
**POLİTEKNİK DERGİSİ**

***JOURNAL of POLYTECHNIC***

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)
URL: http://dergipark.org.tr/politeknik

# Classification of exon and intron regions on DNA sequences with hybrid use of SBERT and ANFIS approaches

## *SBERT ve ANFIS yaklaşımlarının hibrit kullanımı ile DNA dizilimleri üzerindeki ekson ve intron bölgelerinin sınıflandırılması*

***Yazar(lar) (Author(s))***: *Fatma AKALIN[1], Nejat YUMUŞAK[2]*

*ORCID[1]: 0000-0001-6670-915X*

*ORCID[2]: 0000-0001-5005-8604*

**To cite to this article**: Akalın F., and Yumuşak N., "Classification of exon and intron regions on DNA sequences with hybrid use of SBERT and ANFIS approaches", ***Journal of Polytechnic***, 27(3): 1043-1053, (2024).

**Bu makaleye şu şekilde atıfta bulunabilirsiniz:** Akalın F., and Yumuşak N., "Classification of exon and intron regions on DNA sequences with hybrid use of SBERT and ANFIS approaches", ***Politeknik Dergisi***, 27(3): 1043-1053, (2024).

**Erişim linki (To link to this article):** http://dergipark.org.tr/politeknik/archive

**DOI:** **10.2339/politeknik.1187808**

# Classification of Exon and Intron Regions on DNA Sequences with Hybrid Use of SBERT and ANFIS Approaches

## *Highlights*

- *The effect of clustering approaches on cost reduction*
- *Simplification of analysis of exon regions with statistical inferences obtained by repetition frequency of codes*
- *Classification of the DNA structure that has fuzzy configurations, with the ANFIS structure, that is formed by the combination of artificial neural networks and fuzzy inference system*
- *Impact of artificial intelligence approaches on the analysis of increasing genome sequences that have been easily accessed in recent years*

## *Graphical Abstract*

*This study provides a molecular diagnosis to distinguish exon and intron DNA sequences. Symbolic DNA sequences regarding exon and intron regions are given below.*

| *Exon Region* | *Intron Region* |
|---|---|
| *AAACAAACTGAAGCGCTGAAGCAGCGGGTGCAG* | *GTAGGAGAAAGGTCATGGCAGGCCCCCCAG* |
| *AGGAAGCTGGAGCAGGTGTACTACTTCCTGGAG* | *GCTCTGTGCGTGACTCATTGACTGAGTTGAC* |
| *CAGCAAGAGCATTTCTTTGTGGCCTCACTGGAG* | *TCATTAGACCACAGTCCCCAACATGGCCTGG* |
| *GACGTGGGCCAGATGGTTGGGCAGATCAGGAA* | *GTTCCTGGGAGGAACGGGATTATACCCAACA* |
| *GGCATATGACACCCGCGTATCCCAGGACATCGC* | *TAGCATGCAGGGCCCTAAGCAGGGGGTTCCT* |
| *CCTGCTCGATGCGCTGATTGGGGAACTGGAGGC* | *TGTCTTTCCTTGTTGTCAGGACAGTGTAATTT* |
| *CAAGGAGTGCCAGTCAGAATGGGAACTTCTGCA* | *AGCCCCTCTTAATGCTAATGCTCAGGATTTTT* |
| *G* | *TTCCCTATCTGATTTTTCTCCGTAG* |

**Figure.** Symbolic DNA sequences regarding exon and intron regions

## *Aim*

*Classification of exon and intron regions on DNA sequences.*

## *Design & Methodology*

*In this study, the analysis of exon and intron regions in the DNA sequence was carried out with artificial intelligence-based systems.*

## *Originality*

*The clustering approach which is generally preferred for evaluations of textual data was used on DNA sequences. This situation has reduced the computational cost.*

## *Findings*

*As a solution to the increasing amount of data in the field of bioinformatics, an artificial intelligence-based structure was built that offers low cost. Thus, it became easier to investigate situations related to genetics.*

## *Conclusion*

*The exon and intron regions on the DNA structure were classified with an accuracy rate of 88.88%.*

## *Declaration of Ethical Standards*

*The author(s) of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.*

# Classification of Exon and Intron Regions on DNA Sequences with Hybrid Use of SBERT and ANFIS Approaches

*Araştırma Makalesi/Research Article*

**Fatma AKALIN[1*], Nejat YUMUŞAK[2]**

[1]Faculty of Computer and Information Sciences, Information Systems Engineering Department, Sakarya University, Türkiye
[2]Faculty of Computer and Information Sciences, Computer Engineering Department, Sakarya University, Türkiye

(Geliş/Received : 12.10.2022 ; Kabul/Accepted : 06.02.2023 ; Erken Görünüm/Early View :12.03.2023)

**ABSTRACT**

DNA is the part of the genome that contains enormous amounts of information related to life. Amino acids are formed by coding three nucleotides in this genome part, and the encoded amino acids are called codes in DNA. The frequency of the triple nucleotide in the DNA sequence allows for the evaluation of protein-coding (exon) and non-protein-coding (intron) regions. Distinguishing these regions enables the analysis of vital functions related to life. This study provides the classification of exon and intron regions for BCR-ABL and MEFV genes obtained from NCBI and Ensemble datasets, respectively. Then, existing DNA sequences are clustered using pretrained models in the scope of the SBERT approach. In the clustering process, K-Means and Agglomerative Clustering approaches are used consecutively. The frequency of repetition of codes is calculated with a representative sample selected from each cluster. The matrix is created using the frequencies of 64 different codons that constitute genetic code. This matrix is given as input to the ANFIS structure. The %88.88 accuracy rate is obtained with the ANFIS approach to classify exon and intron DNA sequences. As a result of this study, a successful result was produced independently of DNA length.

**Keywords: DNA sequences, exon and intron regions, k-means clustering and agglomerative clustering, SBERT, ANFIS.**

# ANFIS ve SBERT Yaklaşımlarının Hibrit Kullanımı ile DNA Dizilimleri Üzerinde Ekson ve İntron Bölgelerinin Sınıflandırılması

**ÖZ**

DNA, canlılığa ilişkin devasa bilgi barındıran genom parçasıdır. Bu genom parçasındaki üç nükleotidin kodlanması ile aminoasitler oluşur ve kodlanan aminoasitler DNA'da kod olarak isimlendirilir. DNA dizilimindeki üçlü nükleotidin frekansı, protein kodlayan(ekson) ve protein kodlamayan(intron) bölgelere ilişkin analiz imkanı sağlar. Bu bölgelerin ayırt edilmesi yaşama ilişkin hayati fonksiyonların değerlendirilmesini mümkün kılar. Bu çalışma sırasıyla NCBI ve Ensemble veri setlerinden elde edilen BCR-ABL ve MEFV genleri için ekson ve intron bölgelerinin sınıflandırılmasını sağlamıştır. Ardından SBERT yaklaşımı kapsamında önceden eğitilmiş modeller ile mevcut DNA dizilimleri kümelenmiştir. Kümeleme sürecinde K-Means ve Agglomerative Kümeleme yaklaşımları art arda kullanılmıştır. Her bir kümeden seçilen temsili bir örnek ile kodonların tekrarlanma sıklığı hesaplanmıştır. Genetik kodun oluşmasını sağlayan 64 farklı kodonların frekansı kullanılarak matris oluşturulmuştur. Bu matris ANFIS yapısına girdi olarak verilmiştir. ANFIS yaklaşımı ile ekson ve intron bölgelerinin sınıflandırılmasında %88.88 doğruluk oranı elde edilmiştir. Bu çalışmanın sonucunda DNA uzunluğundan bağımsız başarılı bir sonuç üretilmiştir.

