International Research Journal of Multidisciplinary Scope (IRJMS), 2024; 5(4):1034-1046

Original Article | ISSN (O): 2582-631X DOI: 10.47857/irjms.2024.v05i04.01595

LSTM with WSEQ-GAN for Cancer Prediction using DNA Sequence Data

Ravindran U[1], Gunavathi C[2]*

[1]School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, Vellore, India, [2]School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India. *Corresponding Author's Email: gunavathi.cm@vit.ac.in

Abstract

Deoxyribonucleic (DNA) sequence categorization is a significant task in a generic computational setting for biomedical data processing. The sequence information contains the genome information it can retrieve from human chromosome cells. The gene information in the DNA sequence is used to predict the disease, especially cancer diagnosis and therapy. The class samples in the gene expression data are imbalanced. The main objective is to enhance the sequence of samples to make an accurate class prediction. To analyze and categorize the sequence information, which is the challenge task, dominant computational techniques are required. Deep learning (DL) and machine learning (ML) techniques are used for training purposes to process and categorize the genome information. In the data preprocessing stage for converting the sequence information into numerical values, ordinary encoding, one-hot encoding, and k-mer counting techniques are applied. The DNA sequence information contains insufficient samples based on the class labels. To predict better results, the proposed Wasserstein Sequence Generative Adversarial Network (WSEQ-GAN) method is utilized for augmented sequence data, and results are compared with traditional methods like sampling and SMOTE. Traditional ML and DL techniques like Support Vector Machine (SVM), K Nearest Neighbor (KNN), and Long Short-Term Memory (LSTM) are used to train and classify the sequence data. The augmented and non-augmented data using WSEQ-GAN were compared with existing methods. As a result, the proposed WSEQ-GAN with the LSTM network achieved 98% classification accuracy better than existing classification and augmentation techniques.

Keywords: Deep Learning Methods, DNA Sequence, Machine Learning Methods, WSEQ-GAN.

Introduction

DNA, or deoxyribonucleic acid, stores the genetic information for every organism, including humans. The DNA information is fetched from human chromosome cells. DNA values are encoded in the sequence information based on the four nucleotides: adenine (A), guanine (G), cytosine (C), and thymine (T). Each of the nucleotides in the DNA molecule is paired with each other (A with T and G with C), and it forms a sequence for the phosphate and sugar molecules. Figure 1 shows that the DNA base pair is connected to the sugar-phosphate backbone (1). The complete DNA sequence for the human genome is around 6 billion letters. DNA sequencing will continue to expand the volume and complexity of such data sets. For analyzing each genome sequence with other sequences, we require some computational techniques. The challenge has shifted from collecting biological data to extracting useful knowledge from it. The continuous advancement of biological data analysis methods has resulted in the establishment of a difficult new field known as bioinformatics. Bioinformatics is to handle biomedical data analysis using computational methods. The rapid development of data analysis technologies has resulted in a huge number of useful and scalable algorithms. Machine learning and deep learning techniques are used to analyze and compute the huge volume of biomedical data. Analysis of single-cell DNA sequencing data is complicated due to biases and aberrations caused by DNA extraction and whole-genome enhancement, such as mutated imbalance and dropout. The number of class samples in the DNA sequence is imbalanced (2). AI-based ML/DL computational tools are used for genome data to extract valuable information hidden in the large amount of data. Machine learning is a technology that allows machines to learn about a set of data despite being explicitly instructed what to learn. To learn the parameter values based on supervised and unsupervised fashion. The supervised learning approach is to learn the values based on training and testing sets with class labels. The unsupervised learning approach is to recognize patterns in huge amounts of data and make predictions about actual events without

(Received 11thJuly 2024; Accepted 24th October 2024; Published 30th October 2024)

the assistance of humans. The DL algorithm analyzes a dataset and finds patterns and crucial information by modeling how neurons in the human brain interact with one another. It is a computational system that models the brain's ability to balance the significance of some input with other inputs and deals with bias values (3). ML is used for the classification and regression tasks. The supervised learning approach involves classification and making predictions. The unsupervised learning approach is to analyze and cluster the unlabeled data in the dataset. There are some traditional supervised and unsupervised techniques in machine learning: support vector machines (SVM), random forests (RF), logistic regression (LR), K-nearest neighbor (KNN), and neural networks (NN) (4). The DL is used to extract the most significant features from the biological data and to predict a better outcome. There are some traditional DL methods like feed-forward neural networks (FNN), recurrent neural networks (RNN), convolutional neural networks (CNN), and auto-encoders (AE) (5). ML and DL methods are used to identify the relevant information in the genomic data and predict better results. Researchers in the field of genomic data are attempting to reliably identify genetic diseases, determine the primary type of disease and how it will progress, and find disease-causing genome variants.

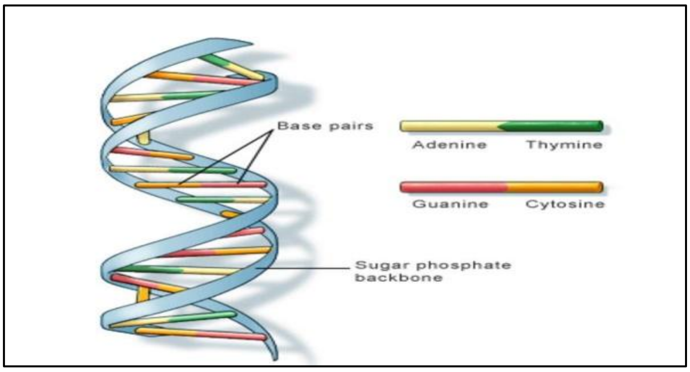

Figure 1: DNA Paired Nucleotides (6)

