A Study of Gene Characteristics and Their Applications using Deep Learning

Prajjwal Gupta1, Saransh Bhachawat2, Kshitij Dhyani3, B.K. Tripathy4

1, 2 School of Computer Science and Engineering, VIT, Vellore, Tamil Nadu - 632014, India

3, 4 School of Information Technology and Engineering, VIT, Vellore 632014, India

1 prajjwal.gupta2019@vitstudent.ac.in, 2 saransh.bhachawat2019@vitstudent.ac.in,

3 emailofkshitijdhyani@gmail.com, 4 tripathybk@vit.ac.in

**Abstract** DNA sequencing deals with figuring out the order of arrangement of the bases in the DNA. These bases are the building blocks of DNA molecules and their arrangement mostly determines the genetic information carried within a DNA segment, therefore sequencing becomes a very important aspect in the field of genomics. Now it becomes ever more important to optimize this process of sequencing and analysis and the field of deep learning has a lot to offer. Autoencoders are artificial neural networks which are trained in an unsupervised manner to obtain feature representation or dimensionality reduction. Now as clustering is difficult to perform for data with large dimensions, autoencoders can be used to reduce the dimension of data by associating each gene cluster with an autoencoder. Genetic algorithms are algorithms which are based on Darwin’s law of evolution and provide a better alternative to traditional clustering algorithms which have been found to have various drawbacks when implemented for genetic data. Drug repositioning is the examination of existing drugs on new disease targets and pharmacogenomics, looking to predict the target’s response to a drug. Deep learning acts as a powerful tool for repositioning drugs by allowing us to perform robust predictions and provide deep insights to drug-disease combinations. This chapter aims to provide the reader with various deep learning models and analysis algorithms which have been employed in some or the other forms for studying gene characteristics and gene development or have the potential to form the basis for ground breaking research for the same.

**Keywords:** Gene clustering, autoencoders, genetic algorithms, RNN, sequencing, drug repositioning.

1 Introduction

The field of genomics has seen an exponential rise of novel deep learning models and genetic algorithms for the analysis, characterization sequencing of genes and also in the development of various unconventional practices such as DNA storage and drug repositioning. Gene data is complex enough that it requires high computational power for processing and analysis. To serve this requirement, there has been a need for development of newer and advanced algorithms and models to tackle this computational problem. Development of new drugs or treatments for a new disease or a variant of an existing disease requires a lot of research, testing and manual work, these efforts can be greatly reduced through drug repositioning with the help of powerful deep learning models. Almost everything which we desire to analyze in the field of genomics, requires us to sequence the DNA, therefore DNA sequencing becomes another center point and an important aspect across a plethora of domains. The primary distinction between Deep Neural Networks (DNN) and other machine learning algorithms lies in their processing of raw natural data. While other machine learning algorithms require external experts for selecting the feature of raw data DNN has its own mechanism to select the features. In DNNs representational learning method is followed. Thus a set of methods are used, which determine the representation of raw data. DNN uses several levels, each level representing a non-linear module and finally composing all of these. The pattern followed is that higher layer modules are more abstract than those in the lower layers and thus the complex functions are learnt.

DNN algorithms have provided us with methods such that problems which were considered to be unsolved by using the other AI techniques could be solved. Different domains of government, business and science could apply DNN techniques because of its capability to handle high dimensional data [36]. For example, in the case of potential drug molecule predictions using several ML techniques are now closed and it has supported in analyzing particle accelerator data.

DNN techniques are applied in numerous application areas now. It has been used in Image processing areas [1]. In [8], the techniques used for the classification of Audio signals have been presented nicely. Among the first few applications of DNN is image retrieval through feature learning. In this direction an architecture is proposed in [10]. Also, in [26] a method for such information retrieval using text based techniques is discussed. Specific parts of human society require the recognition of sign languages and some techniques in this direction are presented in [21]. A useful variant of DNN is the Recurrent Neural Networks (RNN). RNNs are used in the study of sentiment analysis. Which in turn are useful in the study of human behaviour [7]. It goes without saying that the development and advancement of DNN are largely responsible for many advanced research in AI [4]. S-LSTM-GAN, which happens to be a shared neural networks model is proposed in [2] and has been established to be a very efficient method. Generative adversarial networks (GAN) happen to be an important development in the field of DNN, which is used to develop a single image super resolution technique in [3]. Another important variant of DNN is the Convolutional Neural Network (CNN). The concepts associated with these models and their components are discussed in detail in [19] along with their working principles. A difficult but interesting application is to estimate the age and gender from images. However, a successful attempt using Wide ResNet has been proposed in [9].

Deep learning has enabled us to look beyond just the nucleotide sequence and has helped us discover hidden patterns and behaviours within the genomic sequence. Deep Learning models and techniques such as RNN are gaining more and more traction in the field of genomics since they do not require instructions to be specified to them explicitly. This traction has invited innovators to develop novel technologies with high accuracy and applicability, the latest developments are discussed in this chapter.

2 Deep Learning Techniques for Gene Clustering

2.1 Gene Clustering

Gene expression refers to the process by which the information encoded by a gene is used to synthesize a functional gene component/product which often produce and regulate protein as an end product. These proteins in turn define the function of the cell. Thus thousands of genes expressed together determine the function of the cell. At each step in the flow of information from DNA to RNA to proteins, the genes can control and regulate the type and amount of protein to be manufactured. A gene cluster is a group of genes which have similar encoding for polypeptides or proteins and share a common generalized function. They are usually within a few thousand base pairs of each other. Clustering and analysis provides an insight to various genetic and physical interactions, pattern recognition, cross-species analysis, mutation, evolution and development of superior or healthier gene sequences.

Since genetic data has a complex and high dimensional nature with high intra-variance, clustering of such data is difficult and expensive, resulting in low quality gene clusters. Thus there is a need to encode the gene data into smaller dimensions while retaining essential and sensitive information. Thus we can employ the use of Autoencoders to find an optimal feature representation of the gene data to optimize and increase the accuracy of the clustering algorithms.

Also further, traditional methods of sampling and clustering such as K-means produces low quality clusters which are not very useful in the development of high level gene sequences. Newer genetic clustering algorithms like GenClust++ and HEMI++ solve the drawbacks of K-means while producing high quality clusters.

