DNA Classification using Machine Learning for Detecting Genetic Disorders

Amisha Mishra, Shruti Duggal, Snehanshu Banerjee, R. B. Sarooraj

Abstract-Deoxyribonucleic acid is a double- helical molecule composed of two chains that contains genetic instructions. Genetic diseases are caused by changes in pre-existing genes. A genetic abnormality results from the alteration in chromosomes. DNA classification helps to identify genetic disorders in organisms. DNA pattern recognition is a major issue in bioinformatics. DNA is classified into several categories on the basis of Structure, Location, Number of base pairs etc. Traditionally the DNA Molecule is studied by extracting it from the blood sample and is then manually analysed to find out the abnormalities. To increase the accuracy, a machine learning based DNA classification is done which helps in studying the extracted DNA image using various techniques. This consumes minimal amount of time and is more efficient. The image is pre-processed using median filter and canny edge detection. DNA sequences can be recognized correctly and effectively without any uncertainties with the help of Neural Network.The network successfully classifies an image given as input when it is trained with patterns. Thus, we can analyse if a person has a genetic disorder.

Keywords-DNA, Canny edge detection, Neural Network, Genes.

I. INTRODUCTION

A gene is a sequence of nucleotides in DNA which determines a person's appearance and other characteristics. The different DNA sequences made up by genes are called genotypes. Some of the genetic qualities are promptly visible, such as eye colour, while others are not, such as blood type and the risk for certain diseases. To sequence the DNA into different types we classify it.

DNA is classified on the basis of seven categories-

(1) Number of base pairs per turn
(2) Location
(3) Structure
(4) Coiling pattern
(5) Number of strands
(6) Coding and Non-coding DNA
(7) Nucleotide sequence

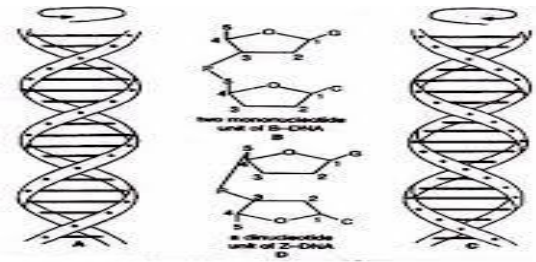

Fig.1 Structure of DNA molecule

Revised Manuscript Received on April 1, 2020.

Amisha Mishra, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.
Shruti Duggal, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.
Snehanshu Banerjee, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.
Mr. R B Sarooraj, Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.

Traditionally, DNA classification was done manually by extracting the DNA from the blood sample using many steps. Then the principle component analysis is done to classify the DNA.

The major Disadvantages are:

- Inaccurate results
- Time consuming
- Not valid pattern detection for multiple images in a short time

The proposed method uses image processing and machine learning techniques. It involves basic four steps-

(1) Pre-processing-The main aim is to supress undesired distortions. It involves using median filter and Canny edge detection.

(2) Dual tree complex wavelet transforms-Wavelets are used to remove noise from two dimensional signals, for example images. The decomposition of the image is done and then it is reconstructed from the modified levels.

(3) GLCM Co-occurrence Matrix-It analyses how frequently a pixel with grey-level value i occurs in any direction to adjacent pixels with the value j.

(4) Neural Network for image classification- The Neural Network contains parameters sometimes also referred to as neurons, receives multiple inputs and produces an output. It performs feature extraction and then classifies the DNA accordingly.

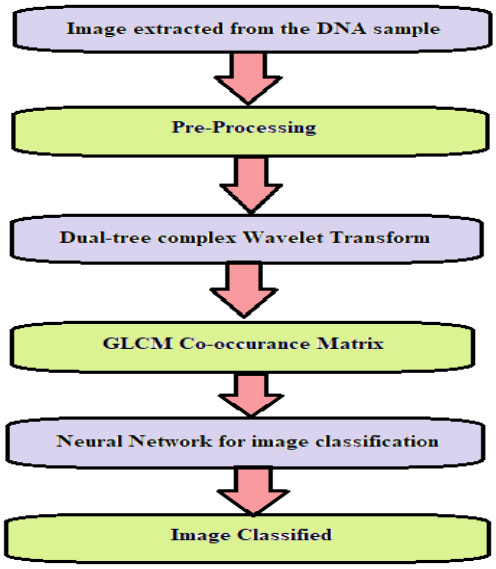

Fig.2 Flow Chart

II. RELATED WORKS

In one of the papers, the DNA pattern recognition is done using sequence compressor.

