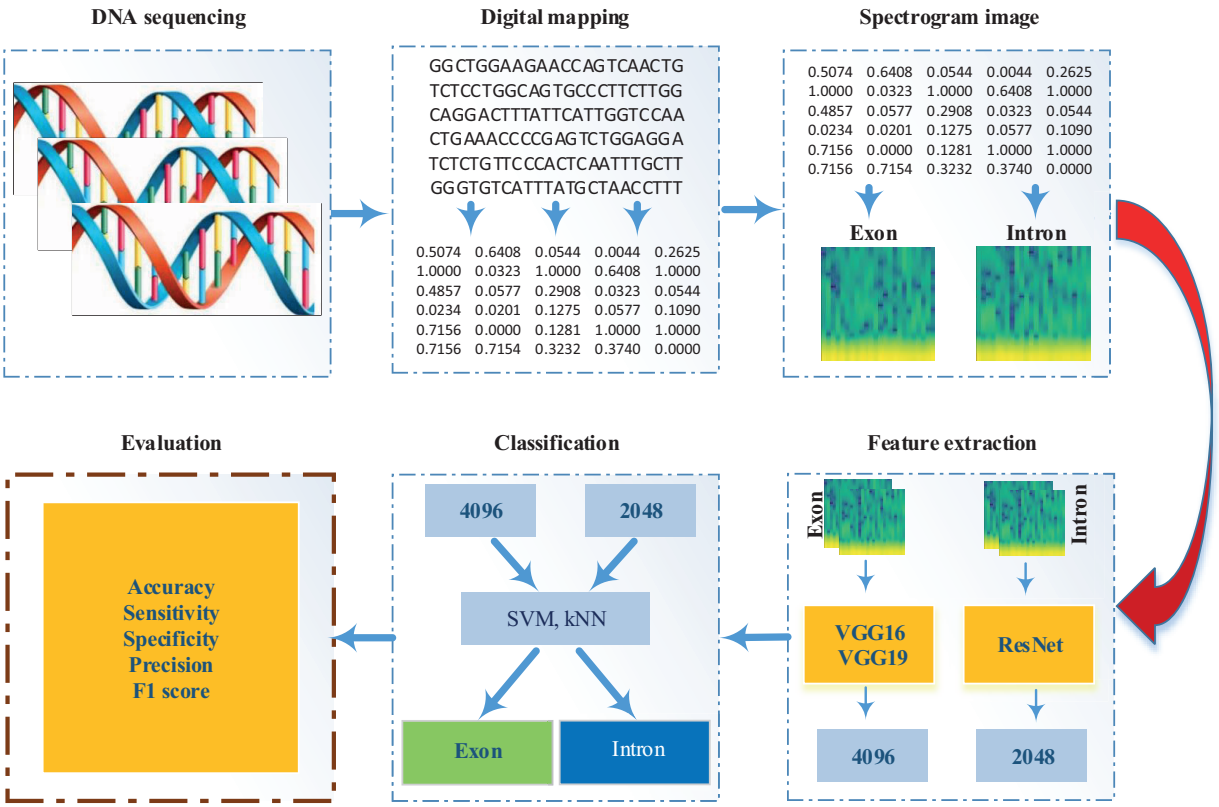

Figure 7. Flowchart of the proposed method.

Table 1. The accuracy values of EIIP, integer, and the entropy-based numerical techniques with VGG16, VGG19, and Resnet models applied with 10-fold cross validation.

Classifier	CNN Models	Integer	EIIP	Entropy-based technique
SVM	VGG16	0.724 ± 0.352	0.777 ± 0.405	0.976 ± 0.076
	VGG19	0.751 ± 0.432	0.741 ± 0.365	0.978 ± 0.074
	ResNet	0.769 ± 0.364	0.705 ± 0.475	0.977 ± 0.103
k-NN	VGG16	0.768 ± 0.213	0.738 ± 0.231	0.934 ± 0.080
	VGG19	0.737 ± 0.242	0.725 ± 0.214	0.940 ± 0.080
	ResNet	0.759 ± 0.189	0.757 ± 0.214	0.960 ± 0.082

Table 2. The best classification results for different kernel functions of SVM.

