Neural Computing and Applications (2022) 34:12587–12599
https://doi.org/10.1007/s00521-022-07121-8

ORIGINAL ARTICLE

A deep learning model for identification of diabetes type 2 based on nucleotide signals

Bihter Das[1]

Received: 24 July 2021 / Accepted: 21 February 2022 / Published online: 12 March 2022

Abstract

In Genome-Wide Association Studies (GWAS), detection of T2D-related variants in genome sequences and accurate modeling of the complex structure of the relevant gene are of great importance for the diagnosis of diabetes. For this purpose, this paper presents a novel strong algorithm to accurately and effectively identify Type 2 Diabetes (T2D) risk variants at high-performance rates. The proposed algorithm consists of five important phases. The first stage is to collect T2D-associated DNA sequences and to digitize them by the Entropy-based technique. The second stage is to transform these digitized DNA sequences into 224 × 224 pixels size spectrum images. The third is to extract a distinctive feature set from these spectrum images using the ResNet and VGG19 architectures. The fourth is to classify the effective feature set using SVM and k-NN methods. The last stage is to evaluate the system with k-fold cross-validation. As a result of the developed algorithm, the performances of the used Convolutional Neural Network (CNN) methods, the Entropy-based technique, and the classifiers were compared in relation. As a result of the study a combination model of the proposed Entropy-based technique, ResNet and Support Vector Machine (SVM) achieved the highest accuracy rate with 99.09%. With this study, the performance of the system in the extraction of epigenetic features and prediction of T2D from spectrogram images was investigated. The results show that the system will contribute to the identification of all genes in diabetes-related tissue and studies on new drug targets.

Keywords Convolutional neural network · Entropy-based technique · DNA sequences · Signal processing · Type 2 diabetes

1 Introduction

With GWAS, significant advances have been made in understanding the genetic makeup of complex human diseases [1]. The main purpose of these studies is to accurately model the complex structure of the disease-related gene regulator [2]. Diabetes is one of these genetic diseases and develops and lasts a lifetime when the gland which is called the pancreas does not produce enough insulin hormone in your body or the insulin hormone cannot be used effectively [3]. The full name of the disease is Diabetes Mellitus. People with diabetes are unable to use glucose, which goes to blood from the food they eat, and their blood sugar levels rise. Type 2 diabetes (T2D) occurs when the body does not produce enough insulin to function properly, or when body cells do not react to insulin [4, 5]. Type 2 diabetes is also called type 2 diabetes mellitus (T2DM). This is known as insulin resistance. T2D is much more common than type 1 diabetes (T1D). In T1D, the body does not produce any insulin. The factors that cause T2D are genetic and environmental factors [6–8]. The basis of T2D is insulin resistance and insulin secretion abnormality. Some patients with diabetes can be misdiagnosed. Not all patients with diabetes are type 1 or type 2. There are also patients with diabetes caused by genetic mutation, all over the world [9–11]. The vast majority of these people are treated unnecessarily using insulin rather than low-dose medication.

Recently, the studies on the correct diagnosis of T2D disease, which is very common, continue intensively. However, studies on disease diagnosis using DNA data set

✉ Bihter Das
bihterdas@firat.edu.tr

[1] Department of Software Engineering, Technology Faculty, Firat University, 23119 Elazig, Turkey

are limited. The conversion of DNA sequences into digital signals and then spectrum images and the detection of diabetes disease is the first with this study. Therefore, this paper aims to develop a novel algorithm based on deep learning to discover the genetic basis and risk mechanisms of T2D. Thus, it will contribute to the identification of all genes in diabetes-related tissue and studies on new drug targets.

1.1 Motivation

In the literature, there are some genome studies on understanding the genetic architecture of diabetes. These studies are microarray-based techniques, deep-learning-based models, machine learning methods, statistical analysis to diagnose T2D. However, these methods either require a laboratory environment or have not been able to clearly demonstrate whether the variants in the genome sequences are a transcript that causes T2D. The most important point that distinguishes this study from others is that it offers a low-cost, high-accuracy deep learning-based hybrid algorithm to detect T2D from nucleotide sequences without the need for a laboratory environment. The proposed algorithm achieves satisfactory sensitivity, precision, and strong robustness in classification.

1.2 Contributions of the paper

The main contributions of this paper are outlined as follows:

- A model using spectrogram images for T2D gene recognition is proposed for the first time, even though deep learning models such as convolutional neural networks (CNNs) are used to predict epigenetic features from T2D-associated DNA sequences.
- The result of the proposed deep learning model provides the best classification accuracy performance of all methods used in studies for T2D detection from DNA sequences in the literature.
- The proposed deep learning model performs at a lower cost and higher accuracy than existing models such as microarray technology, statistical methods, and does not require a laboratory environment in detecting T2D-related variants in genes.

The remainder of this paper is organized as follows. In Sect. 2, the literature review is mentioned. Section 3 contains details of the proposed approach. Section 4 presents the experimental results. Section 5 presents the conclusion.

2 Related work

Genome studies aimed at understanding the genetic architecture of various diseases continue without slowing down. As there is still a significant gap between genetic discoveries and T2D risk mechanisms, current research focuses on learning T2D biological mechanisms. The aim of the ongoing studies is to reveal the effective molecular mechanisms in the emergence of the disease and to use the obtained genetic information to predict the risk of T2D development. There are many studies related to T2D and its genetics in the literature. More than 400 genomic signals associated with T2DM have been identified in some of these studies [8, 12–14]. These signals are often named after their closest genes but it is not known whether the variant is a transcript, which changes the risk of diabetes. They have been used T2D-associated genes, ATAC sequences, T2D variants, mitochondrial DNA (mtDNA) sequences in their studies. In addition, in other studies in the literature, deep learning-based models [12–15], pathway analysis [16], CNN models, statistical analysis, Support Vector Machine Recursive Feature Elimination (SVM-RFE) approach [17] and some machine learning methods have been used to diagnosis T2D from genomic signals [19–23]. Moreover, Ensemble-based methods have been used for the prediction of diabetes [24–28]. Table 1 lists current studies for the detection of T2D-related genes.

3 The proposed algorithm for identification of T2D risk variants

In this section, stages performed for the identification of T2D risk variants were presented in detail. The proposed algorithm consists of 5 important stages. These are dataset collection and preprocessing, creation of images, feature extraction, classification, and evaluation. The flow diagram, which includes all phases and processing steps of the developed approach, is shown in detail in Fig. 1.