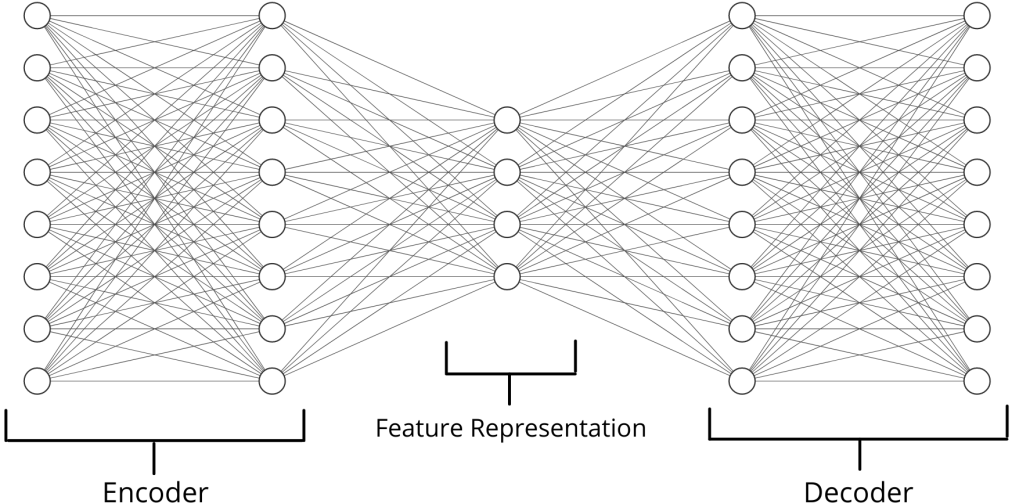
2.2 Autoencoders and novel techniques for gene clustering

Autoencoders are artificial neural networks trained in an unsupervised manner to learn efficient data encodings/ feature representation. An autoencoder essentially performs dimensionality reduction by removing the input signal “noise” [18]. It provides a nonlinear data mapping to a space with lower dimensions (figure 3.1). An autoencoder has two parts: the encoder and decoder. The encoder is the non-linear mapping that we discussed above and the decoder ideally performs accurate reconstruction of the encoded data to the input data provided to the encoder. The aim of the autoencoder is to minimize the reconstruction error.

The autoencoder has to be sensitive enough so that it can accurately build a reconstruction and insensitive enough so that it doesn’t overfit and memorize the training data. By this tradeoff, the model only maintains the variations required to reconstruct the data and redundancies are reduced.

Sensitive to the inputs enough to accurately build a reconstruction.

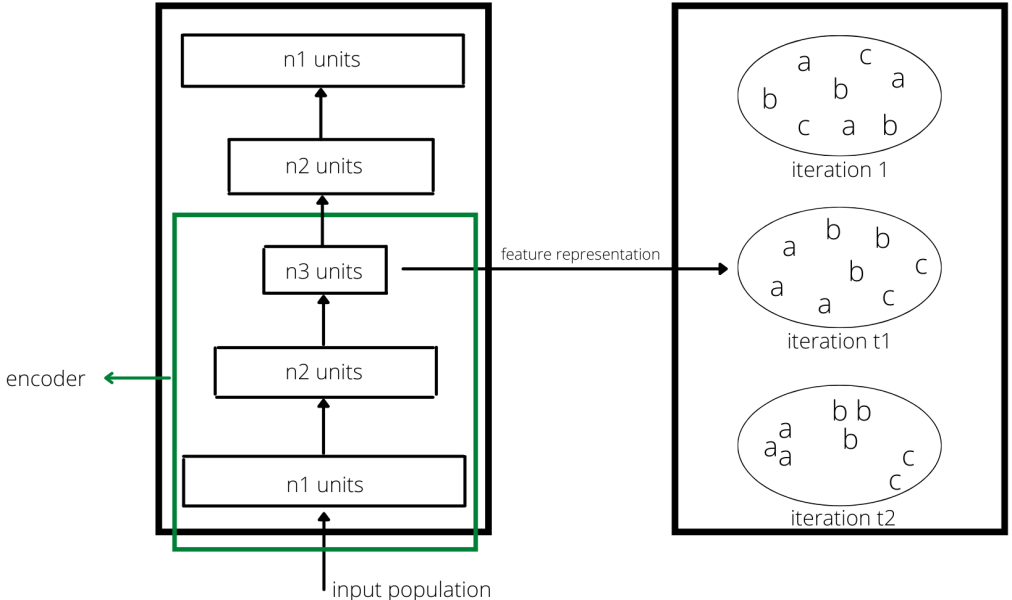
The loss function for such a model consists of two terms: the reconstruction loss, which is the difference between the original input data and reconstructed data, and a regularization term which is used to prevent overfitting. The sensitivity tradeoff can be regulated by multiplying a constant (λ) to the regularization term whose value is set between 0 and 1.



**Fig. 1.** Autoencoder components

Now for clustering encoding can be implemented in two strategies: single autoencoder based clustering, association of each cluster with an autoencoder [16].

The first strategy involves training of the autoencoder model on the entire initial population. But this won’t be an effective strategy as the model would learn only the common features between the ideal clusters and discard the features native to the clusters as noise. Thus the clustering algorithm followed by this type of encoding would produce undesirable results. The second strategy is a more apt one as it involves the association of each cluster with its own autoencoder, i.e. an autoencoder trained on the samples from the cluster itself.



**Fig. 2.** Clustering using autoencoders

The algorithm for autoencoders based clustering [16] is: For dataset X, no. of clusters K (depending on the clustering algorithm, may or may not be required as a user input), hyper-parameter λ and number of iterations T perform:

Initialize cluster 0 (C0) randomly.

For iteration t (t <= T)

-Update the mapping network by minimizing the objective function [reconstruction loss + (λ \* clustering cost function)] with gradient descent (or any other suitable algorithm)

-Update the cluster centers

-Partition X into K clusters and assign samples to respective clusters

-increment t

Now onto clustering algorithms. K-means [5] and Fuzzy C-means are two of the simplest most commonly used algorithms used for clustering. However, they require the user to input the number of clusters k, but for genetic data it is difficult for a user to determine the number of clusters beforehand. Apart from that k-means heavily depends on the quality of initial seeds/samples and a bad set of initial seeds would result in the formation of bad quality clusters. K-means also has a tendency to get stuck in local minima, thus producing poor results.

To overcome these drawbacks, various genetics clustering algorithms have been proposed: AGCUK, GAGR, GenClust [12], HeMI [1], GenClust++ [3] and HeMI++ [2]. Genetic Algorithms (GA) are randomized search and optimization algorithms based on Darwin’s law of evolution: Survival of the fittest. GA primarily involves 5 steps: population initialization, selection, crossover, mutation and elitist operation.

GenClust algorithm produces high quality initial population. But the complexity of the algorithm is O(N2), N being the total number of records in the dataset. Also it requires the user to enter the radius of clusters for the initial population which is a difficult task for the user. HeMI also produces high quality clusters and it does not require any user input for cluster radii while keeping the complexity of initial population selection as low as O(N). GenClust++ is another algorithm for creating high quality of initial population with no user input and low complexity of O(N). It chooses the initial population probabilistically in contrast to random selection used in HeMI. HeMI picks half of the initial population from the set chromosomes obtained through K-means (k ranging from 2 to 10), and the rest of the half obtained from random k. HeMI++ is similar to HeMI but has an added advantage that it learns reasonable properties from the clusters formed and uses this information/knowledge to produce clustering solutions.

