TRANSFER LEARNING AND MACHINE LEARNING FOR GENE PATTERN PREDICTION AND PROBABILITY PREDICTION OF DISEASE

Dr.S.Hemalatha[1], Dr.A.Swaminathan[2], Pravallika Oggu[3], A.S.AronSanura[4].

#1 Professor: Head of the Department, Computer Science and Business Systems,Panimalar Institute of Technology,Chennai-123
#2 Associate Professor , Department, Computer Science and Business Systems,Panimalar Institute of Technology,Chennai-123
#3,4 Student: Computer Science and Business Systems, Panimalar Institute of Technology, Chennai-123

ABSTRACT:

In this paper we are going to use transfer learning and machine learning techniques which help us classify the genes that are prone to the genetic disorder. The First method which has been procedurally used in every predictions to design the predictive model is the feature screening. In this method the fundamental goal of feature screening is to quickly and effectively decrease the feature space's ultrahigh dimensionality to a manageable size while keeping all of the crucial features in tact. After feature screening, more advanced techniques can be used on a smaller feature space for additional analysis, such as parameter estimation. applications, transfer learning—which involves transferring patterns discovered on one dataset to another for the purpose of building prediction models—has proven to be successful.Identifying genes from vast amount of data is very difficult and to make this process easier , many Machine Learning algorithms have been developed .We examine how machine learning may help with early diagnosis, medical imaging interpretation, the discovery and development of new drugs, and other areas in this study.The deep neural network (DNN) is a kind of deep learning (DL) that analyses high-dimensional data using several hidden layers. Because of the large complexity of bioinformatics data, DNN might be a viable model for bioinformatics research. Aside from image and text data from hospital medical files, other types of laboratory data, which are mostly made up of numbers, must be analyzed. However, few research have employed DNN to analyze structured numerical medical data.As an example, we trained our model to predict the genes that cause a genetic condition Sickle cell anaemia.

INTRODUCTION:

Developing a disease risk prediction model is a critical first step in the current push for precision medicine, a new healthcare system that tailors treatments based on patient profiles. Several disease-associated genetic variations have been uncovered during the last several decades that may be used to predict genetic vulnerability utilising whole genome sequencing and genome-wide association studies (GWAS).Even for highly heritable qualities, individual genetic variants often only account for a modest part of the heritability. Since most Non-communicable illnesses with significant effects on public health are polygenic, it is crucial to jointly model these genetic variants in order to create an effective prediction model.

From the whole set of genomic data that is used here , we aim at building our model that best performs to detect the disorder Sickle cell Anemia .Red blood cells (RBCs) in sickle cell disease (SCD) have the shape of a sickle or 1/2-moon rather than the smooth, round form typical of healthy cells. Because of their uneven form, the cells can become stored in blood arteries, resulting in blood flow obstructions throughout the body. Because to their lack of flexibility and uneven form, sickle erythrocytes have a shorter life and consequent anemia, also known as sickle-cell anemia. SCD cannot be cured, but it may be managed to relieve discomfort and avoid complications.A difficult and exhausting procedure for detecting SCD is to analyze a patient's supplementary blood specimens under the eye of a microscope and identify sickle cells from those samples. The procedure we follow here is , from the whole data of genes , we train our model to specifically identify a gene that is responsible to develop Sickle cell , we choose this example to prove that if we train our models in this way we will surely be able to detect all the possible disorders . For this process, the first method used here is Image pre-processing,feature extraction that includes shape,color,texture and RBC classification. We outline our suggested screening rule and deep transfer learning model before analyzing how well our approach performs in terms of prediction and selection.Here we used DNN as the amount of data from which the model has to predict the disorder is huge ,DNN is utilized here because each node in the hidden layer develops the two associations and assesses the significance of the input to estimating the output; why not stack more and more of these on above each other to get the most the hidden layer's perks?

As a result, the deep net contains numerous hidden levels. 'Deep' refers to a model's layers having several layers.