3.1 Phase – 1: data collection and preprocessing

In this first phase of the study, appropriate datasets for DNA gene bank study were obtained. Since these data sets are in the form of textual DNA sequences such as TGACCT ATGCGT …, the pretreatment stage has been applied. Three different methods, which are widely and effectively used in the literature, are used to digitize DNA sequences. In the study, the T2D-associatied gene and normal gene were obtained from the Ensembl genome database [29]. The used gene was BCL11A gene variants with reference sequence ENSG00000119866. The data

Table 1 The studies for the identification of T2D-associated genes in the literature

References	Methods	Datasets	Results
Abdulaima et al. [13]	Deep learning	SNP with T2D	AUC = 96.53%, Sens = 93.91%
Rai et al. [14]	Deep learning model base on U-Net architecture	ATAC sequence with T2D	–
Mattis et al. [15]	Convolutional Neural Network (CNN)	T2D gene ATAG sequence	–
Wang et al. [16]	Ingenuity Pathway Analysis	Diabetic PDAC patients	n = 18, P = 2.6964E-08
Kumar et al. [17]	SVMRFE approach	37 samples of normal human, 34 diabetic humans	AUC = %83.9
Lalrohlui et al. [18]	Statistical analysis	mtDNA sequence for 28 diabetics from Northeast India	ND3 variant 10398A > G was found associated with T2D (OR = 9.489, 95% CI = 1.161–77.54, P value = 0.03)
Liang et al. [19]	SimPo algorithm	Sequences of 24 Type 2 patients with T2D and 47 healthy controls	AUC = 0.902 Sensitivity = 0.837 Specificity = 0.944
Cai et al. [20]	Logistic regression (LR), linear discriminant analysis (LDA), Naive Bayes (NB), and support vector machine (SVM)	Dataset A from Chinese Dataset B from European	SVM on several different experiments The best AUC is 0.97
Malik et al. [21]	LR, SVM, Artificial neural network (ANN)	175 samples half healthy and half diabetic patients	SVM ACC = 84.09
Nilamyani et al. [22]	Recursive feature extraction, Random Forest RFE, SVM	Microarray-based GSE18732 gene for T2D	–
Liu et al. [23]	Independent sample t-test	TRs of diabetes genes, non-diabetes disease genes	P-value of the t-test is 0.557 and 0.422

were converted to digital signals by Entropy-based mapping technique.

3.1.1 The entropy-based mapping technique

Entropy based technique, which is used in the proposed approach, better reflects the complex structure of DNA sequences and digitizes the sequences according to the frequency of repetition of codons. Also, Entropy-based technique provides a wide range of correlation information on the gene sequence. This technique has higher performance than existing numerical mapping techniques in the literature. In the implementation, the performance of this technique was also compared with Electron–Ion Interaction Pseudo Potential (EIIP) and Integer techniques, which are widely used in the literature. The formula of the mapping technique based on fractional Shannon entropy is given in Eq. (1) [30, 31].

$$\mathrm{Sf} = -\sum_{i}[(-p(x_i))^{\propto_i} p(x_i)\log(p(x_i))] \tag{1}$$

$p(x_i)$ represents the repetition frequency of each codon in the given DNA sequence. The alpha (α) value is chosen randomly in many studies, but the alpha value is newly defined in Entropy-based technique. The alpha value is defined as a division of the logarithm of $p(x_i)$ to 1. The formula for the alpha value is shown in Eq. (2).

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

Figure 2 shows the numerical representation of DNA gene sequences by Entropy-based mapping technique.

3.1.2 Electron ion interaction potential (EIIP) mapping technique

In this technique, where bases are defined as the average energy of delocalized electrons [32, 33], bases are represented by the following values, respectively: A = 0.1260, G = 0.0806, C = 0.1340, T = 0.1335 [32–34]. Figure 3 shows the numerical representation of DNA gene sequences by EIIP mapping technique.