**Table 1.** Comparison of novel genetic clustering algorithms [2]

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | Complexity | Tree Index | Silhouette Coefficient Rank\* |
| K-Means | O(N2) | 27.41 | 4.26 |
| K-Means++ | O(N2) | 31.01 | 4.40 |
| GenClust | O(N2) | 5.27 | 3.50 |
| HeMI | O(N) | ꝏ | 5.90 |
| HeMI++ | O(N) | 0.55 | 1.13 |

\*Silhouette coefficient rank computed on multiple datasets as illustrated in the paper.

2.3 HeMI++ algorithm for sensible clusters [2]

HeMI++ algorithm selects only the best chromosomes depending on the fitness value, 50% of which come from a deterministic phase and 50% from random phase. The selection of the initial population in the deterministic phase is done by a k-means algorithm with k (number of clusters) ranging from 2 to 10. Datasets in which the number of clusters is more than 10, the range of k is set from 2 to √n where n is the total number of records in the dataset. The HeMI++ algorithm described in the paper is as follows:

The user has to define/input the following parameters:

a. Number of streams (m)

b. Number of intervals (G)

c. Number of iteration (I)

The first step is to carry out normalization of the data points. Then carry out population initialization for each stream. Now select sensible properties for each cluster. For each iteration carry out the following operations in order: Noise-Based Selection, Crossover Operation, Twin Removal, Three steps Mutation Operation, Health Improvement Operation, Cleansing Operation, Cloning Operation and The Elitist Operation. Carry out “I” number of iterations for each stream. After finishing these operations for each stream, carry out Neighbour Information Sharing. After completing the Neighbour Information Sharing, proceed to the next interval. The steps are the same for each interval. After completion of all the intervals, perform Global Best Selection.

The components of this algorithm are:

i. Normalization: It refers to scaling the attributes of the data so that each attribute is treated equally irrespective of their domain size.

**ii. Number of streams:** HeMI++ uses a multiple stream approach in order to utilize a big population. The population is divided into multiple streams in which each stream contains a small number of chromosomes.

**iii. Population Initialization:** The initial population is selected in two phases: deterministic (p/2) and random (p/2) phase as discussed above. The value of p is set to 20 for HeMI++.

**iv. Selection of sensible properties:** From the initial population, HeMI++ selects ‘p’ number of best chromosomes based on their DB index (goodness of fit). Each generation uses this sensible clustering solution so that chromosomes may not contradict the properties belonging to their cluster.

**v. Noise based selection:** Chromosomes from two generations are compared to select chromosomes for the next genetic operations.

**vi. Crossover Operation:** Crossover operation for chromosomes is performed in pairs. The best chromosome in the population is taken as the first chromosome and the second one is determined probabilistically. To perform the crossover operation, the chromosomes are divided into two segments, and one segment from the first chromosome is swapped with a segment from the second chromosome.

**vii. Twin Removal:** In some cases, genetic operations like mutation and crossover may create a condition wherein there are identical/twin genes in a chromosome. To handle such cases, HeMI++ performs twin removal. If there are twin genes in a chromosome and the length of the chromosome is more than two, then HeMI++ simply removes one of the genes. But if there are twin genes and the length of the chromosome is two, then HeMI++ randomly changes the attribute value of one of the genes.

**viii. Three step mutation operation:** Division, absorption and random changes are performed on chromosomes in this step.

**ix. Health Improvement Operation:** This operation ensures that the health of the chromosomes is consistent or improved in each generation.

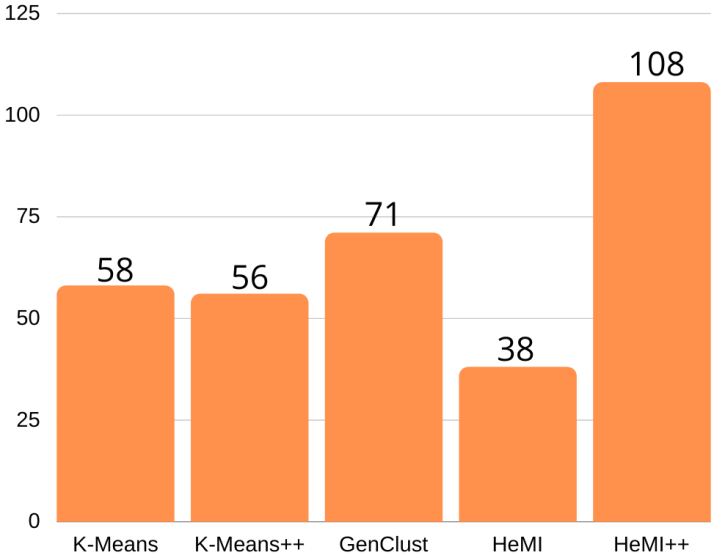
**x. Cleansing Operation:** This operation aims to identify the sensible and non-sensible chromosomes belonging to the population.

**xi. Cloning Operation:** In this operation, the sick chromosomes found in the cleansing operation are replaced.

**xii. The Elitist Operation:** In this operation, the best chromosome throughout the generation is taken and passed to the next generation in order to improve the quality of population in the next generation.

**xiii. Neighbour Information Sharing:** It involves sharing of the best chromosomes among the neighbouring streams at a regular interval like every 10th iteration.

**xiv. Global best selection:** This component is used to find the best chromosome across all streams.



**Fig. 3.** Performance scores of clustering algorithms

3 DNA sequencing using RNN

We are all essentially just carbon atoms, attached with functional groups made up of hydrogen and oxygen. This coherent existence of carbons along with other functional groups gives rise to various kinds of proteins. These proteins along with water, make up what we know as the cell. At the center of this cell lies the nucleus, inside which comfortably sits the key to evolution – the DNA. The information stored inside the DNA is the directing force of protein production, which in turn are vital for physical, mental, biological growth and development. It also carries the hereditary material in almost all organisms, and hence has crucial biological and evolutionary importance. Owing to its vital significance in multiple aspects, it becomes essential that we understand its constituents to attain a better understanding. DNA is made up of four bases, namely adenine (A), cytosine (C), guanine (G), and thymine (T). The sequence of these very bases determines the protein assembly instructions. Therefore sequencing has become the center of almost every development in the fields of biotechnology, molecular biology, forensics, etc. Attempts to sequence genomic data have been made since the mid-1900s, ranging from Sangers sequencing methods to the recent concepts of nanopore sequencing. The latest to join this arsenal of sequencing are the Deep learning techniques. The popularity of Deep Learning techniques in genomics could be accredited to the ability of the DL techniques to learn relevantly new features from the old features without being told what to do. The most accepted architecture amongst all, are the Recurrent Neural Networks, due to their special ability to support feedback connections to previous layers. Deep learning techniques have leaped past simple sequencing of the base sequence, and have dived deeper into finding the intricacies within the genomic data. Novel innovations using DL architectures can be seen emerging in this new space of innovation [34].