**Anahtar Kelimeler: DNA dizilimleri, ekson ve intron bölgeleri, k-ortalama kümeleme ve birleştirici kümeleme, SBERT, ANFIS.**

## 1. INTRODUCTION

Bioinformatics is a modern interdisciplinary field that includes the study of biology, computer science and statistics sciences[1][2]. It obtains the relations and dynamic interactions between biological elements or events[1]. It offers rapid analysis and evaluation for increased genome sequences that have become easily accessible in recent years[3]. Gene prediction and functional annotation are important processes in evaluating these genome sequences. Gene prediction is the localization or restriction of different genes on the genome. Functional annotation assigns the biological function or structural features for decided genes[3]. Microarray gene expression analysis, gene regulatory network inference and gene biomarkers identification are other application areas of bioinformatics[1]. In this framework, different bioinformatics tools and software have been developed in the past[3]. In this study, exon (protein-coding) and intron (non-protein-coding) regions of gene structures[1] expressing a particular part of the DNA molecule in the cells of living organisms are

*Sorumlu Yazar (Corresponding Author)
e-posta : fatmaakalin@sakarya.edu.tr

classified using the proposed artificial intelligence-based system.

A gene refers to a region in DNA[4]. Gene regions of eukaryotic organisms have exon (protein-coding region) and intron (non-protein-coding region) region separation[4]. However, less than %5 of the DNA structure consists of protein-coding sequences. The remaining, about %95 part contains non-protein-coding and untranslated regions[3]. For this situation, which indicates the existence of a complex genome structure[3], the distinguishing of exon and intron regions has been expressed as a challenging problem in [5][6][7][8] studies.

The inference can be made using exon regions to determine the mutated DNA sequences, the coding status of the protein, the regulation of the developmental process, information about which tissues and organs stem cells will turn into, and the conditions of proliferation and death of the cells. In addition, the developmental status of cancer can also be investigated with protein-coding exon regions. Therefore, high accuracy classification of exon and intron regions is a critical issue in making sense of biological processes[9]. However, DNA has a long and double helix structure consisting of nitrogenous bases Adenine (A), Thymine (T), Guanine (G) and Cytosine (C). It stores genetic information here[4]. In [9][10][11][12] studies, DNA sequences are digitized with traditional or proposed numerical mapping techniques to make inferences on this structure consisting of symbolic characters. After this preprocessing step, the classification of the target regions is provided. On the other hand, the [13][14] studies specify that the inference is made depending on the length of DNA sequences. Different studies have been carried out in the literature in order to evaluate DNA sequences. In the [15] study, CNN, CNN-LSTM and CNN-Bidirectional LSTM architectures were applied after label and kmer coding for DNA sequence classification. The maximum accuracy rate was obtained as %93.16 with the kmer encoded CNN approach. The [16] study was performed to predict the gene sequence that causes prostate cancer. First, exon regions were extracted from different prostate gene sequences. Then, kmer coding for these DNA sequences and one-hot encoding for class tags were made. The created bi-LSTM model showed a success rate of %95. The study [3] expressed that splicing site determination and prediction processes in eukaryotic DNA sequences are difficult. In order to find a solution to this problem, a bidirectional Long Short Term Memory (LSTM)-Recurrent Neural Network (RNN) based deep learning model was proposed. This proposed model reached an accuracy rate of %95.5. In addition, in the [17] study that handles the same problem, a deep learning model based on bidirectional Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN) and Gated recurrent unit (GRU) were proposed. The maximum success rate for this model was found to be %96.1. The [18] study aimed the classification human exon and intron sequences. Therefore it expressed the text format of DNA sequences as images using percentage calculation. Then images were classified using CNN models. The achieved average maximum accuracy is %90.11. In the [8] study, which was carried out for the same purpose, statistical information was extracted from DNA sequences with a wavelet-based time series approach. The feature vector was created by using the variance information of the biomarkers that are important in the evaluation of the sequences. Then, exon and intron regions were classified by the optimized support vector machine method. An accuracy rate of %88.95 is achieved on the test dataset. In the field of bioinformatics, genomic signal processing is another preferred approach to obtain relationships, patterns and periodicity between data[19]. In this context, signal processing techniques are used for the analysis of DNA sequences in [11][20][21] studies and exon regions are distinguished.

In this study, in the scope of the SBERT approach, K-Means and Agglomerative Clustering methods were used consecutively on DNA sequences. Thus, the clustered DNA sequences were obtained. Then, representative samples selected on these clusters were classified by the artificial intelligence-based ANFIS method. By this hierarchy, a structure independent of DNA length was created without a preprocessing step.

## 2. MATERIAL AND METHOD

This section presents our proposed approach to classify exon and intron regions. In the direction of our proposed approach, the format of the dataset, clustering process and classification method in our study are explained below.

### 2.1. Dataset

In this study, two public gene banks named NCBI (National Biotechnology Information Bank) [22], and Ensemble [23] are used to distinguish exon and intron DNA sequences. First, 4 different BCR-ABL fusion genes numbered "AM400881.1", "AM600680.1", "AM886138.1", and "EU447303.1" from the NCBI database are used. These genes contain 5 different exon regions and 6 different intron regions. In the genes used, the shortest exon sequence consists of 15 nucleotides; the longest exon sequence consists of 66 nucleotides; the shortest intron sequence consists of 57 nucleotides, and the longest intron sequence consists of 549 nucleotides.

Secondly, DNA sequences numbered "MEFV ENSG00000010331" are obtained from the Ensemble database. 61 exons and 34 intron regions are used from this structure containing DNA sequences in different lengths. The shortest exon sequence consists of 23 nucleotides, the longest exon sequence consists of 554 nucleotides, the shortest intron sequence consists of 165 nucleotides, and the longest intron sequence consists of 468 nucleotides. Examples related to exon and intron DNA sequences are given in Table 1.

**Table 1.** Symbolic DNA sequences regarding exon and intron regions [23]

| Exon Region | Intron Region |
|---|---|
| AAACAAACTGAAGCGCTGAAGCAGCGGGTGCAGAGGAA | GTAGGAGAAAGGTCATGGCAGGCCCCCCAGGCTCTG |
| GCTGGAGCAGGTGTACTACTTCCTGGAGCAGCAAGAGCA | TGCGTGACTCATTGACTGAGTTGACTCATTAGACCAC |
| TTTCTTTGTGGCCTCACTGGAGGACGTGGGCCAGATGGTT | AGTCCCCAACATGGCCTGGGTTCCTGGGAGGAACGG |
| GGGCAGATCAGGAAGGCATATGACACCCGCGTATCCCAG | GATTATACCCAACATAGCATGCAGGGCCCTAAGCAG |
| GACATCGCCCTGCTCGATGCGCTGATTGGGGAACTGGAG | GGGGTTCCTTGTCTTTCCTTGTTGTCAGGACAGTGTAA |
| GCCAAGGAGTGCCAGTCAGAATGGGAACTTCTGCAG | TTTAGCCCCTCTTAATGCTAATGCTCAGGATTTTTTTC |
| | CCTATCTGATTTTTCTCCGTAG |

Modelling of biological features and biomarkers associated with DNA sequences[18], classification of DNA sequences[15] and identification of exon and intron regions [8] are complex and challenging work. Therefore, in this study, a simple hierarchy independent of DNA length was constructed.

