

HEREDITARY DISEASE PREDICTION IN EUKARYOTE DNA

*A Dissertation submitted in partial fulfillment of the
requirements for the award of the Degree of*

BACHELOR OF
ENGINEERING IN
ELECTRONICS AND COMMUNICATION
ENGINEERING BY

SAI MOHITH CHINNARI (1604-18-735-086)

DARAPANENI ANUTHAM SIVA KRISHNA (1604-18-735-083)

MOHAMMAD RIAZ (1604-18-735-079)



Department of Electronics and Communication Engineering

Muffakham Jah College of Engineering and Technology

Banjara Hills,

Hyderabad-500034 (Affiliated
to Osmania University) 2022



MUFFAKHAM JAH
COLLEGE OF ENGINEERING &
TECHNOLOGY

(Est. by Sultan-Ul-Uloom Education Society in 1980)

(Affiliated to Osmania University, Hyderabad)

Approved by the AICTE & Accredited by NBA

CERTIFICATE

This is to certify that the dissertation titled in '**Hereditary Disease Prediction in Eukaryote DNA**' submitted by **Sai Mohith Chinnari (1604-18-735-086)**, **Darapaneni Anutham Siva Krishna (1604-18-735-083)** and **Mohammad Riaz (1604-18-735-079)** partial fulfillment of the requirements for the award of the Degree of **Bachelor of Engineering in Electronics and Communication** is a bonafide record of work carried out by them under my guidance and supervision during the year 2021-2022. The results embodied in this report have not been submitted to any University or Institute for the award of Degree or Diploma.

Mrs. G. Prashanthi

Project Supervisor

Assistant Professor, ECED

MJCET, Hyderabad

Dr. Arifuddin Sohel

Professor and HOD

Department of ECE

MJCET, Hyderabad

EXTERNAL EXAMINER

INTERNAL EXAMINER

8-2-249, Mount Pleasant, Road No.3, Banjara Hills, Hyderabad – 500 034

Phone: 040-23350523, 23352084, Fax: 040-2335 3428,

Website: www.mjcollege.ac.in, e-mail: principal@mjcollege.ac.in

Declaration

We here by declare that the work presented in the dissertation entitled “**Hereditary Disease Prediction in Eukaryote DNA**”, submitted in the partial fulfillment of the requirement for the award of the Degree of Bachelor of Engineering, in the department of Electronics and Communication Engineering, Muffakham Jah College of Engineering and Technology, Hyderabad is an authentic record of our own work carried out under the guidance and supervision of **Mrs. G. Prashanthi**, Assistant Professor, ECED, MJCET.

We have not submitted the matter embodied in this report for the award of any other degree or diploma. This report has not been submitted to another institute or university for the award of any degree or diploma and neither this project is being used by another person or people at any other place.

Date:

Place: Hyderabad

Sai Mohith Chinnari

Mohammad Riaz

Darapaneni Anutham Siva Krishna

ACKNOWLEDGMENT

We would like to acknowledge our indebtedness and render our warmest thanks to our project guide, **Mrs. G. Prashanthi** Assistant Professor, Department of Electronics and Communication Engineering MJCET, who made this work possible. Her immense knowledge, profound experience and professional expertise has enabled us to complete this project successfully. We are grateful for her constant support and encouragement even during the tough times of pandemic. The doubts and the queries we had were attended with a great deal of patience. Her friendly guidance and expert advice has been invaluable throughout all stages of the work.

We would also wish to express our gratitude to **Dr. Kaleem Fatima**, Professor, ECED, MJCET and **Mr. Jaideep Kumar Nag**, Associate Professor, ECED, MJCET for taking regular appraisals and extended discussions. The valuable suggestions have greatly contributed to the improvement of the project.

We would like to thank **Dr. Arifuddin Sohel**, Head of the Department of Electronics and Communication Engineering, MJCET for his generous help, guidance, and encouragement in the successful completion of our project. We also thank the other faculty members and the entire non-teaching staff of the Department of Electronics and Communication Engineering for their valuable suggestions in the project reviews and support throughout our engineering.

Also, we are immensely thankful to our Parents for their over whelming and wholehearted support and understanding without which this project would not have been possible.

Mohammad Riaz
Sai Mohith Chinnari
Darapaneni Anutham Siva Krishna

ABSTRACT

The existing clinical detection techniques of disease genes are rather insufficient in terms of accuracy concern, cost, time consumption, and sometimes harmful also. Hence researchers from different backgrounds are looking for an alternative way. The literature of the DNA nucleotide sequence suggests valuable information in support of the irregularity present in a diseased gene. A mutational disease is caused due to alteration in the normal sequencing of nucleotides in DNA. If those alterations can be recognized, the disease-associated genes can be identified perfectly.

Hereditary disease prediction in eukaryotic DNA using signal processing approaches is an incredible work in bioinformatics. Researchers of various fields are trying to put forth a non-invasive approach to forecast the disease-related genes. As diseased genes are more random than the healthy ones, in this work, a comparison of the diseased gene is made against the healthy ones.

An adaptive signal processing method like functional link artificial neural network-based Levenberg–Marquardt filter has been proposed in this regard. Here, disease genes are discriminated from healthy ones based on the magnitude of mean square error (MSE), which is calculated through the adaptive filter. The performance of the algorithm is inspected by computing some evaluation parameters.

An improved accuracy level compared to the existing methods is the prime concern. Taking the reference gene as healthy, the overall process is accomplished by categorizing the diseased and healthy targets with MSE value at a threshold of 0.012. The proposed technique predicts the test gene sets successfully.

TABLE OF CONTENT

ABSTRACT	v
List of Figures	ix
	1
1. Introduction	1
1.1 Introduction	1
1.1.1 DNA and it's Structure	1
1.1.2 Signal Processing	3
1.1.2.1 Adaptive Signal Processing	3
1.1.3 Synopsis of Previous Work	4
1.2 Aim of Thesis	5
1.3 Motivation of the Thesis	5
1.4 Technical Approach	6
1.5 Organization of Thesis	7
2. Literature Survey	8
2.1 Introduction	8
2.2 Survey of Different Papers	8
2.3 Conclusion	18
3. Methodology	19
3.1 Flow Chart of the Proposed Model	19
3.2 Introduction to MatLab	20
3.3 Introduction to Jupyter	21
3.3.1 Introduction	21
3.3.2 Python	21
3.3.2.1 Pandas	21

3.3.2.2	Decision Tree	21
3.3.2.3	Matplotlib	22
3.3.2.4	Train Test Split	22
4.	EIIP Mapping	23
4.1	Introduction	23
4.2	Methodology	24
4.2.1	Dataset	24
4.2.2	Mapping	24
4.3	Results	25
4.4	Conclusion	26
5.	Adaptive Filter	27
5.1	Introduction	27
5.2	Methodology	28
5.2.1	Functional Expansion	28
5.2.2	Adaptive Filter	29
5.3	Results	31
5.4	Conclusion	32
6.	Mean Square Error	33
6.1	Introduction	33
6.2	Methodology	33
6.2.1	Mean Square Error	33
6.3	Results	35
6.3.1	Diseased Gene (MatLab)	35
6.3.1.1	Output 1	35
6.3.1.2	Output 2	36
6.3.1.3	Output 3	37

6.3.2	Healthy Gene(MatLab)	38
6.3.2.1	Output 1	38
6.3.2.2	Output 2	39
6.3.1.2	Output 3	40
6.3.3	Machine Learning outputs	41
6.3.3.1	Diseased Gene	41
6.3.3.2	Healthy Gene	42
6.3.4	Normalised Mean Square Error Graph	43
6.4	Conclusion	43
7.	Project Codes	44
7.1	Introduction	44
7.2	MATLAB Codes	44
7.2.1	Reading DNA sequences	44
7.2.2	EIIP Mapping	45
7.2.3	Calculation of Mean Square Error	46
7.3	Machine Learning Codes	47
7.3.1	Importing the Libraries and Reading the CSV file	47
7.3.2	Splitting the dataset into input and output	48
7.3.3	Training and Testing the dataset	48
7.3.4	Accuracy	48
8.	Conclusion and Future Scope	49
8.1	Conclusion	49
8.2	Future Scope	49
	References	50

LIST OF FIGURES

Figure No.	TITLE	Page No
Figure 1.1	DNA Structure	1
Figure 1.2	Parts of a Nucleotide	2
Figure 1.3	Representation of Mutated DNA	3
Figure 1.4	Graph for Proposed Method and Existing Method	4
Figure 3.1	Flow Chart for Proposed Method	19
Figure 4.1	EIIP Values for Nucleotides	24
Figure 4.2	Basic Block Diagram	25
Figure 4.3	EIIP Mapping Simulation	25
Figure 5.1	Basic Block Diagram of Adaptive Filter	27
Figure 5.2	Basic Block Diagram of Functional Expansion	28
Figure 5.3	Block Diagram of Adaptive Filter	30
Figure 5.4	Plotting of MSE vs Iterations	31
Figure 6.1	MSE Flow Chart	32

Figure 6.2	Matlab Output- 1 of Diseased Gene	34
Figure 6.3	Plotting of MSE vs Iterations for Output -1 of Diseased Gene	34
Figure 6.4	Matlab Output -2 of Diseased Gene	35
Figure 6.5	Plotting of MSE vs Iterations for Output -2 of Diseased Gene	35
Figure 6.6	Matlab Output -3 of Diseased Gene	36
Figure 6.7	Plotting of MSE vs Iterations for Output -3 of Diseased Gene	36
Figure 6.8	Matlab Output -1 of Healthy Gene	37
Figure 6.9	Plotting of MSE vs Iterations for Output -1 of healthy Gene	37
Figure 6.10	Matlab Output -2 of Healthy Gene	38
Figure 6.11	Plotting of MSE vs Iterations for Output -2 of Healthy Gene	38
Figure 6.12	Matlab Output -3 of Healthy Gene	39
Figure 6.13	Plotting of MSE vs Iterations for Output -3 of Healthy Gene	39
Figure 6.14	Machine Learning Output -1 of Diseased Gene	40
Figure 6.15	Machine Learning Output -2 of Diseased Gene	40
Figure 6.16	Machine Learning Output -1 of Healthy Gene	41
Figure 6.17	Machine Learning Output -2 of Healthy Gene	41
Figure 6.18	Plotting of NMSE vs No of Genes	42

CHAPTER- 1

INTRODUCTION

1.1 Introduction

1.1.1 DNA and it's Structure

Each strand of DNA is a polynucleotide composed of units called nucleotides. A nucleotide has three components: a sugar molecule, a phosphate group, and a nitrogenous base. In genomes, DNA molecules are generally very long, thin polymers with a diameter of 2 nm and a length that can extend to 108–109 nm.[11]

The sugar in DNA's nucleotides is called deoxyribose—DNA is an abbreviation for deoxyribonucleic acid. The bases adenine (A), cytosine (C), guanine (G), and thymine (T) are the four main components that make up DNA. These bases can form pairs and link together to create the double helix structure of DNA. DNA is a macromolecule consisting of two strands that twist around a common axis in a shape called a double helix. Cytosine forms three hydrogen bonds with guanine, and adenine forms two hydrogen bonds with thymine.[11]

Guanine and cytosine make up a nitrogenous base pair because their available hydrogen bond donors and hydrogen bond acceptors pair with each other in space.

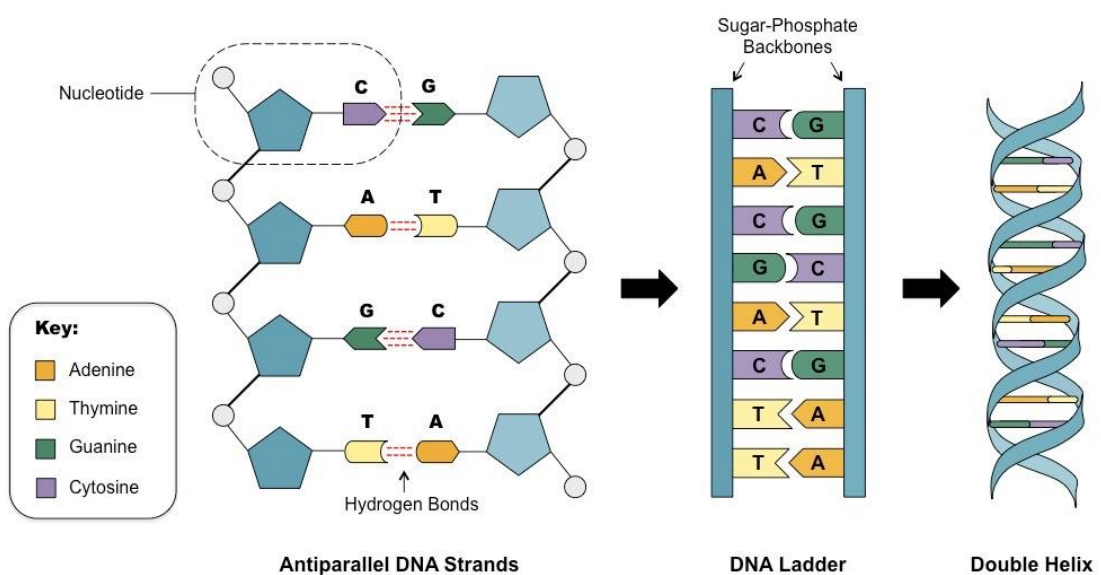


Figure 1.1 – DNA Structure

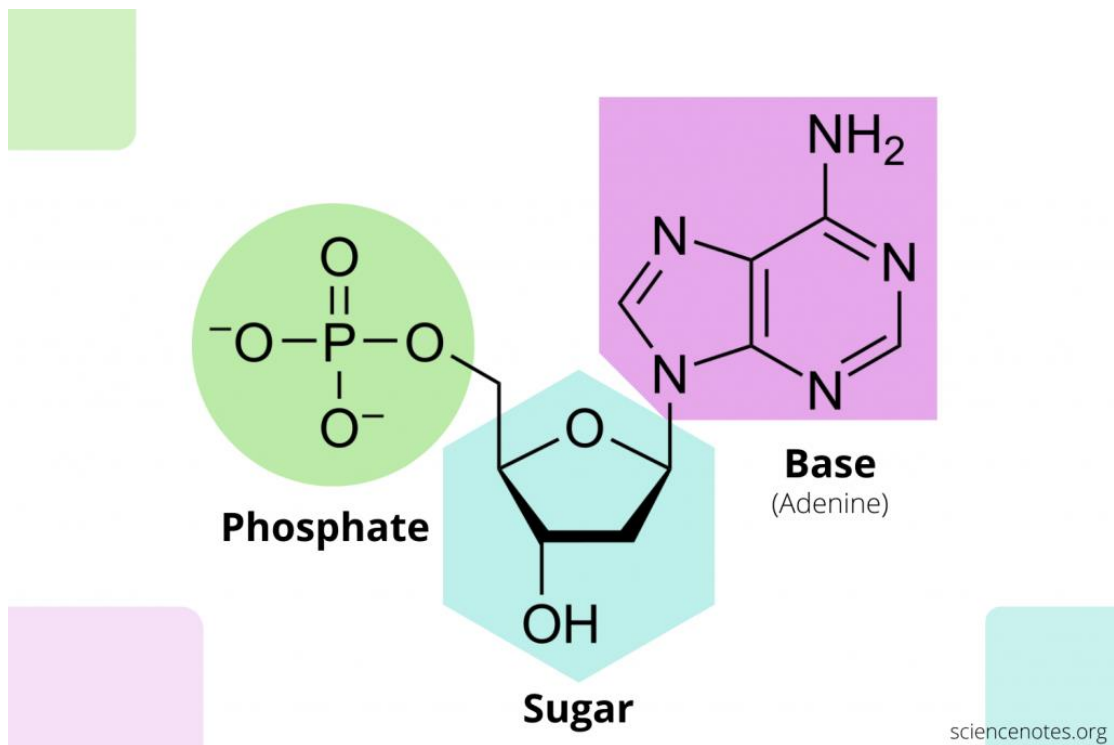


Figure 1.2 – Parts of a Nucleotide

The mutation may be of different types like addition, deletion, duplication, substitution, and many more. As a result, the normal order of the nucleotide sequence gets disturbed and hence, the codons corresponding to amino acid is misrepresented so that the protein coding region (exon) which is responsible for producing amino acids behaves inharmoniously.