CNN Models	Integer	EIIP	Entropy-based technique	Parameters
VGG16	0.633 ± 0.131	0.637 ±0.293	0.960 ± 0.077	(C=1, rbf)
VGG19	0.749± 0.352	0.724 ± 0.330	0.965 ± 0.103	(C=1, poly)
ResNet	0.565 ± 0.276	0.612 ± 0.238	0.967 ± 0.133	(C=1, linear)

k-fold cross-validation method was used. In the study, exon and intron sequences were digitized by three numerical mapping techniques, and the feature extraction was carried out using three different CNN models.

Then, the data was classified with two different classifiers. K-fold cross-validation was used to classify data more objectively. The k value was determined as 10. The results of the classification are shown in Tables 1 and 2 in detail. In Table 1, it is seen that the entropy-based mapping technique is more successful than EIIP and integer techniques. The reason for the higher performance of the entropy-based mapping technique than others that the technique deepens the distinction rate of exon regions. The entropy technique better reflects the complexity of DNA sequences, and makes the digitization according to the repetition frequency of codons in a DNA sequence. Moreover, this mapping technique offers a wide correlation range over the gene sequence based on the repetition frequency of the codons. In addition, very close classification results were gotten in three different CNN models using the obtained data by the entropy-based technique. When the performance of SVM was compared with k-NN, the SVM achieved a better classification accuracy, whereas the k-NN classifier had lower accuracy, but k-NN classified the data with lower standard deviation values than SVM. According to the SVM classification results, all three CNN models have low accuracy in EIIP and integer mapping techniques, and the standard deviation values of EIIP and integer techniques are quite high. This shows that the classification methods cannot make a good classification with the available data. Whereas, according to the classification results of the entropy-based mapping technique in SVM, the accuracy values are quite high, and the standard deviation values are within the acceptable limits compared to the other techniques. When the k-NN classification results were examined, the classification accuracy of the EIIP decreased in VGG16 and VGG19, and increased in ResNet. In the integer mapping technique, while the accuracy values of VGG16 increased, the classification accuracy of VGG19 and ResNet decreased. However, an important point in the k-NN classification is that the standard deviation values are very low compared to the SVM classifier. The accuracy values that belong to the entropy-based mapping technique of the k-NN classifier are higher than the others, and lower than the accuracy values of the SVM classifier. The best classification performance with 97.8% was obtained in the