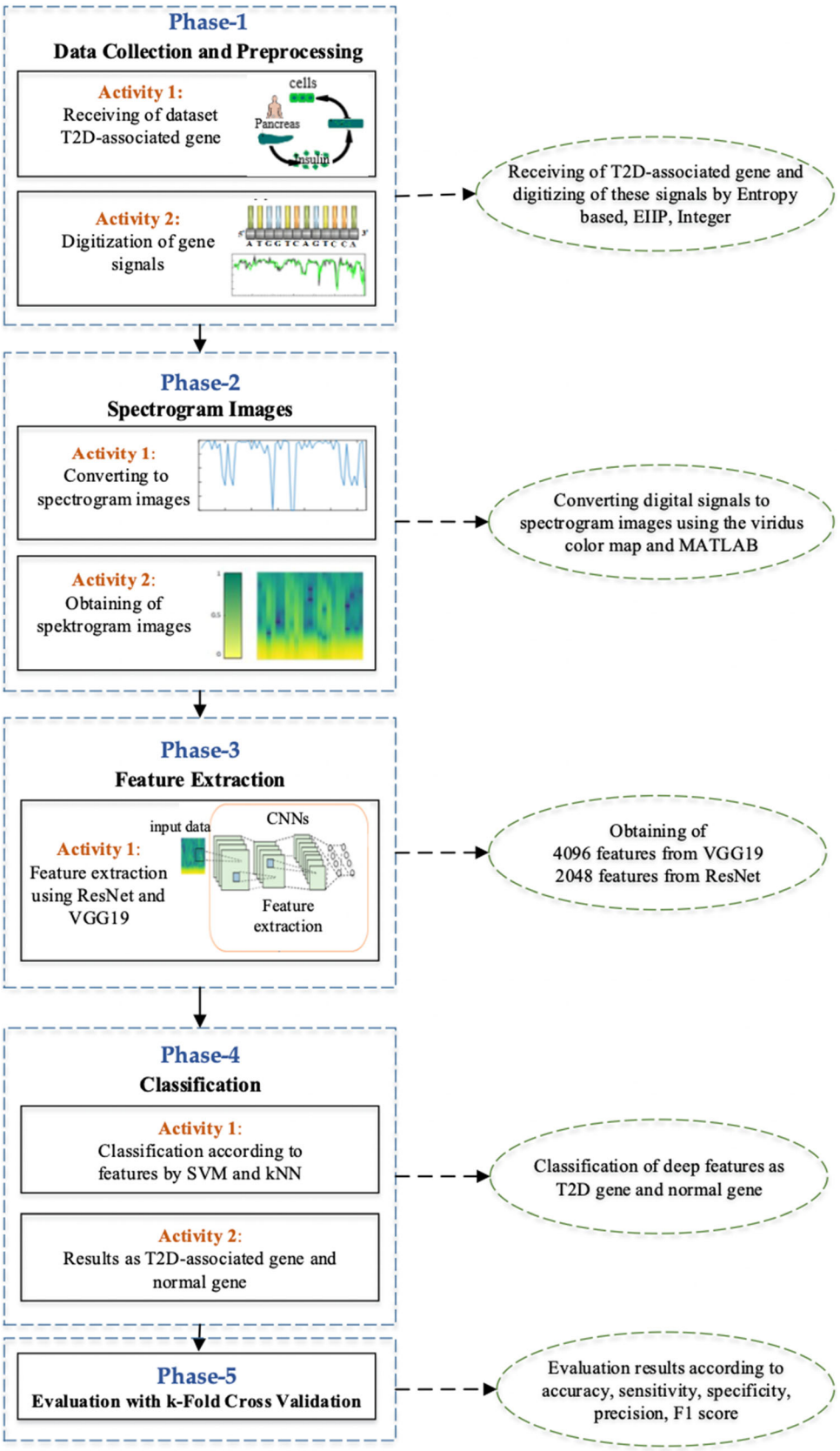


Fig. 1 The flow diagram of the proposed strong algorithm for diagnosis of T2D risk variants from nucleotide signals

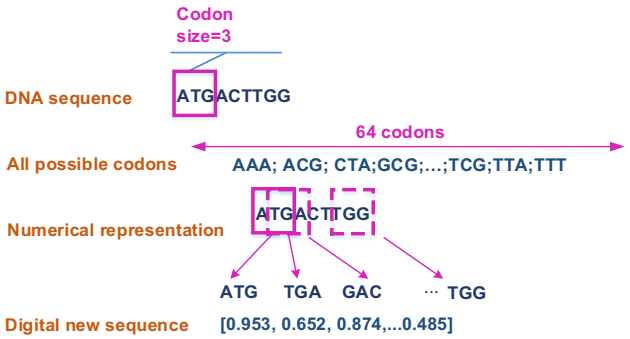


Fig. 2 The numerical representation of the T2D gene by Entropy-based technique

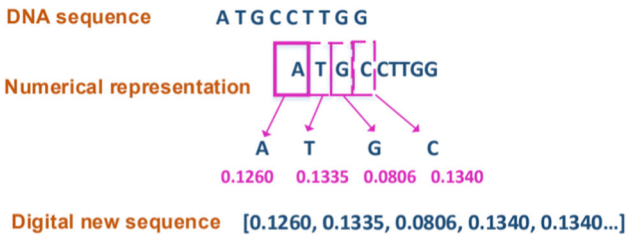


Fig. 3 The numerical representation of the T2D gene by EIIP

3.1.3 Integer technique

In the integer technique, which is a 1-dimensional mapping technique, when T > A and G > C are the bases in that sequence are represented A = 0, C = 1, T = 2 and G = 3, respectively [34, 35]. In a DNA sequence, if purine (A, G) > pyrimidine (C, T), bases are represented as T = 0, C = 1, A = 2, G = 3. In this technique, it is difficult to match these values with mathematical properties [36]. Figure 4 shows the numerical representation of DNA gene sequences by Integer mapping technique.

3.2 Phase – 2: composing of spectrogram images

In this second phase of the proposed approach, the spectrogram images obtained were converted into digital signals. The digitized T2DM-associated gene and normal gene sequences using the entropy-based mapping technique were added consecutively. Both types of gene sequences were taken at a length of 2470 units. Gene sequences are divided into sections by sliding 1-unit using a window width of 100 units in length. Spectrogram images of each 100-unit length gene fragment were generated sequentially. In addition to the Entropy mapping technique, gene sequences were also digitized by two other techniques (EIIP and Integer) for comparison, and the same processing was repeated within those numerical signals. When creating spectrogram images, the window width (Hamming) was set to be 12 ms, the overlap value 8 ms and the number of points(n) 512. Besides, the Viridis color map and MATLAB 2019 were used to obtain spectrogram images. Other processing was carried out using the Keras library in the Python environment. Figure 5 shows the obtained spectrogram images from the T2D digital gene signals using Entropy-based, EIIP, and Integer mapping techniques.

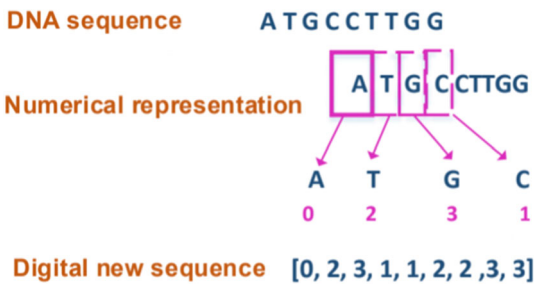


Fig. 4 The numerical representation of the T2D gene by Integer

3.3 Phase – 3: feature extraction

In this phase of the proposed approach, to extract effective features VGG19 and ResNet convolutional neural network models are used.

3.3.1 Convolutional neural network

Convolutional neural networks (CNN) is a deep learning algorithm that can take an input image and distinguish various objects in the image. CNN is mainly an artificial neural network that is used for classifying images, clustering similar features, and object recognition in scenes. Training of some CNN models is not possible on standard computer processors due to the complexity of the model or the size of the data set, so graphics processing units are needed. As a result of very long training, these trained models can be used in various ways to solve different problems. This is called transfer learning. The transfer learning approach is to train a network with a large dataset and copy the first n layers of the trained network to the first n layers of the target network [37]. In pre-trained networks, the first layers are sensitive to horizontal, vertical, diagonal lines. While the next layers focus on more features like corners and edges, the last layers focus on the specific features in the picture [40]. In the transfer learning method, the information in the pre-trained networks with large data sets is transferred to a different field to solve a problem. Even if the data set in the new problem is smaller, the transfer learning enables the network to learn better. Since pre-trained networks are trained with large data sets, they can effectively extract features [38, 39]. In this study,