In order for the DNA sequence information to predict the particular class outcome appropriately, the class samples should be balanced. As a result, sequence augmentation is required to solve the imbalance problem. Traditional augmentation approaches such as sampling and SMOTE. Sampling is a data augmentation strategy used to deal with uneven datasets where the majority class outnumbers the minority class. It adjusts the distribution of classes by raising the number of samples in the minority group. SMOTE is another approach; it generates samples from the minority class. It generates a synthetically or virtually class-balanced training set and then trains the classifier. Two reconstruction techniques in deep learning, like the Variational Autoencoder (VAE) and the Generative Adversarial Network (GAN), are used to reconstruct the data into similar inputs. VAE is used to reduce the input dimensionality and reconstruct the original input. Another approach, GAN, is used to generate fake samples close to the original ones. GAN consists of two adversarial networks: a generator and a discriminator. The generator network is used to generate the fake samples. The discriminator network distinguishes between the real and fake samples. While generating the data samples based on the discriminator feedback, it will update the generator. Traditional GAN architecture is used to generate the image samples. In the generator and discriminator network, they utilized the convolutional layers for generating and classifying the samples. The DNA sequence data is a set of characters for the particular DNA sequence. The GAN technique is used to generate vast amounts of biological data, such as DNA molecule sequences. As a result, we developed the proposed model, WSEQ-GAN, to generate the sequence of characters using the GAN principle. The proposed method, WSEQ-GAN, can generate synthetic sequence samples instead of images using Wasserstein distance. The Wasserstein loss is used to increase model stability while training and includes a loss function that corresponds with sequence quality. The Wasserstein loss value assures that the quality of data samples generated by the GAN network is high, and it also ensures that the data samples generated are more realistic. The benefit of the Wasserstein loss function lies in its ability to train the model and provide a loss value that is correlated with the quality of the sequences that are generated. The generator and discriminator networks utilized the recurrent network for generating and classifying the data instead of the convolution network. This research carries out three stages. First, we generate the sequence data samples based on the original samples. Second, we convert the sequence information into numerical values for the classification task. Finally, the ML and DL techniques are used to classify the DNA sequence information. In the existing studies, the traditional GAN is used to augment the image data samples. Thus, we introduced the WSEQ-GAN method for data augmentation; it is used to generate sequence data close to the original sequence information and to predict better results. The proposed model consists of two adversarial networks: generative and discriminative. The generative sequence model utilizes a recurrent neural network instead of a convolutional neural network. The recurrent convolutional neural network is our discriminator in this research because it is very effective in sequence classification. Deep learning advancements have increased the classification and prediction of target-class information. In terms of outcomes, networks using deep learning differ from machine learning methods. Deep networks are capable of handling massive amounts of data. However, the most time-consuming aspects of deep learning networks are training and retraining, which necessitate high-performance systems. SVM, KNN, and LSTM networks are used to integrate with WSEQ-GAN to predict the DNA sequence class. Researchers are used to classify and predict diseases based on the DNA sequence information available. There are several techniques available for reading and sequencing the DNA information. This research focuses on the classification and prediction of diseases based on DNA sequence information. Gao *et al.* proposed a sequencing tool to generate an unbiased whole-genome circulating tumor DNA (ctDNA) methylation using a small amount of plasma. The aim is to develop extremely specific and sensitive biomarkers for molecular subtyping and the early diagnosis of cancer. In multicenter patient cohorts, a diagnostic signature consisting of 15 ctDNA methylation markers demonstrated high accuracy in the early detection and advanced stages, with AUCs of 0.967 and 0.971. They also managed to identify the types of cancer, such as hepatocellular carcinoma and lung cancer (7). Nurk *et al.* developed a computational technique to identify the thousands of large serine recombines (LSRs) and their DNA attachment in human cells. It increases the LSR diversity by more than 100-fold and enables the prediction of insertion location specificities. They can be classified as genome-targeting, landing pad, or multi-targeting LSRs. It achieved genome-integrating efficiencies of 40–75% (8). Logeshwaran J *et al.* proposed an improvised machine learning approach to analyze tumor sequence patterns in the human genome sequence. It analyzes and monitors the large genetic tumor sequence with the different types of tumors and their sizes (9). Das *et al*. proposed an approach for predicting cancer disease using DNA gene sequences. The proposed approach consists of three distinct numerical techniques for mapping. The 1D signal was converted into the 2D signal. The feature variables are obtained using VGG16, and the images are classified using SVM. The 2D DNA images are classified with a CNN model, which achieved an accuracy of 98.86%. The 1D CNN model achieved an accuracy of 80.36%. The model indicates that it can extract the most significant features using the CNN model and distinguish between normal and tumor liver gene sequences (10). Ritch *et al.* proposed a machine learning framework for classifying DNA repair sequences from ctDNA exomes. Particular types of DNA damage repair (DDR) defects can increase susceptibility to new medicines for prostate cancer. XGBoost-derived models performed well in identifying BRCA2, CDK12, and mismatch repair deficits in metastatic prostate cancer, with AUC values of 0.99, 0.99, and 1.00, respectively (11). Nguyen *et al.* utilized concurrent analysis of associated cancer mutations along with fragment length patterns to discriminate between mutations from numerous sources. The proposed method is used to differentiate between healthy and non-healthy people with hepatocellular carcinoma (HCC). The classification model was used to fragment the features of circulating tumor DNA (ctDNA) for genome sequencing. The model achieved an AUC of 0.88, a specificity of 81%, and a sensitivity of 89% (12). Hamed *et al.* utilized machine learning algorithms to classify the DNA sequence effectively based on its features. The study also investigates how pattern length affects the accuracy and time-based complexity of each approach. The SVM linear model achieved the lowest execution time if the pattern length varied. It also achieved the highest accuracy value of 0.963 and the highest F1 value of 0.97 (13). Senanayake *et al.* proposed the DeepSelectNet method for classifying nanopore sequencing data. DeepSelectNet is a practical solution for improving selective sequencing effectiveness. The proposed method achieved a 12% enhancement in accuracy when compared to the predecessor Squiggle Net deep learning method. It also achieved a precision and recall value of approximately 95% (14). Alshayeji *et al.* proposed a novel approach for combining machine learning and NLP for the classification of genome sequences. The author utilized 19 meta-sequences of genomic data to investigate. In the preprocessing stage, a bag of words and k-mer counting were utilized. The KNN model achieved the highest classification accuracy of 98.6%, a precision value of 98.5%, a recall value of 98.6%, and an F1-score of 98.4% for predicting the tumor DNA sequence samples (15). In this research work, we are predicting the cancer disease of the human genomic DNA sequence using Wasserstein sequence augmentation and traditional classification techniques.

Methodology

The classification in the DNA sequence gene data prediction work consists of extracting features, building classifiers for classification, generating the synthesis sequence data, and selecting optimized classifiers. This research work carried out data augmentation and data preprocessing. In the first stage of data augmentation, the sequence information in the dataset is imbalanced with class labels. So, we require sequence data augmentation techniques for predicting a better outcome. In the second stage of data preprocessing, the sequence information is converted into numerical values for selecting the significant features for classification. The proposed method is Wasserstein Sequence-GAN (WSEQ-GAN) for augmenting the sequence information close to the original sequence information. Figure 2 shows the proposed model for DNA sequence prediction.