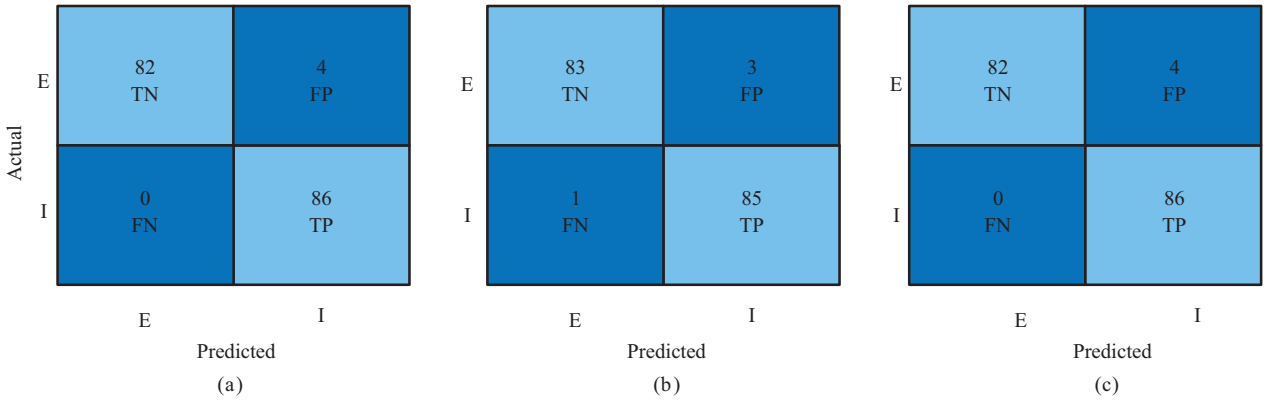

Figure 8. Confusion matrices of exon and intron data classified by SVM. a) with VGG16 b) with VGG19 c) ResNet classification results (E: Exon, I: Intron)

VGG19 model. The VGG19 is a deeper model than the VGG16, but compared to the ResNet, it is a shallower model. When the results are examined, it is seen that increasing the depth of the models will not always give a better result. It could be thought that the VGG19 model has a better feature vector than ResNet. That is why it gives a better result than the ResNet. However, this is different in EIIP for accuracy values of the k-NN classifier. Shortly, it should be kept in mind that which deep learning model gives a better classification result in which the dataset depends on the size of the data, the distinguishing feature of the data, the quality of the data, and the parameters of the model. Moreover, it should not be ignored that there may be differentiation

in the results when the number of data is increased. A confusion matrix is an effective tool that is often used to describe the performance of a classification model on testing data. The confusion is a table that has four different combinations that include matrix, estimation, and actual values. Figure 8 shows the confusion matrix values of the entropy-based numerical mapping technique of the SVM classifier. Although the three CNN models had very similar results, ResNet was better in terms of FN and TP values than VGG19. However, when the accuracy values are examined, it is seen that all three models are successful in extracting features from the dataset. The obtained ROC curves by classification of the entropy-based mapping technique using VGG16, VGG19, and ResNet models are shown in Figure 9. The ROC curve is used to illustrate the diagnostic ability of a classifier system. The area under the ROC curve of a test can be used as a criterion to measure the test's discriminative ability. The area under the ROC curve of the perfect test is 1. An ideal model has AUC close to 1, which means it has a good measure. The maximum AUC value with 99% was obtained in the VGG19 model. The obtained classification results with the developed method look promising. This means that the developed method can be used in practice. Accuracy, sensitivity, specificity, precision and f1 score