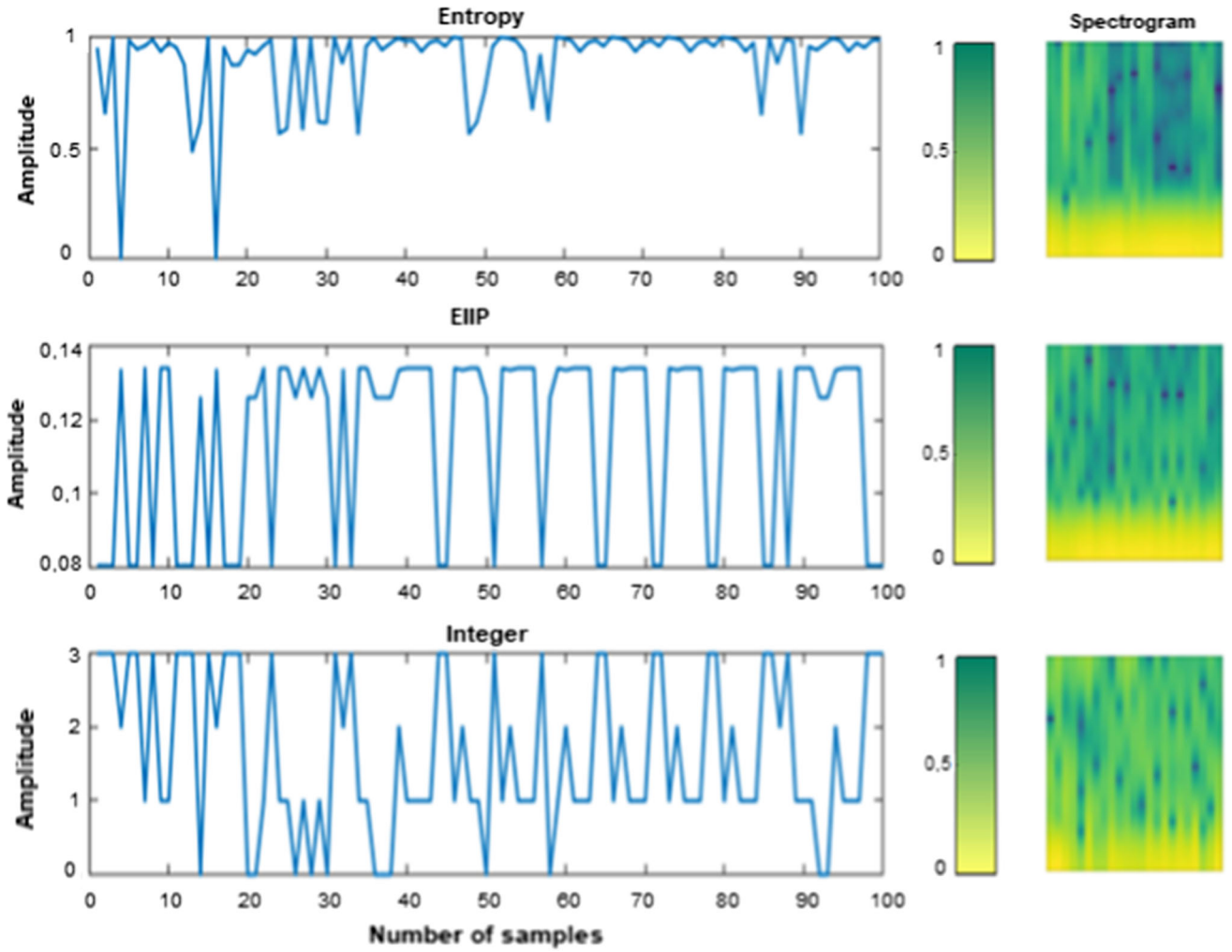


Fig. 5 Graphical representation of the digitized T2DM gene sequence according to Entropy-based technique and others

VGG19 and ResNet models, which are pre-trained models, were used.

(A) *VGG19:* Visual Geometry Group (VGG19) is a convolutional neural network that is 19 layers deep was developed by the Oxford University Visual Geometry Group (VGG) [40]. VGG19 consists of 16 convolution layers and 3 fully connected layers, and 5 max polling and SoftMax as the last layer. This model contains about 144 million parameters [41, 42].

(B) *Residual Neural Network (ResNet):* In recent years, deep learning studies have gained great importance and momentum. LeNet, AlexNet, GoogleNet, VGGNet, and ResNet have been the most important studies in this field, respectively. One of the common views in all these studies is that the number of layers in CNNs is a very important parameter [43, 44]. Increasing the number of layers, in theory, increases the representational capacity in CNNs. This is expected to create more successful network architectures. But this raises problems such as vanishing gradients and optimization difficulty. One of the most important contributions of ResNet is that it can prevent these problems while increasing the depth of the network [45].

3.4 Phase – 4: classification of deep features

In this phase of the proposed approach, SVM and k-NN classifiers are used. Since there are hundreds of studies and dozens of books on these classifiers in the literature, summary information is presented in this section.

3.4.1 Support vector machine classifier

Support Vector Machines is a controlled machine learning algorithm that can be used for classification or regression problems [46]. SVM offers a way to follow a path among

many possible classifiers that will increase your chances of accurately labeling your test data. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. It should be noted that Support Vector Machines work in any number and size; and in these dimensions, they find a similar two-dimensional line [47]. For example, they can classify a hyper plan at higher dimensions, such as generalizing the two-dimensional line and a three-dimensional plane to arbitrary dimensions. Since the hyperplane can act as a linear classifier, the SVM classifier was used to classify spectrogram images of DNA sequences.

3.4.2 k-nearest neighborhood (k-NN) classifier

k-Nearest Neighbor (k-NN) is a type of supervised machine learning algorithm that can be used for both classifications and regression predictive problems [48]. However, it is mainly used for classification predictive problems in the industry. k-NN is a nonparametric algorithm, which means it does not make any assumption on underlying data. It is very important for the k-NN classifier that the training set is large and the k value is selected appropriately [49]. Since the DNA datasets are very large, the k-NN classifier was preferred in this study.

3.5 Phase – 5: evaluation with k-fold cross-validation

In this last phase of the study, the numerical results obtained were evaluated with the k-Fold Cross Validation method. 80% of the data set, which was divided into 5 parts, was used for education and 20% was used for testing. This process was applied for each part separately. In Fig. 6, the K-fold validation diagram for the dataset is shown. The performance of the proposed approach was calculated by averaging of 5 parts. The used parameters are defined as follows; True Positive (TP), the number of correctly identified T2D gene, false negative (FN), the number of incorrectly defined T2DM gene, true negative (TN), the number of the correctly identified normal gene, false positive (FP), false shows the number of identified normal gene.

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}) \times 100$$

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \times 100$$

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN}) \times 100$$

4 Experimental results and discussion

Normal gene and T2D-associated gene sequences of 2470 bases lengths were used in the study. The gene sequences were converted into spectrogram images after digitizing with Entropy-based, EIIP and Integer techniques. Spectrogram images were 875 × 656 pixels in size. These images have been resized to 224 × 224 pixels to been processed by ResNet and VGG19 architectures. These spectrogram images were then given as input to ResNet and VGG19 to extract deep features. While ResNet obtains a 2048 dimensional vector, VGG19 obtains a 4096-dimensional feature vector from each spectrogram image. The feature vectors were classified using SVM. To evaluate the classification results more objectively, the k-fold cross-validation method was used. The k value was determined as 5.