The mutation may be of different types like addition, deletion, duplication, substitution, and many more. As a result, the normal order of the nucleotide sequence gets disturbed and hence, the codons corresponding to amino acid is misrepresented so that the protein coding region (exon) which is responsible for producing amino acids behaves inharmoniously.

This results in different diseases. One example of such a type of mutation is represented in [Figure 1](#). Here, the mutation is due to the repeated addition of trinucleotide (CAG) in the original DNA code for an amino acid sequence. This exceptional type of mutation leads to the popular HD.

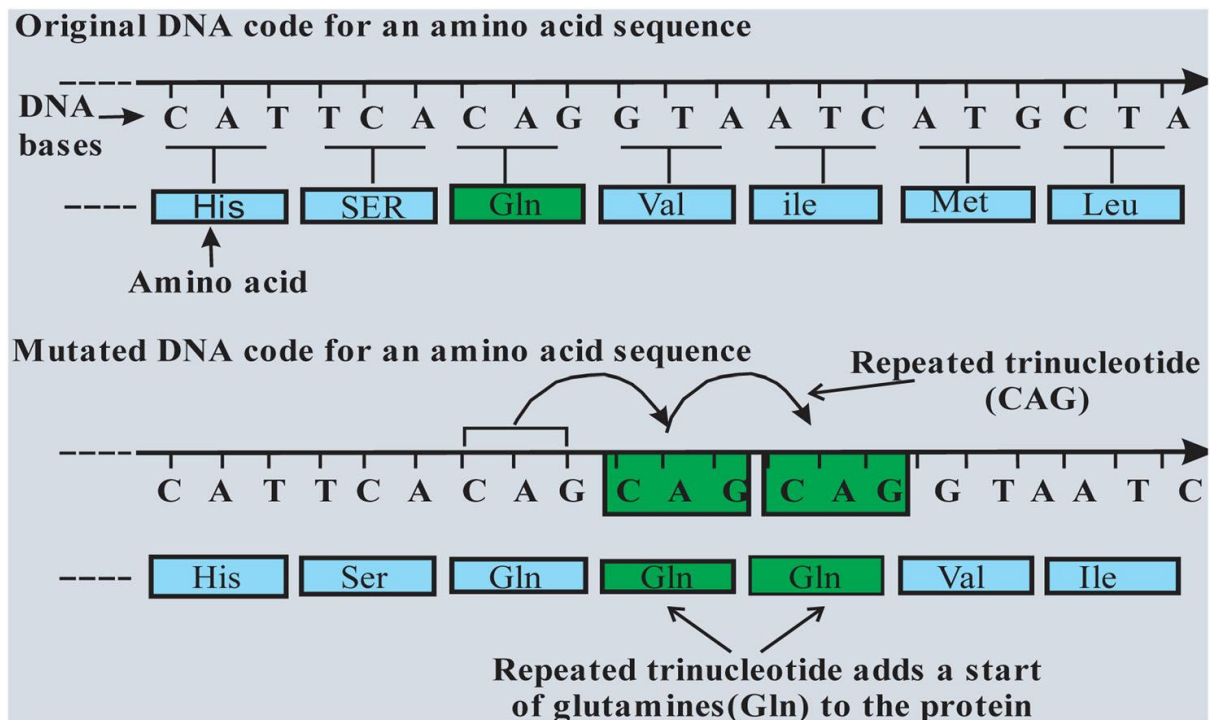


Figure 1.3 – Representation of Mutated DNA

1.1.2 Signal Processing

Signal processing condenses measurements to extract information about some distant state of nature. Signal processing can be described from different perspectives.

To an acoustician, it is a tool to turn measured signals into useful information. Signal processing is the analysis, interpretation and manipulation of signals. Signals of interest include sound, images, biological signals such as ECG, radar signals, and many others. Processing of such signals includes storage and reconstruction, separation of information from noise (e.g., aircraft identification by radar), compression (e.g., image compression), and feature extraction (e.g., converting text to speech).

1.1.2.1 Adaptive Signal Processing

Adaptive signal processing algorithms generally attempt to optimize a performance measure that is a function of the unknown parameters to be identified. The most pervasive of these performance measures are based upon squared prediction errors, although the specific prediction error used in adaptation often depends upon the

particular algorithm. Two broad categories of adaptive signal processing methods are: stochastic and exact. The latter category refers to adaptive filters based upon the actual or exact data signals acquired.

The former category of adaptive techniques known collectively as stochastic methods are based upon derivations using the statistical properties of the data signals. The primary statistical measure used is the ensemble average, or mean, of a squared prediction error function, and this has evolved into wide spread use of the mean squared prediction error as a performance measure. Often this is shortened to simply the mean square error (MSE).

1.1.3 Synopsis of Previous Work

The existing FLANN-based LMS adaptive filter is used to distinguish between healthy and diseased (cancer) genes. This gene detection principle is applied to HD (healthy and disease) dataset in this work. AUC for the proposed technique is higher than that of the existing one, proves its supremacy in identifying disease genes. It is observed that the proposed PSO-tuned Levenberg–Marquardt adaptive filter algorithm gives superior performance. The proposed method could distinguish the diseased genes more correctly resulting in privileged evaluation parameters.[2]

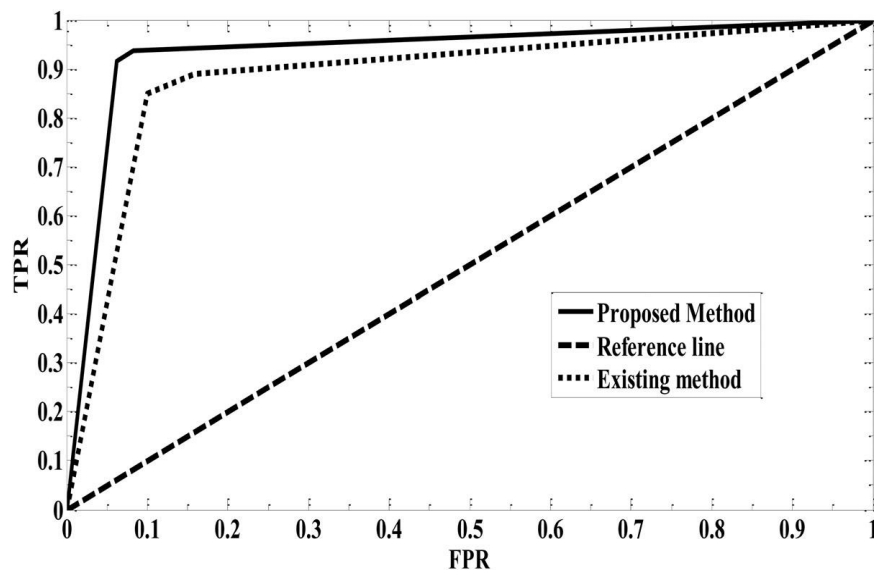


Figure 1.4 – Graph for Proposed Method and Existing Method

Besides, the curve is plotted for receiver operating characteristics (ROC) using TPR and FPR as shown. The number of truly predicted genes is indicated by TPR and the number of false predicted genes is indicated by FPR are ROC curve for the existing and proposed method. NUCLEOSIDES, NUCLEOTIDES AND NUCLEIC ACIDS 1195 evaluated for both proposed and existing algorithms. The area under the ROC curve is more near about 0.5 (reference line) for the existing method than that of the proposed one

1.2 Aim of the Thesis

Aim of the thesis is to introduce a novel approach to detect diseased gene using FLANN based Adaptive filters. A DNA string of characters is not ready to be analysed using DSP tools until it is first converted into numerical values by using mapping techniques such as EIIP representation.

In DNA signal processing, the exact identification and classification of the diseased gene is a great challenge to the researchers. This thesis aims to solve this issue by using adaptive signal processing techniques

1.3 Motivation

The existing clinical detection techniques of disease genes are rather insufficient in terms of accuracy concern, cost, time consumption, and sometimes harmful also. Hence researchers from different backgrounds are looking for an alternative way. The literature of the DNA nucleotide sequence suggests valuable information in support of the irregularity present in a diseased gene. A mutational disease is caused due to alteration in the normal sequencing of nucleotides in DNA. If those alterations can be recognized, the disease-associated genes can be identified perfectly.

Functional link artificial neural network (FLANN) is a single-layered structure with simple operations to conquer the MLP. To boost up the input pattern, nonlinear functions are applied to the FLANN structure. By integrating the FLANN structure with

adaptive filters, improved performance is obtained. FLANN filter offers trigonometric expansion of the input sequences which reduces the error to minimum value at faster rate. Here, diseased genes are distinguished from healthy genes based on amount of normalized mean square error, which is estimated through adaptive filter.

Adaptive filter has its extensive application in identifying an unknown system by comparing its distinctiveness with a known system. In this work, all the diseased as well as healthy genes are regarded as unknown signals and are evaluated against one of the healthy genes which is taken to be the reference signal. The diseased (mutated) genes are detected by the proposed method as they are dissimilar comparing with a healthy one due to the mutation causing the disturbance in a normal run.

1.4 Technical Approach

The project was implemented in four stages. The first stage was to collect the diseased gene and normal gene data from NCBI website, this data works as first input to read the DNA sequence and convert the sequence using mapping techniques.

The second stage was to learn the signal processing techniques in Mat-Lab and understand the technical terms involved in biological sciences.

The Third stage was to implement FLANN based adaptive filter (Mat-lab) to calculate Mean square error (MSE) values using different genes collected from NCBI website

The final stage was to use the collected dataset to train and test in Machine learning (PYTHON) , to calculate performance metrics like accuracy and to plot graphs

1.5 Organization of Thesis

The entire work on our project is split into following chapter

Chapter 1- This chapter deals with the entire basic introduction required for the project along with the technical approach required implementing the project and the motivation behind doing such a project.

Chapter 2- This chapter forms the backbone of our entire project. It contains information regarding all the previous works which were done on similar projects along with their drawbacks and shortcomings. This chapter helps us in deciding the most suitable method available and also helps us in not repeating the same mistakes which were earlier made in the previous papers and projects.

Chapter 3- This includes all the basics and the design methodologies used by us in order to complete the project. This chapter covers all the basics related to Adaptive Signal Processing and working of it.

Chapter 4- This chapter gives a detailed description of EIIP Mapping and the need for conversion of alphabetical DNA sequences into numerical DNA sequences

Chapter 5- This chapter deals with the explanation of Adaptive filters and how FLANN based adaptive filters have been implemented in this model

Chapter 6- This chapter gives a detailed description of Mean Square Error and shows all the simulation results

Chapter 7- This chapter constitutes all the Mat Lab codes and Machine Learning codes that have been implemented for the successful completion of our project

Chapter 8- This chapter gives an overall conclusion for our project and also throws light on the future advancements that can be made on the same project which will be more effective and efficient.

CHAPTER - 2

LITERATURE SURVEY

2.1 Introduction

This chapter gives a detailed literature survey of the related work done in the field of Genomic signal processing for identification of cancer cells using various methodologies. In this chapter, we present a brief description of the most efficient techniques which were used earlier.

2.2 Survey of Different Papers

2.2.1 DSP based entropy estimation for identification and classification of Homo sapiens cancer genes. (Springer-Verlag Berlin Heidelberg 2016) Joyshri Das, Soma Barman

GSP basically processes genes, proteins and DNA sequences using various signal processing methodologies to extract the information hidden in it. Genes are composed of amino acids; therefore, gene identification based on amino acids level is more responsive according to medical research reports. Genes are very random in nature, difficult to extract hidden information resides in it. To reveal the hidden information content of genes, entropy is estimated.

In the present paper, the crucial job of gene identification and classification is attempted. Statistical methods like entropy estimation and mutual information calculation are adopted along with DSP technique. Rayleigh distribution of estimated entropy of gene is treated as identifier of healthy and cancerous Homo sapiens. Once the cancer genes are identified, mutual information estimator based on their minimum entropy is used as classifier to detect different types of cancer genes.

The work is successfully examined on prostate, breast and colon genes collected from NCBI homepage. The work may further be extended in other types of cancer genes.

2.2.2 A non-invasive cancer gene detection technique using FLANN based adaptive filter (Saikat Singha Roy, Soma Barman)

In recent years advancement in cross field technologies lead the world to the new era of genomic research. Several new technologies have been developed for early detection of critical genetic disease like cancer. Since the conventional morphological and clinical tests are invasive in nature and harmful to human body, researchers are intending to find a non-invasive way to predict cancer associated genes. They find accuracy as a major concern to be taken care of in disease gene identification system. Therefore the authors in this paper made an effort to raise the accuracy level compare to existing technique.

Cancer is caused by changes in deoxyribonucleic acid sequence. If those changes can be identified, diseased gene can therefore be recognized exactly. Hence the authors compared the diseased gene with the healthy ones to capture the differences. In order to search such disparity more efficiently and accurately, functional link artificial neural network (FLANN) based adaptive filter with least mean squares algorithm is attempted in the present paper. FLANN filter offers trigonometric expansion of the input sequences which reduces the error to minimum value at faster rate. Here, cancer genes are distinguished from healthy genes based on amount of normalized mean square error, which is estimated through adaptive filter. 131 Genes are used to train the identifier and the proposed technique successfully identifies 369 test dataset. The database is collected from National Center for Biotechnology Information Genbank.