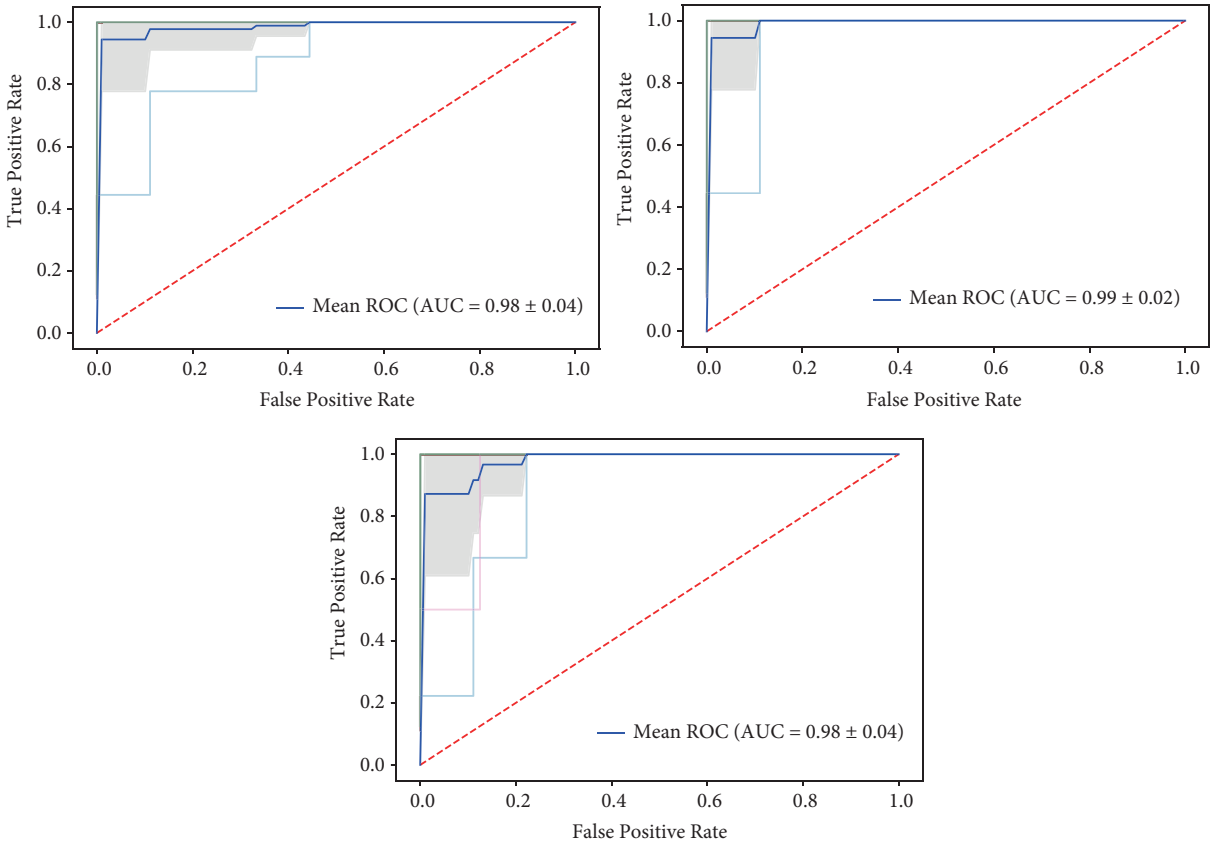

Figure 9. ROC curves for the 10-fold cross validation classification of the entropy-based mapping technique (VGG16, VGG19, and ResNet).

values of the entropy-based numerical technique are given in Table 2. According to the results of performance evaluation of the entropy-based numerical technique, the accuracy was calculated as 97.6% in VGG16, 97.8% in VGG19, and 97.7% in ResNet. It is seen that VGG19 has a better classification performance than others. The accuracy values of the SVM and k-NN classifiers are shown comparatively in Figure 10. The highest