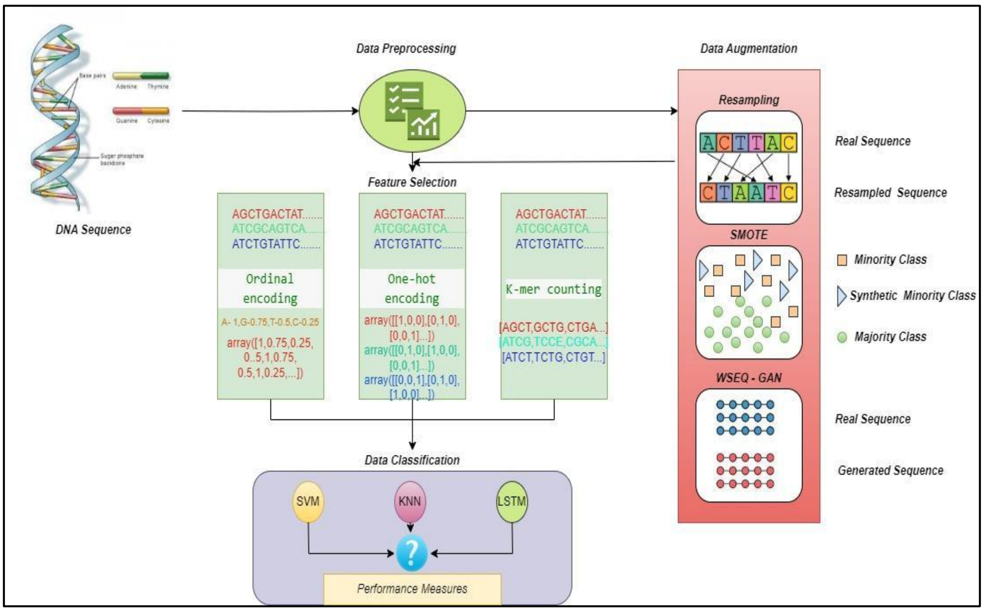

Figure 2: Proposed Work for DNA Sequence Prediction

Dataset

The DNA sequence data was collected from the Cancer Genome Atlas (TCGA) repository. The dataset contains 7 gene classes with 4603 data sequence samples. The human DNA sequence data is primarily used to predict the cancer disease with appropriate classes. The definitions for each of the seven classes in the dataset are (i) G protein-coupled receptors (GPCRs) are receptors on the cell's surface that detect substances from outside the cell and initiate physiological reactions. Its primary application is in the detection of cancer. (ii) Tyrosine kinases (TK) are signaling cascade mediators that regulate a wide range of biological activities, including development, differentiation, metabolism, and the death of cells, in response to stimuli both internal and external. Recent discoveries have linked tyrosine kinases to the pathogenesis of cancer. (iii) Protein tyrosine phosphates (TP) are enzymes that eliminate groups of phosphate from activated tyrosine amino acids in proteins. (iv) Protein tyrosine phosphates (PTPs) have been identified as major targets for a variety of disorders, including cancer, and significant efforts have been undertaken to develop novel PTP inhibitors to combat cancer growth and metastasis formation. (v) Syntase enzymes (SE) are enzymes that connect transfer RNAs to their cognate amino acids during protein translation. (vi) Ion channels (IC) are one of two types of iontophoretic proteins; ion transporters are the other. (vii) Transcription factors (TF) assist in the expression of the appropriate genes in the correct cells of the body at the correct time. Table 1 describes the number of samples presented in the DNA sequence dataset.

Table 1: Class Description of DNA Sequence

Gene Family	Number of Samples	Class Label
GPCR	531	0
TK	534	1
TP	349	2
PTP	672	3
SE	711	4
IC	240	5
TF	1343	6

Feature Selection

To select the most appropriate features in the DNA sequence is used for accurate prediction in the gene classes. To classify the gene data, the sequence information is converted into numerical values. DNA sequence data contains the nucleotide sequence and class label. The sequence information is converted into numerical values in the preprocessing stage. They observed that because categorical variables have a low cardinality, the probability distribution can be created simply using Softmax. Three traditional methods are utilized in this work for selecting the features and converting them into numerical values. The categorical variables must be transformed to binary variables using one-hot encoding, ordinal encoding, or K-mer counting.

Ordinal Encoding

In this approach, each gene expression value must be encoded as an ordinal value. For instance, "A, T, G, C" is transformed into [0.25, 0.5, 0.75, 1.0]. It uses the label encoder technique to transform the categorical values into numerical values. The float-encoded method converts the integer values to floats. The result of this method is an array of vectors for the classification task (16).

One-hot Encoding

In this approach, values are encoded in the vectors and transformed into 2-dimensional arrays. For instance, "A, T, G, C" are transformed into [1,0,0,0], [0,1,0,0], [0,0,1,0],and [0,0,0,1]. It uses the label encoder technique to convert them into numerical values. And it uses an int-encoded method to convert the value into 1. This method returns two-dimensional array vectors for classification purposes (17).

K-mer Counting

In this approach, the long DNA sequence is taken and broken down into the k-mer word length. It simply overlaps the sequence of words. For instance, if we use the word length of 4 (hexamers), "ATGCAAATC" becomes 'ATGC','TGCA','GCAA','CAAA' and soon. To transform the list of hexamer words for each of the genes into string sentences that may be used to construct the Bag of Words model. The count vector is used to transform the 4380 gene values into the uniform length feature vector on the basis of each k-mer word (length 6) in the DNA sequence (18).

Data Augmentation

The gene expression class labels are unbalanced in the DNA sequence. Unbalanced DNA sequence datasets have a distribution that is uneven in observations; that is, one class label has a large number of observations, whereas the other has a small number of observations. To predict more accurate results, we require that the data be balanced. For that, data augmentation is required to generate a synthetic sequence based on the input sequence. Traditional techniques like data re-sampling and the SMOTE technique are utilized in this work. The proposed WSEQ-GAN method is used for the data augmentation process to generate sequence information that is close to the original DNA sequence.

Oversampling (Re-Sampling)

This method is used to increase or decrease the sample size of the minority or majority class. If the dataset is imbalanced, then oversampling and under sampling techniques are used. In oversampling, the data can be replaced by the minority class. In under sampling, the data can be deleted from the majority class. If we use the sampling, some data can be lost, and it cannot predict accurate results. It is a traditional technique for resampling the data, which leads to over fitting and data to resolve the data imbalance issue. Figure 3 shows the structure of oversampling technique (19).