The major disadvantage of this is that due to compression, image quality is reduced. Another method is the robust classification of DNA damage patterns in a single cell using SVM classifier which has a big drawback of low throughput. Other authors have also proposed one double-stranded DNA probes as classifier. This also has a disadvantage of its weakness to noise.

III. METHODOLOGY

In the proposed system an image is taken of the extracted DNA sample. First the image is pre-processed to remove any noise or outliers using techniques such as median filter and canny edge detection. Then the Dual-tree complex wavelet transform is applied. Wavelets are used to remove noise from two dimensional signals, for example images. The decomposition of the image is done and then it is reconstructed from the modified levels. Finally, neural network classification is done. DNA sequences can be recognized correctly and effectively without any uncertainties with the help of Neural Network. The network successfully classifies an image given as input when it is trained with patterns.

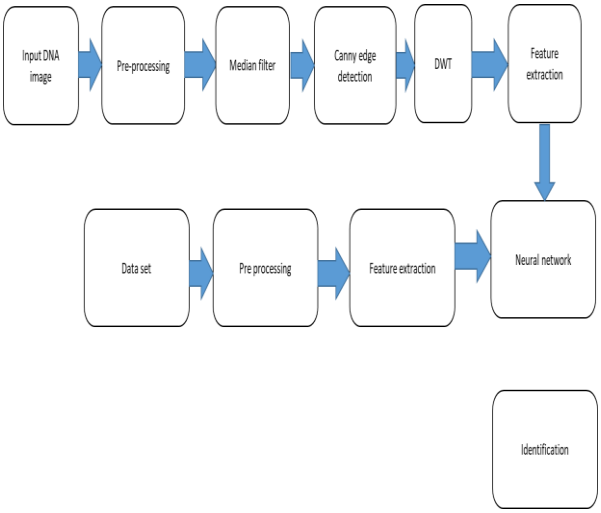

Fig.3 Block Diagram

A. Pre-Processing

Image Pre-processing refers to the operation performed on any image at the lowest level of abstraction. Intensity images act as both input and output in this case. Its main aim is to improve the image by removing unwanted distortions and enhancing features important for further processing. Corruption can be of many forms such as camera misfocus, noise and motion blur. Image restoration is also done but it is different from image enhancement because it emphasizes on that features of the image which make it more pleasing to the viewer, but do not produce realistic data. Pre-processing is done using two techniques Median Filter and Canny Edge Detection.

B. Median Filter

It removes noise from the image. It is a part of pre-processing. A Median Filter selects the median intensity in the window that it operates upon. The pattern of neighbours is known as the "window". It slides, entry by entry, over the entire signal. After that, each entry is replaced by the median of its neighbouring entries. For 1D signals, the window is defined by the first few former and following entries, whereas for 2D data the window must include all entries within a given radius.

$x = (3, 6, 75, 9)$.

The output signal y is:

y_1 = median (3, 6, 75) = 6,

y_2 = median (6, 75, 9) = median (6, 9, 75) = 9,

y_3 = median (75, 9, 3) = median(3, 9, 75) = 9,

y_4 = median (9, 3, 6) = median(3, 6, 9) = 6,

I.e. $y = (6, 9, 9, 6)$

C. Canny Edge Detection

Canny edge detection is done to detect the wide range of edges in an image. Edges isolate the main image from the noise. Thus, it becomes easy to work on the image. The steps to do so are:-

1. Use Gaussian filter to denoise the image.
2. Find out the image's intensity gradients.
3. Perform non-maximum suppression in order to remove forged response of edge detection.
4. Apply double threshold for determining the potential edges.
5. Suppress all other weak edges and those connected to strong edges.