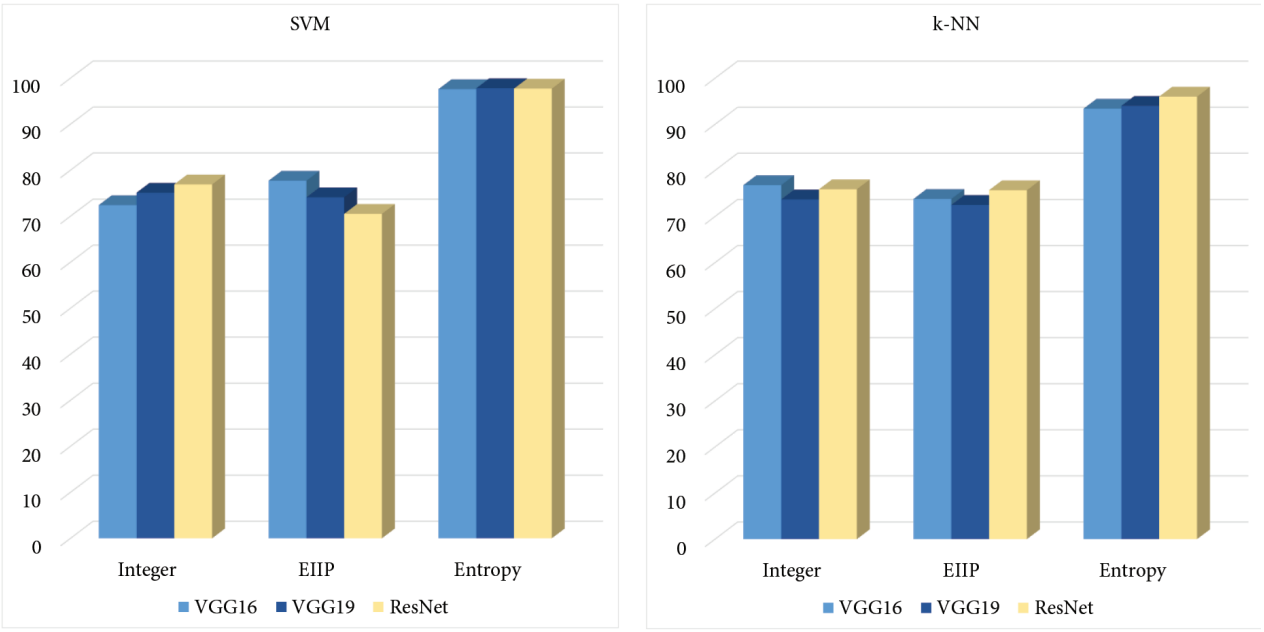

Figure 10. SVM and k-NN classification results of the property vectors obtained by VGG16, VGG19, and ResNet CNN models.

Table 3. Sensitivity, specificity, precision, and f1 score values of VGG16, VGG19, and ResNet models as a result of SVM classification using the integer, EIIP, and entropy-based numerical mapping method.

Techniques	Models	Accuracy (%)	Sensitivity(%)	Specificity(%)	Precision(%)	F1 score
Integer	VGG16	0.724 ± 0.352	72.4 ± 35.2	70.0 ± 46.4	73.3 ± 40.7	72.9 ± 35.4
	VGG19	0.751 ± 0.432	85.3 ± 38.5	65.0 ± 60.2	73.7 ± 45.5	78.2 ± 39.9
	ResNet	0.769 ± 0.364	81.9 ± 30.3	71.9 ± 52.4	77.5 ± 41.2	78.8 ± 32.7
EIIP	VGG16	0.777 ± 0.405	80.6 ± 34.0	74.9 ± 58.1	79.6 ± 42.4	79.2 ± 35.6
	VGG19	0.976 ± 0.076	81.9 ± 28.1	66.2 ± 56.2	74.4 ± 42.9	77.0 ± 32.1
	ResNet	0.741 ± 0.365	74.9 ± 38.4	66.1 ± 66.1	72.9 ± 52.9	72.9 ± 43.4
Entropy	VGG16	0.705 ± 0.475	100 ± 0.00	95.3 ± 15.3	96.0 ± 12.9	97.8 ± 7.00
	VGG19	0.978 ± 0.074	98.9 ± 6.70	96.7 ± 14.2	97.2 ± 11.9	97.9 ± 6.90
	ResNet	0.977 ± 0.103	100 ± 0.00	95.4 ± 20.6	96.4 ± 15.7	98.0 ± 8.90

accuracy in the classification was achieved in VGG19 with the RBF core function of SVM. C, one of the SVM parameters, was examined in range $[10^{-3}, \ldots, 10^{+3}]$. The k parameter of the k-NN classifier was examined in the range $[1, \ldots, 10]$. The k-NN results in Table 1 got the highest results for $k = 1$. The results of several other parameters of SVM in the study are shown in Table 2. Intel Core i5-4200 CPU and 8 GB RAM memory were used for the experimental results. Spectral images were created in Matlab platform. All data were processed using Keras library in Python platform. When the values of both classifiers were compared as shown in Table 1, SVM has better classification performance than k-NN. In Figure 11, the performance measures of all three CNN models in SVM classifier according to all three digitization techniques are shown graphically.