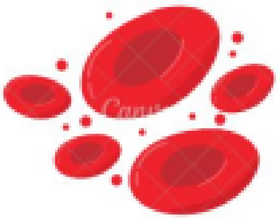


FIGURE 1

A NORMAL BLOOD CELL REPRESENTING HEALTHY HUMAN BEING.

FIGURE 2

A SICKLE CELL REPRESENTING ANEMIA DISORDER.

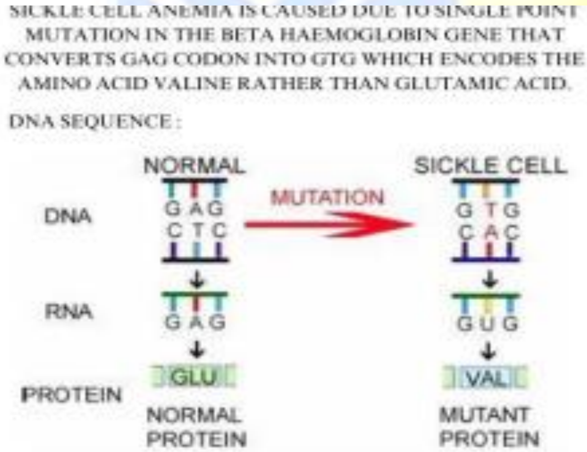

SICKLE CELL ANEMIA IS CAUSED DUE TO SINGLE POINT MUTATION IN THE BETA HAEMOGLOBIN GENE THAT CONVERTS GAG CODON INTO GTG WHICH ENCODES THE AMINO ACID VALINE RATHER THAN GLUTAMIC ACID.
DNA SEQUENCE :
NORMAL
SICKLE CELL
DNA
G A G
C T C
MUTATION
G T G
C A C
RNA
G A G
G U G
PROTEIN
GLU
VAL
NORMAL PROTEIN
MUTANT PROTEIN

FIGURE 3

The cellular makeup of red blood cells, which transport oxygen throughout the human body, is changed. Red blood cells tend to be circular and flexible, enabling them to more readily travel through arteries in the body. Red blood cells with sickle or curve shapes characterize sickle cell disease.

Furthermore, sickle cells create a thick and sticky covering that may hinder or delay blood flow.

LITERATURE REVIEW:

1) Machine Learning Based Computational Gene Selection Models: A Survey, Performance Evaluation, Open Issues, and Future Research Directions.

Nivedhitha Mahendran[1] , one of the authors, explains how gene expression occurs in the following ways: Transcription and translation are two different things.

P. M. Durai Raj Vincent [1], information is transferred from DNA to RNA with the help of enzymes, resulting in the creation of proteins and other biological substances.

2) Machine learning: A powerful tool for gene function prediction in plants.

Elizabeth H. Mahood,The goal of this study is to teach machine learning to experimental plant biologists to predict gene activity in order to identify biological discoveries.

Lars H. Kruse, In this study, we look at detecting Architectural features of sequenced genomes, anticipating linkages between distinct cell properties , and predicting gene mechanism and phenotypes.

Gaurav D. Moghe explains Finally, we propose machine learning-based strategies for increasing functional discovery in plants.

3) A Predictive Model of Gene Expression Using a Deep Learning Framework.

Rui Xie, **Andrew Quitadamo, Jianlin Cheng, and Xinghua Shi**, "Our work shows that deep learning has significant potentials for creating predictive models to better understand biological systems and presents a new application of deep learning in the mining of genomics data."

4) Machine Learning Methods for Cancer Classification Using Gene Expression Data:

Fadi Alharbi and Aleksandar Vakanski, the collection of data strategies for analysis of gene are covered in this work, as well as a list of noteworthy datasets widely used for supervised machine learning in this field. Because of the large number of genes in data sets, the high dimensionality of gene expression data is frequently addressed using proper feature engineering and data pretreatment procedures. The end of the report discusses future research prospects for machine learning-based gene expression analysis for cancer categorization.

