

# **CLASSIFICATION OF CANCEROUS AND NON-CANCEROUS DNA SEQUENCES USING DSP BASED METHOD**

*Project report submitted in partial fulfillment of the requirement for the degree  
of*

**BACHELOR OF TECHNOLOGY**

**IN**

**ELECTRONICS AND COMMUNICATION ENGINEERING**

By

**Anupam Srivastava (171070)**

**Saumya Jajodia (171073)**

**UNDER THE GUIDANCE OF**

**Dr.Sunil Datt Sharma**



**JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY,  
WAKHNAGHAT**

**MAY 2021**

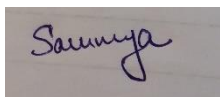
# TABLE OF CONTENTS

<b>CAPTION</b>	<b>PAGE NO.</b>
DECLARATION	4
ACKNOWLEDGEMENT	5
LIST OF ACRONYMS AND ABBREVIATIONS	6
LIST OF FIGURES	8
LIST OF TABLES	9
ABSTRACT	10
<b>CHAPTER-1: INTRODUCTION</b>	<b>11</b>
1.1 DNA	11
1.1.1 TYPES OF DNA BASES	11
1.2 GENES	11
1.3 TYPES FOR CANCER GENES	12
1.3.1 GENETIC MUTATIONS	12
1.4 CANCER BEGINS	13
1.5 SEQUENCING	13
1.5.1 DNA SEQUENCING	14
1.5.2 RNA SEQUENCING	14
1.5.3 GENOME SEQUENCING	15
1.6 DATA SET	16
1.6.1 NCBI	17
1.6.2 FASTA-FORMAT	18

1.7 DIGITAL SIGNAL PROCESSING	18
1.8 GENOMIC SIGNAL PROCESSING	19
<b>CHAPTER-2: LITERATURE REVIEW</b>	<b>20</b>
<b>CHAPTER-3: METHODOLOGY</b>	<b>22</b>
3.1 FOURIER TRANSFORM	22
3.2 SHORT TIME FOURIER TRANSFORM	23
3.3 SPECTROGRAM	24
3.4 PROPOSED METHOD	25
3.4.1 EXAMPLES OF GENES SEQUENCE	26
3.4.2 NUMERICAL MAPPING	27
3.4.3 APPLY SHORT TIME FOURIER TRANSFORM	28
3.4.4 CONCENTRATION MEASURE	30
3.4.5 THERESHOD	31
3.4.6 GENES CLASSIFICATION	32
<b>CHAPTER-4 : RESULT AND OBSERVATION</b>	<b>33</b>
4.1 SPECTOGRAMS FOR EXPERIMENTAL DATA	33
4.2 CONCENTRATION MEASURE CALCULATED FOR EXPERIMENTAL DATA	39
4.3 PERFORMANCE ANALYSIS	40
4.4 PSEUDO CODE	41
<b>CHAPTER-5 : CONCLUSION</b>	<b>42</b>
<b>REFERENCES</b>	<b>43</b>
<b>PLAGIARISM REPORT</b>	<b>45</b>

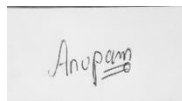
## DECLARATION

We are here by declare that the work report entitled “**Classification of Cancerous and Non-Cancerous DNA Sequences Using Digital Signal Processing Based Method**” in our B.Tech report submitted at **Jaypee University of Information Technology, Wagnaghat, Solan**, is a authentic record of our work that is completed under the guidance of our supervisor **Dr. Sunil Datt Sharma** and we confirm that we have not submitted this work elsewhere for any other degree.



**SAUMYA JAJODIA**

**171073**



**ANUPAM SRIVASTAVA**

**171070**

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.



Signature of the Supervisor

**DR. SUNIL DATT SHARMA**

Date: 18/05/2021

Head of the Department/Project

Coordinator

## **ACKNOWLEDGEMENT**

We have this great opportunity to express our gratitude towards our teacher and our guide Dr. Sunil Datt Sharma (ECE Depart.) for his such an amazing monitoring and guidance so that we can achieve our goal in making and completing our project .The blessing given by him at each point help us throughout this journey .We also want to thank our staff members of JUIT collage for the valuable information provided by them in their respective field Lastly, we thank almighty, our parents and our classmates for their constant encouragement without which this assignment would not have been possible.

## LIST OF ACRONYMS AND ABBREVIATIONS

GSP	Genomic Signal Processing
DNA	Deoxyribonucleic acid
NCBI	National Center for Biotechnology Information
NIH	National Institute of Health
NHGRI	National Human Genome Research Institute
DSP	Digital Signal Processing
DFT	Discrete Fourier Transform
CAD	Computer-aided diagnosis
CEDM	Contrast-Enhanced Digital Mammograms
CGAN	Conditional generative adversarial network
CNN	Convolutional neural network
CRF	Conditional Random Field
CT	Curvelet Transform
DBT	Digital Breast Tomosynthesis
DL	Deep learning
DW	Discrete wavelet
FCN	Fully Convolutional Network
FP	False Positive
IARC	International Agency for Cancer Research
KNN	K-Nearest Neighbor
ML	Machine Learning
ReLU	Rectified Linear Unit
SD-CNN	Shallow-Deep Convolutional Neural Network
SVM	Support Vector Machine
TRW	Tree Re-Weighted
VGG	Visual Geometry Group

WHO	World Health Organization
AI	ARTIFICIAL INTELLIGENCE
NN	NEURAL NETWORK
DWT	DISCERETE WAVELET TRANSFORM
FN	FALSE NEGATIVE
TN	TRUE NEFATIVE
TP	TRUE POSITIVE
CML	Chronic Myelogenous Leukemia
BCR	Breakpoint Cluster Region Protein
SE	Structuring Element

## LIST OF FIGURES

Figure 1.1: Chemical bases present in DNA.[3]	12
Figure 1.2: Genes in humans	13
Figure 1.3: The part of the genetic structure from smallest to largest.[2]	13
Figure 1.3.1: Mutation occurring	14
Figure: 1.6.2.1: Fasta Format[4]	18
Figure 3.1: Fourier Transform	22
Figure 3.2 : Time freq-analysis	24
Figure 3.3: Spectrogram [5]	25
Figure 3.4.1: Proposed hypothesis	26
Figure 3.4.2.1: Numerical conversion for cancerous genes AAQ08976	30
Figure 3.4.2.2: Numerical conversion for non-cancerous genes AF003934	30
Figure 3.4.3.1: Spectrogram image for cancerous gene AAQ08976	31
Figure 3.4.3.2: Spectrogram image for non-cancerous gene AF003934	32



## **LIST OF TABLES**

Table 4.1: Spectrogram images for cancerous genes non-cancerous genes	37
Table 4.2: Concentration measure for non-cancerous genes	43
Table 4.3: Concentration measure for non-cancerous genes	43

## ABSTRACT

The leading cause is cancer now a days. Day to day, cases are increasing and there is still a lack of faster and computational based approach to detect the disease. Earlier, a large proportion of people whose age is above 50, mostly got infected by this disease. But now the situation gets even more worst , it's common in people of every age group. The Worldwide Organization for Exploration on Malignant growth appraises that universally, 1 of every 5 individuals foster malignancy during their lifetime, and 1 out of 8 men and 1 out of 11 ladies bite the dust from the illness. These new gauges recommend that in excess of 50 million individuals are living inside five years of a past malignant growth. As indicated by report, in 2018 there were 17.0 million new malignancy cases and 9.5 million disease passing around the world. By 2040, the worldwide weight is relied upon to develop to 27.5 million new malignancy cases and 16.3 million disease passing just because of the development and maturing of the populace. The extended occurrence of patients with malignant growth in It is found that main cause of cancer is mutation. Mutation is basically a change in DNA sequences. Mutation starts affecting the healthy cells. It leads to the rapid growth of the cells and it is uncontrollable. Gene mutations happens in two cases, one we born with that is we inherited this from our parents but the risk of cancer is low in this cases. Second when it happens after the birth, different but the risk of cancer is low in this cases. Second when it happens after the birth, different are responsible for this. GSP comes under the Digital Signal Processing. Here we applied short time Fourier transform which is a part of DSP for classification. Firstly, we collect the data from NCBI. Data is in the form of genes sequence, as data is in the character format so we use numerical mapping technique to convert data into numerical forms. We use EIIP numerical mapping technique as it gives the best result in comparison to all mapping technique. Mutation is basically a change in DNA sequences. Mutation starts affecting the healthy cells. After this we applied the short time fourier transform and thus obtained the spectrum of each genes sequences and find the threshold value that is the same for each genes sequence. So, finally we calculate the concentration measure for each genes and if the concentration measure greater than threshold value than it is a non-healthy gene that is a cancerous genes and vice-versa.