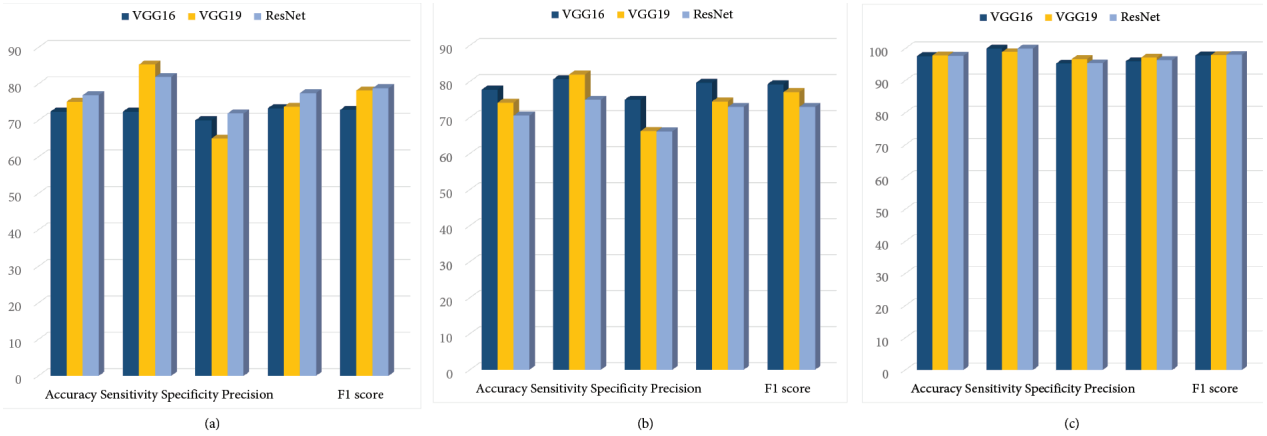

Figure 11. Performance evaluation graphs of VGG16, VGG19, and ResNet models according to a) Integer, b) EIIP, and c) the entropy-based technique.

5. Conclusion

In this study, a novel method in which two approaches are used together is proposed to classify exon–intron regions in DNA sequences. The first of these approaches is the developed entropy-based mapping approach to digitize DNA data. The second one is the automatic feature extraction from the digitized data. The classification of DNA sequences as exon or intron is very important for genome analysis. Therefore, we believe that we have contributed to the improvement of two basic process performances for the classification of exon and intron regions in the DNA sequence by the proposed method. In the future, we plan to develop a new deep learning model for the classification of the obtained data without any pretreatment after the digitization of DNA dataset.

References

[1] Cristea PD. Conversion of nucleotides sequences into genomic signals. Journal of Cellular and Molecular Medicine 2002; 6(2): 279-303. doi: 10.1111/j.1582-4934.

[2] Dougherty ER. Genomic signal processing. IEEE Signal Processing Magazine 2012; 29(3): 124-129.

[3] DeMaria AN. A structure for deoxyribose nucleic acid. JACC: Journal of the American College of Cardiology 2003; 373–374. doi: 10.1016/S0735-1097(03)00800-3.

[4] Koonin EV, Novozhilov AS. Origin and evolution of the genetic code: The universal enigma. IUBMB Life 2009; 61(2): 99-111. doi: 10.1002/iub.146.22

[5] Cristea PD. Genetic signal representation and analysis. In: SPIE Conference Biomedical Optics; Paris, France; 2002. pp. 77-84.