### 2.2. Sentence Bert (SBERT) Model

The increase in the number of data in recent years is important for analyzes that offer fast, accurate and low computational costs, and this situation can be tolerated using approaches in the framework of artificial intelligence[18]. For example, natural language representations are a deep learning area that is widely used in the pre-training process of data[24]. In this context, the BERT (Bidirectional Encoder Representations from Transformers) model is proposed[25]. BERT is a pretrained language representation. It consists of several transformers and encoder layers. Thus, it provides obtaining both token-level and sentence-level language features[24]. Figure 1 shows the structure of the BERT model[26].

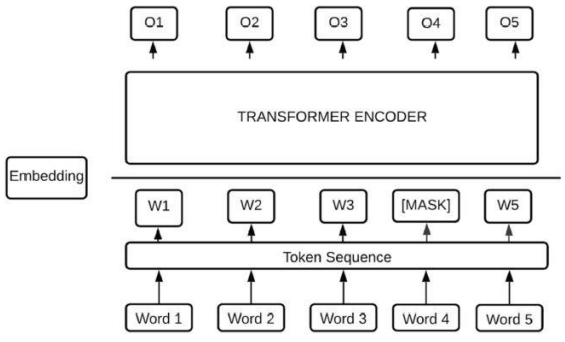


**Figure 1.** Structure of the BERT[26]

BERT successfully manages many natural language processing processes. However, it has weaknesses in applicability to real-world problems[24]. For this reason, the SBERT method, in which the BERT method is modified, is used in the current study[27]. SBERT uses a concatenation architecture to increase the power of sentence similarity-based calculations.

SBERT developed on the Siamese network architecture consists of two BERT networks. In this structure, the pooling process is applied to the output of the BERT network. Thus, fixed-size sentence embeddings are produced. Cosine distance is calculated between two sentence embeddings characterized as similarity ratio u and v[24]. Figure 2 shows the structure of the SBERT model[27].

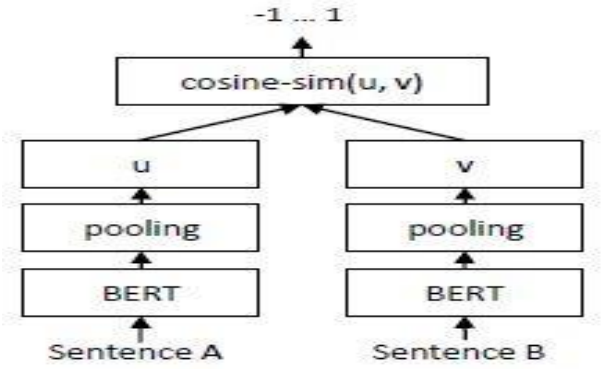


**Figure 2.** Structure of the SBERT model [27]

The mathematical expression of cosine similarity is given in equation 1.

$$\text{SimCos}(d,q) = \frac{\Sigma(P_{(n,d)} * P_{(n,q)})}{\sqrt{\Sigma(P_{(n,d)})^2 * \Sigma(P_{(n,q)})^2}} \qquad (1)$$

P, n, d and q symbols defined in the equation are expressed as document weight, the number of terms, document and query, respectively[26].

Similarity defines that common features in the two cases, the magnitude calculated by distance, and a threshold value determined to decide whether they are similar or not[26]. SBERT performs different tasks such as cosine similarity, semantic textual similarities, semantic searches and paraphrase mining[26]. In this study, unlike the BERT model, the SBERT model, which also performs the clustering task, was used.

### 2.3. Adaptive Neuro Fuzzy Inference System-ANFIS

ANFIS is an artificial neural network technique. It is formed with the combination of artificial neural networks and a fuzzy inference system[28][29]. Takagi-Sugeno ANFIS structure is the preferred fuzzy inference system in this study. The network type of the ANFIS model, which is based on the Takagi Sugeno fuzzy inference

system, resembles a neural network structure in which fuzzy rules are changed with neurons in the hidden layers of the fuzzy inference system.

In fuzzy mathematics, the fuzzifier module maps precise input models to fuzzy sets characterized according to the chosen membership function (triangle, trapezoidal, gaussian, sigmoidal, etc.). Neuro-adaptive techniques can provide a learning strategy for fuzzy rule-based systems. This situation is realized by learning the appropriate rules from the input-output pairs. At the same time, ANFIS has the ability to set membership functions with result parameters from the input output dataset. In this study, the gaussian membership function was chosen[28].

ANFIS provides a result containing a random linear function. Its mathematical expression is expressed in equation 2 [28].

$R^j = \text{IF } x_1 \text{ is } A_1^j \text{ AND } x_2 \text{ is } A_2^j \text{ AND...AND } x_n \text{ is } A_n^j$
$\text{THEN } y^j = a_0^j + a_1^j x_1 + ... + a_n^j x_n$ (2)

$R^j$ given in equation 2 is the $j^{th}$ fuzzy rule. The $k^{th}$ input variable of the N-dimensional input vector is $x_k$, and the fuzzy membership function associated with $x_k$ for the $j^{th}$ fuzzy rule is $A_k^j$. The expression of $a_0^j \ldots\ldots a_n^j$ in the equation defines linear coefficients. In this study, the gaussian membership function is chosen to qualified of the input models as fuzzy rules. A representative ANFIS architecture is given in Figure 3[28].

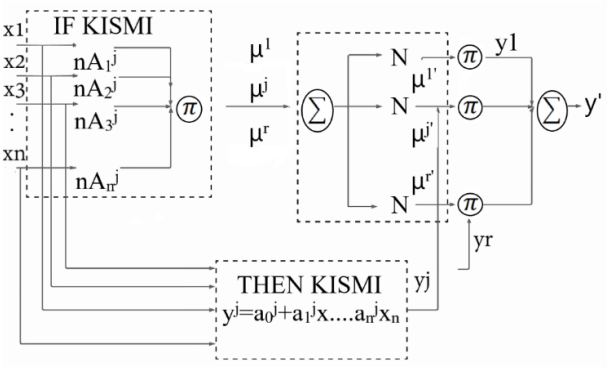


**Figure 3.** Representative ANFIS structure [28]

In the first layer of this structure, the membership function selected as GMF is blurred as in equation 3[28].

$$\eta_{A_k^j} = \exp[-0.5(\frac{x_k - c_k^j}{\sigma_k^j})^2] \quad (3)$$

The $c_k^j$ and $\sigma_k^j$ given in the equation represent the center and width of the $j^{th}$ GMF for the kth input variable. The second and third layers are where the firing power and the normalized version are calculated. The output obtained for each rule with the AND operation in the antecedent part is shown in equation 4[28].

$$\mu^j = \prod_{k=1}^{n} \eta_{A_k^j} \quad (4)$$

The normalized result of the $j^{th}$ rule is given in equation 5[28].

$$\mu'^j = \frac{\mu^j}{\sum_{j=1}^{R} \mu^j} \quad (5)$$

The $\mu'^j$ given in equation 5 is the firing power and decides the amount of contribution for each rule at the final network output as shown in equation 6[28].

$$y' = \sum_{j=1}^{R} \mu'^j y^j \quad (6)$$

The mathematical calculation of the results is done in the fourth layer. In the last step, the defuzzification step in the ANFIS framework is performed with the linear combination of the conclusion part [28].