# **CHAPTER 1**

## **INTRODUCTION**

### **1.1 DNA**

The molecule inside cells that information liable Also called deoxyribo nucleic destructive. DNA particles license this information to be passed beginning with one age then onto the following. DNA is contained a twofold deserted helix held together by delicate hydrogen associations between purine-pyrimidine nucleotide base sets. Also called deoxyribose nucleic destructive. Most DNA is found inside the center of a cell, where it outlines the chromosomes. Chromosomes have proteins considered histones that difficult situation to DNA. DNA has two strands that breeze into the condition of a turning ladder called a helix.[1]

#### **1.1.1 TYPES OF DNA BASES**

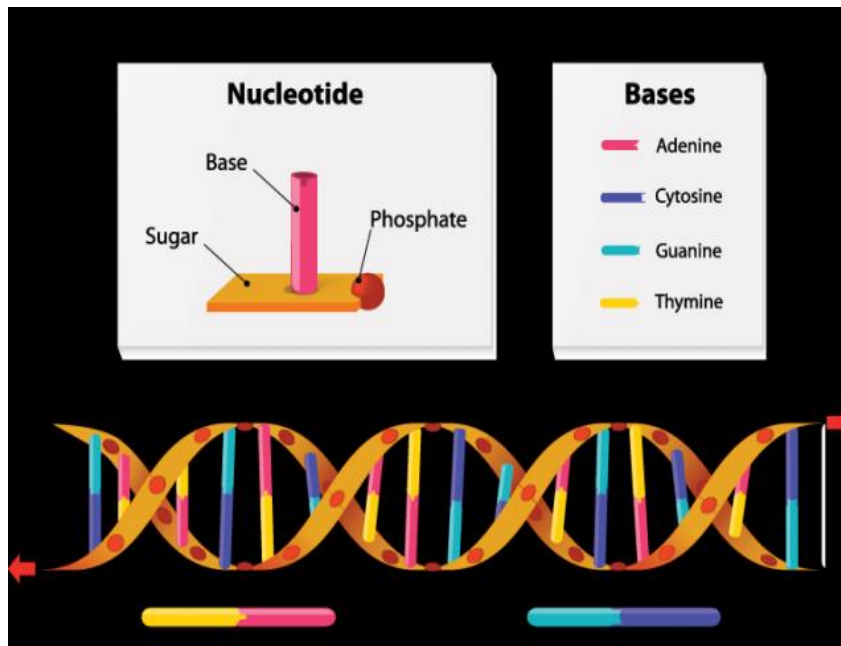
In DNA, all information is stored in code format which are formed from four type of base of chemical are-

A- Adenine

G-Guanine

C-Cytosine

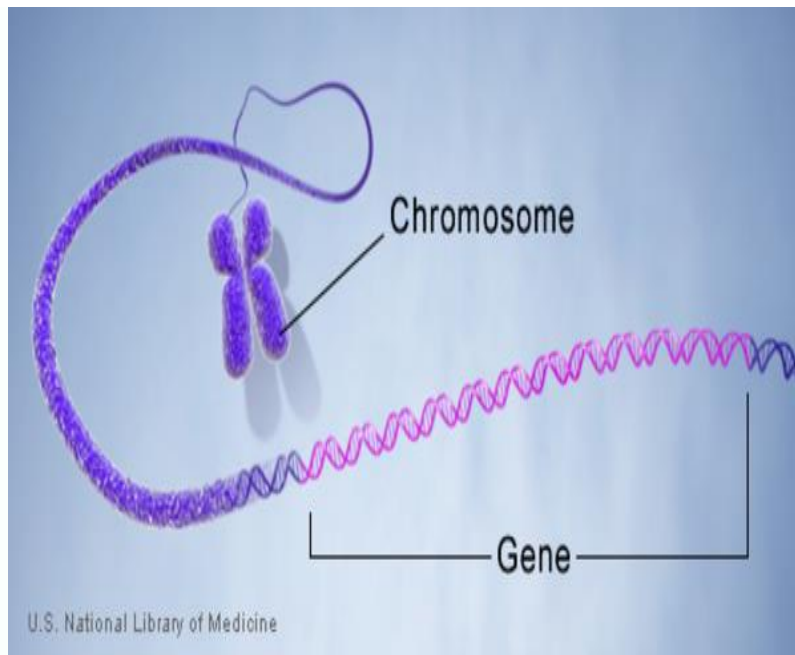
T-Thymine



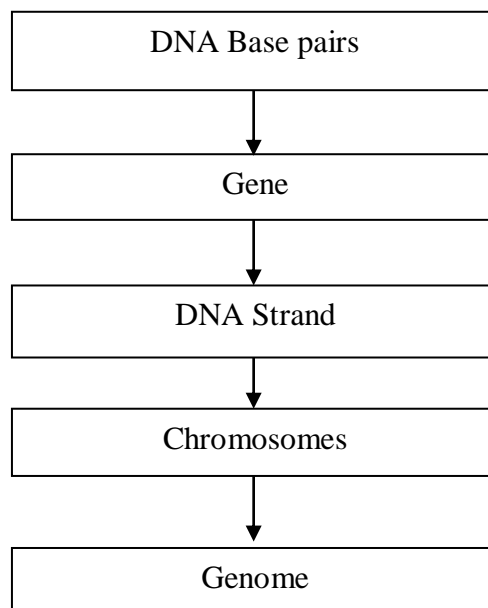
**Figure 1.1:** Chemical bases present in DNA [3]

## 1.2 GENES

Gene, unit of innate data that possesses a fixed position (locus) on a chromosome. Some type of genes are neuclocite and other types of genes which are all available in the genes section .Genes can procure changes in their arrangement, prompting various variations, known as alleles, in the populace. These alleles encode marginally various adaptations of the protine which are the genes.



**Figure 1.2:** Genes in humans



**Figure 1.3:** The part of the genetic structure from smallest to largest.[2]

In figure 1.3, the genetic structure has been described from DNA base pairs to genome. DNA base pairs of nucleotides form the double helix of DNA. The genes are the stretched DNA

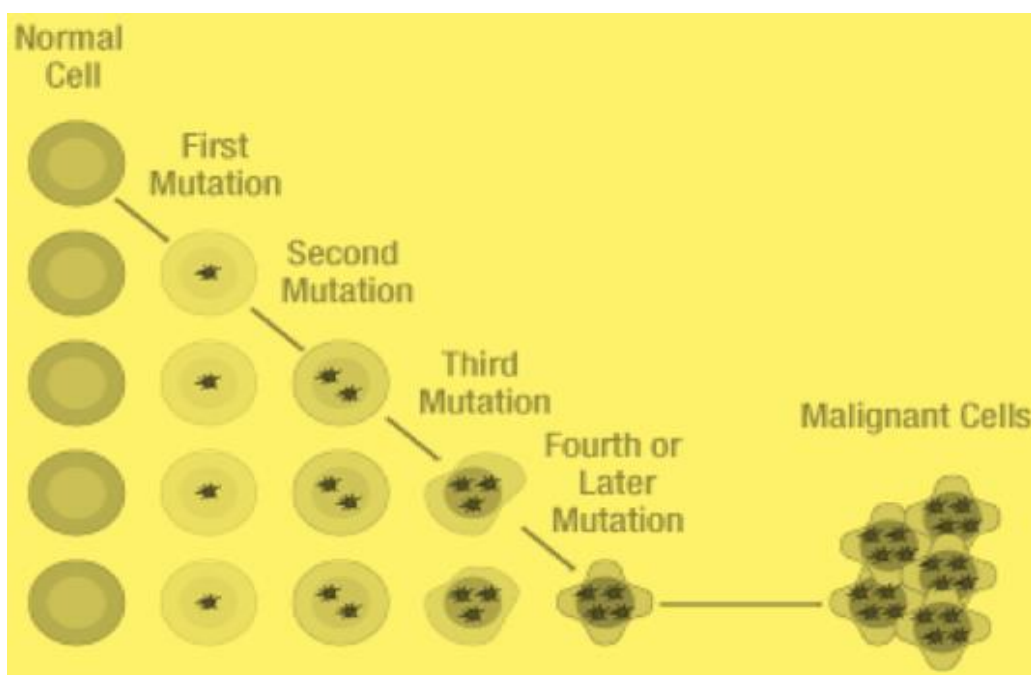
base pairs that contain the information about the functionalities in the living organism. The DNA strands consists the information for the reproduction of the cells. Chromosomes are the organized structure of the DNA within the cell and it contains many genes and Genome contains the hereditary information [2].

### 1.3 REASON FOR CANCER GENES

Nowadays, Cancer is one of the common disease , and we can find this disease in one member of the whole large family. Day to day, cases are increasing and there is still a lack of faster and computational based approach to detect the disease. Earlier, a large proportion of people whose age is above 50, mostly got infected by this disease. But now the situation gets even more worst , it's common in people of every age group.