The accuracy rates of the classifiers are shown in Table 2. In Table 2, it is seen that in both classifiers, the Entropy-based mapping technique is more successful than other numerical techniques. The reason for the high performance of the Entropy-based mapping technique is that better reflection of the complex structure of the gene, and digitization of the sequences according to the frequency of repetition of codons. Besides, Entropy-based technique provides a wide range of correlation information on the

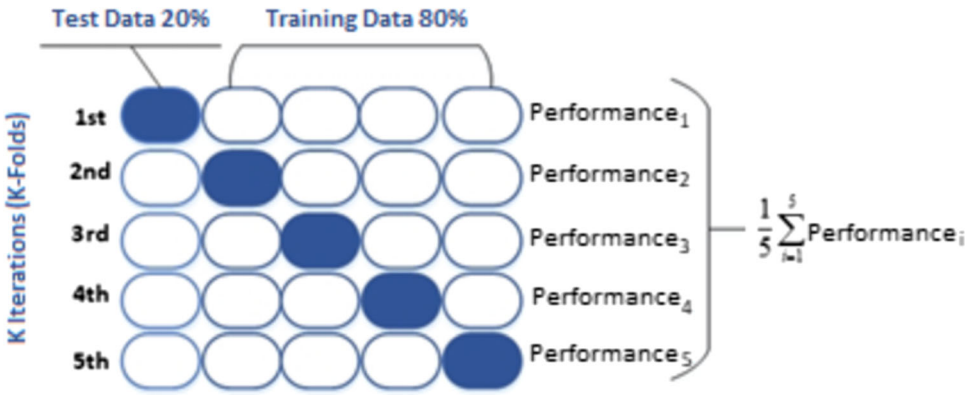


Fig. 6 The graphical representation of test and training data for the 1D-CNN model

Table 2 The accuracy rates of Entropy-based, EIIP, and Integer techniques for VGG19 and ResNet models

	Models	Entropy	EIIP	Integer
SVM	ResNet	**99.09 ± 0.59**	**98.72 ± 0.54**	**98.52 ± 0.92**
	VGG19	98.16 ± 0.76	98.38 ± 0.71	97.81 ± 0.73
k-NN	ResNet	98.58 ± 0.33	98.38 ± 0.30	98.25 ± 0.33
	VGG19	98.48 ± 0.32	98.62 ± 0.41	98.38 ± 0.18

gene sequence. As seen in Table 2, the classification results are very close in the k-NN classifier for three different mapping techniques. When the performances of SVM and k-NN classifiers are compared, the SVM achieved a better classification accuracy. 80% of the data set was used for training and the remaining part was used for testing. Accordingly, 3952 of the total 4940 spectrogram images were used for training and the remaining 988 images were used for testing.

According to the SVM, both CNN models have a lower performance for EIIP and Integer mapping techniques. Whereas the accuracy values of the Entropy based mapping technique for SVM are highest, the standard deviation values are within the acceptable limits. When the k-NN classification results were examined, the classification accuracy of the Entropy-based technique is higher than the other two mapping techniques, but it is lower than the accuracy values of the SVM classifier. Table 3 shows the Sensitivity, Specificity, Precision, F1 score rates of all three mapping techniques for both CNN models in the SVM classifier.

The best classification performance with 99.09% was obtained using Entropy-based technique in the ResNet model. The ResNet is a deeper model than the VGG19. Increasing the depth of the models may not always give a better result. While the VGG19 model has a better feature vector than ResNet, it did not give a better result than ResNet because the performance of a deep learning model changes according to the size of the data, distinguishing feature of the data, quality of the data, and parameters of the model. Nonparametric statistical significance tests also have been applied to the results. Firstly, the data set was evaluated in terms of the homogeneity of variance assumption. The extreme values of Levene's Test are shown in Fig. 7, and the descriptive results of the Levene's Test are shown in Fig. 8. According to the test result, it was seen that the data were not homogeneously distributed and it was certain that nonparametric tests would be applied.

Table 3 The sensitivity, specificity, precision, F1 score values of SVM

	Models	Entropy	EIIP	Integer
ResNet	Sensitivity	**99.95 ± 0.79**	98.62 ± 0.13	98.58 ± 1.38
	Specificity	**99.23 ± 1.21**	98.83 ± 0.70	98.46 ± 0.55
	Precision	**99.23 ± 1.21**	98.83 ± 0.68	98.46 ± 0.56
	F1 score	**99.09 ± 0.58**	98.72 ± 0.54	98.52 ± 0.93
VGG19	Sensitivity	**98.22 ± 0.12**	97.94 ± 0.12	97.33 ± 1.36
	Specificity	**98.10 ± 0.98**	98.83 ± 0.79	98.30 ± 0.98
	Precision	**98.10 ± 0.98**	98.82 ± 0.78	98.29 ± 0.97
	F1 score	**98.16 ± 0.76**	98.37 ± 0.72	97.80 ± 0.74

According to Levene's test results, it was observed that the data were nonparametric. For this, the Kruskal–Wallis test was applied since the number of categories was more than 2 because it was made according to 3 techniques. Krustal Wallis test results are shown in Fig. 9.

The confusion matrix is a special table layout that allows the visualization of an algorithm's performance, typically a controlled learning one. It is an effective tool to measure the performance of classification. Figure 10 shows the confusion matrix values of the Entropy-based technique of the SVM classifier for ResNet. It was seen that Entropy-based technique has better TN and TP values compared to others. These results show that the ResNet architecture is more successful than the other models in feature extraction with Entropy-based technique.

The Receiver Operating Characteristic (ROC) curves of the two techniques with the highest performance for identification of the T2D gene are shown in Fig. 11. The AUC is under the ROC curve. The highest AUC value is 1. AUC value close to 1 indicates that it is the correct classification. The maximum AUC value of 99.09% was obtained in the ResNet model with Entropy-based technique. The obtained classification results with the proposed method are promising. This shows that the proposed method has high applicability in practice.

The proposed approach showed a high-rate performance for T2D disease modeling and achieved an average accuracy of 99.09% in the model where features were extracted with ResNet and classified with SVM. This result showed a high rate of performance in the modeling of T2D diabetes risk variants.