[6] Abo-Zahhad M, Ahmed SM, Abd-Elrahman AS. Genomic analysis and classification of exon and intron sequences using DNA numerical mapping techniques. International Journal of Information Technology and Computer Science 2012; 4(8): 22-36. doi: 10.5815/ijitcs.2012.08.03

[7] Abo-Zahhad M, Ahmed SM, Abd-Elrahman AS. A novel circular mapping technique for spectral classification of exons and introns in human DNA sequences. International Journal of Information Technology and Computer Science 2014; 4: 19-29. doi: 10.5815/ijitcs.2014.04.02

[8] Wang SY, Tian FC, Liu X, Wang J. A novel representation approach to DNA sequence and its application. IEEE Signal Processing Letters 2009; 16(4): 275:278. doi: 10.1109/LSP.2009.2014291.9

[9] Hota MK, Srivastava VK. Performance analysis of different DNA to numerical mapping techniques for identification of protein coding regions using tapered window based short-time discrete Fourier transform. In: International Conference on Power Control and Embedded Systems; Allahabad, India; 2010. pp. 1-4. doi: 10.1109/ICPCES.2010.5698675

[10] Crosby K, Gabbert P. BioSPRINT: classification of intron and exon sequences using the SPRINT algorithm. In: Computational Systems Bioinformatics Conference, CSB, Proceedings; Stanford, CA, USA; 2004. pp. 637-638. doi: 10.1109/CSB.2004.1332540.15

[11] Gupta R, Mittal A, Singh K, Bajpai P, Prakash S. A time series approach for identification of exons and introns. In: 10th International Conference on Information Technology (ICIT) 2007; Roukela, India; 2007. pp. 91-93. doi: 10.1109/ICOIT.2007.4418274

[12] Sahu SS, Panda G. Identification of protein-coding regions in DNA sequences using a time-frequency filtering approach. Genomics. Proteomics Bioinformatics 2011; 9(1-2): 45-55. doi: 10.1016/S1672-0229(11)60007-7

[13] Zhang WF, Yan H. Exon prediction using empirical mode decomposition and Fourier transform of structural profiles of DNA sequences. Pattern Recognition 2012; 45(3): 947-955 doi: 10.1016/j.patcog.2011.08.016.

[14] Sree PK, Rao PSVS, Devi NSSSNU. CDLGP: A novel unsupervised classifier using deep learning for gene prediction. In: 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI); Chennai, India; 2017. pp. 2811-2813. doi: 10.1109/ICPCSI.2017.8392232

[15] Sree PK, Usha Devi NSSSN, Sudheer MS. A robust deep learning mechanism augmented with cellular automata for DNA computing. In: IEEE International Conference Power, Control, Signals and Instrumentation Engineering (ICPCSI); Chennai, India; 2017. pp. 1305-1308. doi: 10.1109/ICPCSI.2017.8391921

[16] Das B, Turkoglu I. A novel numerical mapping method based on entropy for digitizing DNA sequences. Neural Computing and Applications 2018; 29(24): 207-215. doi: 10.1007/s00521-017-2871-5.

[17] Das B. Development of new approaches based on signal processing for disease diagnosis from DNA sequences. PhD, Fırat University, Elazığ, Turkey, 2018.

[18] Karci A. Fractional order entropy: New perspective. Optics 2016; 127(20): 9172-9177.

[19] Grandhi DG, Kumar CV. 2-Simplex mapping for identifying the protein coding regions in DNA. In: IEEE Region 10 Conference; Taipei, Taiwan; 2007. pp. 1-3. doi: 10.1109/TENCON.2007.4429086.

[20] Akhtar M, Epps J, Ambikairajah E. On DNA numerical representations for period-3 based exon prediction. In: IEEE International Workshop On Genomic Signal Processing and Statistics 2007. pp. 1-4. doi: 10.1109/GENSIPS.2007.4365821.