#### 1.3.1 GENETIC MUTATION

Genetic Mutation is one of the main cause of cancer. Mutation is basically a change in DNA sequences. Mutation starts affecting the healthy cells. It leads to the rapid growth of the cells and it is uncontrollable. Gene mutations happens in two cases, one we born with that is we inherited this from our parents but the risk of cancer is low in this cases. It leads to the rapid growth of the cells and it is uncontrollable. Gene mutations .Second when it happens after the birth, different are responsible for this.[2]



**Figure 1.3.1 : Mutation occurring**

## **1.4 CANCER BEGINS**

As, cells are the essential tiny unit that consolidates and offer structure to the human body. Cells work in the structure as they develop after that they gap to newest cells as the body needs this. Generally, cells bite the dust when they get more established and vanished. Now, many cells have their spot and this path goes on. Malignancy starts when hereditary changes meddle with this deliberate cycle. So cells began becoming unmanageable. In this manner they assemble at one spot and structure is called a tumor. A tumor can be carcinogenic or kind and it tends to be threatening moreover. A harmful tumor is dangerous, implies that it can develop and extend to the next body parts. A favorable tumor implies the tumor can develop yet won't spread on the other piece of the body. Some types of cancer do not form a tumor. [3]

## **1.5 SEQUENCING**

Sequencing intends to decide the essential construction of an un-branched bio-polymer. Sequencing brings about an emblematic straight portrayal known as a grouping which compactly sums up a significant part of the nuclear level design of the sequenced particle. There are 3 different types of sequencing they are : [7]

1. DNA Sequencing
2. RNA Sequencing
3. Genomic Sequencing

### **1.5.1 DNA Sequencing**

DNA sequencing, procedure used to decide the nucleotide succession of DNA (deoxyribonucleic corrosive). The nucleotide arrangement is the most crucial degree of information on a quality or genome. The outline contains the guidelines for building a life form, and no comprehension of hereditary capacity or development could be finished without getting this data. Looking at sound and transformed DNA arrangements can analyze various sicknesses including different malignant growths describe neutralizer repertoire, and can be utilized to direct persistent treatment. Having a speedy method to grouping DNA considers quicker and more individualized clinical consideration to be managed, and for additional living beings to be distinguished and classified. [8]

### **1.5.2 RNA sequencing**

RNA-Seq is an as of late created way to deal with transcriptome profiling that utilizes profound sequencing innovations. Studies utilizing this technique have effectively modified our perspective on the degree and intricacy of eukaryotic transcriptomes. RNA-Seq additionally gives an undeniably more exact estimation of levels of records and their isoforms than different techniques. This article depicts the RNA-Seq approach, the difficulties related with its application, and the advances made so far in describing a few eukaryote transcriptomes.

### **1.5.3 Genome sequencing**

Whole genome sequencing provides the most comprehensive collection of an individual's genetic variation. With the falling costs of sequencing technology, we envision paradigm shift from microarray-based genotyping studies to whole genome sequencing. Entire genome sequencing (WGS), in any case called Whole genome sequencing has commonly been used as an investigation instrument, yet was being familiar with focuses the aggregate, or practically the total, of the DNA progression of an animal's genome at a single time. This includes sequencing the total of a natural substance's chromosomal DNA similarly as DNA. Whole genome sequencing has commonly been used as an investigation instrument, yet was being familiar with focuses. Whole genome sequencing has commonly been used as an investigation instrument, yet was being familiar with focuses in 2014. In the destiny of altered drug, whole genome gathering data may be a huge gadget to oversee healing intercession. [9]

The gadget of value sequencing at SNP level is moreover used to pinpoint utilitarian varieties from connection analyzes and improve the data available to experts enlivened by formative science, and hereafter may set up the system for expecting disease weakness and prescription response.

## **1.6 DATA SET**

As the fact now everything is online and easy to access, so this helps in the advancement of many technologies. For example here we take data in the form of genes sequences directly



from NCBI site, now a days no one don't need t to go any hospitals and collect data like old times.

### 1.6. 1. NCBI

NCBI is currently a main hotspot for public biomedical data sets, programming instruments for investigating sub-atomic and genomic information, and exploration in computational science. Today NCBI makes and keeps up more than 40 incorporated information bases for the clinical and mainstream researchers just as the overall population. The information bases at the NCBI/DDBJ/EMBL will contain blunders as the information comes from different sources and the majority of the data sets are just insignificantly curated. However, that remains constant for all huge information bases without manual curation (and surprisingly those are not immaculate).We have taken the nucleotide sequence from the NCBI website and converted into Fasta format then added into our data set for the numerical conversion.[6]

### 1.6.2. Fasta-Format

As the fact now everything is online and easy to access, so this helps in the advancement of many technologies. For example here we take data in the form of genes sequences directly from NCBI site, now a days no one don't need t to go any hospitals and collect data like old times. currently a main hotspot for public biomedical data sets, programming instruments for investigating sub-atomic and genomic information, and exploration in computational science. Today NCBI makes and keeps up more than 40 incorporated information bases for the clinical and mainstream researchers just as the overall population. The information bases at the NCBI/DDBJ/EMBL will contain blunders as the information comes from different sources and the majority of the data sets are just insignificantly curated. However, that remains constant for all huge information bases without manual curation (and surprisingly those are not immaculate).We have taken the nucleotide sequence from the NCBI website and converted into Fasta format then added into our data set for the numerical conversion.

Display Settings:  FASTA Send to:

**insulin precursor [Aplysia californica]**

NCBI Reference Sequence: NP\_001191615.1

[GenPept](#) [Identical Proteins](#) [Graphics](#)

```
>gi|325296757|ref|NP_001191615.1| insulin precursor [Aplysia californica]
MSKFLQSHSANACLLTLLTLASNLDISLANFEHSCNGYMRPHRGLCGEDLHVIISNLCSSLGGNRRF
LAKYMKRDTENVNDKLRGILLNKKEAFSYLTKREASGSITCECCFNQCRIFELAQYCRLPDHHFFSRISR
TGRSNSGHAQLEDNFS
```

**Figure: 1.6.2.1:** Fasta Format [4]

## **1.7 DIGITAL SIGNAL PROCESSING**

Computerized Signal Handling is the way toward addressing signals in a discrete numerical succession of numbers and investigating, adjusting, and separating the data contained in the sign via doing algorithmic tasks and preparing on the signal. To represent this idea, the chart underneath shows how a DSP is utilized in a MP3 sound player. During the account stage, simple sound is contribution through a collector or other source.

The DSP is is the digital signal processing and the very important to all of the detection To represent this idea, the chart underneath shows how a DSP is utilized in a MP3 sound player. all the more proficiently starting with one spot then onto the next (for example remotely coordinating can send discourse and video by means of phone lines). Signs may likewise be upgraded or controlled to improve their quality or give data that isn't detected by people (for example reverberation dropping for PDAs or PC improved clinical pictures[4]. Apart from the so many applications, digital signal processing has also been used in to analyze the genomics data, basically the genomics data is the part of the Big data analysis. The use of digital signal processing tools in the genomics field is also known as genomics signal processing and it is explained in next sub section.

## **1.8 GENOMIC SIGNAL PROCESSING**

Genomic Signal Preparing (GSP) insinuates the usage of cutting edge sign dealing with (DSP) instruments for analyzing genomic data, for instance, DNA groupings. An expected utilization of GSP that has not been totally examined is the estimation of the distance two or three groupings. In this work we present GAFD, a novel GSP game plan free distance computation methodology. We familiarize a DNA progression with signal arranging limit subject to crafted by doublet regards, which fabricates the amount of possible bounty completion regards for the created signal. Besides, we examine the use of three DSP distance estimations as descriptors for masterminding DNA signal areas. Our results show the common sense of using GAFD for figuring course of action distances. DWT (Discrete Wavelet Transform) comes under GSP. GSP has a wide application. Now a days, DWT has

great use, has wide application. We can use it to find various parameters and then these parameter can be taken as input to make any model. Model like using ML (Machine Learning) , AI (Artificial Intelligence) , NN (Neural Network) etc. As the fact now everything is online and easy to access, so this helps in the advancement of many technologies. For example here they take data in the form of genes sequences directly from NCBI site, now a days no one need not to go any hospitals and collect data like old times. Discrete Fourier Transform, after that obtain PSD and apply entropy estimation. Lastly spectrum plot using Rayleigh distribution. In this way we can identify between healthy and non-healthy genes. Now if it is a non-healthy genes, again we need to classify what type of cancer genes it is. So for classification, they applied entropy estimator after that they obtained minimum entropy, then compute DFT, obtain PSD, estimate mutual information, on the basis of that classification happens.[5]