STEP 1:

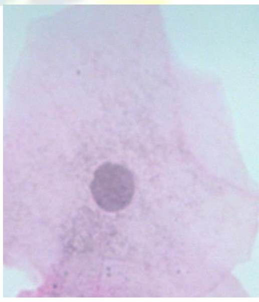
TO TRAIN THE MODELS

Here to train our model to detect the sickle cell disorder , we have selected to train the model by giving the training data set as following :

Based on their physical characteristics, there are three sorts of RBC to take into account. They might be normal , disfigured elongate (sickle cells), or flawed in different ways.

(a) Normal cell

(b) Sickle cell



(c) Other cells

A) IMAGE PRE PROCESSING :

Image pre-processing comes initially before feature extraction. This procedure is required to decrease noise and assure the extracted features' quality. The extraction of individual cells from microscopic pictures is our primary objective.We employed typical image processing methods to produce images for cell extraction. We started by binarizing the original picture. We picked this strategy because it calculates the best threshold value automatically. Second, we deleted tiny objects and decreased picture noise.All of these phases lead to the most crucial one, edge detection..This approach allows us to extract the outlines of every item in the image.We ran into a difficulty when certain items represented overlapping cells. In this situation, two or more cells share the outline of an object. We solved this problem by using an ellipse modification approach and an algorithm for finding important locations.

B) FEATURE EXTRACTION:

The initial process in every ML-based classification problem is feature extraction. Many characteristics have been proposed for detecting red blood cells in microscopic pictures. In our classification exercise, we sought to identify the 3 kinds of cells.Concerning the found cell types, we discovered the most commonly cited characteristics that may separate each kind.The 3 categories: form, color, and appearence. We collected 36 form characteristics, 15 colour features, and 60 appearence features from each cell for a total of 111 features to aid in the categorization of distinct cell types.

C) RBC CLASSIFICATION:

The methods utilised for RBC categorization are presented in this section. First, we discuss the issues we faced in the pre-processing of data and the approaches we utilised to deal with them.
Till now we have trained our model to predict the gene that represent sickle cell gene.But to select the gene from the high dimensional data , we employ the technique DNN.The process is as follows ; Transfer learning method is used to collect the data that has been trained from the classifier algorithm and feature screening method is carried out to nullify the extra layers of values , equations and algorithms has been prepared for DNN testing and further to verify our trained model CV technique for validating our trained model , here 3 fold CV technique is used .

STEP 2:
METHODS AND ALGORITHM:
We screened the gene and found accurate parts (such as genes), and then we used the notion of transfer learning to create a model that directly included data from feature screening.

$D_j = \{(B_{ix}, A_i): i \in I_j$ → (1)

Let Bx = (B1x,......., Bnx) be the genotype for the xth gene region for the n sample, and $B_{ix} = (B_{ix1}, \dots \dots . B_{ixn_x})$ be the genotype → (2) for the n sample. The variations in genes in the area is denoted by nx.
Let A = (A1,...........An) represent the phenotype for the n samples. Ai in a regression event can be a real integer or, in a classification event, an index from 1 to C. The data was then separated into 2 subsets, I1 and I2, with j = 1 and 2.

$D_j = \{(B_{ix}, A_i): i \in I_j$ [same as (1)]

We recommend fitting a basic DNN model for each area first. Then, to establish the predictive value of each region, we propose developing a group-wise feature importance score.Let region x trained on D1 serve as the foundation for the prediction model, fx(Bx). For categorical outcomes, the model fx(Bx) can be conceived of as a conditional probability (P(A|Bx)) or a conditional mean (E[A|Bx]). M(Ai,fx(Bix)) represents a chosen loss function. The GWFI score, denoted as Δx:

$$\Delta x = \sum_{i \in D_2} m_i = \sum_{i = D_2} \times (M(A_i, f_x(B_{ix})) - E[M(A_i, f_x(B'_{ix}))]) \quad \rightarrow (3)$$

B'_{ix} represents permuted data formed by randomly mixing together individual indexes.
As a result;

$$m_i = (M(A_i, f_x(B_{ix})) - E[M(A_i, f_x(B'_{ix}))]) \quad \rightarrow (4)$$

It seems to reason that if area x is not accurate, the difference in loss between observed and permuted data should be near to zero, and so x is expected to be zero.