The performance of the overall system is investigated by measuring sensitivity, specificity and accuracy. The proposed algorithm achieves 86.85% accuracy compared to existing entropy based identifier. Overall system performance of the proposed FLANN based adaptive filter is displayed in receiver operating characteristic curve.

2.2.3 Ross, C. A.; Tabrizi, S. J. Huntington's Disease: From Molecular Pathogenesis to Clinical Treatment. Lancet Neurol.

Huntington's disease is a progressive, fatal, neurodegenerative disorder caused by an expanded CAG repeat in the huntingtin gene, which encodes an abnormally long polyglutamine repeat in the huntingtin protein.

Huntington's disease has served as a model for the study of other more common neurodegenerative disorders, such as Alzheimer's disease and Parkinson's disease. These disorders all share features including: delayed onset; selective neuronal vulnerability, despite widespread expression of disease-related proteins during the whole lifetime; abnormal protein processing and aggregation; and cellular toxic effects involving both cell autonomous and cell-cell interaction mechanisms. Pathogenic pathways of Huntington's disease are beginning to be unravelled, offering targets for treatments.

Additionally, predictive genetic testing and findings of neuroimaging studies show that, as in some other neurodegenerative disorders, neurodegeneration in affected individuals begins many years before onset of diagnosable signs and symptoms of Huntington's disease, and it is accompanied by subtle cognitive, motor, and psychiatric changes (so-called prodromal disease).

Thus, Huntington's disease is also emerging as a model for strategies to develop therapeutic interventions, not only to slow progression of manifest disease but also to delay, or ideally prevent, its onset.

2.2.4 Akhtar, M. Epps, J. Ambikairajah, E. Signal Processing in Sequence Analysis Advances in Eukaryotic Gene Prediction. IEEE J. Sel. Top. Signal Process.

Genomic sequence processing has been an active area of research for the past two decades and has increasingly attracted the attention of digital signal processing researchers in recent years. A challenging open problem in deoxyribonucleic acid (DNA) sequence analysis is maximizing the prediction accuracy of eukaryotic gene locations and thereby protein coding regions.

In this paper, DNA symbolic-to-numeric representations are presented and compared with existing techniques in terms of relative accuracy for the gene and exon prediction problem. Novel signal processing-based gene and exon prediction methods are then evaluated together with existing approaches at a nucleotide level using the Burset/Guigo1996, HMR195, and GENSCAN standard genomic datasets.

A new technique for the recognition of acceptor splice sites is then proposed, which combines signal processing-based gene and exon prediction methods with an existing data-driven statistical method. By comparison with the acceptor splice site detection method used in the gene-finding program GENSCAN, the proposed DSP-statistical hybrid technique reveals a consistent reduction in false positives at different levels of sensitivity, averaging a 43% reduction when evaluated on the GENSCAN test set

2.2.5 Prediction of Cancer cells using DSP techniques. (International conference on Communication and Signal Processing, April 3-5, 2013, India) G.N. Satapathi, Dr. P. Srihari, Senior Member, IEEE, Ch. Aruna Jyothi, S. Lavanya

DSP plays an important role in DNA sequence analysis, cancer diagnosis and gene expression analysis. Digital Signal Processing (DSP) applications have predominantly gained popularity in the study of genomics in recent time. DSP is now a widely applied tool in the segment of DNA sequence analysis, gene expression detection, identification

of coding and non-coding regions and finding out abnormalities present in the coding region.

In this paper researchers applied DFT power spectrum plot to predict protein coding regions of a DNA sequence. The study surveyed DFT power spectrum as a method to predict cancer disease for various databases available in Gene bank.

The result established that this method can be used as an easy tool to predict cancer disease. The filtered power spectrum plots yield high accuracy. The proposed algorithm is tested for several databases of Homosapien chromosomes available in the Gene Bank that yielded in satisfactory results.

Further efforts can be made to improve the accuracy of prediction by using other types of digital filters. The next step in this research can be to generalize this analysis for several other oncogene databases.

2.2.6 DWT based Cancer Identification using EIIP. (2016 Second International Conference on Computational Intelligence & Communication Technology) Shilpi Chakraborty, Vinit Gupta.

With the recent advances in Genomic signal processing (GSP) domain, researchers have been applying digital signal processing (DSP) techniques in raw genomic data for extracting the hidden features and periodicities within the fragments of DNA.

This paper has incorporated Electron ion interaction potential (EIIP) method for mapping of DNA sequence into digital signal and Discrete wavelet transform (DWT) power spectrum methods in the algorithm to predict the abnormalities present in the protein coding region.

Compared with traditional Fourier-based spectral analysis techniques for signal processing, wavelet-based techniques found to be more appealing due to their attractive properties, such as local feature identification, time frequency domain representation, multi-resolution scalability, denoising and compressing of big sample data. Spectral analysis technique has

been applied to raw genomic data for detection of location of exon region and from power spectrum plots of protein coding region, identification of cancer is done.

In future it has great scope in early diagnosis and prognosis of cancer as it depends on the mutation in DNA sequence of gene. It can help in design drugs, in agro-informatics, in synthesize biofuel and in biometrics. From genomic information one can predict how an individual will respond to drugs and based on that development of new drugs will be done. And, without performing biological experiments, one can predict the diseases with the help of signal processing tools which further leads to cost effective experiment and consumes less amount of time.

2.2.7 WHY GENETIC CODE CONTEXT OF NUCLEOTIDES FOR DNA SIGNAL PROCESSING? A REVIEW" (Muneer Ahmad, Low Tan Jung, Al-Amin Bhuiyan)

Protein coding regions are commonly diffused with non-coding regions due to $1/f$ background noise in such a way that a viable discernment between the two regions becomes cumbersome. Commonly employed digital signal processing methodologies lack fundamental genetic code context of nucleotides since these approaches treat DNA signal as normal digital signal that could be processed by traditional DSP tools and techniques.

This paper reviews the prevailing approaches for protein coding regions identification that base on common DSP concepts and highlights the importance of genetic code context to be considered for any computational solution for protein coding regions identification. Nucleotides in a DNA signal carry certain natural characteristics i.e. presence in a distinctive triplet format, maintaining distinct structure, owning and further sharing distribution of densities in codons, fuzzy behaviors, semantic similarities, unbalanced nucleotides' distribution producing a relatively high bias for nucleotides' usage in coding regions etc.

The computational solutions for protein coding regions identification that exploit these fundamental characteristic of nucleotides can significantly suppress the signal noise and hence can better contribute in identification.

A strong correlation has been observed between an enhancement of coding regions identification and the strong $1/f$ background noise. Further, protein coding regions own a fundamental 3-base periodicity which is found absent in non-coding regions, this characteristic of coding regions has been exploited by a large number of researchers to propose various methodologies and tools to resolve issues related with coding regions identification

2.2.8 Functional Link Artificial Neural Network-based Disease Gene Prediction (Jiabao Sun, Jagdish c. Patra and Yong jin Li

Genes that contribute to complex traits pose special challenges that make candidate disease-associated gene discovery more difficult. In this work, Investigation of topological features derived from PPI network to identify the causing genes of four complex diseases: Cancer, Type I Diabetes, Type 2 Diabetes, and Ageing genes has been done.

To-fold cross-validation was used to evaluate the predictive capacity of all possible combinations of these features and found the features with the best predictive ability. The performance of Multi-layer Perceptron (MLP), Functional Link Artificial Neural Network (FLANN), and Support Vector Machines (SVM) has been assessed. It has been found that SVM provides higher accuracy than MLP and FLANN. However, the FLANN has significantly low computation time while its accuracy is comparable to that of SVM and MLP.

Three classifiers have been employed to perform classifications. SVM worked very well. It obtained higher accuracy than MLP and FLANN in the prediction of all the four diseases. With comparable accuracy to SVM and MLP, FLANN is computationally efficient. It worked much faster, because of its simple network structure.

Random selection of genes from the control set as non-disease genes was done. This may not be very appropriate, because there may be some unknown disease genes among the randomly selected genes. With increasing quantity and quality of human protein interaction network, and with more understanding about the disease, the performance of the algorithm should improve.

2.2.9 A Classification Technique for Microarray Gene Expression Data using PSO-FLANN (Jayashree Dev1, Sanjit Kumar Dash2, Sweta Dash 3, Madhusmita Swain)

Despite of an increased global effort to end breast cancer, it continues to be most common cancer deaths in women. This problem reminds that new therapeutic approaches are desperately needed to improve patient survival rate. This requires proper diagnosis of disease and classification of tumor type based on genomic information according to which proper treatment can be provided to the patient.

There exists a no. of classification techniques to classify the tumor types. In this paper we have focused on three different classification techniques: BPN, FLANN and PSO-FLANN and found that the integrated approach of Functional Link Artificial Neural Network (FLANN) and Particle Swarm Optimization (PSO) can better predict the disease as compared to other method.

The performance of PSO-FLANN is better for classification as compared to BPN and FLANN. The performance has been checked on breast cancer data set which is obtained from the UCI machine learning repository website. This method overcomes the nonlinearity of the classification problem.

The FLANN with PSO architecture, because of its simple architecture and computational efficiency may be conveniently

employed in other tasks of data mining and knowledge discovery in databases such as clustering, feature selection, feature extraction, association rule mining, regression, and so on.

2.2.10 PREDICTION OF CANCER CELL USING DIGITAL SIGNAL PROCESSING (S.BARMAN (MANDAL), M.ROY, S.BISWAS, S.SAHA)

Digital Signal Processing (DSP) applications have gained great popularity in the study of genomics in recent time. DSP can be used as a tool in the area of DNA sequence analysis, gene expression detection, identification of coding and non coding regions and also finding out abnormalities present in the coding region. DSP solves this task with great accuracy and less complexity.

According to available medical research reports it has been given to understand that cancer is often caused due to genetic abnormality. In the present article, we present a DFT based approach to analyze the spectral characteristics of cancer cell and non-cancer cell and design a digital IIR low pass filter with Butterworth approximation for better prediction and identification of anomalies in cancer cells. The algorithm is tested for several databases of Homo Sapiens chromosomes available in Gene Bank which gives satisfactory results

DSP nowadays plays an important role in DNA sequence analysis, CANCER diagnosis and gene expression analysis etc. Researchers are using DFT power spectrum plot to predict protein coding regions of a DNA sequence.

Surveyed the DFT power spectrum as a method to predict CANCER disease for various databases available in Gene bank. The result shows this method can be used as an easy tool to predict cancer disease. The filtered power spectrum plots yield high accuracy. The bar plots show prominent results for prediction of cancer

2.3 Conclusion

Each of the above-mentioned papers had their own set of advantages and disadvantages. The advantages of each paper helped us in devising our own methodology on similar terms and each disadvantage provided us with a scope of exploring different algorithms and techniques to tackle their shortcomings and to overcome them. This basic background of the previous related works propelled us to tackle all the possible shortcomings and roadblocks leading to the successful completion of the project.

CHAPTER- 3

METHODOLOGY

3.1 Flow Chart of the proposed Model

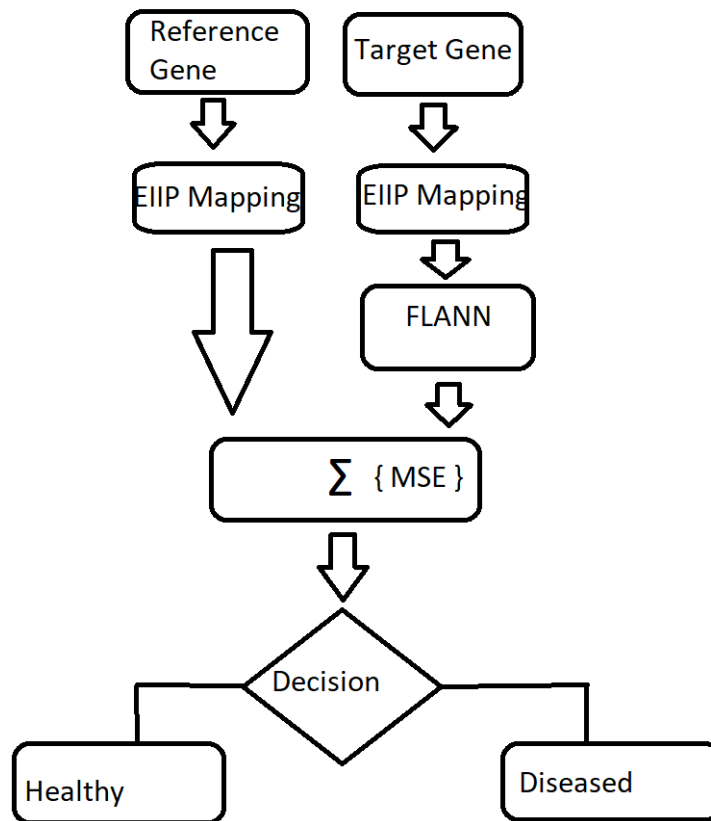


Figure 3.1 – Flow Chart for Proposed Method

Here 2 genes are considered one as Reference gene and other as Target gene

EIIP Mapping is performed for these genes, which is the conversion of Alphabetical DNA sequence into Numerical Sequence

The target Gene EIIP mapping output is fed into FLANN based adaptive filter which expands the sequence by using non-linear trigonometric functions

The output of Reference EIIP mapping and Flann output is compared and Error Sequence is generated.

Based upon the error sequence obtained, a decision is made as to whether the gene is a diseased one or a healthy one.

3.2 Introduction to MATLAB

The name MATLAB stands for MATRIX LABORATORY. MATLAB was written originally to provide easy access to matrix software developed by the LINPACK (linear system package) and EISPACK (Eigen system package) projects.

Millions of engineers and scientists worldwide use MATLAB to analyze and design the systems and products transforming our world. MATLAB is in automobile active safety systems, interplanetary spacecraft, and health monitoring devices, smart power grids, and LTE cellular networks. It is used for machine learning, signal processing, image processing, computer vision, communications, computational finance, control design, robotics, and much more.

The MATLAB platform is optimized for solving engineering and scientific problems. The matrix-based MATLAB language is the world's most natural way to express computational mathematics. Built-in graphics make it easy to visualize and gain insights from data. A vast library of prebuilt toolboxes lets you get started right away with algorithms essential to your domain. The desktop environment invites experimentation, exploration, and discovery. These MATLAB tools and capabilities are all rigorously tested and designed to work together.