## **CHAPTER 2**

### **LITERATURE REVIEW**

Liu Dongwei , Jia Runping , Wang Caifeng , Arunkumar N , K. Narasimhan , M. Udayakumar , V. Elamaranr [1] suggested that the mutations are the main cause of cancer. Mutations are of two types, one is acquired mutations and other is germline mutations . Acquired mutation cause due to not having proper lifestyle and germline mutations cause due to genetic defects. It transfer from parents to children, but its chances is only 5% to 20%. So, identification of mutation is the most important step in diagnosis of cancer. Here research is done on same variant of cell , because it is easy to done on same variants of cells rather than different. Here, Genomic Signal Processing is used which is a part of DSP . GSP includes Discrete Wavelet Transform and firstly DNA bases are converted into numerical form and datasets of DNA bases are taken from NCBI (National Center for Biotechnology Information). After applying numerical mapping , it is passed through DWT and DWT is applied with the help of Haar Wavelet and results we obtained from here in the form of statistical features. Now these parameters are passed through Machine Learning Algorithm, parameters are taken as input. After this whole process, genes sequence are classified as cancerous and non-cancerous genes with the help of SVM (Support Vector Machine) . The result gives 100% accuracy on classification. S.BARMAN (MANDAL) , M.ROY , S.BISWAS, S.SAHA. [2] their main focus is on the prediction of cancer cells using Digital Signal Processing (DSP). Nowadays DSP has gained a lot of popularity and it is highly used in classification techniques. Results obtained from DSP technique gives great accuracy and less complexity. In this paper it is talked about cancerous genes formation that occur due to genetic abnormality. But this reason of cancer cells formation is very less. Here instead of genes, they take the small part of it that is Chromosomes. In this article, they present a Discrete Fourier Transform approach which is a part of DSP. As the fact now everything is online and easy to access, so this helps in the advancement of many technologies. For example here they take data in the form of genes sequences directly from NCBI site, now a days no one need not to go any hospitals and collect data like old times. After, applying this method, we will receive spectrum and again if there is a positive peak in the spectrum, we

will classify as healthy genes whereas if there is a negative peak in the spectrum then it will be non-healthy genes. Here, three types of cancerous genes are taken, they are Breast, Prostate and Colon. For identifier, firstly they divide the genes into 20 numerical indicator sequences, then compute Discrete Fourier Transform, after that obtain PSD and apply entropy estimation. Lastly spectrum plot using Rayleigh distribution. In this way we can identify between healthy and non-healthy genes. Now if it is a non-healthy genes, again we need to classify what type of cancer genes it is. So for classification, they applied entropy estimator after that they obtained minimum entropy, then compute DFT, obtain PSD, estimate mutual information, on the basis of that classification happens. Safaa M. Naieem , Mai S. Mabrouk , Mohamed A. Eldosoky, Ahmed Y. Sayaed [3] In this paper basically it is given that there are two ways to classify and distinguish healthy and non-healthy genes. One is more like a manual method where any skilled person requires to examine and gives the result based on blood, urine and stools samples. But there is a great chances of error in this, Second is totally computational method. Examination is done on 51 healthy genes and 51 non-healthy genes that is they are different types of cancer genes. The size of each DNA sequences are around 400 nucleotides. Three approaches are used here, first is unsupervised learning method, and the second and the third are supervised learnings method . For numerical mapping EIIP method is used here and after that feature extraction is done. Accuracy were for 67.5% , 87.5% and 95%. So, the most accurate result comes when trainable cascade-forward backpropagation network is used. Joyshri Das Soma Barman [4] the field sign handling, another territory of examination has been presented specifically genomic signal preparing (GSP). GSP fundamentally measures qualities, proteins and DNA arrangements utilizing different sign handling procedures to extricate the data covered up in it. As some hereditary anomalies transform into malignancy infections, legitimate arrangement and examination of qualities and proteins may prompt another skyline in disease in GSP . In genomic signal preparing, precise finding or classification of unhealthy quality is an incredible test to the specialists. Discrete Fourier Transform, after that obtain PSD and apply entropy estimation. Lastly spectrum plot using Rayleigh distribution. In this way we can identify between healthy and non-healthy genes. Now if it is a non-healthy genes, again we need to classify what type of cancer genes it is. Consequently, in the current paper, the significant employment of quality identified and classified is endeavoured. Now an answer for these issue, factual strategies like entropy assessment and common data estimation is received alongside DSP procedure. Rayleigh appropriation of assessed entropy of quality is treated as classifier of sound and dangerous for human beings. When the disease qualities are identified, common

data assessor dependent on their base entropy is utilized as classifier to distinguish various kinds of malignancy qualities. DFT based approach is used to classify genes with the help of spectral characteristics

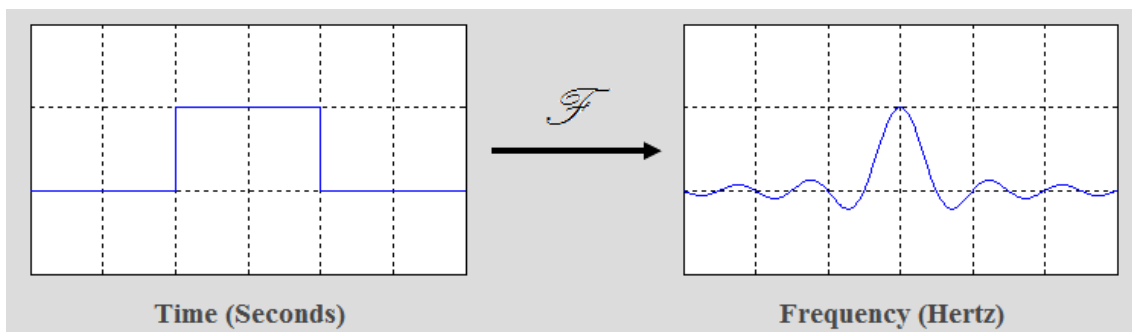
## CHAPTER 3

### MATERIAL METHOD

In this chapter the digital signal processing tools and proposed hypothesis have been presented.

#### 3.1 FOURIER TRANSFORM

It is a wave form that is recomposed of of a sinousdle wave for to depict the wave for for a certain reason of for the mathematical problem . It characterizes an especially valuable class of time-recurrence circulations which determine complex abundancy versus time and recurrence for any sign. We are fundamentally worried here with tuning the STFT boundaries for the accompanying application. It characterizes an especially valuable class of time-recurrence dispersions which determine complex versus time and recurrence for any sign.[1 1]



**Figure 3.1:** Fourier Transform

### 3.2. SHORT TIME FOURIER TRANSFORM

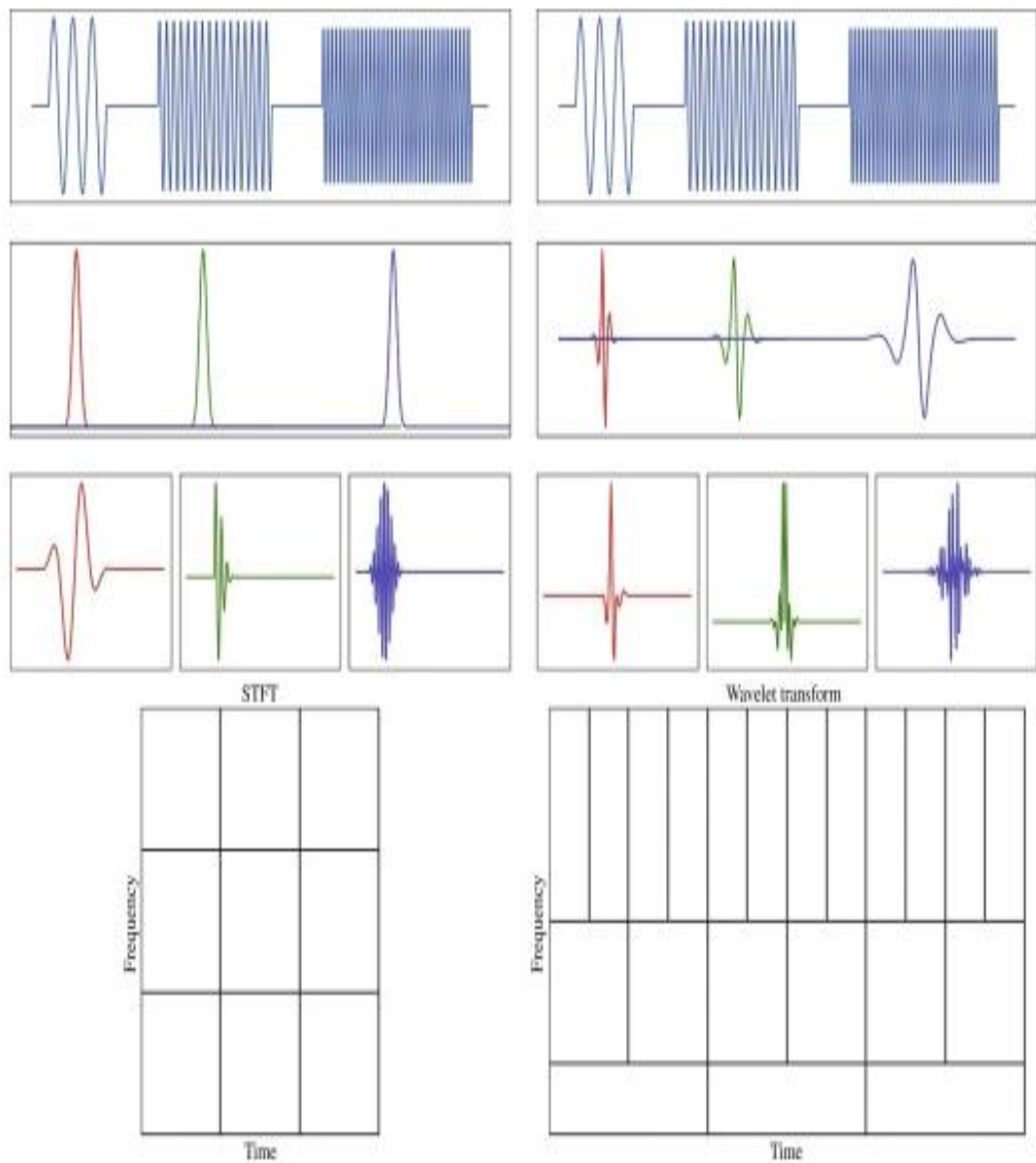
The brief timeframe Fourier change (STFT) permits us to perform time-recurrence investigation. It is utilized to produce portrayals that catch both the neighborhood time and recurrence content in the sign. Like the Fourier change, the STFT actually depends on fixed premise capacities; in any case, it utilizes fixed-size time-moved window capacities  $w(n)$  to get a change of the flag and can be communicated a where  $m$  is the measure of shift. [13]