The edge detection is done in the following way:

Any edge direction falling within the range of (0 to 22.5 &157.5 to 180 degrees) is set to 0 degrees. Edges falling in the range (22.5 to 67.5 degrees) is set to 45 degrees. Other edges between (67.5 to 112.5 degrees) is set to 90 degrees. Finally, any edges falling within (112.5 to 157.5 degrees) is set to 135 degrees.

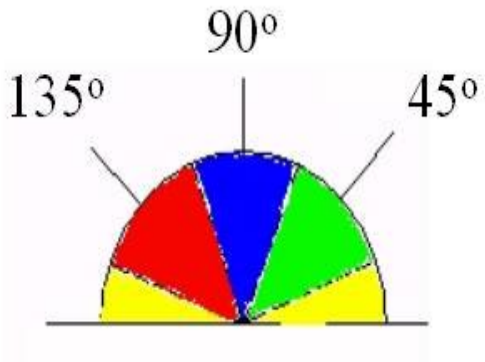

Fig.3 Edge Detection

Thus, the edges following in the area are set to one value for example 0. This reduces the number of edges making the working on the image easier.

D. Dual-Tree Complex Wavelet Transform

The dual-tree complex wavelet transform (CWT) is better than the discrete wavelet transform (DWT), because it has significant added properties. It is directionally selective in two and higher dimensions and also shift invariant. Wavelets are used to remove noise from two dimensional signals like images. The decomposition of the image is done and then the reconstruction of the image from the modified levels is the final step.

E. GLCM Co-Occurrence Matrix

The GLCM calculates how often a pixel with grayscale intensity value i occurs to adjacent pixels of the value j horizontally, vertically, or diagonally.

GLCM direction of Analysis:

1.Horizontal (0°)
2.Vertical (90°)
3.Diagonal: Bottom left to top right (-45°) or Top left to bottom right (-135°)

It has the following properties:

- Contrast: Measures local disparities and texture of shadow depth in the GLCM.
- Correlation: Measures the joint probability existence of the defined pixel pairs.
- Homogeneity: Measures the proximity of elements in the GLCM to the GLCM diagonal.

F. Neural Network Classifier

The neural network is a set of functions, also called parameters. It works by training the computer from analyzed data. Each parameter, which is also referred to as neurons, is a function receiving one or multiple inputs and producing an output. Those outputs are then passed on to the next layer of neurons. This process continues until the last layer of neurons has received its input. The terminal neurons produce the output which is the final result for the model. So, the network successfully classifies an image given as input.

All NN networks have four layers:

1. The input layer contains only a single neuron for each predictor variable. Typically, N-1 neurons are used where N stands for the number of categories.
2. The hidden layer has one neuron for every case in the training data set. That neuron holds the values of the predictor variables along with its target value.
3. The pattern layer is different for both NN networks and GRNN. NN networks have one pattern neuron for each category of the target variable.
4. The decision layer is the last layer. In NN networks, it compares the weighted votes for each target category and uses the largest vote to predict the target category.

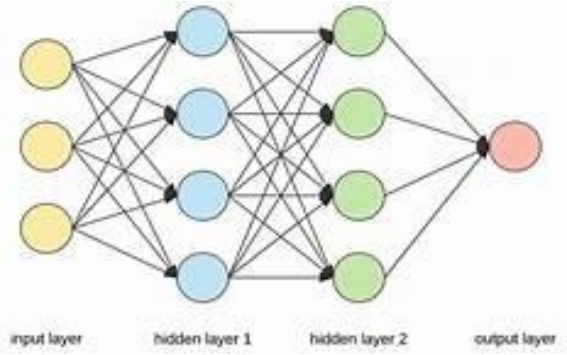

Fig.4 Neural Network Architecture

First the image is taken from the extracted DNA sample. Then the image is pre-processed to remove the noise and outliers. It is done by the median filter and canny edge detection. After this the dual-tree complex wavelet transforms GLCM are applied. These help to enhance the image for feature extraction. The data set contains training examples which help the neural network to classify the type of DNA.