There are some studies dealing with the identification of T2D from DNA sequences. In these studies, several methods have been used such as deep learning [13–15], statistical analysis [18], machine learning methods [19–21]. Table 4 shows a comparison of our method with the methods proposed in the literature with regard to performance results of liver cancer DNA sequences classification.

The proposed framework has higher performance than the existing techniques. The proposed study is different in two aspects from the other studies in the literature about the

Fig. 7 The extreme values of Levene's test

Extreme Values

Output				Case number	Value
DNA seq.	Normal	Highest	1	7716	3,00
			2	7717	3,00
			3	7718	3,00
			4	7719	3,00
			5	7720	3,00[a]
		Lowest	1	10288	,00
			2	10285	,00
			3	10284	,00
			4	10281	,00
			5	10277	,00[b]
	Diabetes	Highest	1	5144	3,00
			2	5145	3,00
			3	5146	3,00
			4	5148	3,00
			5	5149	3,00[a]
		Lowest	1	7701	,00
			2	7698	,00
			3	7697	,00
			4	7696	,00
			5	7693	,00[b]

a.Only a partial list of cases with the value 3,00 are shown in the table of upper extremes.

b. Only a partial list of cases with the value ,00 are shown in the table of lower extremes.

identification of T2D-associated DNA sequences. The first is the use of Entropy-based mapping technique, which is a new method for digitizing DNA sequences, and the second is the transferring of digitized signals, which were converted to spectrogram, into 2-dimensional space and the feature extraction using pre-trained CNN models. The proposed study showed good performance compared to other studies in the literature and achieved an average accuracy of $99.09 \pm 0.59\%$.

It is difficult to understand the complexity of the T2D-associated gene and to detect disease-causing variants without the need for a laboratory environment. With the proposed approach, a satisfactory accuracy rate and low-cost system is presented. Thus, a successful and effective alternative solution approach has been developed for the identified problem. On the other hand, pre-trained CNN models can effectively extract features from small datasets as they are pre-trained with large image data. The size of the gene sequences used in this study and the size of the spectrogram image used to provide an adequate training set for designing a new CNN model were the limitations of the study.

Descriptive

Output			Statistic	Std. Error
DNA seq.	Normal	Mean	.7799	.00977
		95% Confidence Interval for mean — Lower bound	.7607	
		95% Confidence Interval for mean — Upper bound	.7990	
		5% Trimmed Mean	.6999	
		Median	.7023	
		Variance	.736	
		Std. Deviation	.85796	
		Minimum	.00	
		Maximum	3.00	
		Range	3.00	
		Interquartile Range	.87	
		Skewness	1.293	.028
		Kurtosis	.846	.056
	Diabetes	Mean	.8381	.00973
		95% Confidence Interval for mean — Lower bound	.8190	
		95% Confidence Interval for mean — Upper bound	.8571	
		5% Trimmed Mean	.8457	
		Median	.730	
		Variance	.736	
		Std. Deviation	.85444	
		Minimum	.00	
		Maximum	3.00	
		Range	3.00	
		Interquartile Range	.87	
		Skewness	1.286	.028
		Kurtosis	.937	.056

Case Processing Summary

Output		Cases					
		Valid		Missing		Total	
		N	Percent	N	Percent	N	Percent
DNA seq.	Normal	7717	100.0%	0	.0%	7717	100.0%
	Diabetes	7717	100.0%	0	.0%	7717	100.0%

Ranks

	Techniques	N	Mean Rank
DNA seq.	Entropy	5142	8783,92
	EIIP	5146	10300,57
	Integer	5146	4068,84
	Total	15434	

Test Statistics [a,b]

	DNA seq.
Chi-Square	5502,477
Df	2
Asymp. Sig.	,000

a. Kruskal Willis Test

b. Grouping Variable: Techniques

Fig. 8 The descriptive and case processing results of the Levene's test

Fig. 9 The results of the Krustal Wallis test

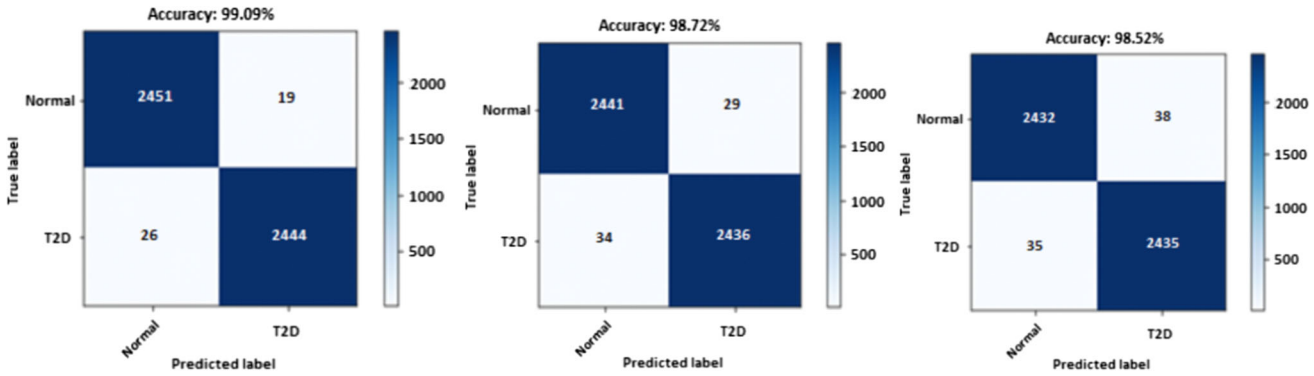
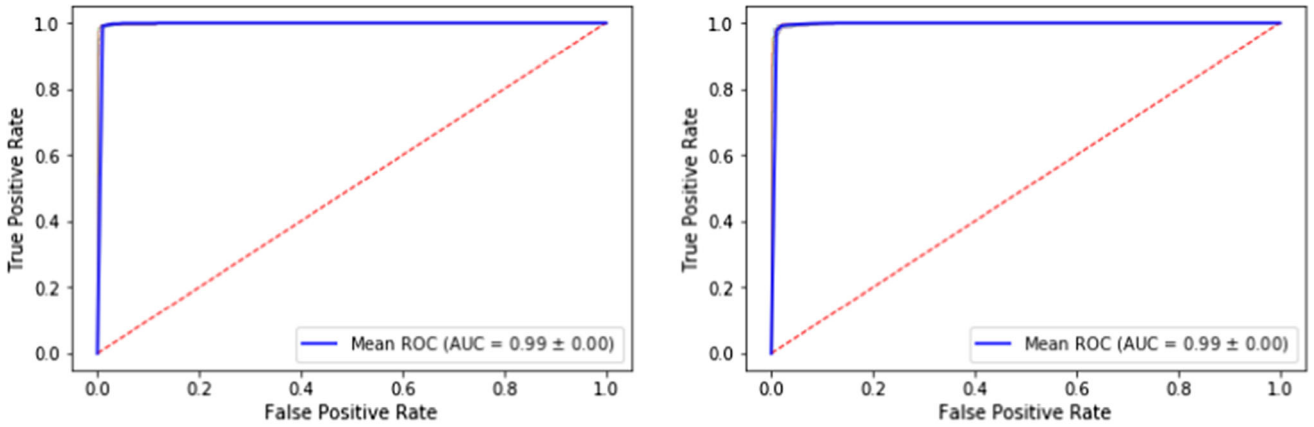