## 3. PROPOSED METHODOLOGY

In this study, a hybrid structure is proposed which clustering and classification processes are used consecutively to classify exon and intron DNA sequences. In this context, firstly, the K-Means pretrained clustering model was applied to the existing DNA sequences. The number of clusters was determined as 2, and the all-MiniLM-L12-v2 pretrained model was used. Then, these two clusters with exon and intron regions were given as input to the Agglomerative Clustering method. Unlike the K-Means method, the Agglomerative Clustering method performs the clustering process using the threshold value. Thus, clusters below the determined threshold value are combined[30]. The paraphrase-albert-small-v2 pre-trained model was used for the Agglomerative Clustering method. The pretrained models used in the study were decided making various trials. All informations related pretrained models are available at [31].

The clusters obtained after the clustering process on the BCR-ABL gene that is used 7 exons and 6 intron regions and the MEFV gene that is used 61 exons and 34 intron regions are depicted in detail in Figure 4 and Figure 5.

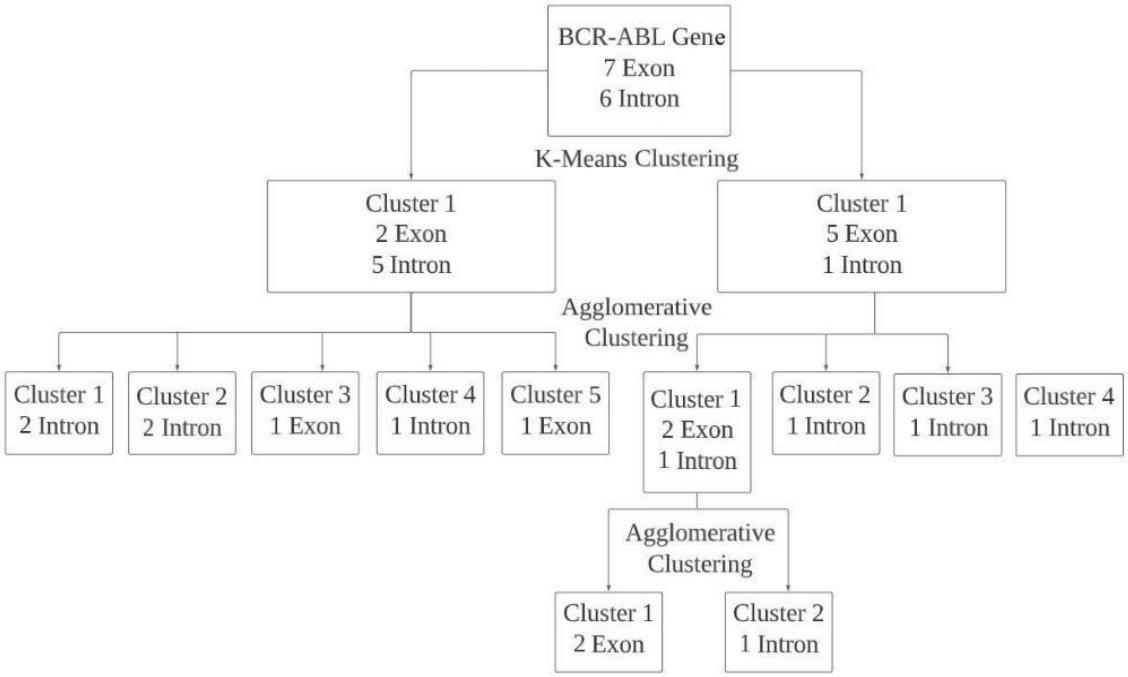


**Figure 4.** Clustering process performed on the BCR-ABL gene with consecutive use of K-Means and Agglomerative Clustering algorithms

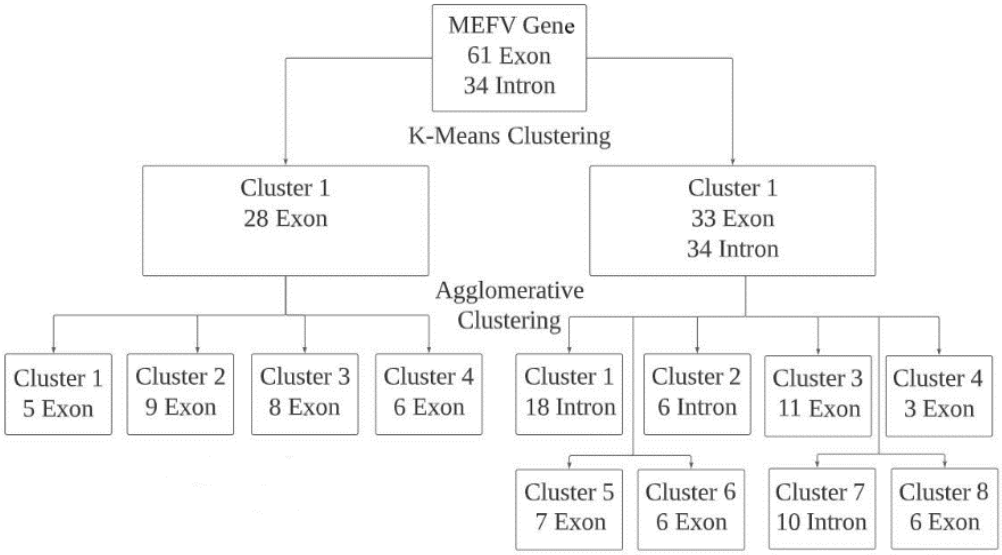


**Figure 5.** Clustering process performed on the MEFV gene with consecutive use of K-Means and Agglomerative Clustering algorithms

The Agglomerative Clustering method, which was used after the K-means Clustering method, was applied until it was stated that the reached clusters contained data related to the same class. This situation is explained by reproducing the same cluster output with the Agglomerative method applied to the last clusters. Then, a representative sample was selected from each cluster. Finally, classification was carried out by extracting features from the DNA sequences in these selected samples. The flow chart of this proposed approach is given in Figure 6.

After the clustering methods applied to the exon and intron DNA sequences, the final cluster outputs that could not be further divided for the determined threshold value were obtained. Then, the frequencies of triple nucleotides called as codes in each DNA sequence were calculated. Thus, a labelled matrix structure consisting of 22 rows (10-BCR-ABL gene + 12-MEFV gene) and 64+1 columns was created. The matrix's 40% and 60% were separated as test and training sets, respectively. In the last step, classification process was provided with ANFIS structure.