3.1 Applications of DNA Sequencing

DNA Sequencing has become a center point of innovation across many areas of biology since the sequence of bases in the DNA determines almost all characteristics showcased by a majority of the organisms. As a result, it has countless applications across many domains.

3.1.1 Molecular Biology

Molecular Biology studies the composition of cellular molecules and has special interests in nucleic acids and proteins. To study these very nucleic elements we gravely require fast sequencing technologies. This helps us understand and identify phenotypes, associate certain types of gene change with diseases, also to understand the effect of medicine and drugs on specific gene types. Deep learning has become popular amongst molecular biologists since it helps them dive deeper into the biological secrets of the DNA sequence.

3.1.2 Virology

Virology has gained special popularity after the pandemic onset. Viruses enjoy benefits due to their ultra-small size, this renders them invisible to a light microscope, and dealing with them extremely difficult. This makes sequencing the only tool which gives us insights into the structure and behavior of viruses. Sequencing methodologies proved especially useful to sequence the genome of the Coronavirus. Deep learning techniques have allowed us to understand further depths from the obtained sequences.

3.1.3 Forensics

Forensics is often the turning point of many criminal cases, with faster and more robust sequencing techniques, we can attain speedy results on DNA profiling and paternity testing. This has resulted in the resolution of many pending criminal cases, and also the release of a few wrongly accused. Deep learning techniques have accelerated the speed of advancements by multiple folds.

3.1.4 Metagenomics

Metagenomics aims to identify all the different types of microbes present in a microbial ecosystem. Faster and next-generation sequencing techniques have enabled us to identify new and different microbes within a microbiome. Many Deep learning models are being proposed which perform taxonomical classification with higher accuracy than the state of the art genus identification tools. Deep Microbes is one such DLM that showed an accuracy of around 89.40% [6].

3.1.5 Medicine

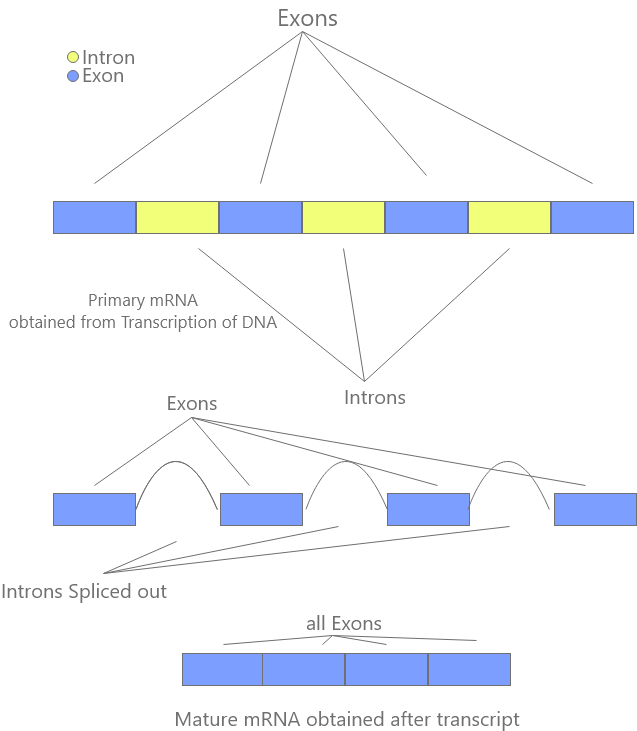
DNA majorly controls the physical mental and biological characteristics portrayed by all living organisms. Therefore through DNA sequencing, we can get insights into how our body might react to certain medicinal compounds; moreover, it helps us identify genetic disease in the early stages. Additionally, we can sequence the DNA of the bacterial genomes, which might give us insights into developing more effective antibiotics and treat disease more effectively.

3.2 Latest advances in DNA Sequences using Deep Learning

Deep learning has identified various new vertices in DNA sequencing, which has developed a whole new level of understanding of biological processes. We have gained deeper insights into working at cellular levels and microscopic interactions.

3.2.1 Splice Junction Prediction in DNA Sequence Using Multi-Layer RNN Model

Genes are made of up DNA and hence are the driving force for protein synthesis. At times a single gene can take charge of synthesizing multiple proteins, this is due to the post-transcriptional modification. This post-transcriptional modification removes intron as shown in figure 4.1 enzymatically and this site of removal is known as a splice junction.



**Fig. 4.** Splicing of Genes

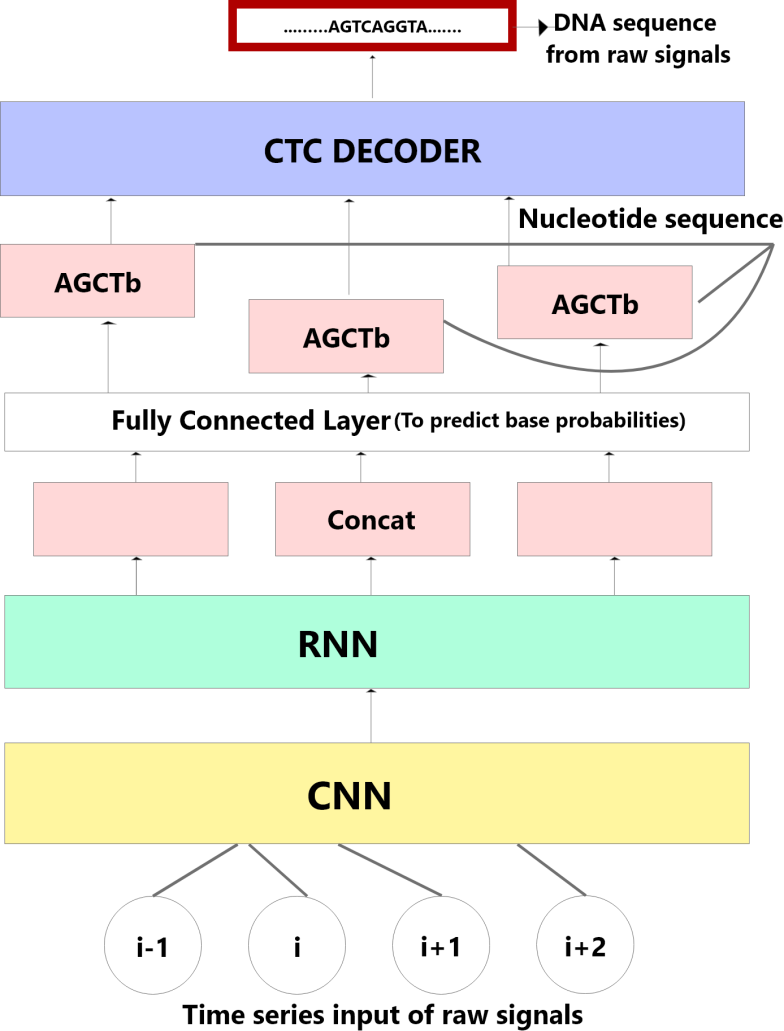
These splice joints are of special concern since they assist in understanding the protein that is being synthesized and its nature, function, and characteristics. These splice junctions can be predicted by sequencing the DNA and then analyzing it, and the Deep Learning technique utilizing RNN seems to perform this prediction with high accuracy [13]. A three-layered RNN model utilizing the Molecular Biology data set by Murray et al. achieved an accuracy of 99.95%.