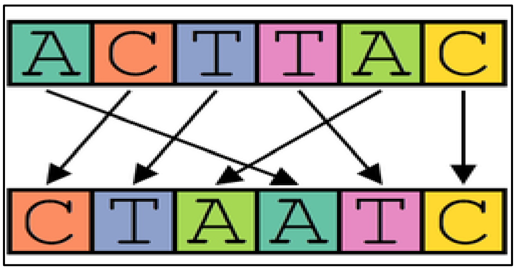

Figure 3: Oversampling Technique

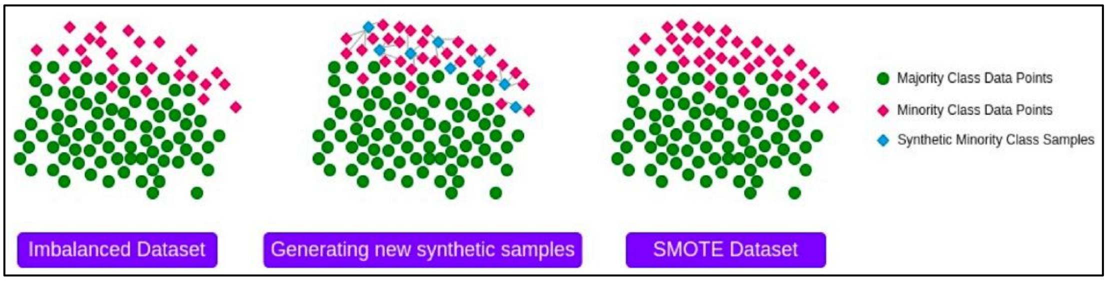
Majority Class Data Points
Minority Class Data Points
Synthetic Minority Class Samples
Imbalanced Dataset
Generating new synthetic samples
SMOTE Dataset

Figure 4: SMOTE Technique

Synthetic Minority Oversampling Technique (SMOTE)

The SMOTE technique is used to oversample the data in the minority class. SMOTE generates new instances by using previous information. SMOTE evaluates minority class instances and uses k nearest neighbors to locate a random nearest neighbor, following which an artificial instance is generated at random in feature space. Most class samples are not considered while generating synthetic samples. Figure 4 shows the structure of oversampling technique (20).

Wasserstein Sequence – Generative Adversarial Network (WSEQ – GAN)

The basic GAN approach consists of two adversarial networks: a generator (g(x)) and a discriminator (d(x)). The synthetic samples are created using the generator network. To differentiate between the original and fake samples using the discriminator network. To create the synthetic sequence based on the feedback of the discriminator. The generator's goal is to minimize cost value, whereas the discriminator's is to maximize. Traditional GAN methods are used to generate synthetic image samples close to the original samples (21). Consequently, sequence information is discrete because it can be represented by vectors, which are continuous real values. The proposed method, WSEQ-GAN, is used to create synthetic data sequence samples from the GAN network. In SEQ-GAN, the recurrent neural network is employed in the generative model. It maps the input sequence $x_1, x_2, \ldots x_i$ into the hidden sequence $h_1, h_2, \ldots h_i$ by updating the generator value recursively (22). The generator network's inputs are initialized at random using the uniform approach. The activation functions in the hidden layers are ReLU, and the activator for the output layer is softmax. The hidden function h_i of the hidden layer in the generator network is shown in equation 1. The output function o_g of the generator network utilizes the softmax layer for sequence distribution, as shown in equation 2.

$$h_i = g(h_{i-1}, x_i) \quad [1]$$

$$o_g = softmax(Wh_i + b) \quad [2]$$

In SEQ-GAN, the recurrent CNN is employed in the discriminator model. First, we adopt a bidirectional recurrent structure, which produces considerably less noise than a normal window-based neural network, in order to collect as much relevant data as possible when learning word representations (23). In the discriminator network, the inputs are fed from the generated sequence and the real sequence. All hidden layers use the ReLU activator; even the output layer uses the tanh function. The output function o_d of the discriminator network is shown in equation 3. Finally, tanh activation is used to output the probability that the input sequence is real.

$$o_d = tanh\,(Wx_i + b) \quad [3]$$

The SEQ-GAN method is a primary augmentation approach because it delivers more consistency. SEQ-GAN is used to generate the synthetic sequence data based on the distribution of the original sequence information. In the genome DNA sequence dataset, the real DNA sequence information is trained using a parameterized generator model g_θ to produce the sequence $V_{1:T} = (v_1, \ldots v_t, \ldots v_T)$, $v_t €V$, where V is the vocabulary of candidate sequences. At each time step t, the state s represents the current created tokens ($v_1,...v_{t-1}$), and the action a represents the next token v_t to select. The model $g_\theta(v_t|V_{1:t-1})$ is stochastic, but the state transition is stable after an action is chosen. We also train a parameterized discriminator model d_Φ to provide suggestions to improve the generator's performance. $d_\Phi(V_{1:T})$ is a probability indicating how likely a sequence $V_{1:T}$ is from the real sequence data or not. Based on generator G, discriminator D's gradient loss value is inconsistent. It also provides the loss value for the real/fake sequence for the entire sequence. Therefore, the use of traditional SeqGAN for sequence generation has been restricted by the discrete form of text sequence. The Monte Carlo (MC) search algorithm-based realistic sequence can be provided by the GAN at each time step. Starting with the root node, the MC search method builds child nodes for every possible combination. Every child node's value is evaluated (24). The proposed WSEQ-GAN method generates the sequence data based on the Wasserstein loss value. In the proposed method, the synthetic sequence data is generated based on the realistic sequence data for better predictions. The recurrent neural network is utilized in the generator and discriminator networks instead of the CNN for generating the synthetic sequence. The traditional SEQ-GAN method utilizes Jensen-Shannon divergence. It will locally saturate the discriminator, and the gradients will vanish. This divergence does not give the generator the freedom to generate the data samples (i.e., lack of diversity). To address the vanishing and exploding gradient problem of back propagation through time, we use Long Short-Term Memory (LSTM) cells to construct the update function in the generator network (25). To overcome this problem, the Wasserstein distance was utilized in the SEQ-GAN approach. The objective function of the Wasserstein distance is to be more stable and to avoid mode collapse. Figure 5 illustrates the generation of synthetic sequence samples from the generator g_θ and the discriminator d_Φ, which discriminates between real sequence data and fake sequence data.