MATLAB helps you take your ideas beyond the desktop. You can run your analyses on larger data sets and scale up to clusters and clouds. MATLAB code can be integrated with other languages, enabling you to deploy algorithms and applications within web, enterprise, and production systems. For the implementation of our project, we have used the MATLAB R2019a version which supported all our requirements.

There are various tools in MATLAB that can be utilized for image processing, such as Simulink, GUI etc. Simulink contains various toolboxes and image processing toolbox is one such example. Simulink is used for simulation of various projects. GUI is another important tool in MATLAB. It can be designed either by manual programming which is tedious task or by using guide.

MATLAB is used for performing EIIP Mapping and Evaluating mean square error and gives us the average mean square error for 1000 iterations

3.3 Introduction to JUPYTER

3.3.1 Introduction

The Jupyter Notebook App is a server-client application that allows editing and running notebook documents via a web browser. The Jupyter Notebook App can be executed on a local desktop requiring no internet access (as described in this document) or can be installed on a remote server and accessed through the internet.

In addition to displaying/editing/running notebook documents, the Jupyter Notebook App has a “Dashboard” (Notebook Dashboard), a “control panel” showing local files and allowing to open notebook documents or shutting down their kernels.

3.3.2 Python

Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly procedural), object oriented and functional programming. It is often described as a "batteries included" language due to its comprehensive standard library.

Python is used to identify whether the gene is healthy or diseased based upon the Mean square error by `train_test_split` method

3.3.2.1 Pandas

Pandas is an open-source library that is made mainly for working with relational or labeled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. This library is built on top of the NumPy library. Pandas is fast and it has high performance & productivity for users.[14]

3.3.2.2 Decision Tree

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

3.3.2.3 Matplotlib

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack[14]

One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

3.3.2.4 Train Test Split

The `train_test_split` function is for splitting a single dataset for two different purposes: training and testing. The testing subset is for building your model. The testing subset is for using the model on unknown data to evaluate the performance of the model.

Train size - This parameter sets the size of the training dataset. There are three options: None, which is the default, Int, which requires the exact number of samples, and float, which ranges from 0.1 to 1.0.

Test size - This parameter specifies the size of the testing dataset. The default state suits the training size. It will be set to 0.25 if the training size is set to default.

Using the same dataset for both training and testing leaves room for miscalculations, thus increases the chances of inaccurate predictions.

CHAPTER- 4

EIIP MAPPING

4.1 INTRODUCTION

The pivotal problem of gene identification in eukaryotes is distinguishing exons, from introns and intergenic regions. A number of coding measures like single and polynucleotide bias differences, spectral differences etc which exist between these regions have been utilized for this purpose in various gene finding algorithms. But simultaneous improvement of sensitivity and selectivity of these algorithms is still a challenge and so the hunt for new coding measures is to be continued.

The existing method of locating exons by genomic signal processing technique employing four binary indicator sequences, one for each nucleotide, depends on the period three peaks observed in the power spectrum of the exon regions and which do not exist in non-coding regions. [9]

One of the major constraints of genomic research is the availability of gene data in suitable format to process. But in recent years, due to progress of microarray technology, an extensive amount of information rich genomic data is available in public domain like NCBI, NHGRI etc. Therefore these data now become easily accessible and come in the form of discrete sequences. The gene sequence consists of four nucleotides A, C, T and G and hence called nucleotide sequence according to FASTA representation[2]

This form of alphabetic representation is called nucleotide sequence. The discrete nature of genomic signals ensures the application of digital signal processing to process them accurately after numerical conversion. Proper choice of mapping technique makes easier to identify the abnormality in the target gene.[6]

4.2 METHODOLOGY

4.2.1 DATASET

The datasets required for the proposed method are collected from the gene bank of the NCBI website. Since our discussion about HD, healthy and disease (HD affected) nucleotide sequences were collected for analysis. Out of which some are trained using the proposed algorithm and the rest are for testing. The collected datasets are in the form of four characters and before applying a signal processing method it must be converted into a numeric sequence.[12]

4.2.2 MAPPING

The EIIP is defined as the average energy of delocalized electrons of the nucleotide. Table shows the corresponding numerical values of four nucleotides. The EIIP values of each nucleotide are used to convert the alphabetical sequence into corresponding numerical sequence. For Example, consider a short length DNA sequence

$$X(n) = \{CTGAACCTGGG...\} \text{ ---(1)}$$

Nucleotide	EIIP Value
Adenine [A]	0.1260
Guanine [G]	0.0806
Cytosine [C]	0.1340
Thymine [T]	0.1335

Figure 4.1 – EIIP Values for Nucleotides

The collected gene set is a DNA sequence comprised of four characters (A, C, G, and T) arranged in a particular manner. The resulting numerical sequence corresponding to (1) is in the form of (2).

$$x[k] = [0.1340 \ 0.1340 \ 0.1260 \ 0.1340 \ 0.1335 \ 0.1335 \ 0.0806 \ 0.0806 \ 0.1260 \ 0.1335 \ 0.1260 \ 0.1340]. \text{ ---(2)}$$

Thus, all the collected nucleotide sequences are mapped into the corresponding numerical sequences.[2]

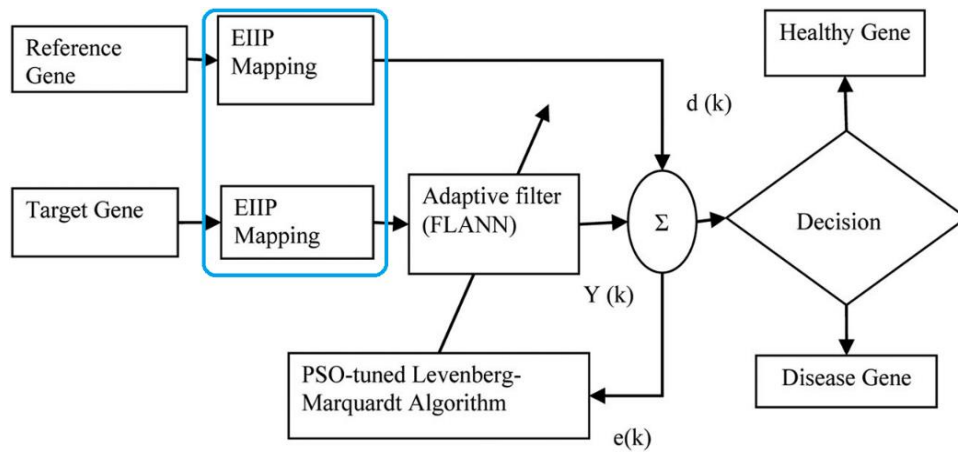


Figure 4.2 – Basic Block Diagram

4.3 RESULTS

Simulation of EIIP Mapping is performed in MATLAB and corresponding Numerical sequence is obtained

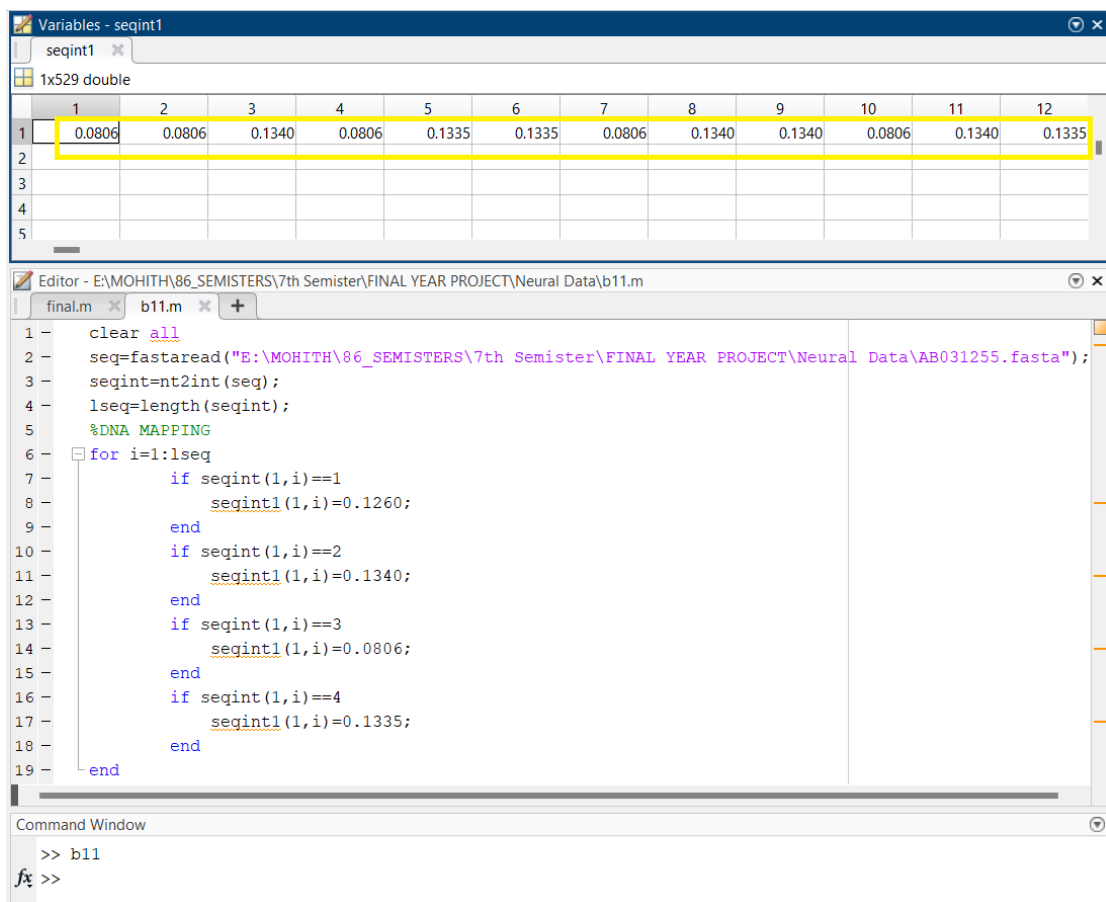


Figure 4.3- EIIP Mapping Simulation

4.4 CONCLUSION

Thus, all the collected nucleotide sequences are mapped into the corresponding numerical sequences. To be processed further, all the numeric sequences are provided as the input signal to the FLANN-based PSO-tuned Levenberg–Marquardt adaptive filter

CHAPTER- 5

ADAPTIVE FILTER

5.1 INTRODUCTION

An adaptive filter is a digital filter that has self-adjusting characteristics. It is capable of adjusting its filter coefficients automatically to adapt the input signal via an adaptive algorithm. Adaptive filters play an important role in modern digital signal processing (DSP) products in areas such as telephone echo cancellation, noise cancellation, equalization of communications channels, biomedical signal enhancement, active noise control (ANC), and adaptive control systems.

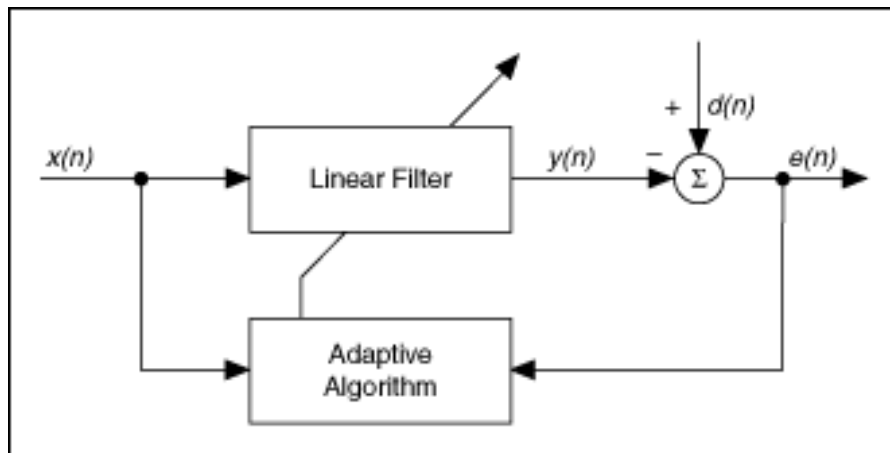


Figure 5.1 – Basic Block Diagram of Adaptive Filter

Where

$x(n)$ is the input signal to a linear filter at time n

$y(n)$ is the corresponding output signal

$d(n)$ is an additional input signal to the adaptive filter

$e(n)$ is the error signal that denotes the difference between $d(n)$ and $y(n)$

5.2 METHODOLOGY

5.2.1 FUNCTIONAL EXPANSION

To realize the nature of gene specifically the input gene sequence is expanded using non-linear trigonometric functions, which are the subset of a complete set of orthonormal basis functions spanning an n-dimensional representation space. This result, mapping of input gene pattern into a larger pattern space. Therefore, the combination of overall functional expansion possesses clear idea about the gene characteristics[2]

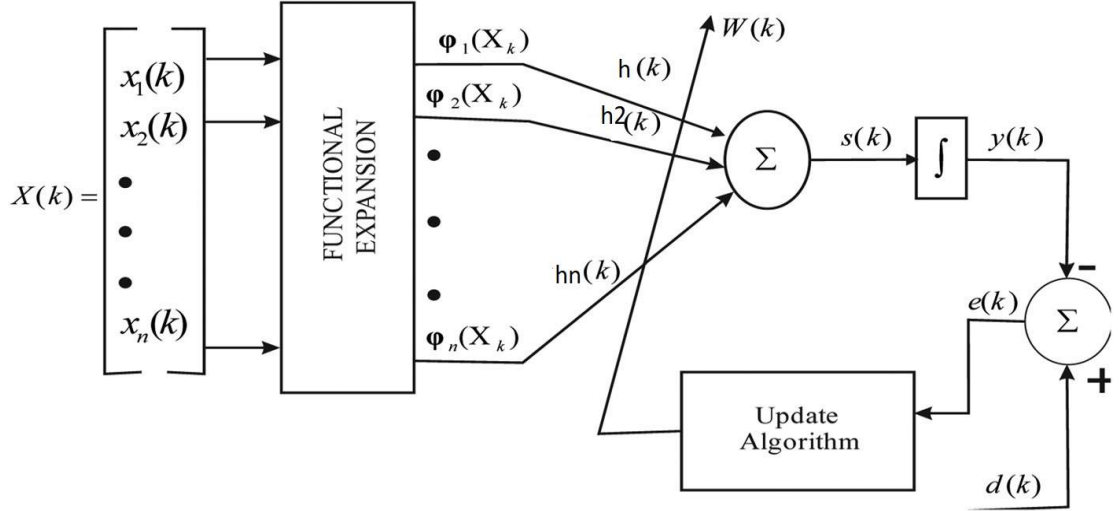


Figure 5.2 – Basic Block Diagram of Functional Expansion

$$x(k) = [1, \sin\{ax(k)\}, \cos\{nx(k)\}, \sin\{2nx(k)\}, \cos\{2nx(k)\}]A$$

$$h(k) = [h_0(k), h_1(k), h_2(k), \dots, h_{2n+1}(k)]^T$$

Hence the estimated output of the identification model is computed as:

$$y(k) = v(k) h(k)$$

5.2.2 ADAPTIVE FILTER

At first the weight of the adaptive filter is set to zero. With this initial condition the adaptive filter measures the NMSE between input and desired signal. Now the weight of the adaptive filter is adjusted to reduce the NMSE.