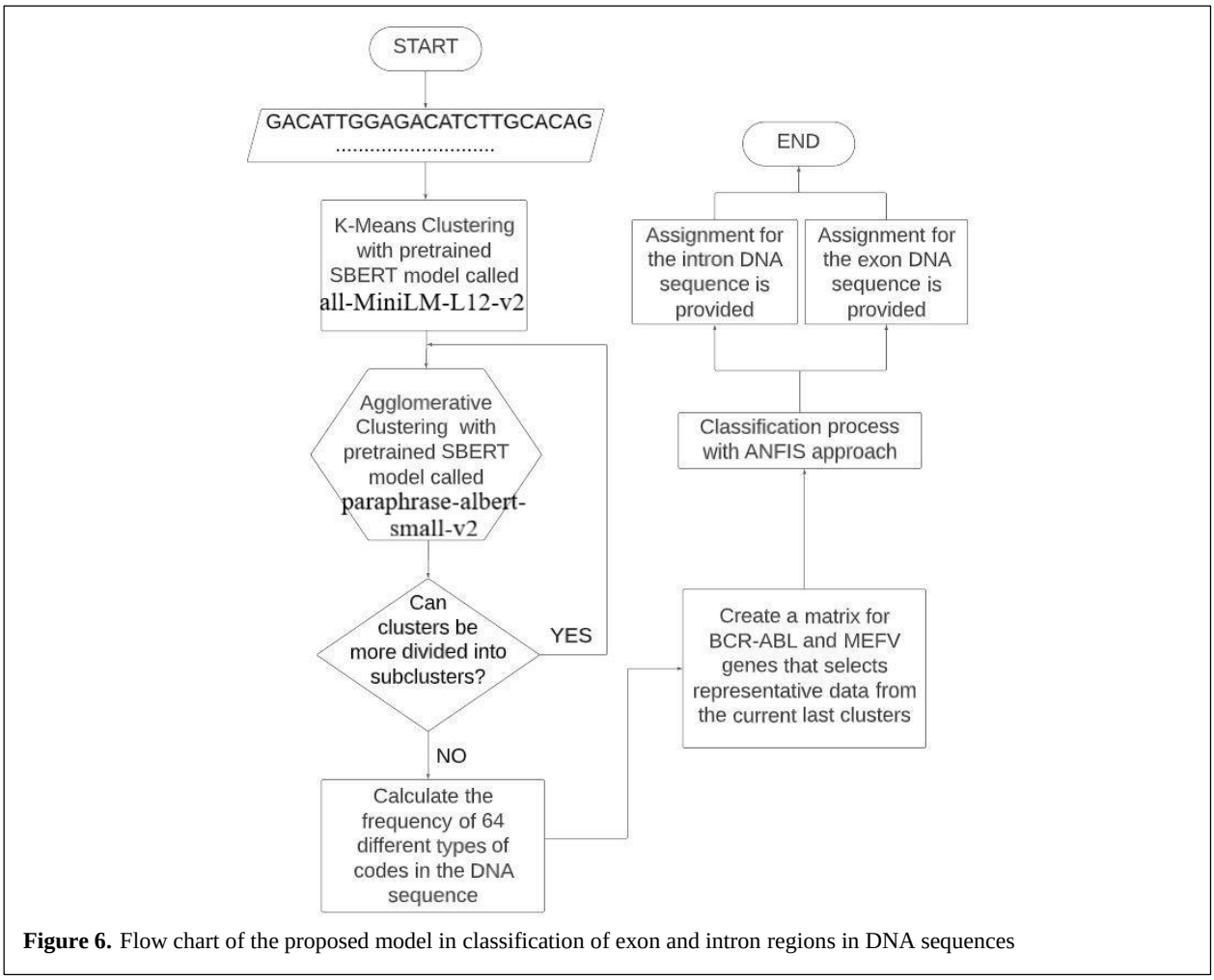


**Figure 6.** Flow chart of the proposed model in classification of exon and intron regions in DNA sequences

## 4. DISCUSSION AND RESULTS

Amino acids that play a role in protein synthesis are formed by the coding of three nucleotides. Triple nucleotides in DNA are called code. During mRNA synthesis in DNA, these triple nucleotides are converted to codons in RNA. Code and codon structures are complementary to each other, and there are 64 different codons in RNA. The genetic code consists of codes[32], and the repetition frequency of codes enables analysis of protein-coding regions. For this reason, a representative sample was selected from the last clusters obtained after the clustering process, and classification was carried out. Clustering performed in the first stage of this study offers a process that reduces the computational cost despite the increasing number of data. This situation explains that it is classified with 10 different samples of the BCR-ABL gene, which was studied on 13 separate regions, and with 12 different samples of the MEFV gene, which was studied on 95 separate regions. In the first stage, the analysis that reduces the computational cost was provided for the increasing genome sequences that have been easily accessed in recent years [3].

In the second stage of the study, the repetition frequency of 64 different codes was calculated in order to identify the protein-coding regions. Then, a total of 22 rows and 64+1 columns labelled matrix consisting of 10 rows of 64 columns for the BCR-ABL gene and 12 rows of 64 columns for the MEFV gene was obtained and given as input to the ANFIS structure. Despite the fuzzy configuration of the DNA structure, ANFIS is a structure that can map clear input patterns with its fuzzing module[28]. Therefore it will assist in the identification of exon and intron regions. The confusion matrices of outputs produced on the data tested after the training are given in Figure 7.

| | | TRAIN DATASET | | | | TEST DATASET | |
|---|---|---|---|---|---|---|---|
| | | Predicted | | | | Predicted | |
| | | EXON | INTRON | | | EXON | INTRON |
| Actual | EXON | 7 | 0 | Actual | EXON | 5 | 0 |
| | INTRON | 0 | 6 | | INTRON | 1 | 3 |

**Figure 7.** Confusion matrices about to training dataset and test dataset

After the training process, an accuracy rate of %100 and %88.88 was achieved on the training dataset and test dataset, respectively. Accuracy rate (Acc.) identicates the ratio of correctly predicted exon and intron regions to all images. Its mathematical expression is given in equation 7.

$$Acc = (TP+TN) / (TP + FP + TN + FN) \quad (7)$$

However, in order to evaluate the performance of the proposed hybrid structure correctly, the criteria of True Negative, False Positive, False Negative, True Positive, Sensitivity, Specificity, Precision and F score should also be examined. Explanations about the evaluation criteria are given below.

True Negative (TN) is that the DNA sequence decided as exon by experts is decided as exon also by the ANFIS model. True Positive (TP) is that the DNA sequence decided as intron by experts is decided as intron also by the ANFIS model. False Positive (FP) expresses the characterization of the exon region as an intron region by the ANFIS model. False Negative (FN) expresses the characterization of the intron region as an exon region by the ANFIS model. Specificity (Spec.) is the ratio of exon DNA sequences correctly predicted by the ANFIS model to all exon DNA sequences[14]. Its mathematical expression is given in equation 8.

$$Spec = TN / (TN+FP) \quad (8)$$

Sensitivity (Sens.) is the ratio of intron DNA sequences correctly predicted by the ANFIS model to all intron DNA sequences[14]. Its mathematical expression is given in equation 9.

$$Sens = TP / (TP+FN) \quad (9)$$

Precision (Prec.) expresses the ratio of intron DNA sequences correctly predicted by the ANFIS model to all intron predictions[14]. Its mathematical expression is given in equation 10.

$$Prec = TP / (TP+FP) \quad (10)$$

The F score (F Sc.) is the harmonic mean of precision and sensitivity values. It provides that extreme values are considered[14]. Its mathematical expression is given in equation 11.

$$F\ Sc = (2* Precision * Sensitivity) / (Precision + Sensitivity) \quad (11)$$

The TN, FP, FN, TP, Precision, Sensitivity, Specificity and F score evaluation criteria reached by the trained model in the training dataset and test dataset were given in Table 2.