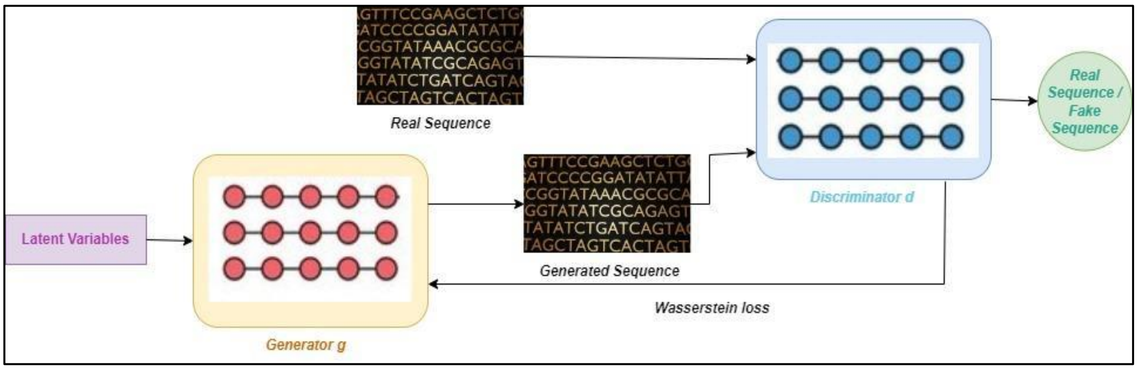

Figure 5: WSEQ-GAN

Using the discriminator d_φ as a reward function allows for continuous updating and iterative improvement of the generative model g_θ. Once we obtain a set of more accurately generated sequences using the Wasserstein distance, we will train the generator model as shown in equation 4. The Wasserstein distance helps to improve the stability of learning the parameters and to overcome the problem of mode collapse.

$$W = E_{x\varepsilon pdata}\left[d_\Phi(x)\right] - E_{z\varepsilon p(z)}\left[d_\Phi[g_\theta(z)]\right] \qquad [4]$$

When a new discriminator model is obtained, we are prepared to update the generator. The proposed strategy is based on improving a parameter to directly maximize the long-term payoff. Furthermore, WSEQ-GAN provides a gradient descent (Wasserstein loss function) that is directly related to the veracity and correctness of the generated sample data (26, 27). And it is one of the most effective and efficient remedies for GAN deterioration (loss). As a result, the vanishing gradient and mode collapse issues are effectively handled.

Data Classification

To classify the DNA sequence data using traditional classifiers like SVM, KNN, and LSTM. It is used to predict the gene class labels in the DNA sequence data. The classification task was carried out before and after the data augmentation task.

Support Vector Machine (SVM)

It is a supervised machine learning model used for classification and regression purposes. The basic goal of the SVM method is to find a hyper plane that distinguishes between data points of various classes. The hyper plane is targeted so that the biggest margin separates the classes under investigation. The SVM algorithm depicts each data item as a point in n-dimensional space (in which n is the number of features), with each value of the feature representing the value of a certain coordinate. SVM can handle data that is not linearly separable by utilizing a kernel approach to move the data into a higher-dimensional space where it can be separable linearly (28).

K-Nearest Neighbor (KNN)

It is a supervised learning model for classification and regression tasks. The basic goal of the KNN technique is to anticipate the features of a data point using the features of its neighbors. It is used to select the cluster of the new data point based on the distance between each cluster. The KNN algorithm estimates the values of newly collected data points using "feature similarity". This means that the new point is allocated a value based on how closely it resembles the training points. It is employed to get the necessary precision and accuracy for an unknown function (29).

Long Short-Term Memory (LSTM)

LSTM is a recurrent neural network that is used to process sequence information. It specializes in recognizing long-term dependencies, making it perfect for sequence prediction tasks. The recurrent neural network operations are selective read, selective forget, and selective write. The selective read captures the current input information and previous state information. The selective forget keeps only the relevant information and the remaining information will be deleted. The selective write writes only the related information in the particular state. The architecture of the LSTM network is shown in Figure 6.

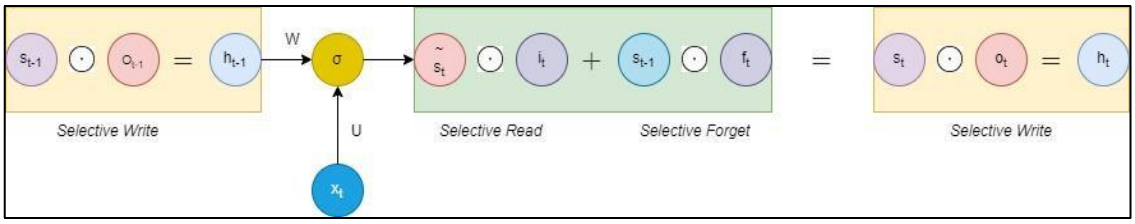

Figure 6: Working Mechanism of LSTM Network

The LSTM process is based on three gates: the input, output, and forgets gates. In the figure, the weight and bias parameters utilized in the input, output, and forget gates are W_i, W_f, W_o, b_i, b_f, and b_o. The input gate i_t stores current information and previous information in the hidden neuron h_{t-1}, and the bias value b_i is stored in the particular state, as represented in equation 5. The forget gate f_t stores the relevant information and tells the state to throw away the irrelevant information; it is represented in equation 6. The output gate o_t provides the information needed to activate the final layer, as represented in equation 7. Equation 8 shows that the hidden information is created from the input and the previous time step t. Equation 9 shows that we are receiving the hidden information from the previous state. Equation 10 shows that the output of the RNN network is represented as RNN_{out} (ht) (30, 31).

$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i) \quad [5]$$

$$f_t = \sigma(W_f h_{t-1} + U_f x_t + b_f) \quad [6]$$

$$o_t = \sigma(W_o h_{t-1} + U_o x_t + b_o) \quad [7]$$

$$\sim s_t = \sigma(W h_{t-1} + U x_t + b) \quad [8]$$

$$s_t = f_t \cdot s_{t-1} + i_t \cdot \sim s_t \quad [9]$$

$$h_t(rnn_{out}) = o_t \cdot +\sigma(s_t) \quad [10]$$

LSTM is used to predict the class labels in the DNA sequence information. To extract the DNA structural features using long short-term memory to precisely predict enhancement elements in genomics data (32). LSTMs are more flexible than GRUs since they contain more gates and parameters. It also performs complex tasks like detecting patterns. Because of their diverse cell states, LSTMs are able to store and output numerous kinds of data. LSTMs are significantly better at dealing with long-term dependence. This is related to their ability to retain information over long periods of time. LSTMs are quietly less sensitive to the vanishing gradient problem (33, 34).

Results and Discussion

The proposed WSEQ-GAN model's performance has been evaluated using traditional classifiers like SVM, KNN, and LSTM. The proposed method WSEQ-GAN provides better augmentation results than sampling and the SMOTE technique. Table 2 shows that the prediction on DNA sequence samples using different classification techniques with sequence augmentation. The LSTM classifier achieved better classification accuracy, precision, recall, and F1-score compared with SVM and KNN. The K-mer feature selection technique achieved better results compared with ordinal encoding and hot encoding.