Because observed data is expected to be less than permuted data, the alternative is that $\Delta x < 0$.
As a result, $\Delta x < 0$. assesses the accuracy of area x, with a less negative value represent more predictive power. To determine if area x is predictive, we recommend checking whether x is much less than 0, and we employ

$$H_0: \Delta_x \geq 0 \quad versus \quad H_1: \Delta_x < 0 \quad \rightarrow (5)$$

In the □$_0$, we assume that mi where
i belongs to D2 is drawn from the same distribution and that var(mi) < ∞ as |D2| → ∞ . Consequently,we

have

$\Delta_x \sim N(0, \sigma_x^2)$ → (6)

where D2 is the empirically calculated cardinality of D2.Despite having high generalisation properties, When there are too many parameters it may lead to overfitting of training data(i.e., D1). Therefore, it is suggested that the validation data (i.e., D2) be used to analyse the variations in loss.We employ the concept of 3-fold CV to determine by arbitrarily dividing the data into the two subsets (D1 and D2) at one time might lead to chance discovery.where ix is determined by applying Eq 1 to the information from the ith CV. Regarding each ix

$\Delta_x = \frac{1}{x} \Sigma_i^x \Delta_{ix}$ → (7)

Asymptotically, $\Delta_{ix} \sim N(0, \sigma_x^2)$ we have , $\Delta_x \sim N(0, \frac{i}{x^2} \Sigma_i^x \quad \sigma_{ix}^2)$ → (8) and hence, it is possible to assess area x's predictive significance.Even having more cross-validation folds might strengthen test statistics,they can also add to the workload. Our analyses revealed that when X is over 20, there isn't much of a difference. As a result, we default to X = 20. Our suggested feature selection approach differs from prior screening criteria in two key ways:By using the fitted model fk(.), we remove the need to train a new model , which makes it simple to calculate the loss difference. The key benefit of this approach is its effectiveness, which is crucial for complicated DNN models.The downstream prediction job is well-aligned with our recommended screening technique. The suggested GWFI score is designed to quantify the accurate capacity for each area.Unlike many previous approaches, which consider variable screening and predictive modelling as separate processes,The values of p are significant to prediction, and can therefore offer useful advice for feature selection. The proposed method catches features with difficult consequences (such as interactions), and also greatly simplifies the model.

Prediction modelling: -

We aim to create a graph in this study,When a supplementary node is also introduced to hold the miniscule impacts and DNNs produced from feature screening are coupled using the transfer learning concept.