If the unknown input is a healthy gene, the estimated NMSE will reduce faster compared to a cancerous input gene. Since two similar type genes are more biologically related, the estimated error between two healthy genes is less compared to the error estimated between one healthy and one cancer gene.

Therefore depending on the amount of estimated error, the adaptive filter decides whether the unknown input gene is healthy or cancerous. These steps are shown. The adaptive filter with zero initial conditions is chosen for gene identification because of its simplicity and inherent stability. The LMS algorithm is employed to adjust the coefficients of the FIR estimator to reduce mean squared error (MSE), which is a measure of the accuracy of the estimated signal. The estimation error can be found as:

$$e(k) = d(k) - y(k)$$

$$d(k) = h(k) * x(k)$$

where $e(k)$ is the error value, $d(k)$ is the known gene signal, $y(k)$ is the adaptive filtered response of unknown gene signal, $h(k)$ is the weight vector, $x(k)$ is the unknown gene signal, and $*$ denotes convolution.

Generally, the tuning of the learning parameter is done using hit and trial method. But because the learning rates are not most appropriate, the best dynamic performance might not be achieved.

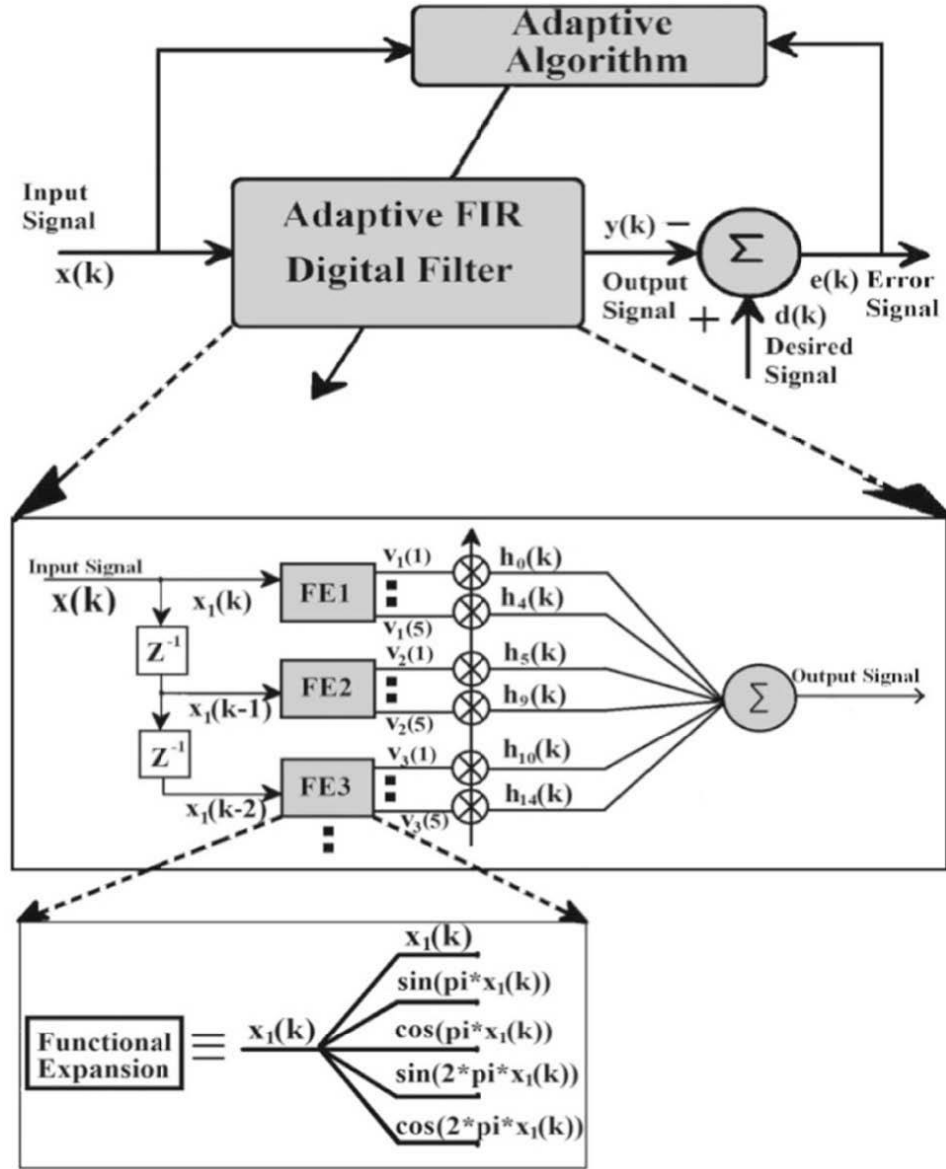


Figure 5.3 – Block Diagram of Adaptive Filter

Therefore, authors have introduced the “PSO-tuned algorithm”[7] to achieve the best dynamic performance of the model which provides us the optimal learning rates. For fast convergence, the optimum value is selected by applying the “PSO-tuned algorithm” as shown in the following section. Because of this reason, a function called cost function in is defined which will be optimized in the “PSO algorithm” to obtain the best value of g . Here the “cost function” is defined as follows:

$$f = \frac{1}{L} \sum_{i=1}^L (e[k])^2,$$

5.3 RESULT

Algorithm based adaptive filter is used as a tool to identify the healthy and cancerous nature of genes. Mean of healthy gene is considered as reference which is treated as desired signal for adaptive algorithm. Rest genes are treated as unknown signal or target and are filtered out using adaptive filtering technique. According to basic concept of adaptive algorithm, it is desired to minimize the error between reference and target. The number of iterations for the algorithm is set to 1000 to minimize the error. Since after 1000 iterations the NMSE values of both target and reference gene become stable

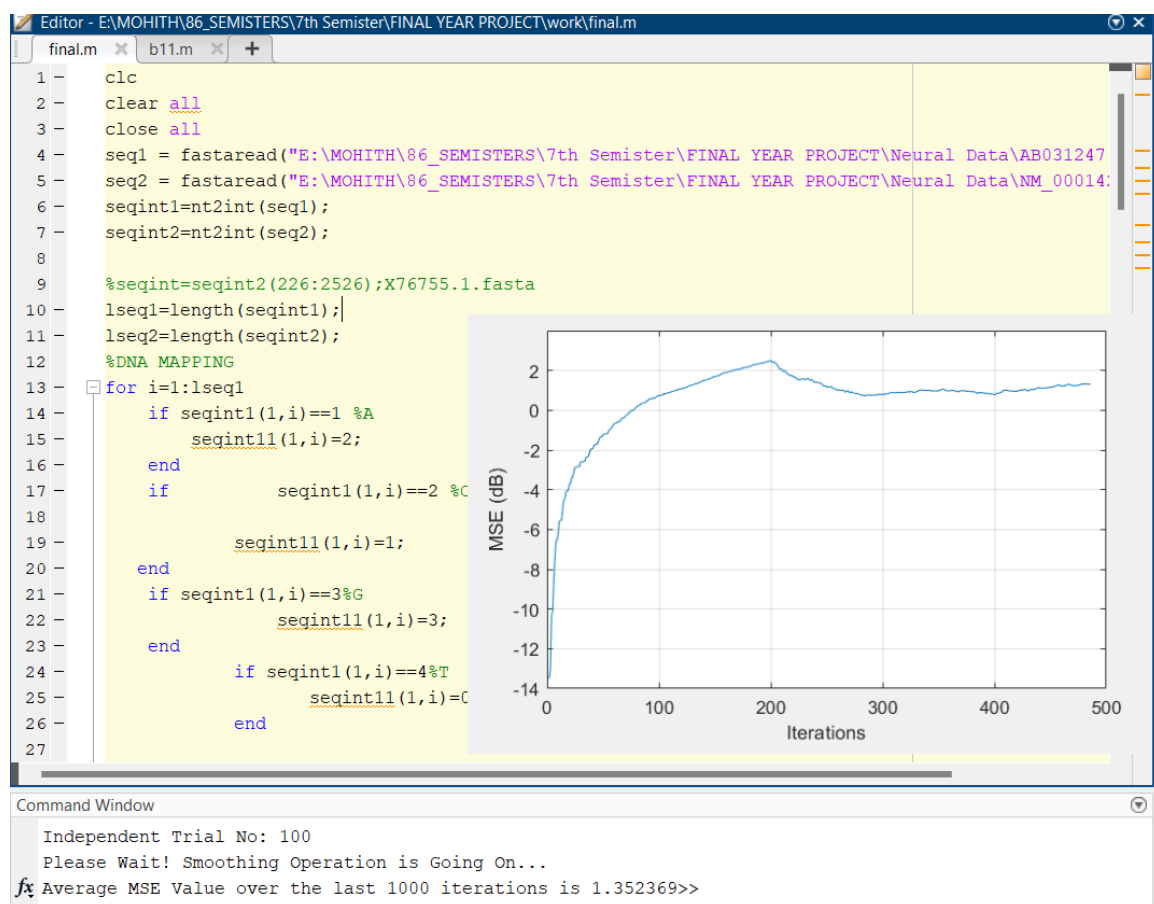


Figure 5.4 – Plotting of MSE vs Iterations

5.4 CONCLUSION

The FLANN based adaptive filtering is proved to be very promising technique for gene identification and can successfully be used for separating the diseased genes from the healthy ones. The random variation of cancer gene is reduced after 2000 iterations when average NMSE is evaluated.

CHAPTER- 6

MEAN SQUARE ERROR

6.1 Introduction

In statistics, the mean squared error or mean squared deviation of an estimator measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value. MSE is a risk function, corresponding to the expected value of the squared error loss

6.2 Materials and Methods

6.2.1 MEAN SQUARE ERROR

The MSE formula is represented as follows:

$$MSE = \sum_{k=1}^L \frac{|Y_k - E_k|^2}{L},$$

Where:

L is the length of the signal anticipated, Y_k is the desired signal, and E_k is the estimated test signal of the k-th element of the signal.

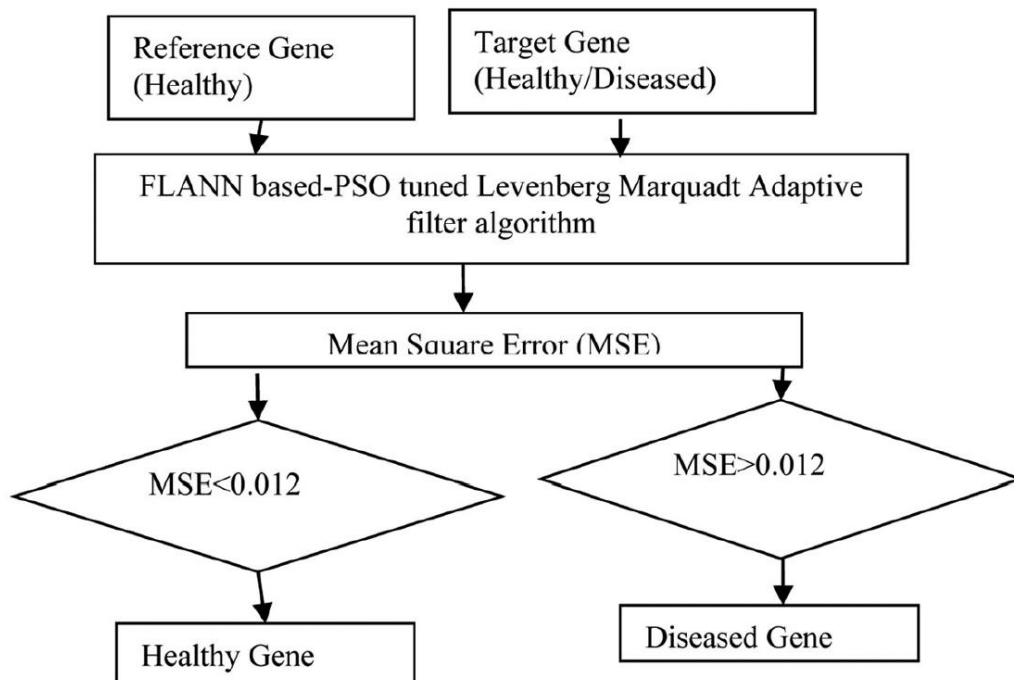


Figure 6.1 – MSE Flow Chart

During the entire iterations, it is observed that in case of a healthy gene, the error decreases much faster than the diseased one. This entails that as the target gene is like reference, that is, healthy, error attains zero in a small number of iterations, whereas the target gene is unlike, that is, diseased one; more iterations are taken to reach to zero. Hence the MSE acts as a signature for the disease gene identification.[2]