Table 2: Predictive Results for DNA Sequence

Classifiers	**Feature Selection Methods**	**Augmentation Techniques**	**Accuracy**		**Precision**		**Recall**		**F1-Score**	
			BA	**AA**	**BA**	**AA**	**BA**	**AA**	**B**	**AA**
SVM	Ordinal Encoding	Sampling	0.71	0.83	0.72	0.84	0.73	0.87	0.72	0.85
		SMOTE	0.83	0.90	0.81	0.87	0.79	0.84	0.80	0.85
		WSEQ-GAN	0.87	0.91	0.89	0.94	0.88	0.93	0.88	0.93
	One-hot Encoding	Sampling	0.81	0.85	0.74	0.88	0.78	0.89	0.76	0.88
		SMOTE	0.83	0.88	0.87	0.91	0.87	0.91	0.87	0.91
		WSEQ-GAN	**0.88**	**0.90**	**0.89**	**0.94**	**0.88**	**0.93**	**0.88**	**0.93**
	K-Mer Counting	Sampling	0.91	0.92	0.96	0.98	0.89	0.98	0.92	0.98
		SMOTE	0.93	0.94	0.94	0.95	0.91	0.96	0.92	0.95
		WSEQ-GAN	**0.89**	**0.95**	**0.88**	**0.93**	**0.89**	**0.94**	**0.88**	**0.93**
KNN	Ordinal Encoding	Sampling	0.73	0.84	0.86	0.87	0.81	0.84	0.83	0.85
		SMOTE	0.75	0.86	0.83	0.86	0.82	0.86	0.82	0.86
		WSEQ-GAN	**0.86**	**0.90**	**0.87**	**0.91**	**0.87**	**0.93**	**0.87**	**0.92**
	One-hot Encoding	Sampling	0.81	0.93	0.78	0.87	0.78	0.84	0.76	0.85
		SMOTE	0.83	0.94	0.88	0.97	0.89	0.96	0.90	0.92
		WSEQ-GAN	**0.89**	**0.95**	**0.89**	**0.94**	**0.88**	**0.93**	**0.88**	**0.93**
	K-Mer Counting	Sampling	0.82	0.89	0.92	0.93	0.78	0.89	0.83	0.91
		SMOTE	0.84	0.93	0.93	0.94	0.76	0.87	0.83	0.90
		WSEQ-GAN	**0.86**	**0.91**	**0.89**	**0.94**	**0.87**	**0.93**	**0.88**	**0.93**
LSTM	Ordinal Encoding	Sampling	0.78	0.83	0.81	0.88	0.79	0.85	0.79	0.86
		SMOTE	0.81	0.84	0.83	0.87	0.82	0.89	0.82	0.88
		WSEQ-GAN	**0.88**	**0.94**	**0.89**	**0.95**	**0.89**	**0.95**	**0.89**	**0.95**
	One-hot Encoding	Sampling	0.87	0.89	0.84	0.86	0.85	0.88	0.84	0.86
		SMOTE	0.84	0.90	0.86	0.92	0.87	0.93	0.86	0.92
		WSEQ-GAN	**0.89**	**0.95**	**0.91**	**0.94**	**0.91**	**0.93**	**0.91**	**0.93**
	K-Mer Counting	Sampling	0.89	0.94	0.89	0.91	0.88	0.89	0.88	0.90
		SMOTE	0.88	0.93	0.91	0.94	0.89	0.93	0.90	0.93
		WSEQ-GAN	**0.91**	**0.98**	**0.94**	**0.97**	**0.93**	**0.97**	**0.93**	**0.97**

Table 2 compares the outcomes of the traditional classifiers, like SVM, KNN, and LSTM. It also states that the before augmentation (BA) and after augmentation (AA) of the sequence data. It measures the performance of accuracy, precision, recall, and F1-Score values. It clearly shows that the sequence data after augmentation gives better results than before augmentation. The WSEQ-GAN method with LSTM can achieve better results before and after data augmentation when compared to other classification and augmentation techniques. The proposed WSEQ-GAN augmentation technique exceeds both sampling and the SMOTE technique by showing an average variation of 4.11% in accuracy, 3.72% in precision, 5% in recall, and 4.33% in F1-score. The machine learning and deep learning classifiers are used to classify the sequence data before and after augmentation. The LSTM can achieve 98% accuracy and 97% precision, recall, and F1-score value for augmented data. It also achieved 91% accuracy, 94% precision, and 93% recall and F1-value for non-augmented data when compared to SVM and KNN.

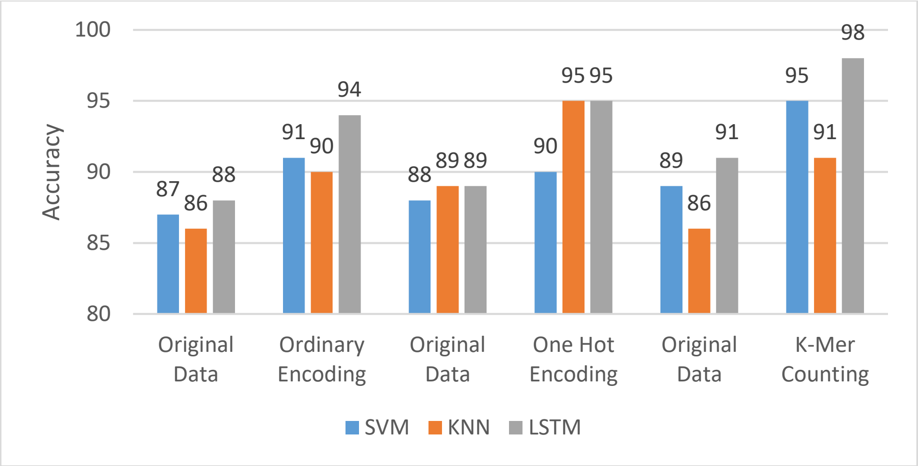

Figure 7: Accuracy Value obtained from Original and Augmented Data

Figure 7 depicts the accuracy value obtained after augmentation techniques using different classifiers and feature selection techniques. The proposed WSEQ-GAN method produces better results than the sampling and SMOTE techniques. The WSEQ-GAN with LSTM classifier and k-mer counting technique achieved the highest accuracy of 98%. From the existing analysis of the accuracy value obtained from previous study (13), it is noted that the accuracy outcome difference is 1.7% in the proposed method. The existing (32) DNA sequence classification using ML techniques achieved the highest accuracy value of 90.9%; the variation in the proposed model is 7.1% enhancement.