$$X(k, m) = \sum_{n=0}^{N-1} x(n + m) w(n) W_N^{nk} \quad (1)$$

Here  $k, m = 0, 1, 2, \dots, N-1$

Not with standing, the STFT has better transient and recurrence confinement properties contrasted and the Fourier change. Be that as it may, since the result of worldly and recurrence goal is steady (due to the old style Heisenberg's vulnerability standard), the produced highlights can't accomplish quick limitation of both time and recurrence. Furthermore, because of utilizing a fixed window length and fixed premise works, the STFT actually can't catch occasions with various spans or when the sign contains quick (sharp) occasions.

The wavelet change is among the broadly utilized procedures for extricating highlights from biomedical signs. The wavelet change attempts to relieve the impediments of the STFT and make a superior showing. It begins by characterizing unique premise capacities called "mother wavelets." Mother wavelets are not confined to a solitary group of capacities (e.g., occasional capacities as the case in the FT). Furthermore, the fundamental capacities have both transient and recurrence parts. This permits us to produce a progression of variable-sized wavelet works; each has a portion of the time-recurrence range. The STFT partitions the time-recurrence space into similarly measured network, while for the wavelet change, we see a coarse-to-fine (i.e., nonuniform) portrayal of the sign.[14]



**Figure 3.2:** Time frequency-analysis

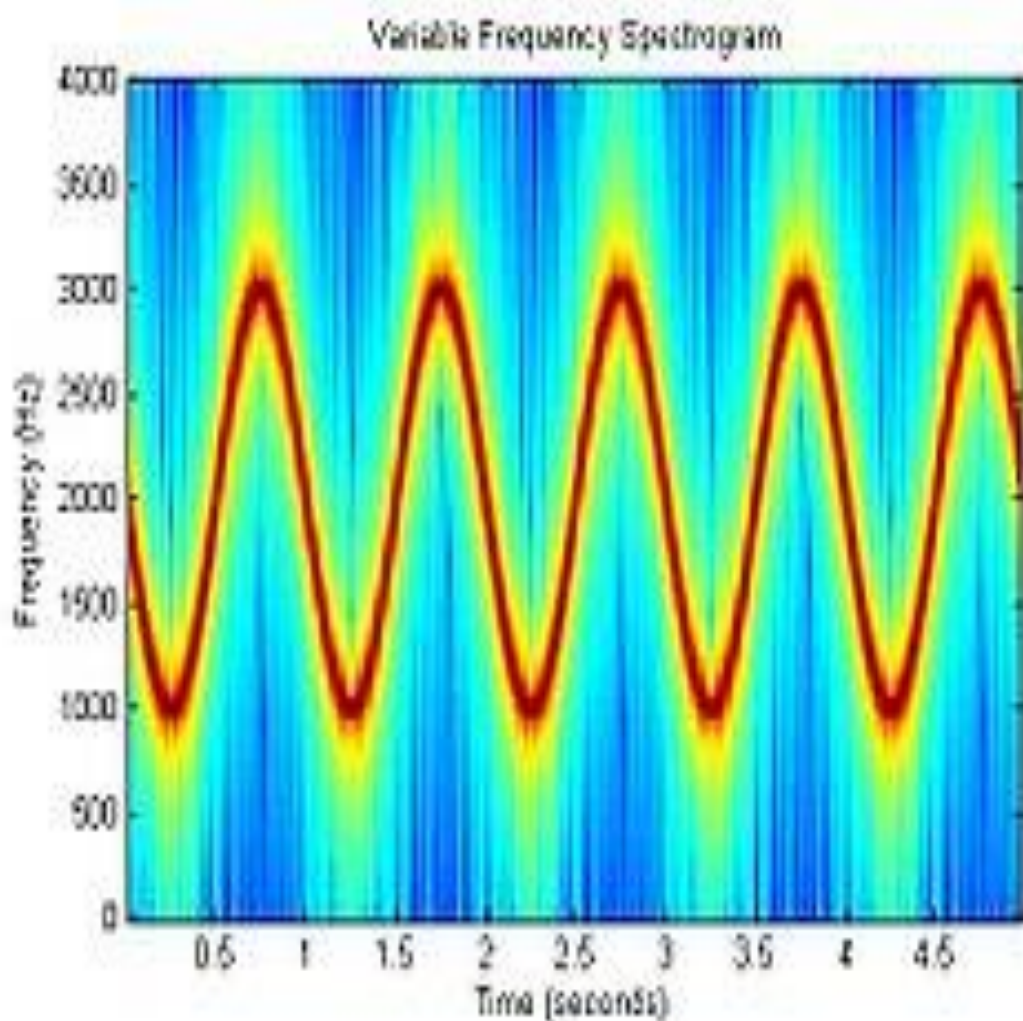
### 3.3 SPECTROGRAM

Spectrogram utilizing a Brief time frame Fourier Change (STFT).  $S = \text{spectrogram}(X)$  returns the spectrogram of the sign determined by vector  $X$  in the lattice  $S$ . Of course,  $X$  is partitioned into eight portions with half cover, each portion is windowed with a Hamming window. The number of recurrence focuses used to figure the discrete Fourier changes is equivalent to the limit of 256 or the following force of two more noteworthy than the length



of each section of X. On the off chance that X can't be isolated precisely into eight fragments, X will be shortened likewise. [15]