3.2.2 Deep learning method for identification of short viral sequences from metagenomes

Viruses and their high mortality effects are something we have all become familiar with after the Covid-19 epidemic. Therefore identifying the viral sequences and understanding their characteristics from those sequences is essential. An RNN based deep learning model is discussed named VinSeeker, which exhibits higher accuracy compared to other standard virological sequencing models for short sequences. The data set on which the VinSeeker was trained, contained sequences of well-known Virus genomes. To fully prove the technology the proposed mechanism was also trained and tested on a CAMI dataset and also a dataset involving the human gut metagenome [7]. RNN based VinSeeker outperformed similar sequencing models with a greater AUROC score and precision.

3.2.3 Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning

Nanopore Sequencing is a rapidly growing and advancing technology. It possesses the capabilities to not only sequence long DNA/RNA fragments at high speeds but also to generate entire genome assemblies, spot modified base sequences in a DNA, and recognize and classify microbes in a metagenome. The technology at the base root employs a bed of membrane wells which each contains a nanopore. Nanopores have an active electrical current running through them. These nanopores are further augmented with a few tethers which allow the DNA strands to be linked with the nanopore opening. Once linked, the nucleic acid-containing DNA strands are run through the nanopore, this causes the calculated disruptions within the electric current running through the nanopore, and the resulting electric signal obtained is then decoded to get the DNA sequence. This decoding process is where the Chiron steps in, it is a hybrid neural network that coherently integrates an RNN and a CNN along with a CTC decoder (Connectionist Temporal Classification) as shown in figure 4.2.



**Fig. 5.** Chiron- an RNN and CNN hybrid deep learning model architecture

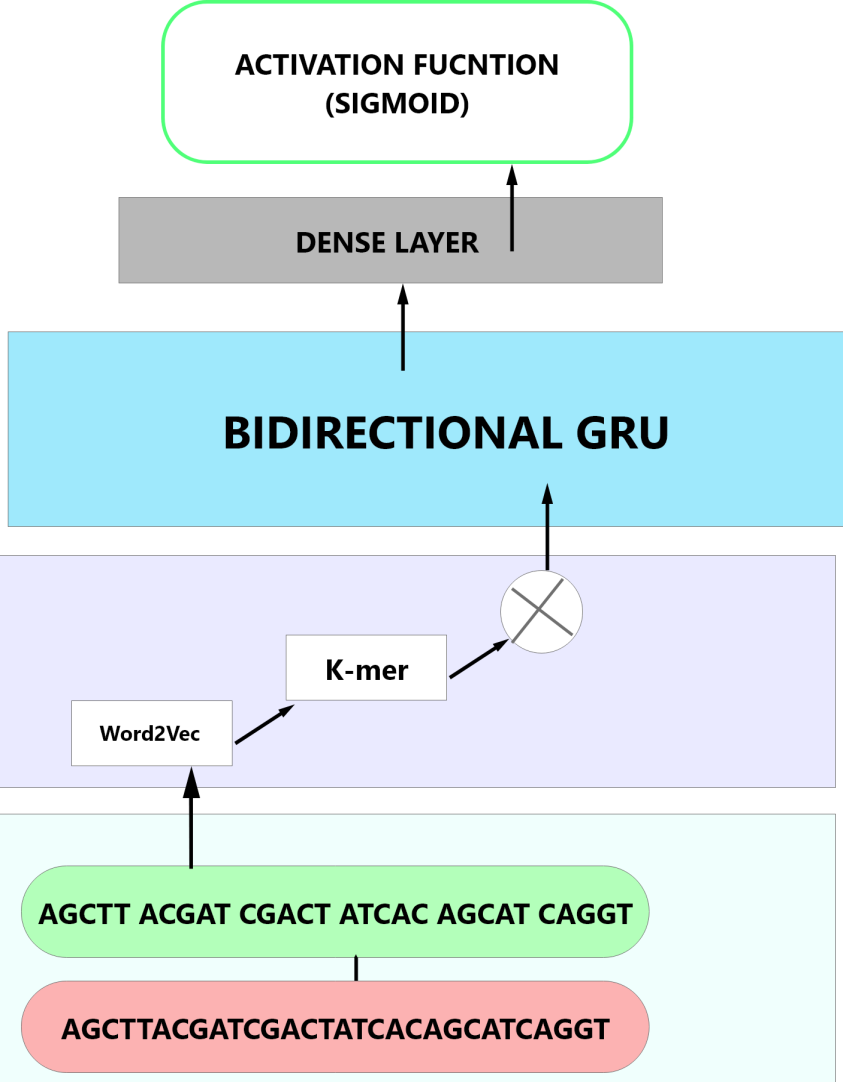
This was trained and tested on multiple genomes namely Lambda virus, E. coli, and human, and showed to perform better than the other market standards [17].

3.2.4 Biren: predicting enhancers with a deep-learning-based model using the DNA sequence alone

Transcription is a very vital process, as it is responsible for the expression of the genes and controls the extent of development of many characteristics in different types of cells. This transcription process is amplified when enhancer sequences attach themselves to specific proteins. These enhancer sequences have high importance since they majorly control and regulate the gene expression process. Therefore predicting and identifying these enhancer sequences can be very beneficial and could give us insight into many challenges being faced by the bioinformatics industry. However, accurate enhancer prediction methods are absent.

A recent solution being proposed utilizes a hybrid of Convolution neural networks and a modified Recurrent Neural Network (RNN). Regular RNN’s support only forward states progression. The proposed solution employs a concept of the bidirectional flow of input information. Bidirectionality allowed the network to extract features from the DNA sequence more efficiently. This was the Neural Network can process input data in both positive and negative time direction. This modified RNN model is called the Bidirectional recurrent neural network (BRNN) and the proposed technology is being called BiRen. BiRen essentially aims to directly identify enhancer sequences from a DNA sequence. This is achieved by deploying hybrid architecture of CNN and BRNN. The CNN half employs the DeepSEA model which encodes the original DNA Sequence into label vectors. The BRNN half predicts the probability of enhancer sequences after learning from the features obtained from the label vectors. Both the CNN and BRNN work in harmony and utilize the strengths of each other to showcase excellent potential. BiRen showed better performance than other such enhancer prediction models with an AUC of 0.956 [20].