6.3 RESULTS

6.3.1 DISEASED GENE

6.3.2.1 Output-1

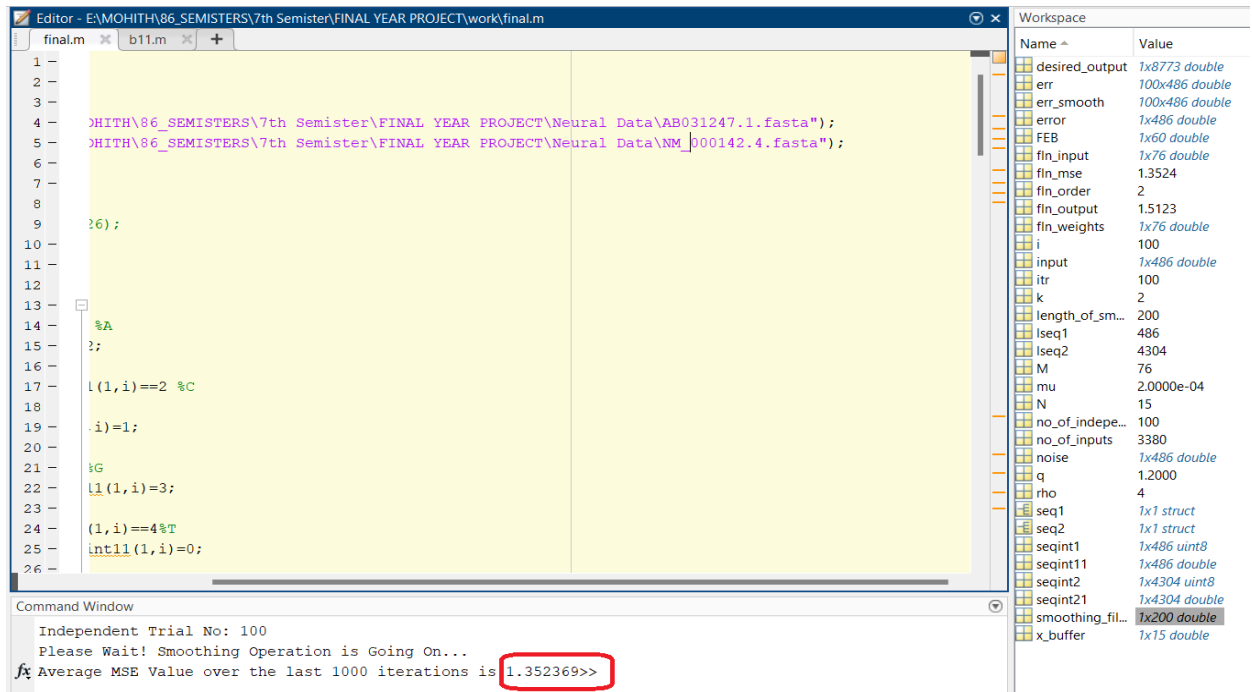


Figure 6.2 – Matlab Output- 1 of Diseased Gene

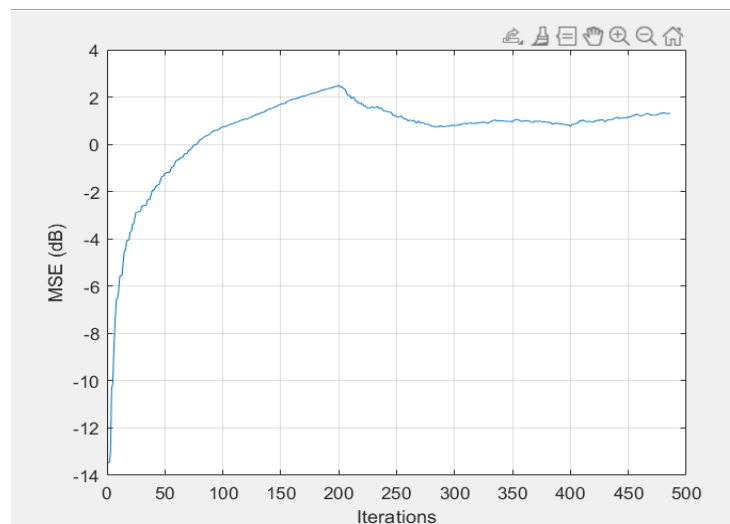


Figure 6.3 – Plotting of MSE vs Iterations for Output -1 of Diseased Gene

A result of **1.35** was obtained after simulation which when compared with the MSE value indicates that the target gene is Diseased.

A graph between MSE values and number of iterations is also plotted and shown.

6.3.2.2 Output-2

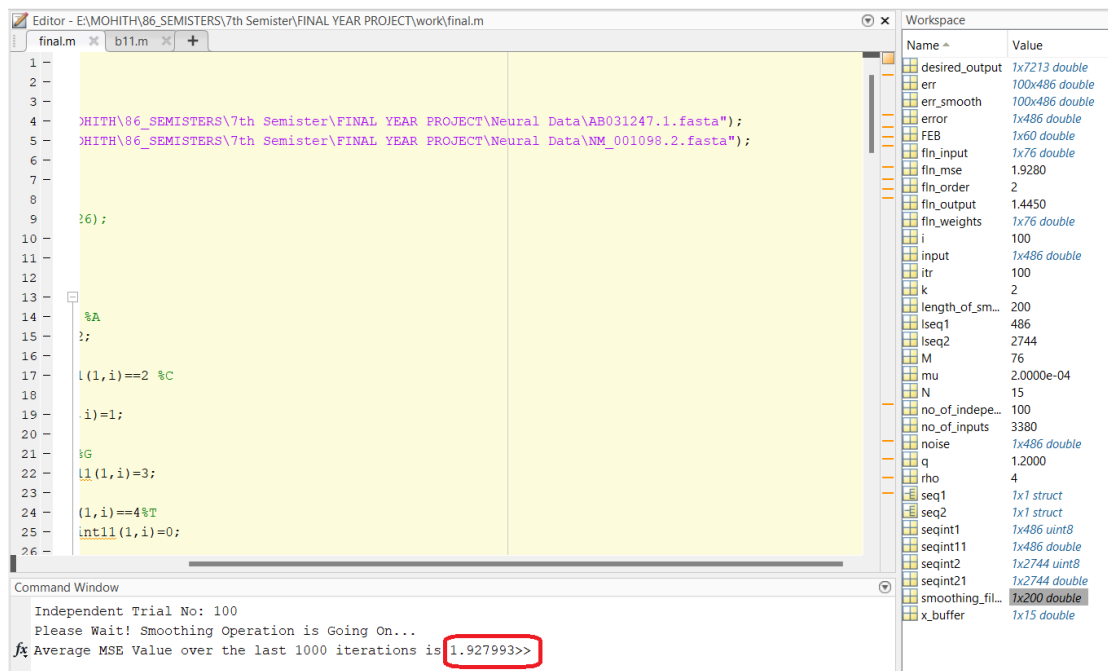


Figure 6.4 – Matlab Output -2 of Diseased Gene

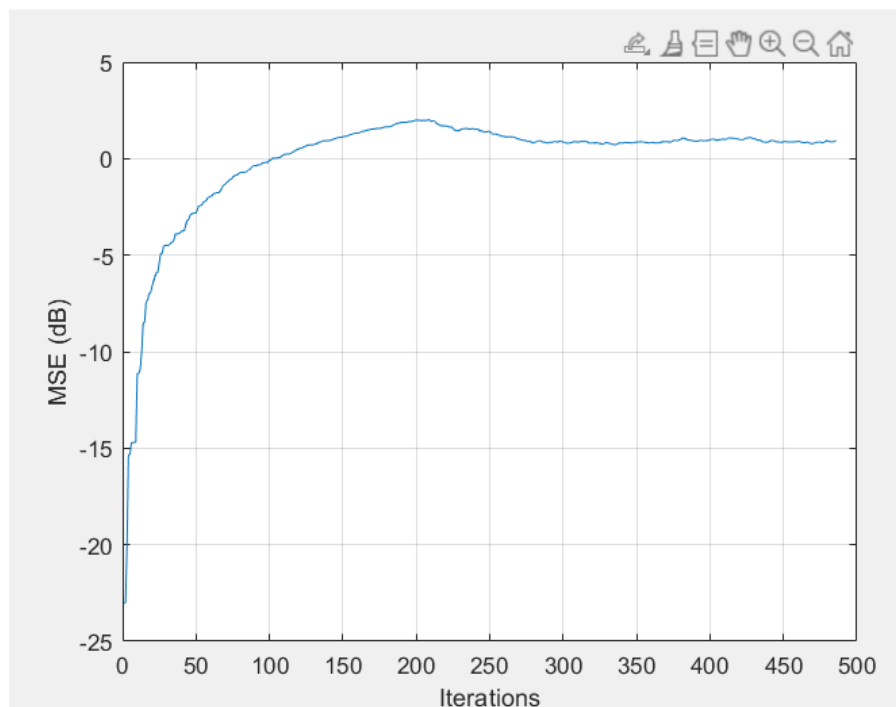
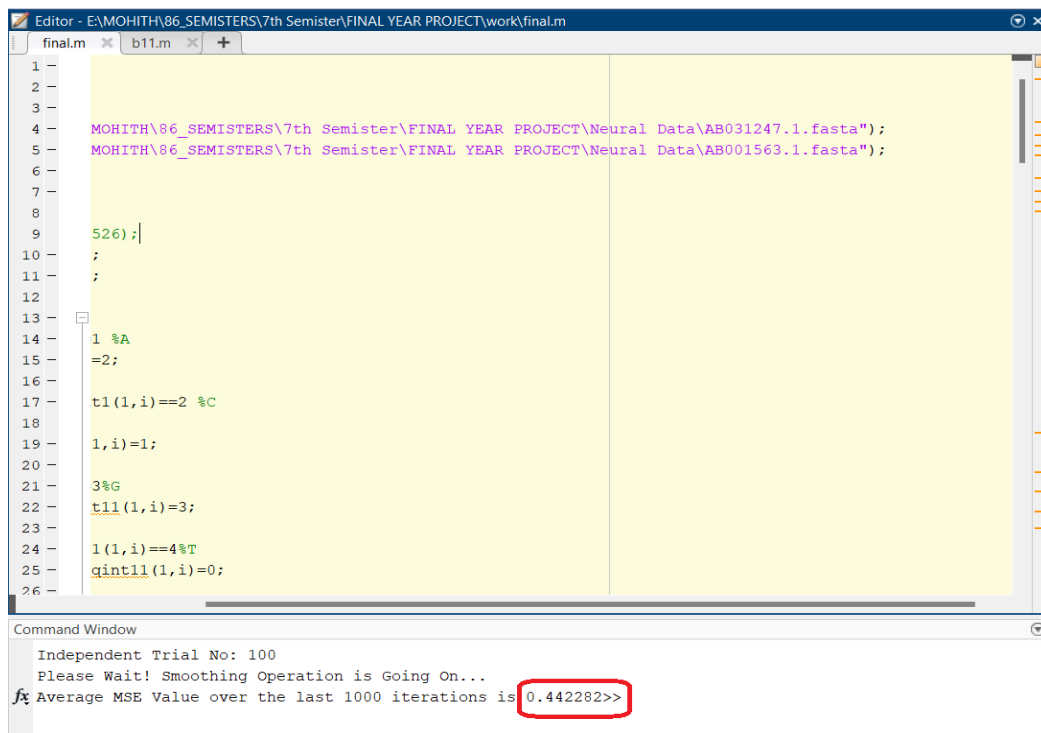


Figure 6.5 – Plotting of MSE vs Iterations for Output -2 of Diseased Gene

A result of **1.92** was obtained after simulation which when compared with the MSE value indicates that the target gene is Diseased.

A graph between MSE values and number of iterations is also plotted and shown.

6.3.2.3 Output-3



```

1
2
3
4 MOHITH\86_SEMISTERS\7th Semester\FINAL YEAR PROJECT\Neural Data\AB031247.1.fasta");
5 MOHITH\86_SEMISTERS\7th Semester\FINAL YEAR PROJECT\Neural Data\AB001563.1.fasta");
6
7
8
9 526);|
10 ;
11 ;
12
13
14 1 %A
15 =2;
16
17 t1(1,i)==2 %C
18
19 1,i)=1;
20
21 3%G
22 t11(1,i)=3;
23
24 1(1,i)==4%T
25 qint11(1,i)=0;
26

```

Command Window

Independent Trial No: 100
Please Wait! Smoothing Operation is Going On...
fx Average MSE Value over the last 1000 iterations is **0.442282>>**

Figure 6.6 – Matlab Output -3 of Diseased Gene

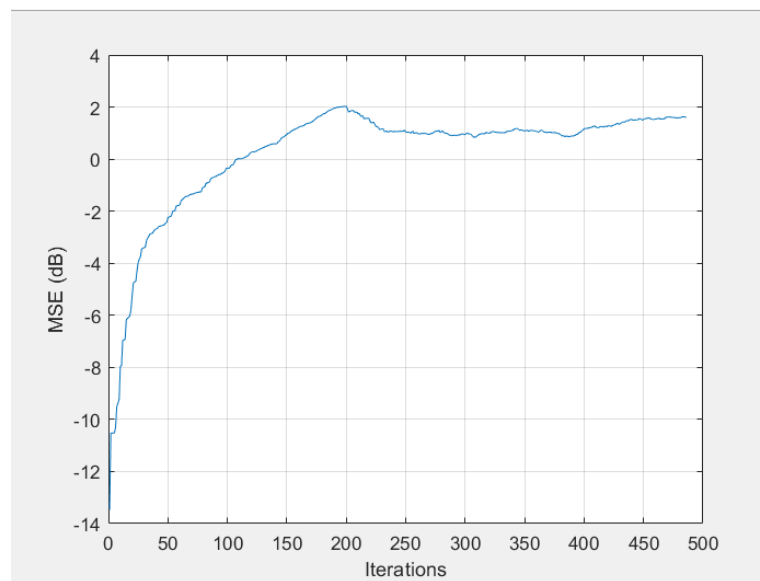


Figure 6.7 – Plotting of MSE vs Iterations for Output -3 of Diseased Gene

A result of **0.44** was obtained after simulation which when compared with the MSE value indicates that the target gene is Diseased.

A graph between MSE values and number of iterations is also plotted and shown.