**Table 2.**Evaluation criteria about recognition of exon and intron regions

| Datasets/Criteria | TN | FP | FN | TP | Prec. | Sens. | Spec | Acc. | F Sc. |
|---|---|---|---|---|---|---|---|---|---|
| Train | 7 | 0 | 0 | 6 | %100 | %100 | %100 | %100 | %100 |
| Test | 5 | 0 | 1 | 3 | %100 | %75 | %100 | %88.88 | %85.71 |

These ratios in table 2 show that strong estimates are produced in classifying exon and intron regions accurately. Also, it is shown in Figure 8 that the exon and intron regions predicted by the model produce high detection accuracy on both the training and test data set.intron regions predicted by the model produce high detection accuracy on both the training and test data set

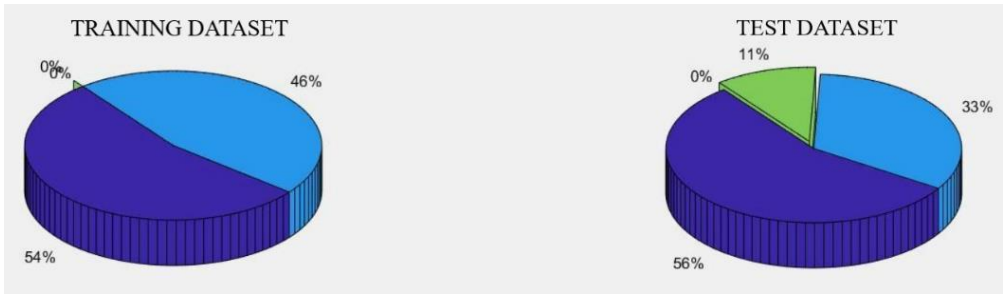


**Figure 8.** Accuracy percentages of exon and intron regions predicted by the constructed model

These percentiles in Figure 8 show that approximately 54% and 56% correct predictions were made for the accurate detection of exon regions by the model in the training set and test set, respectively. On the other hand, approximately 46% and 33% correct predictions were made for the accurate detection of intron regions by the model in the training set and test set, respectively. In the training set, no determination of exon regions as introns and intron regions as exons was made by the model. However, a region in the test set that should have been defined as an intron was expressed as an exon and constituted the incorrectly predicted 11% percentile. The description of this situation is expressed in the graphs in Figure 9 for each sample in the training set and test set.

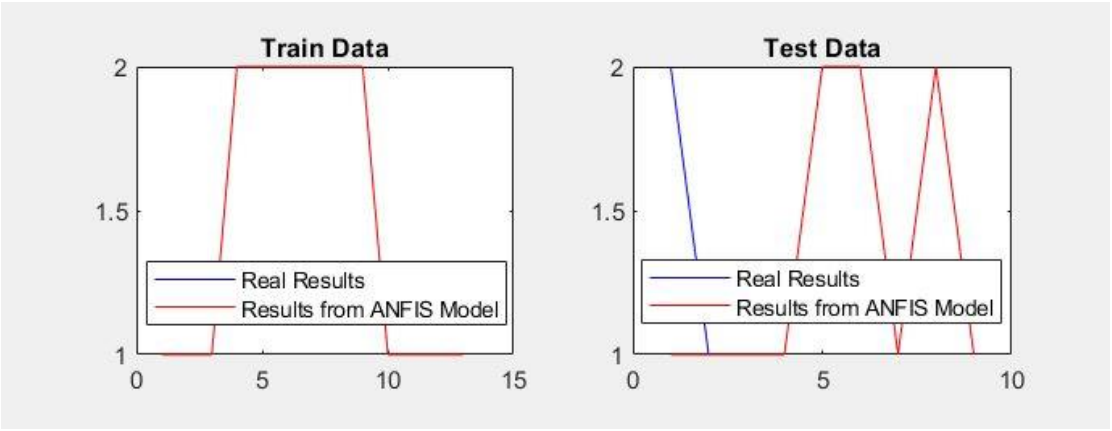


**Figure 9.** Comparison of actual values and outputs obtained by the ANFIS model for the training dataset and test dataset

Figure 9 compares actual values and results produced by the ANFIS model. In the first graph in Figure 9, the training dataset has a 100% success rate, and all the results are overlapped. However, in the second graph using the test dataset, between outputs produced by the ANFIS model and the actual values made one false detection.

In addition, in this study, the error value between the results obtained by the ANFIS model and the actual values were calculated for the training dataset and test dataset.

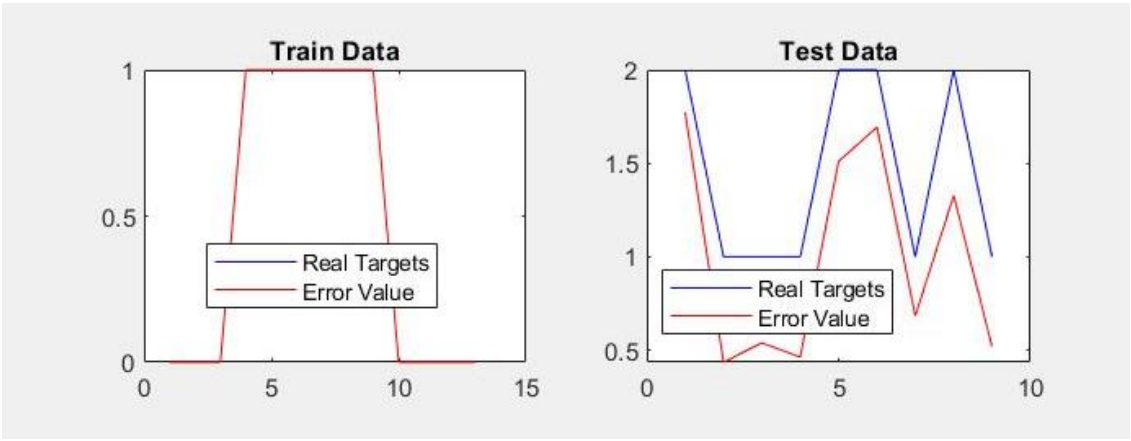


**Figure 10.** Comparison of the error values and the actual values for the training and test dataset

The error values expressed in Figure 10 are obtained by subtracting the fuzzy value produced by the ANFIS model from the actual output value. Fuzzy values reached by ANFIS are expressed with two separate fuzzy values for one sample in the scope of one hot encoding method. The trained model assigns a higher fuzzy number for the region where it produces the highest recognition accuracy among two separate classes than for the other region. The class is determined for the fuzzy value with the higher value. The error value is obtained by subtracting this final fuzzy value from the real value. A minimum error value indicates perfect success. In this framework, a successful classification was made for the training dataset in the first graph given in Figure 10. On the other hand, the fuzzy outputs produced for the first example of the intron class in the second graph were obtained as 0.2271 (first region) and 0.1546 (second region), respectively. Since the value of 0.2271 is greater than the value of 0.1546 among these outputs, the intron region is defined as an exon.