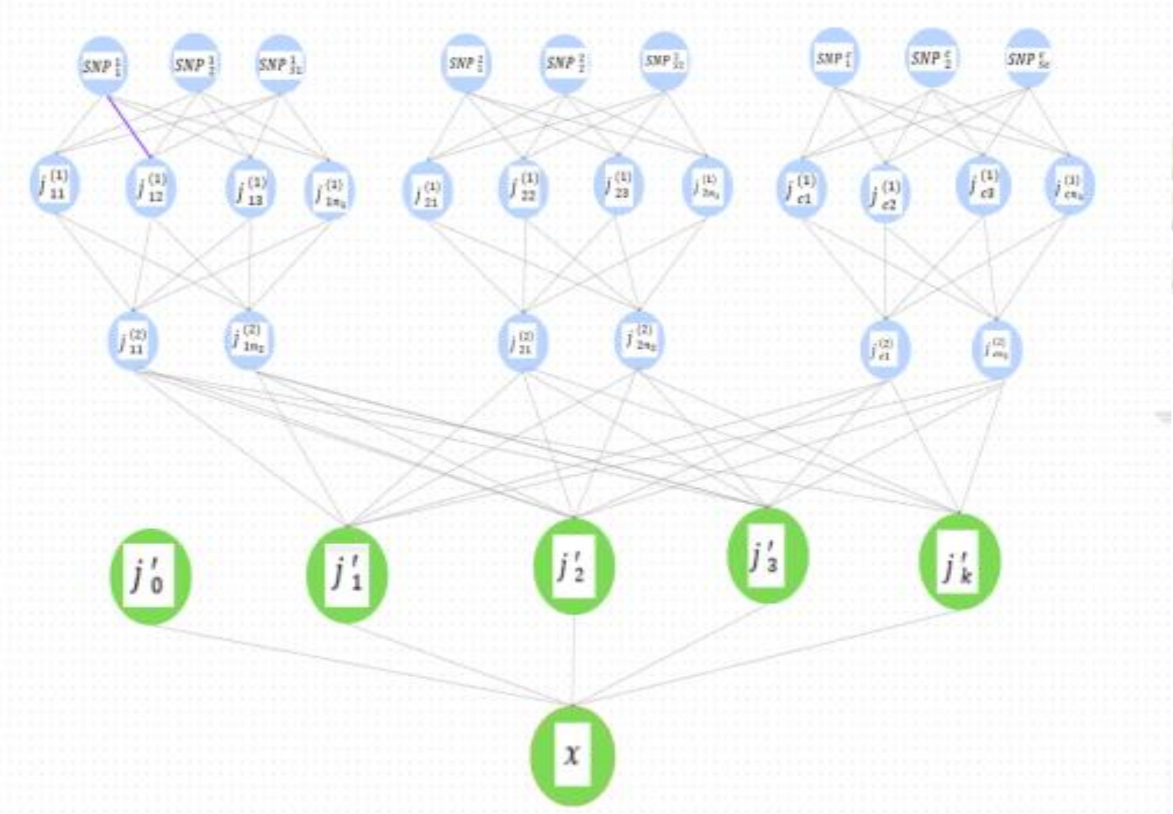


FIGURE 4 presents an example of our suggested concept in illustrated form.The background node (j'_0) is constructed by fitting a gBLUP model and is designed to capture the microscopic impacts . Because the target function in feature screening is the same as the aim function in the final prediction task, the networks trained during the feature screening phase are informative for the final prediction job. As a consequence, we use the concept of transfer learning to create the final prediction model. To simulate the cumulative effects of these selected genes, we suggest using the final hidden layers of pre-trained models acquired through feature screening as input and stacking the most recently added hidden layers on top.

Let $f_x^j(B_x; \alpha_x)$ represent the DNN model's final hidden layer trained on gene z, where z is a vector model parameters and c is the no. of genes chosen based on our proposed GWFI score mentioned in the preceding section.Our suggested prediction model seems as follows:

$$E(A|B)=l(f_1^j(B_1;\alpha_1), \dots\dots, f_c^j(B_c;\alpha_c);\beta_1) + l'(j_0';\beta_0) \quad \rightarrow (9)$$

where $j(.;\beta_0)$ is a function with parameter β_0 for the background and $j(.;\beta_1)$is a function with parameter β_1 for the extra hidden layers.

Similarly to transfer learning, we simply calculate parameters for the supplementary layer (β_0) and newly added hidden layers (β_1). Due to the necessity to estimate (β0, β1, α1, and c), the suggested transfer learning model's parameter count was significantly lower than that of a DNN with the same architecture. Similar to the feature screening procedure, parameter estimation uses a conventional loss function and optimisation approach .

BLUE CIRCLE: The DNN models and associated parameters were derived through feature screening.