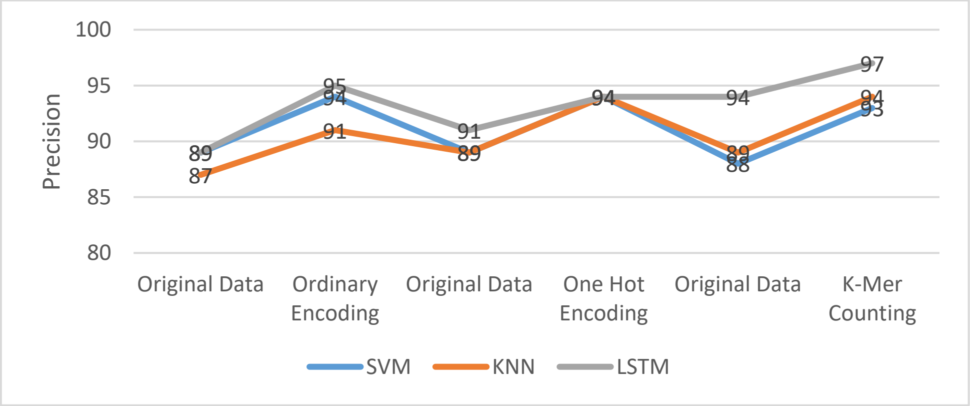

Figure 8: Precision Value Obtained using WSEQ-GAN

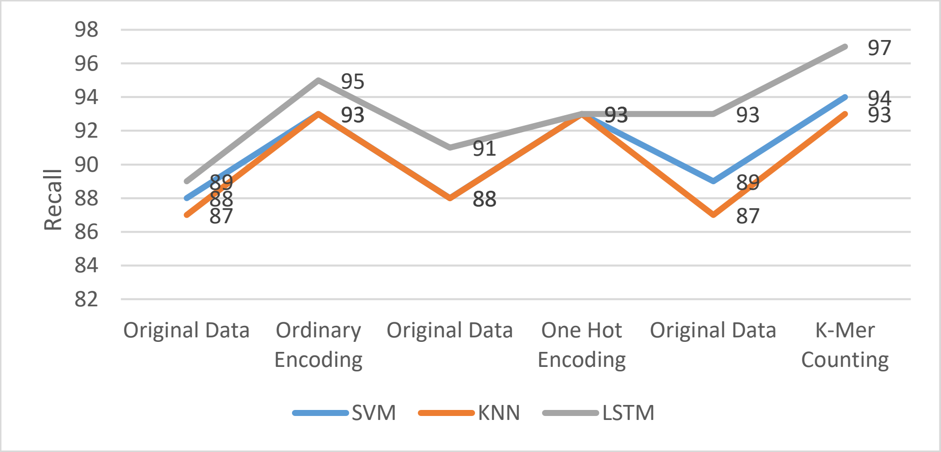

Figure 9: Recall Value Obtained using WSEQ-GAN

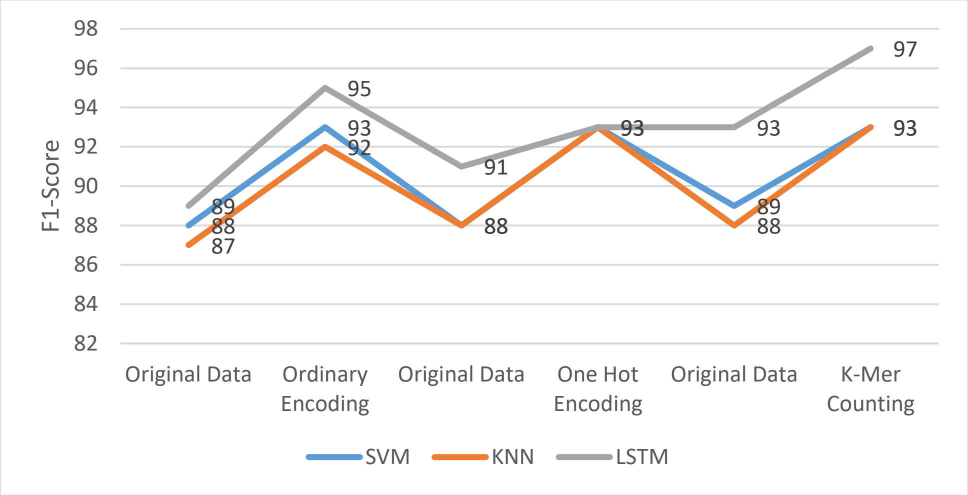

Figure 10: F1-Score Value obtained using WSEQ-GAN

Figure 8, 9 and 10 depicts the precision, recall, and F1-score values obtained from the proposed method, WSEQ-GAN, with different classifiers. The proposed augmentation method with an LSTM classifier achieved better results for DNA sequence data. It achieved 97% precision, recall, and F1-Score values. The recall value will ensure increasing the dataset size. From the existing analysis (12), the recall value is 89%; to compare our proposed model, it improved around 8% of the recall value. In the existing sequence data classification (14), the recall and precision values achieved 95%; while comparing our proposed method, it achieved around 97% and enhanced around 2%. The result shows that the proposed WSEQ-GAN method can utilize critical applications like cancer for predicting a better outcome and to solve the data imbalance problem in the sequence data.

Conclusion

The genome information of all living organism molecules is encoded in the DNA sequence. The four nucleotides A, T, C, and G are the basic building blocks of a DNA sequence. DNA sequence analysis is a challenging task for analyzing the huge amount of sequence information based on a variety of classes. In this research, feature selection techniques are applied to convert the sequence information into numerical values. The proposed method WSEQ-GAN is applied to generate a fake sequence similar to the original sequence to solve the data imbalance problem. It is compared with traditional augmentation techniques like sampling and SMOTE. The proposed WSEQ-GAN augmentation method shows a significant improvement in the results. For classification tasks, LSTM is used to classify the augmented and non-augmented sequence information. It is compared with traditional ML approaches like KNN and SVM. The proposed WSEQ-GAN along with the LSTM classifier and the K-mer feature selection technique achieved better classification results. DNA sequencing is useful in a variety of fields; it can be mainly dominant in the healthcare sector. The sequence information is used to identify the specific disease and the drug discovery. Sequence information is employed in this research to predict accurate diseases with better results. Subsequent studies employing ML and DL techniques for DNA genome sequencing will focus on creating medications for particular kinds of samples.

Abbreviation

Nil.

Acknowledgement

This work was supported by Vellore Institute of Technology, Vellore, India.

Author Contributions

Ravindran U: Conceptualization, Methodology, and Manuscript preparation, Gunavathi C: Guidance, Investigation, and Validation of the Manuscript.

Conflict of Internet

The authors declare that there is no conflict of interest in the research work.

Ethics Approval

Not applicable.

Funding

None.

References

1. Roth SC. What is genomic medicine? Journal of the Medical Library Association: JMLA. 2019 Jul;107(3):442.
2. Luquette LJ, Bohrson CL, Sherman MA, Park PJ. Identification of somatic mutations in single cell

DNA-seq using a spatial model of allelic imbalance. Nature communications. 2019 Aug 29;10(1):3908.