One sample detected incorrectly by the model training performed for randomly generated training dataset and test datasets using different DNA sequences indicates that the study is generally successful. In addition, the difference between the 0.2217 and 0.1546 fuzzy outputs is not high. It is thought that this disadvantage can be solved by increasing the data in the training dataset and carrying out stronger training. All these processes were carried out in the Matlab program and Takagi-Sugeno inference system was used.

Different approaches were used in the literature to classify exon and intron regions. In this context, DNA sequences were digitized to provide inferences on DNA sequences consisting of symbols in [9][10][11][12][20][21] studies. Then, after the analysis process using signal processing approaches, exon and

intron regions were analyzed with the changes seen in the signals. But, in this present study, digitization preprocessing was not performed for symbolic DNA sequences. After the clustering process, a statistical evaluation was carried out with the code frequencies of the sequences selected to represent the clusters. At the same time, this study indicates that it provided an analysis independent of the length of DNA sequences compared to [13][14] studies. Thus, powerful outputs will be produced on DNA sequences of different lengths. On the other hand, RNN and LSTM methods, which are considered in the scope of natural language processing methods, were used in [15][16][3][17] studies for the analysis of DNA sequences. These approaches are generally used in text analysis and evaluation of time series. However, it has powerful units to make successful decisions[33]. Therefore, in this study, SBERT structure based on natural language processing was used in the analysis of DNA sequences.

In addition, there are many studies in which artificial intelligence and fuzzy logic approaches are used as a hybrid to produce clear output for solving critical problems. In this direction, ANFIS-based approaches were used to obtain a successful performance output in the [34] study, to classify cancer genes in the [35] study, and to differentiate malignancies in the [36] study. Successful inferences were made with the ANFIS model also used in this study.

In this study, all the methods used for successfully detecting exon and intron regions are related to each other, and they are chosen in accordance with our main purpose. In this framework, SBERT is used to derive semantically meaningful sentence embeddings. This approach, which can perform semantic similarity comparison, clustering, and information retrieval processes, has produced successful outputs in clustering DNA sequences. Then, the %88.88 accuracy rate independent of the length of DNA sequences was obtained with the ANFIS classification process performed in the identification of protein-coding regions. Thus, a strong and hybrid classification process was realized by combining the learning ability of artificial intelligence on data and the inference power of fuzzy logic on data[29]. In addition to the studies [37][38][39][40][41] carried out to obtain successful outputs using hybrid structures, successful and powerful outputs optimized with the hybrid structure proposed in this study were produced.

This study will provide to determine the exon regions, which are an essential indicator in illuminating many vital events (regulations in growth and development processes, proliferation and termination states of the cells, informations related to the changes experienced by stem cells, data on whether genes are mutated or not, and assessments of the development of cancer). [32][9]. At the same time, the determination of exon regions that it is expressed as a difficult problem by researchers [5][6][7][8], which increases the importance of this study. Finally, this study, which offers a solution to the increasing amount of data, has performed a fast, accurate and low-cost analysis. Therefore, it is thought that this study will contribute to the literature.

## 5. CONCLUSION

This study provided a molecular diagnosis to distinguish exon and intron DNA sequences. First, exon and intron DNA sequences were clustered with the K-Means, and Agglomerative Clustering approaches applied consecutively in the scope of the SBERT approach. Then, a representative DNA sequence was selected from these clusters, and the frequency of repetition of the codes that formed the genetic code was calculated for the analysis of protein-coding regions. A matrix was created with the frequency of 64 different codes calculated. Finally, the classification of exon and intron regions was provided using the ANFIS structure. This structure provides rapid analysis and evaluation for increased genome sequences that have become easily accessible in recent years. Thus the computational cost is reduced. In addition, a system that is independent of the length of the DNA sequences was created without the digitization step of DNA sequences. The hybrid use of the SBERT and the ANFIS approach offers a new perspective to the study. Because clustering is generally preferred in analyzes related to textual data for SBERT approach[42]. However, it was used in the biomedical field in this study. In the future, it is aimed to use a hybrid with a new optimization algorithm to be explored as used in the [43] article on the ANFIS model. Thus, parameter optimization will be achieved, and predictions will be improved.

## DECLARATION OF ETHICAL STANDARDS

The authors of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

## AUTHORS' CONTRIBUTIONS

**Fatma AKALIN:** Formation of the idea, literature review, writing the article and examining the results

**Nejat YUMUŞAK:** Idea generation, literature review, spelling and content checking

## CONFLICT OF INTEREST

There is no conflict of interest in this study.

## REFERENCES

[1] Raza K., 'Fuzzy logic based approaches for gene regulatory network inference', ***Artificial Intelligence in Medicine***, 97: 189–203, (2019).

[2] Zheng P., Wang S., Wang X., and Zeng X., 'Editorial: Artificial Intelligence in Bioinformatics and Drug Repurposing: Methods and Applications', ***Frontiers in Genetics***, 13: 1–4, (2022).

[3] Singh N., Nath R., and Singh D.B., 'Splice-site identification for exon prediction using bidirectional

LSTM-RNN approach', ***Biochemistry and Biophysics Reports***, 30, (2022).

[4] Kar S. and Ganguly M., 'Study of effectiveness of FIR and IIR filters in Exon identification: A comparative approach', ***Materials Today: Proceedings***, 58: 437–444, (2022).

[5] Barman S., Saha S., Mandal A., and Roy M., 'Prediction of protein coding regions of a DNA sequence through spectral analysis', ***2012 International Conference on Informatics, Electronics and Vision, ICIEV 2012***, 12–16, (2012).

[6] Das L., Das J. K., and Nanda S., 'Detection of exon location in eukaryotic DNA using a fuzzy adaptive Gabor wavelet transform', ***Genomics***, 112: 4406–4416, (2020).

[7] Das L., Nanda S., and Das J. K., 'An integrated approach for identification of exon locations using recursive Gauss Newton tuned adaptive Kaiser window', ***Genomics***, 111: 284–296, (2019).

[8] Gupta R., Mittal A., Singh K., Bajpai P., and Prakash S., 'A Time Series Approach for Identification of Exons and Introns', ***10th International Conference on Information Technology (ICIT 2007)***, 91–93, (2007).

[9] Das B. and Türkoglu I., 'Sayisal haritalama teknikleri ve Fourier dönüsümü kullanılarak DNA dizilimlerinin sınıflandirilmasi', ***Journal of the Faculty of Engineering and Architecture of Gazi University***, 31(4): 921–932, (2016).

[10] Hota M. K. and Srivastava V. K., 'Performance analysis of different DNA to numerical mapping techniques for identification of protein coding regions using tapered window based short-time discrete Fourier transform', ***ICPCES 2010 - International Conference on Power, Control and Embedded Systems***, (2010).

[11] Dessouky A. M., Taha T. E., Dessouky M. M., Eltholth A. A., Hassan E., and Abd El-Samie F. E., 'Non-parametric spectral estimation techniques for DNA sequence analysis and exon region prediction', ***Computers and Electrical Engineering***, 73: 334–348, (2019).