GREEN CIRCLE:The background node (j_0'), which records extremely modest changes, and the newly added hidden layers, which represent the combined impact of certain genes. The input values of the supplementary node and the newly added hidden layers are computed.

X result is the SCA GENE ,to further verify , we use CV i.e 3-fold cross validation technique to verify the obtained result.

FLOWCHART AND TESTING:-

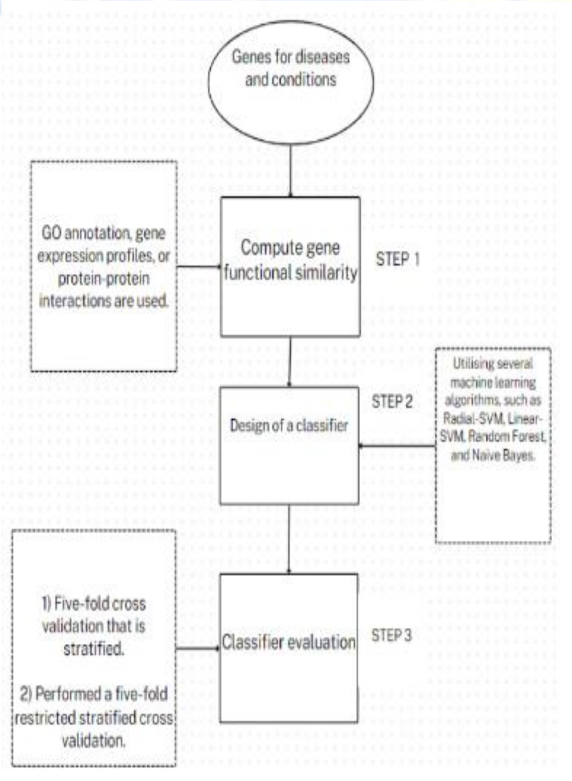

Genes for diseases and conditions
GO annotation, gene expression profiles, or protein-protein interactions are used.
Compute gene functional similarity
STEP 1
Utilising several machine learning algorithms, such as Radial-SVM, Linear-SVM, Random Forest, and Naive Bayes.
Design of a classifier
STEP 2
1) Five-fold cross validation that is stratified.
2) Performed a five-fold restricted stratified cross validation.
Classifier evaluation
STEP 3

FIGURE 5

Principal elements of the process for predicting disease genes.For a given collection of genes, functional similarities are calculated. Functional similarity matrices are used to establish the criteria that separate illness genes from non-disease genes using a variety of machine learning techniques.Utilised are two assessment strategies: held-out restricted stratified three-fold cross-validation and stratified.

A visual illustration of the SCA gene prediction process.