IV. RESULT

The proposed system was tested on DNA images taken from various samples and the classification of the DNA pattern was done successfully. First the input DNA image is fed into the system. After that, the noise and unwanted parts are removed from the image. Then, many techniques and functions are applied to the image. Finally, the neural network is used to classify the type of DNA.

V. CONCLUSION

The DNA pattern classification is a major problem in the medical field. DNA of a human being contains genes which are responsible for inherent features of a person such as appearance and also susceptibility to certain diseases. This project solves the problem of DNA classification and reveals important information about the person like his/her likelihood to catch a certain infection or develop a genetic disease. Thus, it is an efficient, accurate and time-saving method of DNA classification, thereby acting as an invaluable aid to medical science.

ACKNOWLEDGEMENT

This project consumed a lot of hard work and dedication and it would not have been possible to complete it successfully without the proper guidance and suggestions of many individuals. So, we would like to extend our gratitude to all of them.

Firstly, we would like to thank our computer science department for providing us with such an opportunity. In addition, special thanks to our guide Mr. Sarooraj who guided us throughout and provided us with all the support we needed.

Lastly, the recommendations of our classmates helped us throughout and we are thankful to all of them.

REFERENCES

1. Mandel P, Metais P. (1948). Les acides nucleiques du plasma sanguin chez l'homme [in French]. C R Seances Soc Biol Fil 142:241-243.
2. otezatu I, Serdyuk O, Potapova G, Shelepov V, Alechina R, Molyaka Y, Anan'ev V, Bazin I, Garin A, Narimanov M, Melkonyan H, Umansky S, Lichtenstein AV. (2000). Genetic analysis of DNA excreted in urine: a new approach for detecting specific genomic DNA sequences from cells dying in an organism. Clin Chem 46:1078-1084.
3. Sriram KB, Relan V, Clarke BE, Duhig EE, Windsor MN, Matar KS, et al. (2012). Pleural fluid cell-free DNA integrity index to identify cytologically negative malignant pleural effusions including mesotheliomas. BMC Cancer 12:428.
4. Liimatainen SP, Jylhv J, Raitanen J, Peltola JT, Hurme MA. (2013). The concentration of cell-free DNA in focal epilepsy. Epilepsy Res 105(3):292-8
5. Stroun M, Lyautey J, Lederrey C, Olson-Sand A, Anker P. (2001). About the possible origin and mechanism of circulating DNA: Apoptosis and active DNA release. Clin Chim Acta 313(1-2):139- 42.
6. Jahr S, Hentze H, Englisch S, Hardt D, Fackelmayer FO, Hesch RD, et al (2001). DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells. Cancer Res 61(4):1659-65
7. Chiu RWK, Chan KCA, Gao Y, Lau VYM, Zheng W, et al. (2008). Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. Proc Natl Acad Sci U S A 105: 20458-20463.

AUTHORS PROFILE

Amisha Mishra was born in Aurangabad, Maharashtra, India in 1998. She is expected to complete her Bachelor of Technology (B-Tech) in the branch of Computer Science & Engineering from SRM Institute of Science & Technology in 2020. Her research interests include IoT, Big data and Cloud Computing.

Shruti Duggal, was born in Ghaziabad, Delhi, India in 1998. She is expected to complete her Bachelor of Technology (B-Tech) in the branch of Computer Science & Engineering from SRM Institute of Science & Technology in 2020. Her research interest include Cloud computing, IoT and AI.

Snehanshu Banerjee was born in Bilaspur, Chhattisgarh, India in 1999. He is expected to complete his Bachelor of Technology (B-Tech) in the branch of Computer Science and Engineering from SRM Institute of Science and Technology in 2020. His research interest includes Machine Learning, and Android App development.

Mr. R B Sarooraj born in Tamil Nadu, completed his Bachelor of Engineering in the branch of Information Technology from Anna University, Chennai 2009. He completed his Master of Engineering n Computer and Communication Engineering from Anna University, Chennai in 2012. He's working as Assistant Professor in the Department of Computer Science and Engineering at SRM Institute of Science and Technology, Chennai from 2013 till date. His research interests include Data Mining and Analysis of Algorithms.