Fig. 10 Confusion matrices of all three techniques (Entropy, EIIP, Integer, respectively) for ResNet model in SVM



(a) Entropy-based mapping technique **(b)** EIIP

Fig. 11 ROC curves of the ResNet model in SVM

Table 4 Comparison of methods used for liver cancer DNA sequences classification

Authors	Methods	Dataset	Performance
Abdulaima et al. [13]	Deep learning	SNP with T2D	96.53%
Kumar et al. [17]	Machine learning-based approach	T2D sequences	83.9%
Lalrohlui et al. [18]	Statistical analysis	mtDNA sequences with T2D	95%
Liang et al. [19]	SimPo algorithm	T2D sequences	90.2%
Malik et al. [21]	LR, SVM, Artificial neural network (ANN)	T2D sequences	84.09%
The proposed method	**CNN-based algorithm**	**T2D sequences**	**99.09%**

5 Conclusion

In this study, a novel approach based on deep learning is developed for the discovery of the genetic architecture of T2D. A combination of Entropy-based technique and CNN models is used to extract distinctive features from TD2-associated gene signal. The T2D genome signals are digitized by the Entropy-based technique. The proposed approach is the first study in literature in which the T2D-associated gene sequence is digitized and the spectrogram images are obtained and classified by CNNs. Deep learning methods have been seldom used in genomics, but the proposed approach differs from other methods by using together with the Entropy-based technique, which is for numerical representation of the T2D gene, and deep learning, which is used for effective feature extraction. The proposed approach showed high performance for T2D disease modeling and achieved an average accuracy with 99.09% in the model in which features were extracted with

ResNet and classified with SVM. It also contributes to GWAS on the genetic development of T2D.

Funding There is no funding source for this article.

Declarations

Conflict of interest The author declares that there are no known competing financial interests or personal relationships that could appear to influence the work reported in this paper.

References

1. Ho DSW, Schierding W, Wake M, Saffery R, O'Sullivan J (2019) Machine learning SNP based prediction for precision medicine. Front Genet. https://doi.org/10.3389/fgene.2019.00267
2. Imani M, Ghoreishi S, F. (2020) Optimal finite-horizon perturbation policy for inference of gene regulatory networks. IEEE Intell Syst. https://doi.org/10.1109/MIS.2020.3017155
3. Guariguata L, Whiting DR, Hambleton I, Beagley J, Linnenkamp U, Shaw JE (2014) Global estimates of diabetes prevalence for 2013 and projections for 2035. Diabetes Res Clin Pract 103:137–149
4. Arikoglu H, Kaya DE (2015) Tip 2 diyabetin moleküler genetik temeli; Son gelişmeler. Genel Tıp Dergisi 25:147–159
5. Defronzo RA, Ferrannini E, Groop L, Henry RR, Herman WH, Holst JJ et al (2015) Type 2 diabetes mellitus. Nat Rev Dis Primers 1:15019. https://doi.org/10.1038/nrdp.2015.19
6. Morris AP (2018) Progress in defining the genetic contribution to type 2 diabetes susceptibility. Curr Opin Genet Dev 50:41–51
7. Das KW, Elbein SC (2006) The Genetic basis of type 2 diabetes. Cell Sci 2:100–131
8. Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM, Rayner NW et al (2018) Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat Genet. https://doi.org/10.1038/s41588-018-0241-6
9. Vinuela A, Varshney A, van de Bunt M, Prasad RB, Asplund OB, Bennett A et al (2019) Influence of genetic variants on gene expression in human pancreatic islets-implications for type 2 diabetes. BioRxiv. https://doi.org/10.1101/655670
10. Varshney A, Scott LJ, Welch RP, Erdos MR, Chines PS, Narisu N et al (2017) Genetic regulatory signatures underlying işlet gene expression and type 2 diabetes. Proc Natl Acad Sci 114:2301–2306. https://doi.org/10.1073/pnas.162119214
11 Kleinberger JW, Pollin TI (2015) Personalized medicine in diabetes mellitus: current opportunities and future prospects. Ann N Y Acad Sci 1346:45–56. https://doi.org/10.1111/nyas.12757
12. Awotunde JB et al (2021) Chapter Nine—Prediction and classification of diabetes mellitus using genomic data. In: Sangaiah AK, Mukhopadhyay S (eds) Intelligent IoT systems in personalized health care. Academic Press, pp 235–292
13. Abdulaimma B, Fergus P, Chalmers C, Montañez C (2020) Deep learning and genome-wide association studies for the classification of type 2 diabetes. In: içinde 2020 international joint conference on neural networks (IJCNN), Tem, pp 1–8. https://doi.org/10.1109/IJCNN48605.2020.9206999
14. Rai V et al (2020) Single-cell ATAC-Seq in human pancreatic islets and deep learning upscaling of rare cells reveals cell-specific type 2 diabetes regulatory signatures. Mol Metab 32:109–121. https://doi.org/10.1016/j.molmet.2019.12.006
15. Mattis KK, Gloyn LA (2020) From Genetic association to molecular mechanisms for Islet-cell dysfunction in type 2 diabetes. J Mol Biol 432:1551–1578. https://doi.org/10.1016/j.jmb.2019.12.045
16. Wang K, Zhou W, Meng P, Wang P, Zhou C, Yao Y, Wu S, Wang Y, Zhao J, Zou D, Jin G (2019) Immune-related somatic mutation genes are enriched in PDAGs with diabetes. Transl Oncol 12(9):1147–1154
17. Kumar A, JeyaSundaraSharmila D, Singh S (2017) SVMRFE based approach for prediction of most discriminatory gene target for type II diabetes. Genom Data 12:28–37. https://doi.org/10.1016/j.gdata.2017.02.008
18. Lalrohlui F, Zohmingthanga J, Hruaii V, Kumar NS (2020) Genomic profiling of mitochondrial DNA reveals novel complex gene mutations in familial type 2 diabetes mellitus individuals from Mizo ethnic population, Northeast India. Mitochondrion. https://doi.org/10.1016/j.mito.2019.12.001
19. Liang F et al (2020) Insulin-resistance and depression cohort data mining to identify nutraceutical related DNA methylation biomarker for type 2 diabetes. Genes Dis. https://doi.org/10.1016/j.gendis.2020.01.013
20. Cai L, Wu H, Li D, Zhou K, Zou F (2015) Type 2 diabetes biomarkers of human gut microbiota selected via iterative sure independent screening method. PLoS ONE. https://doi.org/10.1371/journal.pone.0140827
21. Malik S, Khadgawat R, Anand S et al (2016) Non-invasive detection of fasting blood glucose level via electrochemical measurement of saliva. Springerplus 5:701. https://doi.org/10.1186/s40064-016-2339-6
22. Nilamyani N, Lawi A, Thamrin SA (2018) A preliminary study on identifying probable biomarker of type 2 diabetes using recursive feature extraction. In: 2018 2nd East Indonesia conference on computer and information technology (EIConCIT), pp 267–270. https://doi.org/10.1109/EIConCIT.2018.8878565
23. Liu ZY, Ding XP, Bian HJ (2008) Comparisons of properties of tandem repeats associated with beteen diabetes genes and non-diabetes disease genes. In: 2nd international conference on bioinformatics and biomedical engineering, iCBBE 2008, pp 436–440. https://doi.org/10.1109/ICBBE.2008.107
24. Reddy SS, Sethi N, Rajender R, Mahesh G (2020) Extensive analysis of machine learning algorithms to early detection of diabetic retinopathy. Mater Today Proc. https://doi.org/10.1016/j.matpr.2020.10.894
25. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I (2017) Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J 15:104–116. https://doi.org/10.1016/j.csbj.2016.12.005
26. Sikder N, Masud M, Bairagi AK, Arif ASM, Nahid A-A, Alhumyani HA (2021) Severity classification of diabetic retinopathy using an ensemble learning algorithm through analyzing retinal images. Symmetry 13:670
27. Islam MT, Raihan M, Aktar N, Alam MS, Ema RR, Islam T (2020) Diabetes mellitus prediction using different ensemble machine learning approaches. In: 2020 11th international conference on computing, communication and networking technologies (ICCCNT), pp 1–7
28. Islam MT, Raihan M, Farzana F, Aktar N, Ghosh P, Kabiraj S (2020) Typical and non-typical diabetes disease prediction using random forest algorithm. In: 2020 11th International conference on computing, communication and networking technologies (ICCCNT), pp 1–6
29. "Ensembl Genbank". Available: https://www.ensembl.org/index.html. Accessed 04 Apr 2020
30. Das B, Turkoglu I (2018) A novel numerical mapping method based on entropy for digitizing DNA sequences. Neural Comput Appl 29:207–215. https://doi.org/10.1007/s00521-017-2871-5