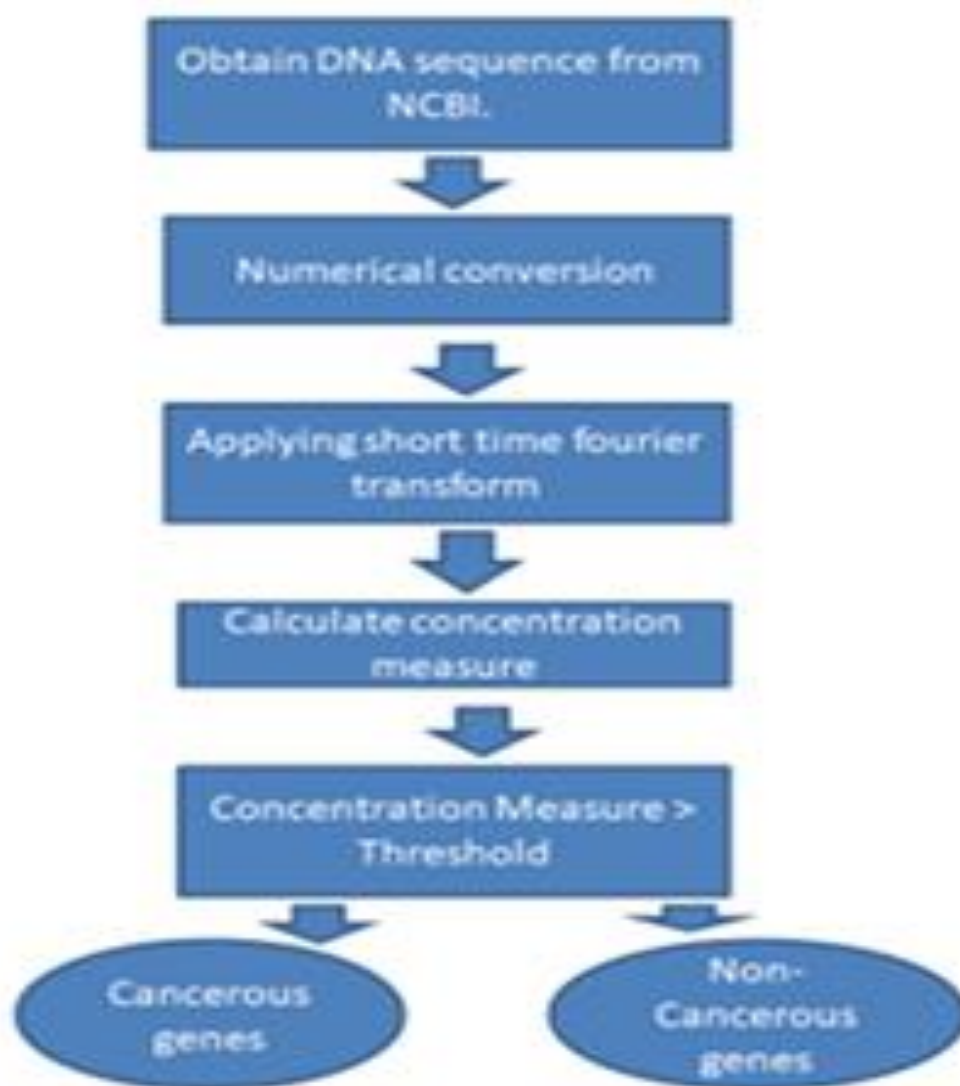
$S = \text{spectrogram}(X, \text{WINDOW})$  when WINDOW is a vector, separates X into portions of length equivalent to the length of WINDOW, and afterward windows each portion with the vector indicated in WINDOW. In the event that WINDOW is a whole number, X is partitioned into sections of length equivalent to that whole number worth, and a Hamming window of equivalent length is utilized. On the off chance that WINDOW isn't indicated, the default is utilized.



**Figure 3.3:** Spectrogram [5]

### 3.4 PROPOSED METHOD

In this section the steps for the proposed hypothesis has been has been described in details with the help of the example sequence.



**Figure3.4.1:** Proposed hypothesis

#### 3.4.1 Example- sequence of the genes for the experiments

**Healthy gene:**

1 AF003934.1

FASTA format we got from NCBI

AGCGTTTAAACTTAAGCTTGGAGTTATTTCCACCATGCCCCGGGCAAGAACTCAGGACGCTGA  
ATGGCTCT  
CAGATGCTCCTGGTGTGCTGGTGTCTCGTGGCTGCCGCATGGGGGCGCCCTGTCTCTGGC  
CGAGGCGA  
GCCGCGCAAGTTTCCCGGGACCCTCAGAGTTGCACTCCGAAGACTCCAGATTCCGAGAGTTG  
CGGAAACG  
CTACGAGGACCTGCTAACCAGGCTGCGGGCCAACCAGAGCTGGGAAGATTGGAACACCGACC  
TCGTCCCG  
GCCCTGCAGTCCGGATACTCACGCCAGAAGTGC GGCTGGGATCCGGCGGCCACCTGCACCT  
GCGTATCT  
CTCGGGCCGCCCTTCCCTGAGGGGCTCCCCGAGGCCTCCCGCCTTACCGGGCTCTGTTCCGG  
CTGTCCCC  
GACGGCGTCAAGGTCGTGGGACGTGACACGACCGCTGCGGCGTCAGCTCAGCCTTGCAAGAC  
CCCAGGCG  
CCCGCGCTGCACCTGCGACTGTGCGCCGCCGCGTGCAGTCGGACCAACTGCTGGCAGAATC  
TTCGTCCG  
CACGGCCCCAGCTGGAGTTGCACTTGC GGCCGCAAGCCGCCAGGGGGCGCCGCAGAGCGCGT  
GCGCGCAA  
CGGGGACCACTGTCCGCTCGGGCCCCGGGCGTTGCTGCCGTCTGCACACGGTCCGCGCGTCGC  
TGGAAGAC  
CTGGGCTGGGCCGATTGGGTGCTGTGCGCCACGGGAGGTGCAAGTGACCATGTGCATCGGCGC  
GTGCCCGA  
GCCAGTTCCGGGCGGCAAACATGCACGCGCAGATCAAGACGAGCCTGCACCGCCTGAAGCCC  
GACACGGT  
GCCAGCGCCCTGCTGCGTGCCCGCCAGCTACAATCCCATGGTGCTCATTCAAAGACCGACA  
CCGGGGTG  
TCGCTCCAGACCTATGATGACTTGTAGCCAAAGACTGCCACTGCATATGAACTAGTACTAA  
GCCGAATT  
CTGCAGATATCC

Same like we have taken all the other healthy genes fasta format

2 AF335477.1

3 AJ459782.1

4 AJ459784.1

5 AJ512346.1

6 BT006816.1

7 NM\_017436.4

8 NM\_018713.2

**Non healthy gene**

1 NM\_012278.2

FASTA format we got from NCBI

```
CCATTCGGCAGCCAGACTCCTTGAAATACCCTTTCAGTAATCATTCAACCAACGCTTCCATG
TCTCTACT
CTGTTCGTAACAAAGGCTGTGGGCAGCACTTTGACCCTAATACCAACCTTCCTGATTCCTGTT
GCCATCAC
CCTGGGGTCCCAATCTTCCATGATGCACTTAAGGGTTGGTCCTGCTGCCGAAAGCGAACTGT
AGATTTCT
CTGAGTTCTTAAACATCAAGGGCTGTACTATGGGACCACACTGTGCTGAGAAGCTTCCTGAG
GCCCCTCA
ACCTGAAGGCCCTGCTACAAGCAGTTCACCTCAGGAGCAAAAACCTCTGAATGTGATTCCAA
AGTCAGCA
GAGACCTTGCGCCGGGAGAGGCCCAAGTCAGAGTTGCCTCTGAAGCTGCTGCCGCTAAATAT
ATCCCAAG
CCCTGGAAATGGCATTGGAACAGAAGGAATTAGACCAGGAACCTGGGGCAGGACTTGACAGT
CTGATCCG
GACTGGTTCAGCTGCCAGAACCCAGGATGTGATGCTGTTTACCAAGGCCCTGAGAGTGATG
CTACTCCA
TGTACCTACCACCCAGGAGCACCCCGATTCCATGAGGGGATGAAGTCTTGGAGCTGTTGTGG
CATCCAGA
CCCTGGATTTTGGGGCATTCTTGGCACAACCAGGGTGCAGAGTCGGTAGACATGACTGGGGG
AAGCAGCT
CCCAGCATCTTGCCGCCATGATTGGCACCAGACAGATTCCTTAGTAGTGGTACTGTATATG
GCCAGATT
CCACTTCCTGCGTTTAACTGGGTGAAGGCCAGTCAAACCTGAGCTTCATGTCCACATTGTCTT
TGATGGTA
ACCGTGTGTTCCAAGCACAGATGAAGCTCTGGGGGGTTCATAAACGTGGAGCAGAGCTCTGTC
TTCTTGAT
GCCATCTCGGGTTGAAATCTCCCTGGTCAAGGCTGACCCAGGATCCTGGGCCAGCTGGAGC
ACCCTGAT
GCACTAGCTAAGAAGGCTAGGGCAGGGGTTGTGTTAGAGATGGATGAGGAAGAATCTGACGA
TTCAGATG
ATGATCTGAGCTGGACAGAGGAGGAGGAAGAGGAGGAAGCAATGGGGGAATAGTGACACCAG
ACAGTTGA
TGTCTAGATAGGACCTCAATGATTCCCTTAGAATCTTAGATACCAGGATATTGTTGGCCATG
TGGCATCA
TTGAGCAGCAGGAGGCTGAAGGAGGGGAGAACAAAATTGTCCAACCATGCTGTTTTTTTCC
CTTAAATA
AATCTTGTATTCTTCAGTTTCAAAAAAAAAAA
```