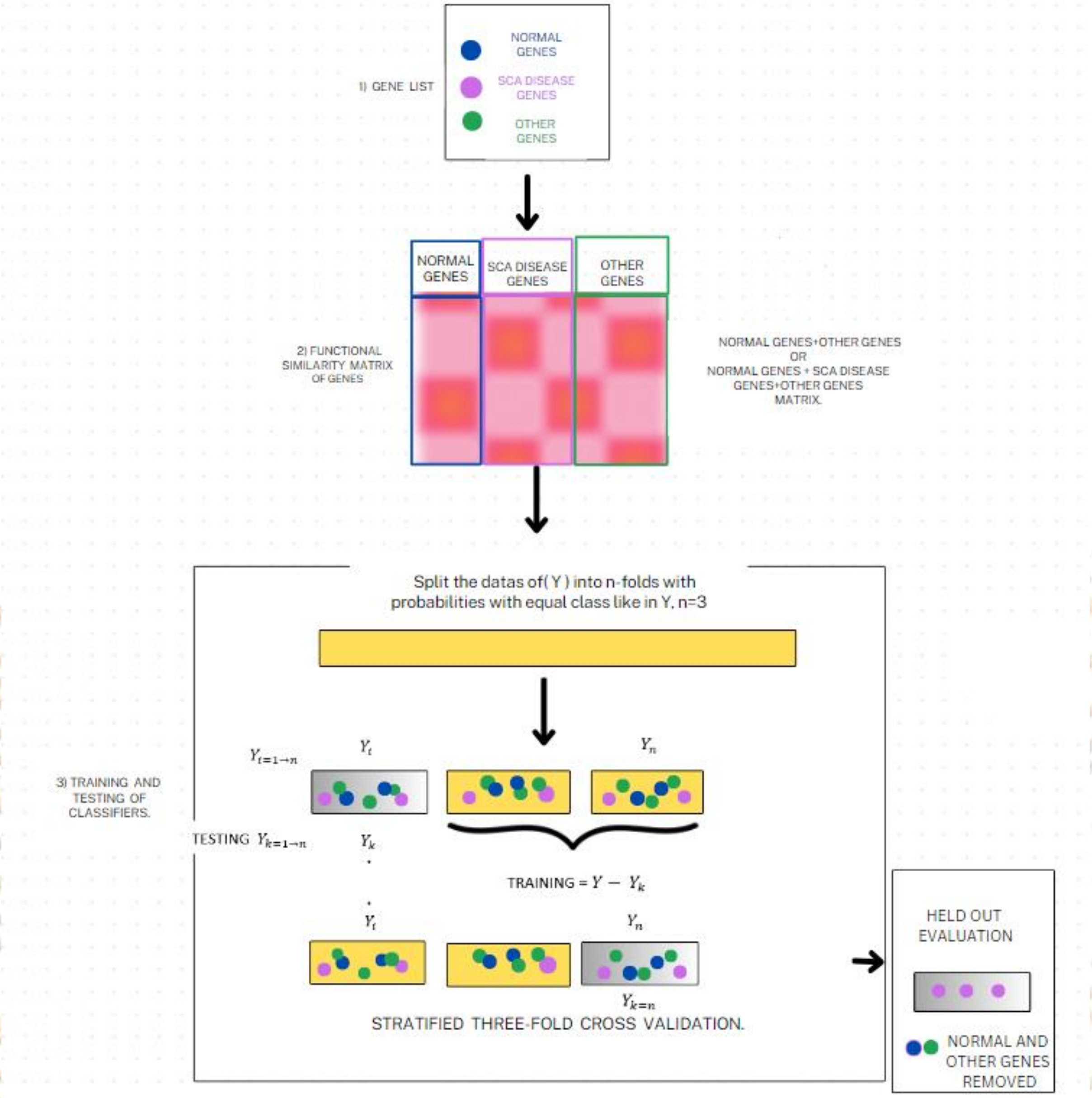

1) GENE LIST
NORMAL GENES
SCA DISEASE GENES
OTHER GENES
NORMAL GENES
SCA DISEASE GENES
OTHER GENES
2) FUNCTIONAL SIMILARITY MATRIX OF GENES
NORMAL GENES+OTHER GENES OR NORMAL GENES + SCA DISEASE GENES+OTHER GENES MATRIX
Split the datas of(Y) into n-folds with probabilities with equal class like in Y, n=3
$Y_{t=1\to n}$
Y_t
Y_n
3) TRAINING AND TESTING OF CLASSIFIERS.
TESTING $Y_{k=1\to n}$
Y_k
TRAINING = $Y - Y_k$
Y_t
Y_n
$Y_{k=n}$
STRATIFIED THREE-FOLD CROSS VALIDATION.
HELD OUT EVALUATION
NORMAL AND OTHER GENES REMOVED

FIGURE 6

A. To implement the recommended method, other genes and SCA genes with varied levels of evidence were used.
B. The functional similarities for Normal genes + other genes and Normal + SCA genes + other genes were determined using three different semantic similarity metrics.
C. Functional genes were analysed using four different machine learning algorithms. ML classifiers were tested using stratified and held-out restricted stratified 3-fold CV. To test the classifier, only SCA genes were chosen.
THUS , It is the SCA genes that are obtained as result and it is confirmed .Now our model is trained to predict the gene that represents the SICKLE GENE .