6.3.2 HEALTHY GENE

6.3.2.1 Output-1

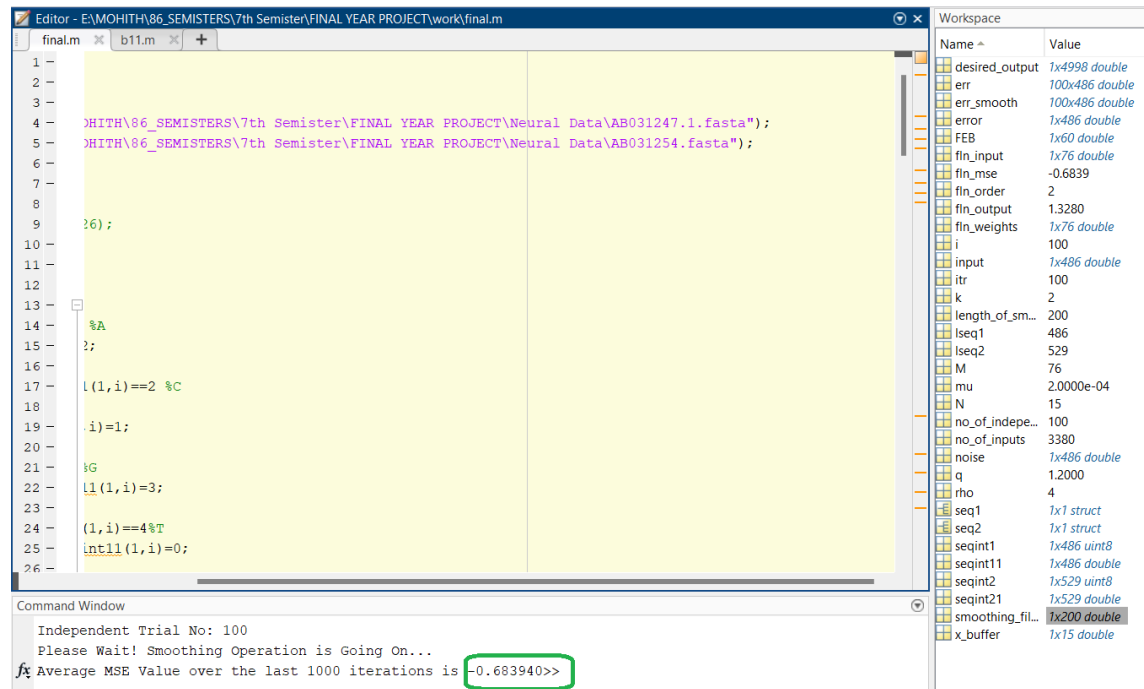


Figure 6.8 – Matlab Output -1 of Healthy Gene

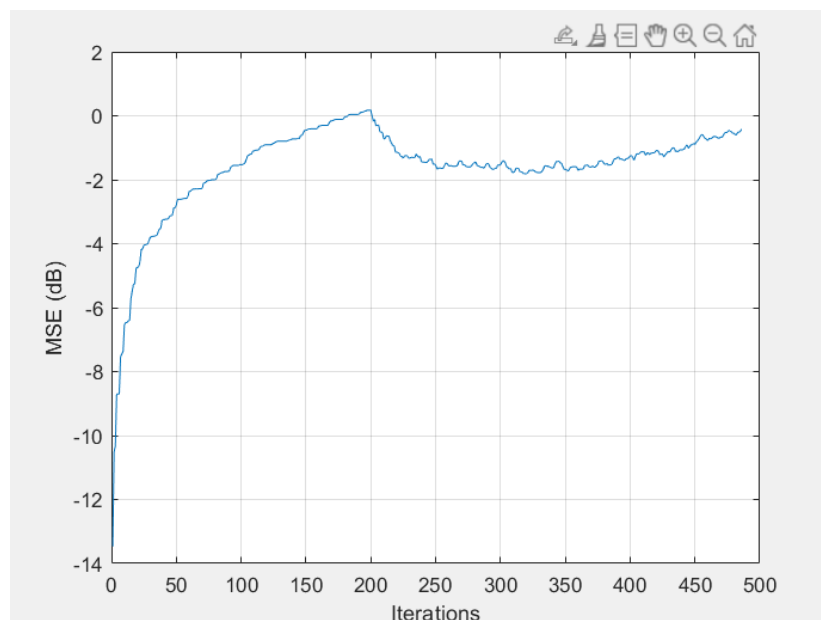


Figure 6.9 – Plotting of MSE vs Iterations for Output -1 of healthy Gene

A result of **-0.68** was obtained after simulation which when compared with the MSE value indicates that the target gene is Healthy.

A graph between MSE values and number of iterations is also plotted and shown.

6.3.2.2 Output-2

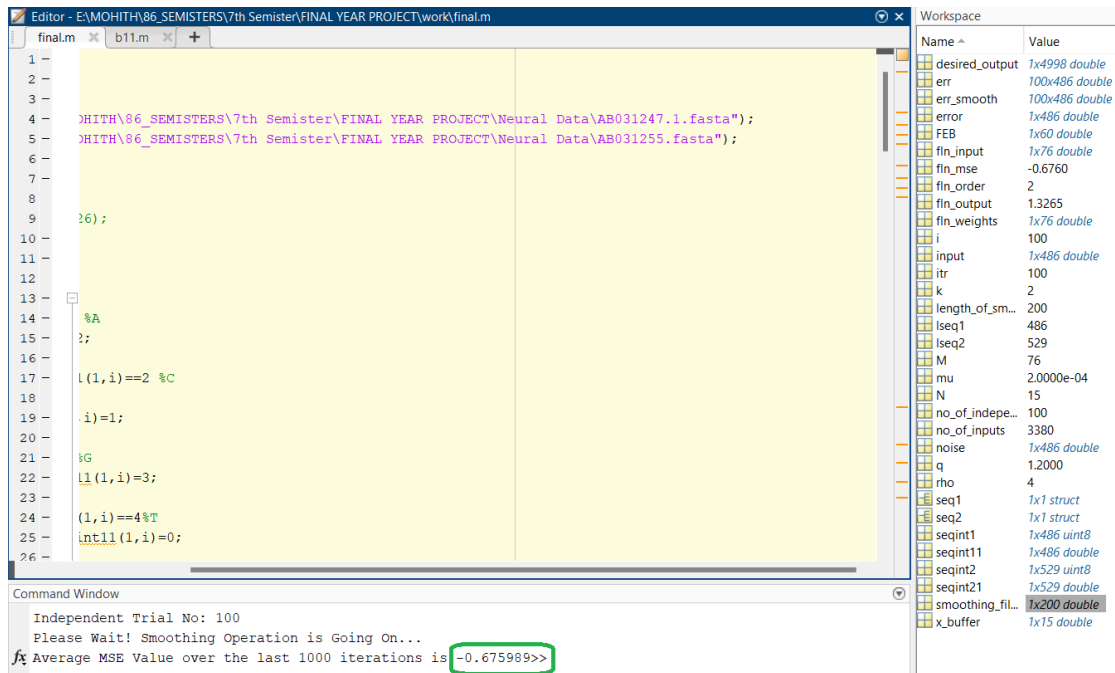


Figure 6.10 – Matlab Output -2 of Healthy Gene

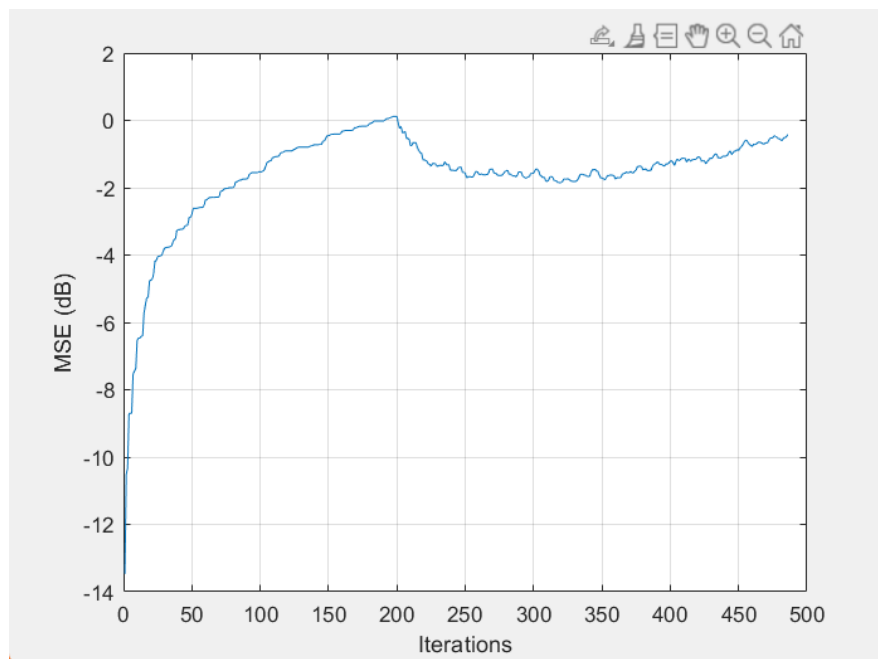


Figure 6.11 – Plotting of MSE vs Iterations for Output -2 of Healthy Gene

A result of **-0.67** was obtained after simulation which when compared with the MSE value indicates that the target gene is Healthy.

A graph between MSE values and number of iterations is also plotted and shown.

6.3.2.3 Output-3

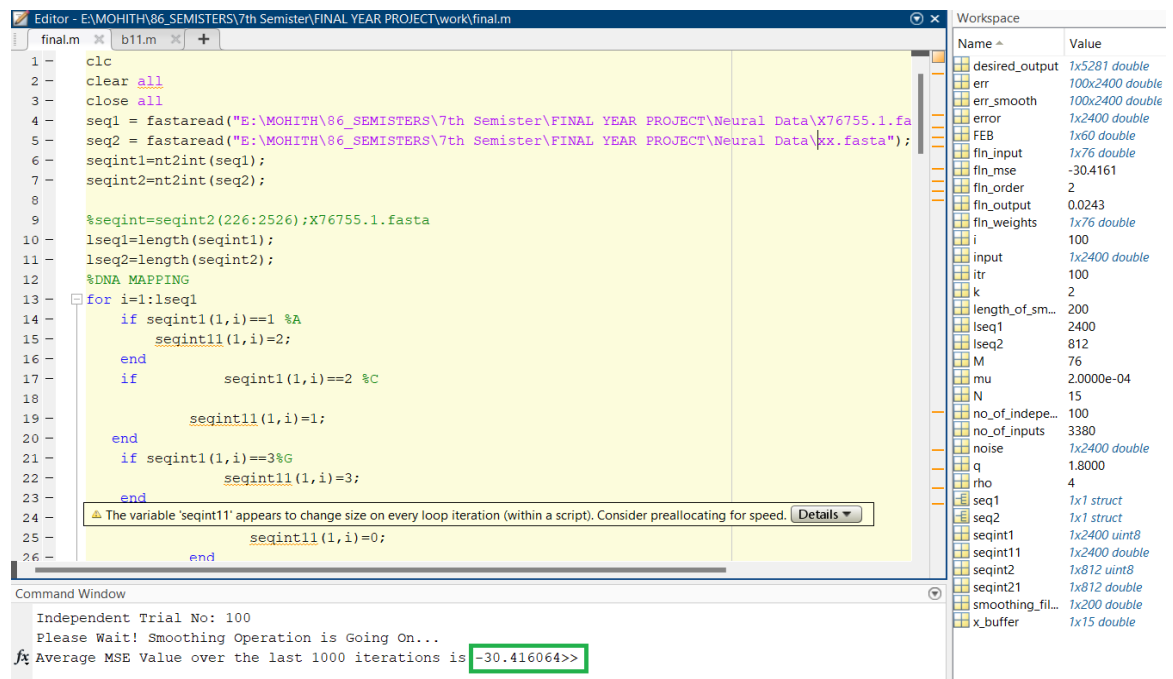


Figure 6.12 – Matlab Output -3 of Healthy Gene

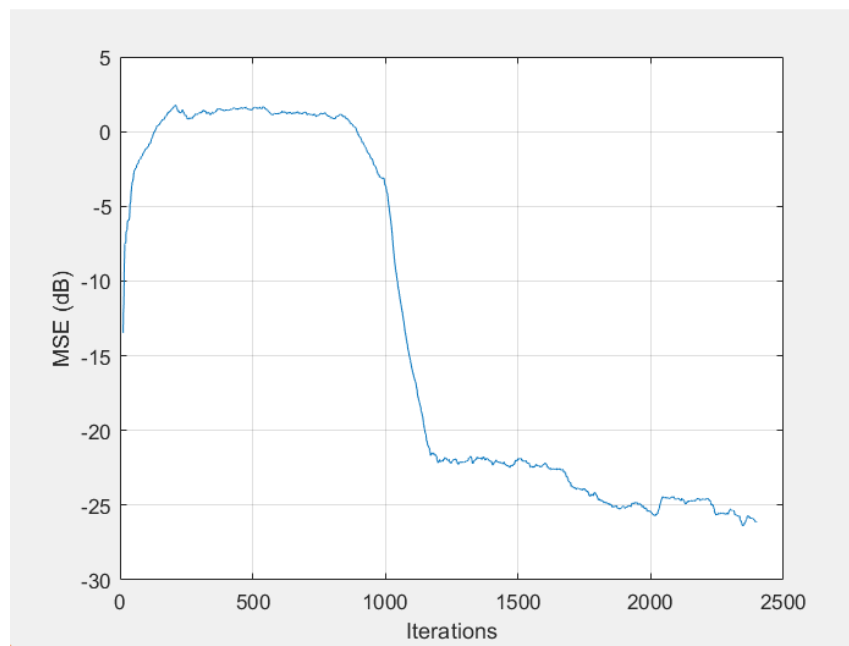


Figure 6.13 – Plotting of MSE vs Iterations for Output -3 of Healthy Gene

A result of **-0.67** was obtained after simulation which when compared with the MSE value indicates that the target gene is Healthy.

A graph between MSE values and number of iterations is also plotted and shown.

6.3.3 Machine Learning Outputs

After training the model with the dataset, a random input is given to the model and the code makes a decision as to whether the given input is healthy or diseased

6.3.3.1 Diseased Gene

```
In [4]: #Training and Testing the Dataset
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.2)
model=DecisionTreeClassifier()
model.fit(X_train, y_train)
prediction=model.predict([[16,0.040549]])
prediction

array([1], dtype=int64)
```

```
In [22]: # prediction of type of gene
if(prediction==1):
    print("The given Gene is Diseased")
else:
    print("The given Gene is Healthy")

The given Gene is Diseased
```

Figure 6.14 – Machine Learning Output -1 of Diseased Gene

```
In [9]: #Training and Testing the Dataset
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.2)
model=DecisionTreeClassifier()
model.fit(X_train, y_train)
prediction=model.predict([[16,0.080549]])
prediction

array([1], dtype=int64)
```

```
In [10]: # prediction of type of gene
if(prediction==1):
    print("The given Gene is Diseased")
else:
    print("The given Gene is Healthy")

The given Gene is Diseased
```

Figure 6.15 – Machine Learning Output -2 of Diseased Gene

6.3.3. Healthy gene

```
In [15]: #Training and Testing the Dataset
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.2)
model=DecisionTreeClassifier()
model.fit(X_train, y_train)
prediction=model.predict([[16,0.101000]])
prediction
```

```
array([0], dtype=int64)
```

```
In [16]: # prediction of type of gene
if(prediction==1):
    print("The given Gene is Diseased")
else:
    print("The given Gene is Healthy")
```

```
The given Gene is Healthy
```

Figure 6.16 – Machine Learning Output -1 of Healthy Gene

```
In [18]: #Training and Testing the Dataset
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.2)
model=DecisionTreeClassifier()
model.fit(X_train, y_train)
prediction=model.predict([[21,0.121]])
prediction
```

```
array([0], dtype=int64)
```

```
In [16]: # prediction of type of gene
if(prediction==1):
    print("The given Gene is Diseased")
else:
    print("The given Gene is Healthy")
```

```
The given Gene is Healthy
```