[21] Holden T, Subramaniam R, Sullivan R, Cheung E, Schneider C et al. ATCG nucleotide fluctuation of Deinococcus radiodurans radiation genes. Proc. SPIE Instruments, Methods, and Missions for Astrobiology X 2007; 6644: 669417. doi: 10.1117/12.732283

[22] Zhuang Z, Shen X, Pan W. A simple convolutional neural network for prediction of enhancer-promoter interactions with DNA sequence data. Bioinformatics 2019; 35(17). 2899-2906. doi: 10.1093/bioinformatics/bty1050

[23] Hasan MJ, Islam MMM, Kim JM. Acoustic spectral imaging and transfer learning for reliable bearing fault diagnosis under variable speed conditions. Measurement 2019; 138: 620-631. doi: 10.1016/j.measurement.2019.02.075

[24] Alom MZ, Taha TM, Yakopcic C, Westberg S, Sidike P et al. A state-of-the-art survey on deep learning theory and architectures. Electronics 2019; 8(3): 292. doi: 10.3390/electronics8030292

[25] Ullah I, Hussain M, Qazi EH, Aboalsamh H. An automated system for epilepsy detection using EEG brain signals based on deep learning approach, Expert Systems with Applications 2018; 107: 61-71. doi: 10.1016/j.eswa.2018.04.021

[26] Gopalakrishnan K, Khaitan SK, Choudhary A, Agrawal A. Deep convolutional neural networks with transfer learning for computer vision-based data-driven pavement distress detection. Construction and Building Materials 2017; 157: 322-330. doi: 10.1016/j.conbuildmat.2017.09.110.

[27] Zhu F. Estimating left ventricular volume with ROI-based convolutional neural network. Turkish Journal of Electrical Engineering and Computer Sciences 2018; 26(1): 23-34.

[28] Rizvi MdAI, Deb K, Khan MdI, Kowsar MMdS, Khanam, T. A comparative study on handwritten Bangla character recognition. Turkish Journal of Electrical Engineering Computer Sciences 2019; 27: 3195-3207. doi: 10.3906/elk-1901-4813

[29] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition; Las Vegas, NV, USA; 2016. e-ISSN: 1063-6919 doi: 10.1109/CVPR.2016.90

[30] Wu Z, Shen C, Hengel AV. Wider or deeper: revisiting the ResNet model for visual recognition. Pattern Recognition 2019; 90: 119-133. doi: 10.1016/j.patcog.2019.01.006

[31] Khazaee A, Ebrahimzadeh A. Classification of electrocardiogram signals with support vector machines and genetic algorithms using power spectral features. Biomedical Signal Processing and Control 2010; 5(4): 252-263. doi:10.1016/j.bspc.2010.07.006

[32] Toraman S, Girgin M, Ustundag B, Turkoglu I. Classification of the likelihood of colon cancer with machine learning techniques using FTIR signals obtained from plasma. Turkish Journal of Electrical Engineering and Computer Science 2019; 27 (3): 1765-1779. doi:10.3906/elk-1801-259

[33] Osuna E, Freund R, Girosi F. Support Vector Machines Training and Applications. Massachusetts Institude of Technology 1997; 1602: doi: 10.1.1.41.418

[34] Pal M, Mather PM. Support Vector classifiers for land cover classification. ArXiv 2008; doi: 10.1080/01431160802007624.

[35] Kavzoglu T, Colkesen I. A kernel functions analysis for support vector machines for land cover classi cation. International Journal of Application of Earth Observation Geoinformation 2009; 11(5): 352-359.

[36] Panda AK, Rapur JS, Tiwari R. Prediction of flow blockages and impending cavitation in centrifugal pumps using Support Vector Machine (SVM) algorithms based on vibration measurements. Measurement 2018; 130: 44-56. doi: 10.1016/j.measurement.2018.07.092.

[37] Das R, Sengur A. Evaluation of ensemble methods for diagnosing of valvular heart disease. Expert Systems with Applications, 2010; 37(7): 5110-5115,. doi: 10.1016/j.eswa.2009.12.085.

[38] Duda RO, Hart PE, Stork DG. Pattern Classification. 2nd ed. New York, NY, USA: Wiley, 2001.

[39] Das R. A comparison of multiple classification methods for diagnosis of Parkinson disease. Expert Systems with Applications 2010; 37(2): 1568-1572. doi: 10.1016/j.eswa.2009.06.040.