ALGORITHM:

Prediction model based on deep neural networks.
Input: the screening threshold γ and the genetic variations organised into C areas (such as genes).
Output: f(B) prediction model.
Feature Screening (DNN-screen) in Step 1:
1: Input: Training data on the result x and genetic variations categorised into C regions.
2: Output: Neural network models for each area and the GWFI scores S
= ($\square_1, \square_2, \ldots\ldots. \square_\square$) for each region.
3: for 1=i to C do
4:Construct a neural network model fi(Bi;αi) .
5: Based on fi(Bi;αi), calculate the GWFI score Si.
6: end for.
Step 2: DNN-transfer accurate model
1: Input: DNN models for each area, fs(B), the result x from training data, S, the group-wise feature significance score, and C regions' worth of genetic variations.
2: Output: The complete f(B) prediction model.
3: Assume that the input set and model, fs and Bs, are empty.
4: for i← 1 to C do
5: If Si< γ then
6:Bs = [Bs, Bi]
7: Concatenate the built-in model:
fs = [fs,fi(Bi;αi)]
8: end if
9: end for
10: Network design a) stack hidden layers js(.;β1) on top of fs(Bs;α); as

$$f^s(B_s;\alpha)$$

$$j^s(.;\beta_1)$$

f(B;(α,β)) as the final network by concatenating the background layer b(B;α0).

$$\text{b}(\text{B};\beta_0), \quad f^s(B_s;\alpha)$$

$$j^s(.;\beta_1)$$

11: Estimate parameters while maintaining fixed parameters.

REFERENCE :

https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0208626

https://www.nature.com/articles/s41598-020-74921-0#Sec7

https://www.sciencedirect.com/science/article/pii/S109836002300816X

FINDINGS AND CONCLUSION :

JOURNAL FOR ENGINEERING RESEARCH TIJER OPEN ACCESS JOURNAL

The final results are , we have established this process by two methods , we have trained our model by giving the three samples normal cell, sickle cell,and other cells.After training our model we used DNN technique and algorithms to train the model to detect the sickle gene from a huge pool of genetic data , since we are using huge amount of data here , DNN technique is the best option to train our model , after training our model the obtained results are in such a way that our model can easily distinguish normal and sickle cells also it is trained to identify sickle gene from a huge set of genetic data . To perfectly test our model we used 3-fold cross validation technique to verify the result . Hence we were able to successfully develop a model that can differentiate the cells and also interpret the defective gene . Here to compare with other neural networks,such as ANN,CNN,RNN

ANN:Each layer of an artificial neural network (ANN) is made up of many perceptrons or neurons. Because inputs are exclusively processed in the forward direction, ANN is also known as a Feed-Forward Neural Network.
This sort of neural network is one of the most basic neural network versions. They convey data across several input nodes until it reaches the output node in a single path. The network may or may not include hidden node layers, which helps to explain how it works.

CNN: Convolutional neural networks (CNN) are one of the most often utilised models nowadays. This neural network computational model employs a variant of multilayer perceptrons and includes one or more convolutional layers that can be either totally linked or pooled. These convolutional layers provide feature maps that record a portion of an image, which is then divided into rectangles and sent out for nonlinear processing.

RNN:RNNs (recurrent neural networks) are more complicated. They preserve the output of processing nodes and feed the result back into the model (the information was not sent in only one direction). The model is said to learn to anticipate the outcome of a layer in this manner. Each node in the RNN model serves as a memory cell, calculating and executing operations. If the network's prediction is incorrect, the system self-corrects and continues to backpropagate towards the correct prediction.The reason why we used the **DNN** model is that deep learning models learn patterns which the tree model cannot solve, so deep learning actually well predicts different diseases.