Another proposed model based on a Bidirectional Gated Recurrent Unit network with k-mer embedding called KEGRU showed promise in identifying the transcriptional factors binding site. The KEGRU model divides the DNA sequence into k-mer sequences. Each k-mer sequence is treated as a word and fed to the word2vec algorithm for being vectorized.



**Fig. 6.** Deep learning model architecture of KEGRU for predicting enhancers

This data is then used by the deep Bidirectional Gated Recurrent Unit for the learning of features. Once the features are learned, the classification into TF and non-TF binding sites is performed as shown in figure 3. KEGRU model possesses an advantage over other models crediting to its robustness [14].

3.2.5 A Deep Learning Model for Predicting NGS Sequencing Depth from DNA Sequence

Next-generation sequencing (NGS) techniques are vastly based around parallelism's distributive nature. This has resulted in most NGS protocols being centered on the breaking of genome sequences into shorter fragments. These shorter genome fragments are then read individually and later computationally overlapped to make meaningful sequences. This results in each nucleotide being sequenced multiple times over and over. This is where the concept of Sequencing Depth comes into the picture, the number of iteration of reading, a nucleotide goes under is known as the Sequencing Depth. Sequencing depths have a great impact on sequencing cost and also on the accuracy of the models being used. Hence understanding the sequencing depth can sometimes be very essential for a project.

To solve this challenge, a deep learning model is being conceptualized which takes the DNA sequence and probability of parity as input and predicts with accuracy the sequencing depth of the different NGS panels. It is built upon a recurrent neural network to facilitate the capture of both short-range and long-range interaction using gated recurrent units (GRUs) within the DNA. The proposed deep learning model (DLM) was tested on 2 different NGS panels with an accuracy of 99% and 93% [22].

3.2.6 Deep Recurrent Neural Network for Protein Function Prediction from Sequence

DNA stores the instructions to produce various kinds of proteins and these proteins, in turn, help in carrying out various functional processes. However, understanding the exact functions of these numerous proteins has been a long-standing challenge.

A deep learning model based on recurrent neural networks appears to solve this by converting it into a problem of classification, where classes represent the protein function. The RNN model is powered by a long short term memory (LSTM) unit, and trained on a UniProt dataset [8]. The proposed model seems to show several advantages over similar tools and has several potential benefits.

3.3 Future Scope

The extent and the speed at which we can sequence the DNA plays a crucial role across multiple disciplines. Therefore it has been attracting innovators and research to develop ever faster sequencing technologies and techniques. The future is very bright as new discoveries are being uncovered in the field of biology that might just be the next big break for DNA sequencing [15].

3.3.1 DNA Storage

Every day millions of bytes of data are being stored and processed, at this rate, there is an inevitable data explosion just around the corner. There is not enough metal on earth to sustain the upcoming data needs. However, there are a few technologies that could help us contain this data explosion, one such being DNA storage. We know that DNA already stores information and instructions about protein synthesis proteins. Therefore it has become a very real possibility to store data in the DNA as well. DNA has a much greater storage density and can store quantities of magnitude much greater than any present technology on the planet [4]. Storing the data is just the first half of the story, we need fast sequencing techniques to be able to read data at high speeds, since without high-speed reading capabilities, storing large data in DNA is like throwing it down an inaccessible well. The increasing popularity of DNA sequencing and the in-pour of funds in the Bioinformatics industry could result in the entrance of DNA based storage devices in the market very shortly.

3.3.2 Exploration of Genome Diversity

Countless genomes are waiting to be sequenced and we have just barely scratched the surface. With concepts such as Metagenomics picking up attention, we are regularly gaining insights into microbial diversity. Countless biomes possess rich genomic diversity and gaining a large scale understanding of these different genomic sequences can be beneficial in understanding why the DNA functions the way it functions. This will also give information about protein structure determination and an overall better understanding of the environment we see around ourselves.

3.3.3 Developmental Biology

We are already capable of cultivating and nurturing many plants and animals in labs under delicate and mindful conditions. However, we lack knowledge of the actual development process on a cellular level. Some events and processes such as the gene expressions can be observed and analyzed in a more broad and general way by editing the genes. This editing requires sequencing techniques and through editing we can observe the effects of each edit and develop a better understanding. Many deep learning solutions have emerged in the past couple of years which bring promising possibilities to the table [35].

3.3.4 Extensive Monitoring of Nucleic acids

Technologies such as nanopore sequencers allow us to sequence data at high speeds for a specific genome. A disseminated, extensive, well-coordinated network of nanopore sequencers could result in a real-time monitoring system of nuclide acids. This could result in innovative applications such as real-time air and water quality testing, food tasting, and human body testing.

4 Deep learning for repositioning of drug and pharmacogenomics

Time and again new diseases and disorders are discovered all over the globe. The HIV, Covid-19 pandemic and Cushing syndrome are some examples of diseases which have affected numerous people all over the world. We are constantly in the process of developing cures for these diseases but the process of finding and developing a new cure is often very time consuming and requires very large amounts of funds, therefore making us unable to provide quick cures for these diseases and also leads to side-lining and neglecting of the rare diseases. To overcome these issues drug manufacturers often use an alternative method called drug repositioning or repurposing. Drug repositioning is the examination of existing drugs on new disease targets and pharmacogenomics potentially looking to predict the target’s response to a drug, potentially saving many lives and time by reducing the cost and time to produce a new drug. Drug repositioning has been a booming field of research in the past year as the need for faster drug production echoed throughout the planet. Repositioning of drugs is also significantly cheaper than building your own drug from the scratch and the safeties of the repositioned drugs are pre-determined from previous preclinical tests. Viagra is a famous example of a repositioned drug where although its initial use was meant for curing heart related diseases, now it is widely used to cure erectile dysfunction. Another example of this is the “Thalidomide” medicine which was highly detrimental when it was released initially, was successfully repositioned and transformed into an effective cancer drug therapy [9].

During drug repositioning features of drugs are taken along with tests taken across on patients diagnosed with a disease. This is then passed into a Deep Learning model which uses various algorithms such as autoencoders, variational autoencoders, deep walking etc. These algorithms provide us with the architecture to determine and compare similarities between the drug features, drug-disease, drug-side-effect, drug-target, etc. It is through these similarities that the model predicts potential drugs for the target diseases. A large abundance of such data, increases the likelihood of predicting the drug. With Deep Learning models becoming extremely powerful with predictive power, drug repositioning is becoming more and more feasible by the day.