Figure 6.17 – Machine Learning Output -2 of Healthy Gene

6.3.4 NORMALISED MEAN SQUARE ERROR (NMSE) GRAPH

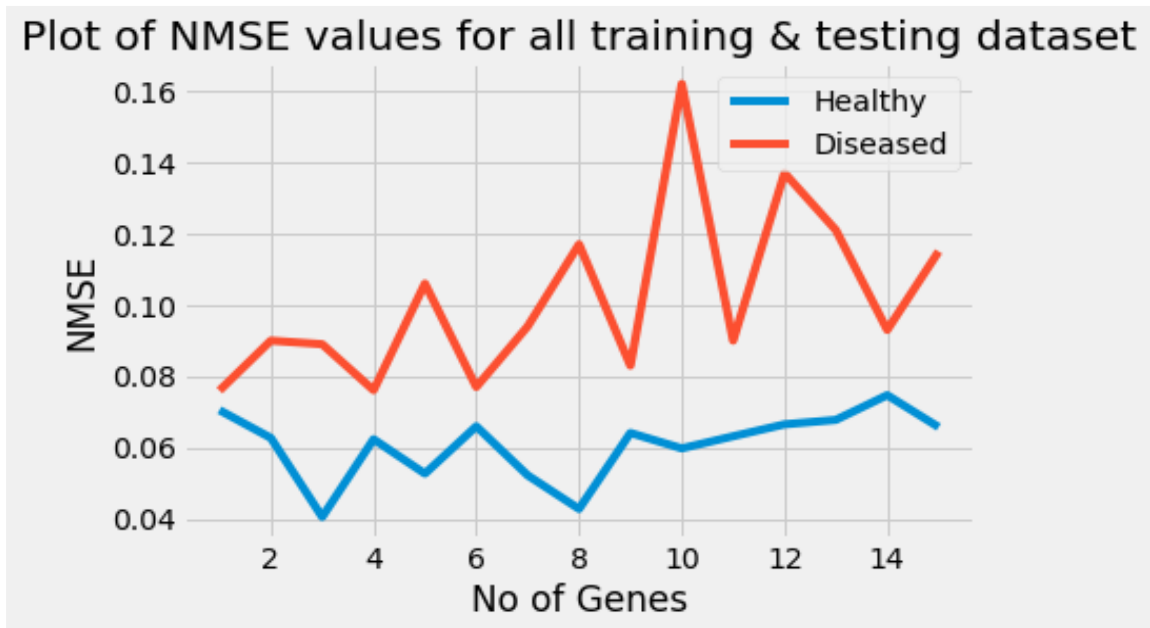


Figure 6.18 – Plotting of NMSE vs No of Genes

6.4 CONCLUSION

One of the collected genes (healthy) is taken as the reference signal and is referred to as the desired signal for the proposed algorithm. Others are taken as the target (unknown) signal and are filtered out employing the proposed algorithm to minimize the error between the reference and the target signal.

For healthy gene the error decreases much faster than the diseased gene. This implies when target gene is healthy in nature like reference, error reaches zero within few iterations, whereas for cancer target gene, it takes large number of iterations to reach zero. Therefore this NMSE can act as a signature for diseased gene identification.

CHAPTER- 7

PROJECT CODES

7.1 Introduction

This chapter gives us the detailed Mat Lab codes for all the simulation results which were shown in the previous chapter (**Chapter 6**). This chapter helps us to understand the exact operations of all the various blocks which were involved in the successful implementation of our project.

7.2 Matlab Codes

7.2.1 Reading DNA sequences

```
clc

clear all

close all

seq1 = fastaread("E:\MOHITH\86_SEMISTERS\7th Semester\FINAL YEAR
PROJECT\Neural Data\AB031247.1.fasta");

seq2 = fastaread("E:\MOHITH\86_SEMISTERS\7th Semester\FINAL YEAR
PROJECT\Neural Data\AB031254.fasta");

seqint1=nt2int(seq1);

seqint2=nt2int(seq2);

lseq1=length(seqint1);

lseq2=length(seqint2);
```

7.2.2 EIIP Mapping

%DNA MAPPING FOR SEQ-1

```
for i=1:lseq1

    if seqint1(1,i)==1 %A

        seqint11(1,i)=0.1260;

    end

    if seqint1(1,i)==2 %C

        seqint11(1,i)=0.1340;

    end

    if seqint1(1,i)==3 %G

        seqint11(1,i)=0.0806;

    end

    if seqint1(1,i)==4 %T

        seqint11(1,i)=0.1335;

    end

end
```

%DNA MAPPING FOR SEQ-2

```
for i=1:lseq2

    if seqint2(1,i)==1 %A

        seqint21(1,i)=0.1260;

    end

    if seqint2(1,i)==2 %C

        seqint21(1,i)=0.1340;

    end

    if seqint2(1,i)==3 %G
```

```

        seqint21(1,i)=0.0806;

    end

    if seqint2(1,i)==4%T
        seqint21(1,i)=0.1335;
    end

end

```

7.2.3 Calculation of Mean Square Error

```

no_of_independent_trials = 100;

for itr=1:no_of_independent_trials

    clc;

    disp(['Independent Trial No: ',num2str(itr)])

    no_of_inputs = 3380;

    input=seqint11;

    N=15;

    fln_order =2;

    x_buffer=zeros(1,N);

    M = (2*fln_order+1)*N + 1;

    fln_weights=zeros(1,M);

    mu=0.0002;

    noise = awgn(input,30)-input;

    for i=1:length(input)

        x_buffer=[input(i) x_buffer(1:end-1)];
    end
end

```

```

q = 1.5 * input(i) - 0.3*input(i)^2 ;

if q>0

    rho = 4;

else

    rho=0.5;

end

desired_output=[seqint21,zeros(1,4469)];

FEB=[];

for k =1:fln_order

    FEB=[FEB, sin(pi*k*x_buffer), cos(pi*k*x_buffer)];

end

fln_input= [1,x_buffer,FEB];

fln_output= fln_weights * fln_input';

error(i)= desired_output(i) - fln_output;


fln_weights=fln_weights + 2 * mu * error(i) * fln_input;

end

err(itr,:)=error.^2;

end

disp(['Please Wait! Smoothing Operation is Going On...'])

length_of_smoothing_filter = 200;

smoothing_filter_coeff =

(1/length_of_smoothing_filter)*ones(1,length_of_smoothing_filter);

```



```

for i=1:itr

    err_smooth(i,:) = filter(smoothing_filter_coeff,1,err(i,:));

end

figure;

plot(10*log10(mean(err_smooth))); xlabel('Iterations');ylabel('MSE (dB)'); grid on;

fln_mse=(10*log10(mean(mean(err(end-1000:end))))));

fprintf('Average MSE Value over the last 1000 iterations is %f', fln_mse);

```

7.3 Machine Learning Codes

7.3.1 Importing the libraries and reading CSV File

```

import pandas as pd

from sklearn.tree import DecisionTreeClassifier

from sklearn.model_selection import train_test_split

from sklearn.metrics import accuracy_score

gene_data=pd.read_csv("pro100.csv",index_col=0,usecols=[0,1,3,4])

gene_data

```

7.3.2 Splitting the Dataset into input and output

#Inputs

```

X =gene_data.drop(columns=['Decision'])

print(X)

```

#Outputs

```

y= gene_data['Decision']

```

```
print(y)
```

7.3.3 Training and Testing the Dataset

```
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.2)
```

```
model=DecisionTreeClassifier()
```

```
model.fit(X_train, y_train)
```

```
prediction=model.predict(X_test)
```

```
prediction
```

7.3.4 Accuracy

```
score=accuracy_score(y_test,prediction)
```

```
print(f"Accuracy: {score}")
```

CHAPTER- 8

CONCLUSION AND FUTURE SCOPE

8.1 Conclusion

Diseased natures of genes are identified by measuring NMSE using adaptive algorithm and decision is taken about the gene's nature depending on the values of NMSE. According to concept of adaptive algorithm, NMSE between two healthy genes is less compared to one healthy and one cancer gene, when estimated after certain iterations. The value of NMSE reflects the differences between healthy and cancer genes and acts as identification criterion.

8.2 Future Scope

With the growing demand for disease-related research and with a detailed idea on the nucleotide sequence, the proposed approach can be a better non-invasive technique that may prove itself as a robust technique to fight with many deadly diseases. Other evolutionary algorithms can be included in the proposed process to perform better results. The technique may be tested for other diseases like cancer, mutational disease, and many more for predicting the approach as a versatile one.

REFERENCES

1. DSP based entropy estimation for identification and classification of Homo sapiens cancer genes. (Springer-Verlag Berlin Heidelberg 2016) Joyshri Das, Soma Barman
2. A non-invasive cancer gene detection technique using FLANN based adaptive filter (Saikat Singha Roy, Soma Barman)
3. Ross, C. A.; Tabrizi, S. J. Huntington's Disease: From Molecular Pathogenesis to Clinical Treatment. Lancet Neurol.
4. Akhtar, M.; Epps, J.; Ambikairajah, E. Signal Processing in Sequence Analysis Advances in Eukaryotic Gene Prediction. IEEE J. Sel. Top. Signal Process.
5. Prediction of Cancer cells using DSP techniques. (International conference on Communication and Signal Processing, April 3-5, 2013, India) G.N. Satapathi, Dr. P. Srihari, Senior Member, IEEE, Ch. Aruna Jyothi, S. Lavanya
6. DWT based Cancer Identification using EIIP. (2016 Second International Conference on Computational Intelligence & Communication Technology) Shilpi Chakraborty, Vinit Gupta.
7. Why genetic code context of nucleotides for DNA signal processing? a review (Muneer Ahmad, Low Tan Jung, Al-Amin Bhuiyan)
8. Functional Link Artificial Neural Network-based Disease Gene Prediction (Jiabao Sun, Jagdish c. Patra and Yong jin Li)
9. A Classification Technique for Microarray Gene Expression Data using PSO-FLANN (Jayashree Dev1, Sanjit Kumar Dash2, Sweta Dash 3, Madhusmita Swain)

10. Prediction of cancer cell using digital signal processing (S.Barman (mandal), M.Roy, S.Biswas, S.Saha)
11. Travers, Andrew & Muskhelishvili, Georgi. (2015). DNA structure and function. The FEBS journal. 282. 10.1111/febs.13307.
12. <https://www.ncbi.nlm.nih.gov/nucleotide/>
13. <https://in.mathworks.com/products/matlab.html/>
14. <https://www.geeksforgeeks.org/best-python-libraries-for-machine-learning/>

Appendix A

Details of Project and Relevance to Environment, Safety, Ethics & Cost

Mapping with POs and PSO with Justification

Title of Project	Roll No of Students	Project Supervisor	Relevance to Environment	Relevance to Human Safety	Relevance to Ethics	Hardware or Software	Type (Application, Product, Research, Review) Application Based Product
HEREDITARY DISEASE PREDICTION IN EUKARYOTE DNA	160418735086 160418735083 160418735079	Mrs. G. Prashanthi	2	-	-	Software Project	

Mapping with PO and PSO with justification

Implementation Details:	This Project uses Adaptive Signal Processing Techniques, Numerical representation methods, NMSE method to classify the normal genes and cancer genes. The main aim of this project is detect whether the gene is healthy or diseased using Adaptive signal processing without involving any biological treatment.
PO1	Application of basic mathematics and applications of concepts of engineering fundamentals and digital signal processing for the design of algorithms for our codes.
PO2	Research of previous literature, study and analysis of weights calculation and Biological sciences.
PO3	Program developed with the purpose of being either personal or public software tool in mind.

PO4	Research based knowledge and previous research methods used to identify the diseased genes and healthy genes by using datasets that are available on NCBI website.
PO5	Modern and cost-effective software tools are used to develop and run this program. Mat Lab R2021a and Python Jupyter notebook has been used.
PO6	Project aims to help in no medical or biological testing needed to identify diseased genes at an early state.
PO7	The program runs on MATLAB and JUPYTER NOTEBOOK which are easily available software tools thereby requiring no other additional resources.
PO8 PO9	Work was equally divided and deadlines were properly met. A three-member team was involved in the research of the biological study and DNA mapping methods and study of signal Processing which helped in improving team working skills of students.
PO10	Project report was submitted and PowerPoint presentation was given to improve communication (Spoken and Written) skills.
PO11	Project management fundamentals were utilized as the students had worked together and the work load of learning and implementing the new concepts were equally shared.
PO12	With the advancement in the technology, Project helps in building the lifelong learning skills in the student to face real world challenges and provide solutions.
PSO1	NA
PSO2	NA
PSO3	The project deals with the concept of Adaptive Signal Processing.
PSO4	The students have designed the code for detecting diseased genes using Adaptive Signal Processing in MATLAB 2021a and JUPYTER NOTEBOOK

PROGRAM OUTCOMES

PO1: Engineering knowledge: Apply the knowledge of mathematics, science, engineering fundamentals, Biological sciences, and an engineering specialization to the solution of complex engineering problems.

PO2: Problem analysis: Identify, formulate, research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, biological sciences and engineering sciences

PO3: Design/development of solutions: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

PO4: Conduct investigations of complex problems: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

PO5: Modern tool usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.

PO6: The engineer and society: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

PO7: Environment and sustainability: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

PO8: Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice. 66

PO9: Individual and teamwork: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

PO10: Communication: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

PO11: Project management and finance: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

PO12: Life-long learning: Recognize the need for and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

PROGRAM SPECIFIC OUTCOMES

PSO1: The ECE Graduates will be Equipped with knowledge of complete design flow from specification to silicon in areas of both digital and analog VLSI Design and will be able to work in IC Design companies.

PSO2: The ECE Graduates will be Equipped with microprocessor and microcontroller based system design skills and can work as design and verification engineers in the area of Embedded Systems Design

PSO3: The ECE Graduates will be able to apply engineering knowledge for design and implementation of projects pertaining to signal processing and Communications

PSO4: The ECE Graduates will be Equipped with necessary soft skills, aptitude, and technical skills to work in the software industry and IT sector.

Appendix B

Gantt Chart

TASK	PROGRESS	START	END
Study of DNA Structure	100%	11/10/21	11/16/21
Study of Research Papers	100%	11/17/21	11/22/21
A non-invasive cancer gene detection technique using FLANN based adaptive filter	100%	11/17/21	11/17/21
Akhtar, M. Epps, J. Ambikairajah, E. Signal Processing in Sequence Analysis Advances in Eukaryoti	100%	11/18/21	11/18/21
WHY GENETIC CODE CONTEXT OF NUCLEOTIDES FOR DNA SIGNAL PROCESSING	100%	11/19/21	11/20/21
Functional Link Artificial Neural Network-based Disease Gene Prediction	100%	11/21/21	11/22/21
ElIP Mapping	100%	11/24/21	12/6/21
Theretical Calculations	100%	11/24/21	11/27/21
MATLAB implementation	100%	11/28/21	12/6/21
Working with adaptive Filter	100%	12/10/21	1/15/22
Theretical Calculations	100%	12/10/21	12/18/21
MATLAB implementation	100%	12/19/21	1/15/22
Creating MSE Dataset	100%	2/10/22	2/23/22
Understanding Machine Learning	100%	3/26/22	4/8/22
Finding Performance Metrics	100%	4/9/22	4/29/22
Thesis	100%	5/1/22	5/16/22