[12] Roy M. and Barman S., 'Spectral analysis of coding and non-coding regions of a DNA sequence by Parametric method', ***Proceedings of the 2010 Annual IEEE India Conference: Green Energy, Computing and Communication, INDICON 2010***, 7–10, (2010).

[13] Singh A. K. and Srivastava V. K., 'The three base periodicity of protein coding sequences and its application in exon prediction', ***2020 7th International Conference on Signal Processing and Integrated Networks, SPIN 2020,*** 64: 1089–1094, (2020).

[14] Akalın F. and Yumuşak N., 'DNA genom dizilimi üzerinde dijital sinyal işleme teknikleri kullanılarak elde edilen ekson ve intron bölgelerinin EfficientNetB7 mimarisi ile sınıflandırılması', ***Journal of the Faculty of Engineering and Architecture of Gazi University***, 37(3): 1355–1371, (2022).

[15] Gunasekaran H., Ramalakshmi K., Rex Macedo Arokiaraj A., Kanmani S. D., Venkatesan C., and Dhas C. S. G., 'Analysis of DNA Sequence Classification Using CNN and Hybrid Models', ***Computational and Mathematical Methods in Medicine***, (2021).

[16] Abass Y.A., Adeshina S.A., Agwu N.N., Boukar M.M., Department of Computer Science, 'Analysis of Prostate Cancer DNA Sequences Using Bi-Directional Long Short Term Memory Model', ***2021 16th International Conference on Electronics Computer and Computation (ICECCO)***, 21–26, 2021.

[17] Canatalay P. J. and Ucan O. N., 'A Bidirectional LSTM-RNN and GRU Method to Exon Prediction Using Splice-Site Mapping', ***Applied Sciences***, 12(9), (2022).

[18] Nasr F.B., Oueslati A. E., 'CNN for human exons and introns classification', ***2021 18th International Multi-Conference on Systems, Signals & Devices***, 249–254, (2021).

[19] Chakraborty S. and Gupta V., DWT based cancer identification using EIIP, ***Proceedings - 2016 2nd International Conference on Computational Intelligence and Communication Technology, CICT 2016***, 718–723, (2016).

[20] Marhon S. A. and Kremer S. C., 'Protein coding region prediction based on the adaptive representation method', ***Canadian Conference on Electrical and Computer Engineering***, 000415–000418, (2011).

[21] Li J. et al., 'Integrated entropy-based approach for analyzing exons and introns in DNA sequences', ***BMC Bioinformatics***, 20, (2019).

[22] **https://www.ncbi.nlm.nih.gov/**,'NCBI'.

[23]**https://www.ensembl.org/Homo_sapiens/Gene/Sequence?db=core;g=ENSG00000103313;r=16:3242027-3256633**, 'Ensemble'.

[24] Wang T., Shi H., Liu W., and Yan X., 'A joint FrameNet and element focusing Sentence-BERT method of sentence similarity computation', ***Expert Systems with Applications***, 200, (2022).

[25] Devlin J., Chang M. W., Lee K., and Toutanova K., 'BERT: Pre-training of deep bidirectional transformers for language understanding', ***NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference***, 4171–4186, (2019).

[26] Santander-Cruz Y, et al., 'Semantic Feature Extraction Using SBERT for Dementia Detection' ***brain sciences***, (2022).

[27] Reimers N. and Gurevych I., 'Sentence-BERT: Sentence embeddings using siamese BERT-networks', ***arXiv***, 3982–3992, (2019).

[28] Mahdevari S. and Khodabakhshi M. B., 'A hybrid PSO-ANFIS model for predicting unstable zones in underground roadways', ***Tunnelling and Underground Space Technology incorporating Trenchless Technology Research***, 117, (2021).

[29] Karaboga D. and Kaya E., 'Estimation of number of foreign visitors with ANFIS by using ABC algorithm', ***Soft Computing***, 24:7579–7591, (2020).

[30]**https://www.sbert.net/examples/applications/clustering/README.html,** 'SBERT-Clustering'

[31] **https://www.sbert.net/docs/pretrained_models.html**, 'SBERT-Pretrained Models'

[32] Bihter DAŞ, 'DNA dizilimlerinden hastalik tanilanmasi için işaret işleme temelli yeni yaklaşımların geliştirilmesi', Fırat Üniversitesi Fen Bilimleri Enstitüsü Yazılım Mühendisliği Anabilim Dalı, ***Doktora Tezi***, (2018).

[33] Sak H., Senior A, and Beaufays F., 'Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition', ***arXiv***, (2014), [Online]. Available: http://arxiv.org/abs/1402.1128.

[34] Precup R. E., Bojan-Dragos C. A., Hedrea E. L., Roman R. C., and Petriu E. M., 'Evolving Fuzzy Models of Shape Memory Alloy Wire Actuators', ***Romanian Journal of Information Science and Technology***, 24(4): 353–365, (2021).

[35] Mishra P. and Bhoi N., 'Cancer gene recognition from microarray data with manta ray based enhanced ANFIS technique', ***Biocybernetics and Biomedical Engineering***, 41(3): 916–932, (2021).

[36] Akalın F., and Yumuşak N., 'Lösemi hastalığının temel türlerinden ALL ve KML malignitelerinin graf sinir ağları ve bulanık mantık algoritması ile sınıflandırılması', ***Journal of the Faculty of Engineering and Architecture of Gazi University,*** 38(2): 707–719, 2023.

[37] Zhu M. and Lai Y., 'Improvements Achieved by Multiple Imputation for Single-Cell RNA-Seq Data in Clustering Analysis and Differential Expression Analysis', ***Journal of Computational Biology***, 29(7): 634–649, (2022).

[38] Radpour V. and Soleimanian Gharehchopogh F., 'A Novel Hybrid Binary Farmland Fertility Algorithm with Naïve Bayes for Diagnosis of Heart Disease', ***Sakarya University Journal of Computer and Information Sciences***, 5(1), 2022.

[39] Ibrahim M. H., 'WBBA-KM: A Hybrid Weight-Based Bat Algorithm with K-Means Algorithm For Cluster Analysis', ***Journal of Polytechnic***, 25(1): 65–73, 2022.

[40] M. E. BAYRAKDAR and A. ÇALHAN, 'Optimization of Ant Colony for Next Generation Wireless Cognitive Networks', ***Journal of Polytechnic***, 24(3): 779–784, 2021.

[41] Garip Z., Çimen M. E., and Boz A. F., 'Fotovoltaik Modellerin Parametre Çıkarımı İçin Geliştirilmiş Bir Kaotik Tabanlı Balina Optimizasyon Algoritması', ***Journal of Polytechnic***, 25(3): 1041–1054, 2022.

[42] Alghobiri M., Mohiuddin K., Khaleel M. A., Islam M., Shahwar S., and Nasr O., 'A Novel Approach of Clustering Documents: Minimizing Computational Complexities in Accessing Database Systems', ***International Arab Journal of Information Technology***, 19(4), 617–628, (2022).

[43] Konar M., 'Redesign of morphing UAV's winglet using DS algorithm based ANFIS model', ***Aircraft Engineering and Aerospace Technology***, 91(9): 1214–1222, (2019).