2 NM\_024533.4

3 NM\_030754.4

4 NM\_032044.3

5 NM\_138937.2

6 NM\_001127380.2

7 NM\_001159352.1

### 3.4.2 Numerical Mapping

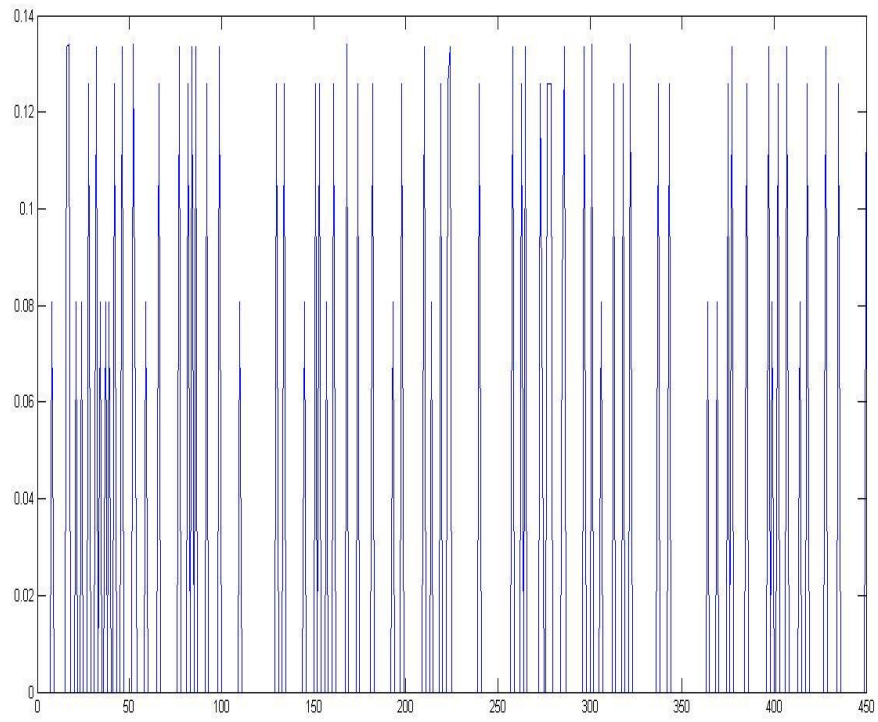
As data in the form genes sequences that is in the character format initially, so it's important to convert datas into numerical form. Here, we applied EIIP (Electron-ion interaction pseudo potentials) numerical mapping technique as we applying this we will get the accurate results. DSP to apply method examining the dna arrangement for coding protiene area ID, character upsides of DNA succession must appropriately planned mathematical qualities. In the writing there exist numerous strategies like parallel pointer successions (BIS), complex marker arrangements, genuine number planning, tetrahedron planning, planning, ,electronion collaboration EIIP marker grouping. **EIIP** Values for different dna bases:

$$A = 0.1260$$

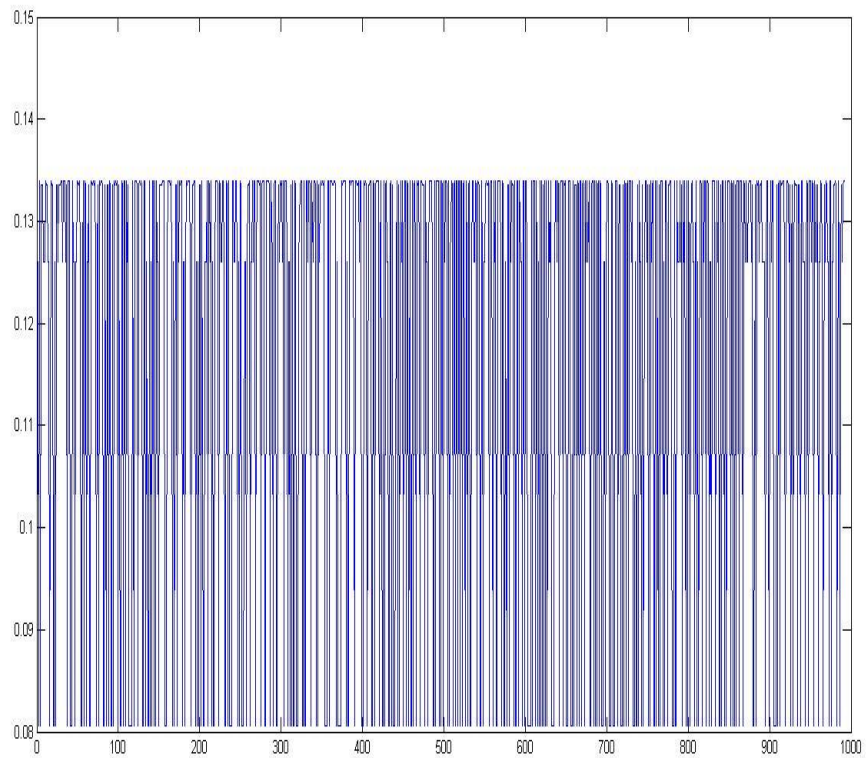
$$T = 0.1335$$

$$C = 0.0806$$

$$G = 0.1340.$$



**Figure 3.4.2.1:** Numerical conversion for cancerous genes AAQ08976

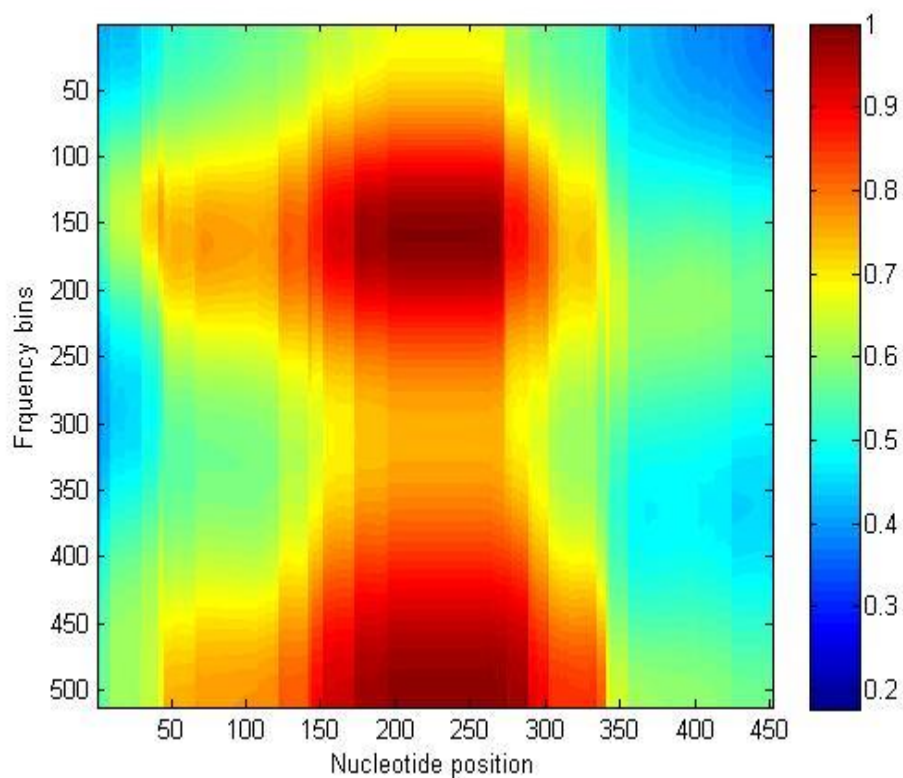


**Figure 3.4.2.2:** Numerical conversion for non-cancerous genes AF003934

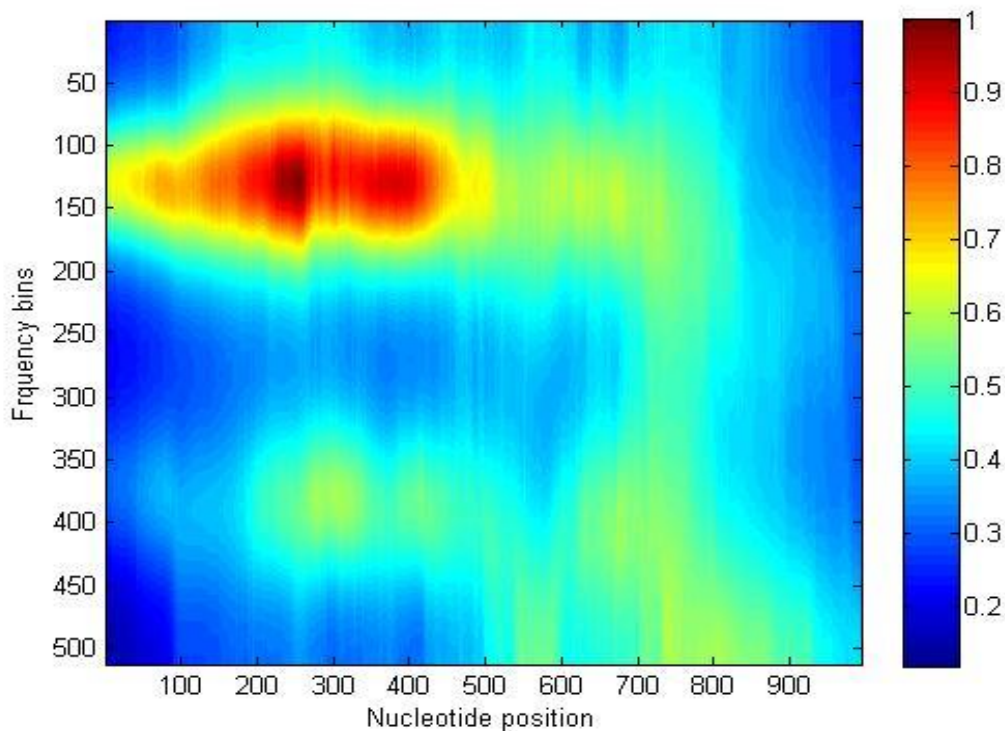
### 3.4.3 APPLY THE SHORT TIME FOURIER TRANSFORM

Spectrogram utilizing a Brief time frame Fourier Change (STFT).  $S = \text{spectrogram}(X)$  returns the spectrogram of the sign determined by vector  $X$  in the lattice  $S$ . Of course,  $X$  is partitioned into eight portions with half cover, each portion is windowed with a Hamming window. The number of recurrence focuses used to figure the discrete Fourier changes is equivalent to the limit of 256 or the following force of two more noteworthy than the length of each section of  $X$ . On the off chance that  $X$  can't be isolated precisely into eight fragments,  $X$  will be shortened likewise.[20]