4.1 Applications of Drug Repositioning

Drug repositioning has been cited as one of the best alternatives to manufacturing new drugs. It reduces cost and is a very time efficient process. It also provides a much higher safety than manufacturing a new drug. Drug repositioning is being used to battle various diseases which require immediate solutions or are very costly and rare to develop a new drug. Cancer and Covid-19 are some of the diseases where drug repositioning is being used for a cure [19].

4.2 Novel Deep learning methods for Drug Repositioning

As more and more companies are getting attracted towards drug repositioning there has been a significant rise in the research for more powerful deep learning models. These models incorporate a number of deep learning algorithms together.

Some of these algorithms are [10]:

i) Logistic regression and kernel regression: Logistic regression model is used to estimate a binary dependent variable and predict the probability of a certain class. A Kernel regression on the other hand is used to estimate the expectation of a random variable based on certain conditions.

ii) Random forest: A random forest is an ensemble method for prediction, classification, estimation etc. It works by constructing multiple decision trees and giving the output of the class which is the mode or mean of the prediction outputs of individual trees.

iii) Autoencoder: An autoencoder is an artificial network used to obtain a feature representation of high dimensional data (as discussed in section 3.2).

iv) Variational autoencoders: A variational autoencoder is used to generate the probability of the latent vector in a confined space to represent the high dimensional data.

v) Support Vector machine: Support vector machine algorithm is used to find an optimal hyper plane to classify different data points.

A Deep learning model for repositioning of drugs needs to follow certain steps (figure 1) in order to identify new scope for existing drugs. The following five steps fulfil this requirement [11]:

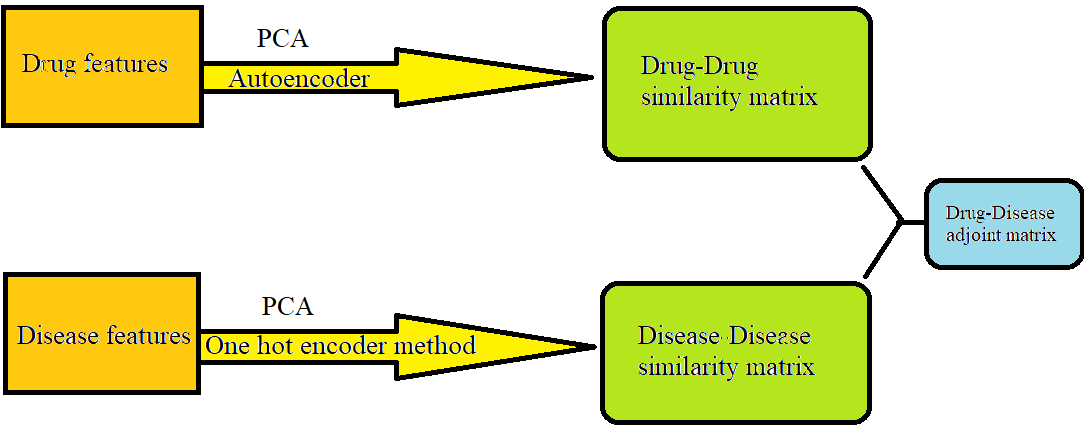
i) Representing drug features: In the first step i.e. the drug feature representation we take as input data of various drug features such as the chemical structure and use neural networks such as autoencoders to perform PCA and dimensionality reduction and obtain an optimal feature representation.

ii) Transforming disease features: In this step we aim to identify disease-disease similarities. First we convert the categorical data to a form that can be understood by the model using the one hot encoder technique. After the conversion of the data we determine the disease-disease similarities using PCA.

iii) Using drug features to construct the drug-drug similarity matrices: We use the extracted drug features from the previous steps to compare the properties/features of different drugs and construct a similarity matrix.

iv) Using drug features to construct the disease-disease similarity matrices: We use the extracted disease features from the previous steps to compare the properties/pharmacogenomics of different diseases and construct a similarity matrix.

v) Using drug-drug similarity and disease-disease similarity to construct drug-disease association matrices: We use the information from the existing drug and disease matrices to form an association matrix where we once again compare the similarities between the drugs and diseases to form drug, disease pairs.



**Fig. 7.** Steps followed by a data learning model.

Let us now discuss a novel deep learning model based on the methods mentioned above [21]:

The deepDR neural network architecture has been developed to facilitate repositioning of drugs by integrating various networks such as the drug-drug, drug-disease, drug-side-effects etc. The architecture uses autoencoders to determine the high-level features of the drugs from the network. Then the low-dimensional representations of these features are encoded and decoded through a variational autoencoder. Thus it finally approves drugs for some shortlisted candidates which were not originally approved for these drugs.

5 Future Scopes

As discussed above drug repositioning is a widely supported alternative to novel drug manufacturing. Having already seen successful results such as thalidomide, Viagra, etc., drug repositioning has become an accepted method at various drug research institutions. Drug repositioning also saw a significant boost in research in the year 2020 due to the covid-19 pandemic. It is seen as a good method to find a cure for covid-19 due to the urgent requirement of a vaccine. It has also been used for finding cures for various other diseases including different types of cancers. Drug repositioning not only helps manufacturers save money in building a new drug, it also recycles the existing drug in the market. During the current circumstances drug repositioning can safely be called the next big step in the drug manufacturing industry and so we can also expect better and more advanced deep learning models to be built to assist with these requirements.

6 Conclusions

In this chapter we saw how gene analysis on different levels can form the basis for extraction of characteristics on higher levels. We talked about clustering the genes based on existing sequences and gene expressions and how genes taken from these clusters can be sequenced in a specific manner to find applications in various fields. Such a method can be employed in developing mutations, predicting successful cross breeding sequences, development of new drugs or repositioning the drugs to treat some other diseases. Further refinement of the approaches or development of new ones can significantly help the medical world in the development of newer treatment practices and drugs in a lesser amount of time. Some deep learning models like a RNN can also pave the way to the commercial scaling of DNA storage which could be the next ground breaking industry in the upcoming years.

7 References

[1] Adate A, Tripathy BK (2018) Deep learning techniques for image processing. In: Machine Learning for Big Data Analysis Berlin, Boston, De Gruyter, pp. 69-90

[2] Adate A, Tripathy BK (2019) S-lstm-gan: Shared recurrent neural networks with adversarial training. In: Proceedings of the 2nd International Conference on Data Engineering and Communication Technology, Springer, Singapore, pp. 107-115

[3] Adate A, Tripathy BK (2019) Understanding single image super resolution techniques with generative adversarial networks. In: Advances in Intelligent Systems and Computing, Springer, Singapore, Vol. 816, pp. 833–840

[4] Adate A, Tripathy BK, Arya D, Shaha A (2020) Impact of deep neural learning on artificial intelligence research. In: Deep Learning Research and Applications, De Gruyter Publications, Vol. 7, pp. 69–84

[5] Beg, A. H., Islam, M. Z., & Estivill-Castro, V. (2016). Genetic algorithm with healthy population and multiple streams sharing information for clustering. Knowledge-Based Systems, 114, 61-78.