31. Daş B (2018) Development of new approaches based on signal processing for disease diagnosis from Dna sequences, Fırat University, PhD Thesis, 2018
32. Grandhi DG, Kumar CV (2007) 2-Simplex mapping for identifying the protein coding regions in DNA. In: TENCON 2007-2007 IEEE reg. 10 conf., pp 1–3. IEEE
33. Chakraborty S, Gupta V (2016) DWT Based cancer identification using EIIP. In: 2016 second international conference on computational intelligence communication technology (CICT), pp 718–723. https://doi.org/10.1109/CICT.2016.148
34. Akhtar M, Epps J, Ambikairajah E (2007) On DNA numerical representations for period-3 based exon prediction. In: 2007 IEEE international workshop on genomic signal processing and statistics, pp 1–4. IEEE
35. Cristea PD (2002) Conversion of nucleotides sequences into genomic signals. J Cell Mol Med 6:279–303. https://doi.org/10.1111/j.1582-4934.2002.tb00196.x
36. Cristea PD (2005) Representation and Analysis of DNA sequences. Genomic signal processing and statistics. Eurasip B Ser Signal Process Commun 15–66
37. Yosinski J, Clune Y, Lipson BH (2014) How transferable are features in deep neural networks?. Adv Neural Inf Process Syst. http://arxiv.org/abs/1411.1792
38. Ozcan T, Basturk A (2019) Transfer learning-based convolutional neural networks with heuristic optimization for hand gesture recognition. Neural Comput Appl 31:8955–8970. https://doi.org/10.1007/s00521-019-04427-y
39. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), pp 818–833
40. Ullah I, Hussain M, Qazi E-H, Aboalsamh H (2018) An automated system for epilepsy detection using EEG brain signals based on deep learning approach. Expert Syst Appl 107:61–71. https://doi.org/10.1016/j.eswa.2018.04.021
41. Gopalakrishnan K, Khaitan SK, Choudhary A, Agrawal A (2017) Deep Convolutional Neural Networks with transfer learning for computer vision-based data-driven pavement distress detection. Constr Build Mater 157:322–330. https://doi.org/10.1016/j.conbuildmat.2017.09.110
42. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 [cs]
43. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition
44. Reddy N, Rattani A, Derakhshani R (2018) Comparison of deep learning models for biometric-based mobile user authentication. In: 2018 IEEE 9th international conference on biometrics theory, applications and systems (BTAS), pp 1–6. https://doi.org/10.1109/BTAS.2018.8698586
45. Chen Z, Cen J, Xiong J (2020) Rolling bearing fault diagnosis using time-frequency analysis and deep transfer convolutional neural network. IEEE Access 8:150248–150261. https://doi.org/10.1109/ACCESS.2020.3016888
46. Dilmen E, Beyhan S (2017) A novel online LS-SVM approach for regression and classification. IFAC-PapersOnLine 50(1):8642–8647. https://doi.org/10.1016/j.ifacol.2017.08.1521
47. Khairandish MO, Sharma M, Jain V, Chatterjee JM, Jhanjhi NZ (2021) A Hybrid CNN-SVM threshold segmentation approach for tumor detection and classification of MRI brain images. IRBM. https://doi.org/10.1016/j.irbm.2021.06.003
48. Baby Saral G, Priya R (2021) Digital screen addiction with KNN and -Logistic regression classification. Mater Today Proc. https://doi.org/10.1016/j.matpr.2020.11.360
49. Wang Y, Pan Z, Dong J A new two-layer nearest neighbor selection method for kNN classifier—ScienceDirect. https://www.sciencedirect.com/science/article/pii/S0950705121008662. Accessed 07 Feb 2022