3. Botes M. Regulating scientific and technological uncertainty: The precautionary principle in the context of human genomics and AI. South African Journal of Science. 2023 Jun;119(5-6):1-6.
4. Ayodele TO. Types of machine learning algorithms. New advances in machine learning. 2010. https://api.semanticscholar.org/CorpusID:53061796
5. Ravindran U, Gunavathi C. A survey on gene expression data analysis using deep learning methods for cancer diagnosis. Progress in Biophysics and Molecular Biology. 2023 Jan 1; 177:1-3.
6. Molnar C, Gair J. Concepts of biology. BCcampus; 2015. (https://openlibrary-repo.ecampusontario.ca/jspui/handle/123456789/345)
7. Gao Y, Zhao H, An K, Liu Z, Hai L, Li R, Zhou Y, Zhao W, Jia Y, Wu N, Li L. Whole-genome bisulfite sequencing analysis of circulating tumour DNA for the detection and molecular classification of cancer. Clinical and Translational Medicine. 2022 Aug;12(8):e1014.
8. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, Aganezov S. The complete sequence of a human genome. Science. 2022 Apr 1;376(6588):44-53.
9. Logeshwaran J, Adhikari N, Joshi SS, Saxena P, Sharma A. The deep DNA machine learning model to classify the tumor genome of patients with tumor sequencing. International Journal of Health Sciences. 2022 Jul;6(S5):9364-75.
10. Das B, Toraman S. Deep transfer learning for automated liver cancer gene recognition using spectrogram images of digitized DNA sequences. Biomedical Signal Processing and Control. 2022 Feb 1;72:103317.
11. Ritch EJ, Herberts C, Warner EW, Ng SW, Kwan EM, Bacon JV, Bernales CQ, Schönlau E, Fonseca NM, Giri VN, Maurice-Dror C. A generalizable machine learning framework for classifying DNA repair defects using ctDNA exomes. NPJ Precision Oncology. 2023 Mar 13;7(1):27.
12. Nguyen VC, Nguyen TH, Phan TH, Tran TH, Pham TT, Ho TD, Nguyen HH, Duong ML, Nguyen CM, Nguyen QT, Bach HP. Fragment length profiles of cancer mutations enhance detection of circulating tumor DNA in patients with early-stage hepatocellular carcinoma. BMC cancer. 2023 Mar 13;23(1):233.
13. Hamed BA, Ibrahim OA, Abd El-Hafeez T. Optimizing classification efficiency with machine learning techniques for pattern matching. Journal of Big Data. 2023 Jul 25;10(1):124.
14. Senanayake A, Gamaarachchi H, Herath D, Ragel R. DeepSelectNet: deep neural network based selective sequencing for oxford nanopore sequencing. BMC bioinformatics. 2023 Jan 28;24(1):31.
15. Alshayeji MH, Sindhu SC. Viral genome prediction from raw human DNA sequence samples by combining natural language processing and machine learning techniques. Expert Systems with Applications. 2023 May 15;218:119641.
16. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. bioinformatics. 2007 Oct 1;23(19):2507-17.
17. Choong AC, Lee NK. Evaluation of convolutionary neural networks modeling of DNA sequences using ordinal versus one-hot encoding method. In2017 International Conference on Computer and Drone Applications (IConDA) 2017 Nov 9 (pp. 60-65). IEEE..
18. Melsted P, Pritchard JK. Efficient counting of k-mers in DNA sequences using a bloom filter. BMC bioinformatics. 2011 Dec;12:1-7.
19. Hu J, He X, Yu DJ, Yang XB, Yang JY, Shen HB. A new supervised over-sampling algorithm with application to protein-nucleotide binding residue prediction. PloS one. 2014 Sep 17;9(9):e107676.
20. García-Pedrajas N, Pérez-Rodríguez J, García-Pedrajas M, Ortiz-Boyer D, Fyfe C. Class imbalance methods for translation initiation site recognition in DNA sequences. Knowledge-Based Systems. 2012 Feb 1;25(1):22-34.
21. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. Advances in neural information processing systems. 2014;27. https://doi.org/10.48550/arXiv.1406.2661
22. Goodfellow I, Bengio Y. courville A. Deep learning. 2016;1(2). https://doi.org/10.4258/hir.2016.22.4.351
23. Lai S, Xu L, Liu K, Zhao J. Recurrent convolutional neural networks for text classification. InProceedings of the AAAI conference on artificial intelligence 2015 Feb 19;29(1). https://doi.org/10.1609/aaai.v29i1.9513
24. Yu L, Zhang W, Wang J, Yu Y. Seqgan: Sequence generative adversarial nets with policy gradient. InProceedings of the AAAI conference on artificial intelligence 2017 Feb 13;31(1). https://doi.org/10.1609/aaai.v31i1.10804
25. Hochreiter S. Long Short-term Memory. Neural Computation MIT-Press. 1997.
26. Xiao Y, Wu J, Lin Z. Cancer diagnosis using generative adversarial networks based on deep learning from imbalanced data. Computers in Biology and Medicine. 2021 Aug 1;135:104540.
27. Ravindran U, Gunavathi C. Cancer Disease Prediction Using Integrated Smart Data Augmentation and Capsule Neural Network. IEEE Access. 2024 Jun 10.
28. Zou C, Gong J, Li H. An improved sequence based prediction protocol for DNA-binding proteins using SVM and comprehensive feature analysis. BMC bioinformatics. 2013 Dec;14:1-4.
29. Muflikhah L, Widodo N, Mahmudy WF. Prediction of Liver Cancer Based on DNA Sequence Using Ensemble Method. In2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI). IEEE. 2020 Dec 10:37-41.
30. Zhang Y, Qiao S, Ji S, Li Y. DeepSite: bidirectional LSTM and CNN models for predicting DNA–protein binding. International Journal of Machine Learning and Cybernetics. 2020 Apr;11:841-51.

31. https://www.youtube.com/watch?reload=9&v=9TFnjJkfqmA
32. Seth D, Dharmanshu Mahajan KP, Khanna R, Chugh G. Gene Family Classification Using Machine Learning: A Comparative Analysis. InInternational Conference on Data Analytics & Management 2023 Jun 23 (pp. 343-360).
33. Singh N, Nath R, Singh DB. Splice-site identification for exon prediction using bidirectional LSTM-RNN approach. Biochemistry and Biophysics Reports. 2022 Jul 1;30:101285.
34. Kaur A, Chauhan AP, Aggarwal AK. Prediction of enhancers in DNA sequence data using a hybrid CNN-DLSTM model. IEEE/ACM transactions on computational biology and bioinformatics. 2022 Apr 13;20(2):1327-36.