[6] Beg, A. H., Islam, M. Z., & Estivill-Castro, V. (2020, July). HeMI++: A Genetic Algorithm based Clustering Technique for Sensible Clusters. In 2020 IEEE Congress on Evolutionary Computation (CEC) (pp. 1-8). IEEE.

[7] Baktha K, Tripathy BK (2017) Investigation of recurrent neural networks in the field of sentiment analysis. In: Proceedings of IEEE International Conference on Communication and Signal Processing, pp. 2047-2050

[8] Bose A, Tripathy BK (2020) Deep learning for audio signal classification. In: Deep Learning Research and Applications, De Gruyter Publications, pp. 105–136

[9] Debgupta R, Chaudhuri BB, Tripathy BK (2020) A Wide ResNet-Based Approach for Age and Gender Estimation in Face Images. In: Proceedings of International Conference on Innovative Computing and Communications, Springer, Singapore, pp. 517-530

[10] Garg N, Nikhitha P, Tripathy BK (2014) Image retrieval using latent feature learning by deep architecture. In: Proceedings of IEEE International Conference on Computational Intelligence and Computing Research, pp. 1-4

[11] Islam, M. Z., Estivill-Castro, V., Rahman, M. A., & Bossomaier, T. (2018). Combining K-Means and a genetic algorithm through a novel arrangement of genetic operators for high quality clustering. Expert Systems with Applications, 91, 402-417.

.[12] Koch, J., Gantenbein, S., Masania, K., Stark, W. J., Erlich, Y., & Grass, R. N. (2020). A DNA-of-things storage architecture to create materials with embedded memory. Nature biotechnology, 38(1), 39-43.

[13] Krishna, K., & Murty, M. N. (1999). Genetic K-means algorithm. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 29(3), 433-439

[14] Liang, Q., Bible, P. W., Liu, Y., Zou, B., & Wei, L. (2020). DeepMicrobes: taxonomic classification for metagenomics with deep learning. NAR Genomics and Bioinformatics, 2(1), lqaa009.

[15] Liu, F., Miao, Y., Liu, Y., & Hou, T. (2020). RNN-VirSeeker: a deep learning method for identification of short viral sequences from metagenomes. IEEE/ACM Transactions on Computational Biology and Bioinformatics.

[16] Liu, X. (2017). Deep recurrent neural network for protein function prediction from sequence. arXiv preprint arXiv:1701.08318.

[17] Low, Z. Y., Farouk, I. A., & Lal, S. K. (2020). Drug Repositioning: New Approaches and Future Prospects for Life-Debilitating Diseases and the COVID-19 Pandemic Outbreak. Viruses, 12(9), 1058. https://doi.org/10.3390/v12091058

[18] Luo, H., Li, M., Yang, M., Wu, F. X., Li, Y., & Wang, J. (2020). Biomedical data and computational models for drug repositioning: a comprehensive review. Briefings in bioinformatics.

[19] Maheshwari K, Shaha A, Arya D, Rajasekaran R, Tripathy BK (2019) Convolutional Neural Networks: A Bottom-Up Approach. In: Deep Learning Research and Applications 7:21-50

[20] Moridi, M., Ghadirinia, M., Sharifi-Zarchi, A., & Zare-Mirakabad, F. (2019). The assessment of efficient representation of drug features using deep learning for drug repositioning. BMC bioinformatics, 20(1), 1-11.

[21] Prakash V, Tripathy BK (2020) Recent Advancements in Automatic Sign Language Recognition (SLR). In: Computational Intelligence for Human Action Recognition, CRC Press, pp. 1-24

[22] Rahman, M. A., & Islam, M. Z. (2014). A hybrid clustering technique combining a novel genetic algorithm with K-Means. Knowledge-Based Systems, 71, 345-365

[23] Sarkar, R., Chatterjee, C. C., Das, S., & Mondal, D. (2019, March). Splice junction prediction in DNA sequence using multilayered RNN model. In International Conference on Emerging Trends in Engineering (pp. 39-47). Springer, Cham.

[24] Shen, Z., Bao, W., & Huang, D. S. (2018). Recurrent neural network for predicting transcription factor binding sites. Scientific reports, 8(1), 1-10.

[25] Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., & Waterston, R. H. (2019). Publisher Correction: DNA sequencing at 40: past, present and future. Nature, 568(7752), E11-E11

[26] Singhania U, Tripathy BK (2020) Text-Based Image Retrieval Using Deep Learning. In: Encyclopedia of Information Science and Technology, Fifth Edition, IGI Global, USA, pp. 87-97

[27] Song, C., Liu, F., Huang, Y., Wang, L., & Tan, T. (2013, November). Auto-encoder based data clustering. In Iberoamerican congress on pattern recognition (pp. 117-124). Springer, Berlin, Heidelberg.

[28] Teng, H., Cao, M. D., Hall, M. B., Duarte, T., Wang, S., & Coin, L. J. (2018). Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. GigaScience, 7(5), giy037.

[29] Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P. A. (2008, July). Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th international conference on Machine learning (pp. 1096-1103).

[30] Xue, H., Li, J., Xie, H., & Wang, Y. (2018). Review of drug repositioning approaches and resources. International journal of biological sciences, 14(10), 1232.

[31] Yang, B., Liu, F., Ren, C., Ouyang, Z., Xie, Z., Bo, X., & Shu, W. (2017). BiRen: predicting enhancers with a deep-learning-based model using the DNA sequence alone. Bioinformatics, 33(13), 1930-1936.

[32] Zeng, X., Zhu, S., Liu, X., Zhou, Y., Nussinov, R., & Cheng, F. (2019). deepDR: a network-based deep learning approach to in silico drug repositioning. Bioinformatics, 35(24), 5191-5198.

[33] Zhang, J., Yordanov, B., Gaunt, A., Wang, M., Dai, P., Chen, Y. J., ... & Zhang, D. (2020). A Deep Learning Model for Predicting NGS Sequencing Depth from DNA Sequence.

[34] Dutta, P., Patra, A. P., & Saha, S. (2021). DeePROG: Deep Attention-based Model for Diseased Gene Prognosis by Fusing Multi-omics Data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

[35] Tasaki, S., Gaiteri, C., Mostafavi, S., & Wang, Y. (2020). Deep learning decodes the principles of differential gene expression. *Nature machine intelligence*, *2*(7), 376-386.

[36] Biswas, R., Vasan, A., & Roy, S. S. (2020). Dilated deep neural network for segmentation of retinal blood vessels in fundus images. *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, *44*(1), 505-518.