Spectrogram we obtained after applying Fourier transform:



**Figure 3.4.3.1:** Spectrogram image for cancerous gene AAQ08976



**Figure 3.4.3.2:** Spectrogram image for non-cancerous gene AF003934

### 3.4.4 Concentration measure

Concentration measure = L4 norm of spectrogram / L2 norm of spectrogram

Concentration measure also known as energy of the spectrogram.

### 3.4.5 THRESHOLD

In this hypothesis, Fixed threshold has been selected manually with the help of the values of the concentration measure.

Here, by experiment we got the value of Threshold= $1.3462 \times 10^{-6}$ .

### 3.4.6 Genes classification

Classification of the genes have been done using threshold. To decide whether given gene is healthy gene or non-healthy gene. Concentration measure values have been compared with threshold. If **concentration measure value > Threshold**, then it is healthy/ non-cancerous gene. Else, it is non-healthy gene



# CHAPTER 4

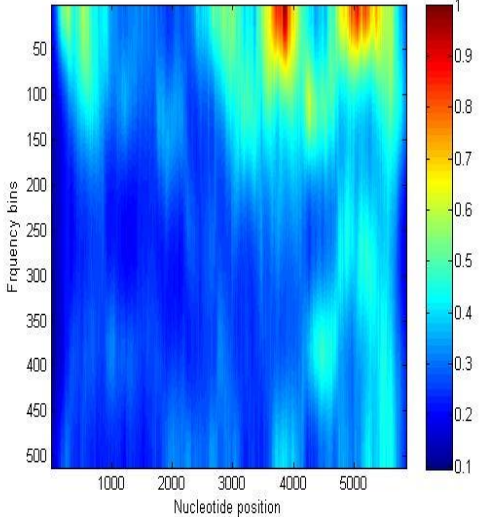
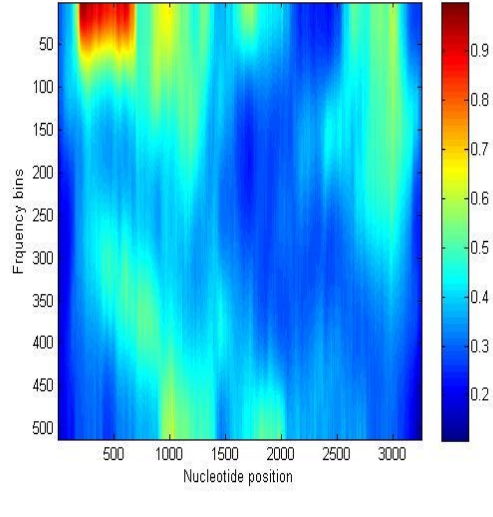
## Results and Observation

In this chapter, observations based on the experimental results have been discussed. The experiment has been performed on 12 cancerous and 12 non cancerous genes.

### 4.1. Spectrograms for experimental data

Spectrograms have been plotted for the cancerous and non cancerous genes and these tabulated in table 4.1

**Table 4.1:** Spectrogram images for cancerous genes non-cancerous genes

Spectrogram images for cancerous genes	Spectrogram images for non-cancerous genes
 <p>A spectrogram plot for a cancerous gene. The y-axis is labeled 'Frequency bins' and ranges from 50 to 500. The x-axis is labeled 'Nucleotide position' and ranges from 1000 to 5000. A color scale on the right indicates intensity from 0.1 (dark blue) to 1.0 (dark red). The plot shows a complex pattern of vertical bands with some high-intensity (red) regions.</p> <p data-bbox="223 1668 702 1736">Figure 1: For AAQ08976</p>	 <p>A spectrogram plot for a non-cancerous gene. The y-axis is labeled 'Frequency bins' and ranges from 50 to 500. The x-axis is labeled 'Nucleotide position' and ranges from 500 to 3000. A color scale on the right indicates intensity from 0.2 (dark blue) to 0.9 (dark red). The plot shows a complex pattern of vertical bands with some high-intensity (red) regions.</p> <p data-bbox="869 1668 1404 1736">Figure 2: For AAQ08976</p>

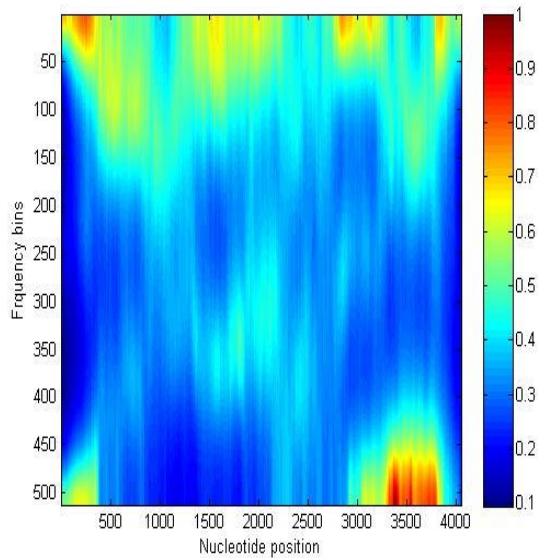


Figure 3: For AB489153

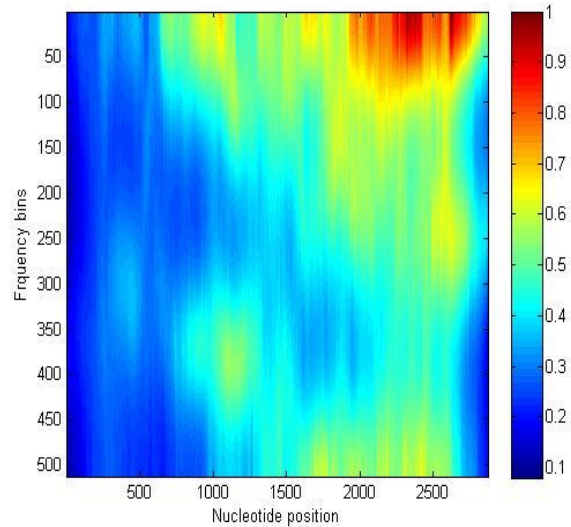


Figure 4: For AAQ08976

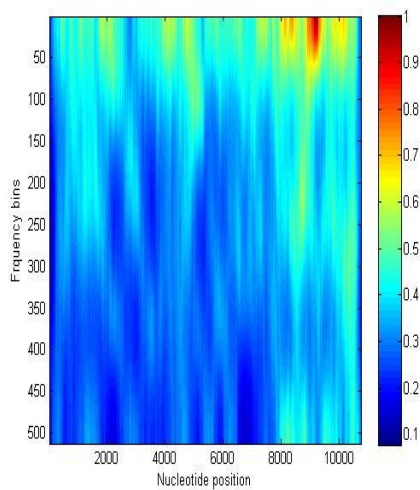


Figure 5: For AB489154

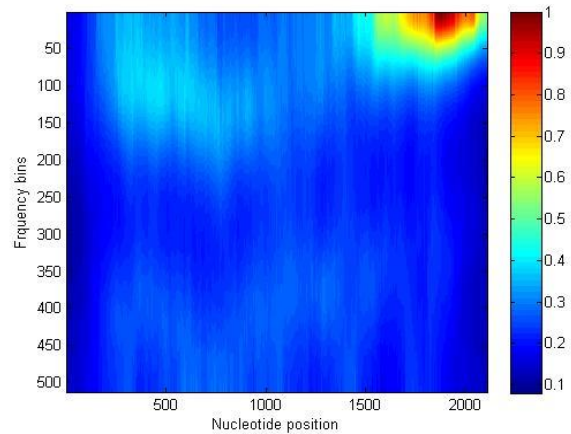


Figure 6: For AAQ08976

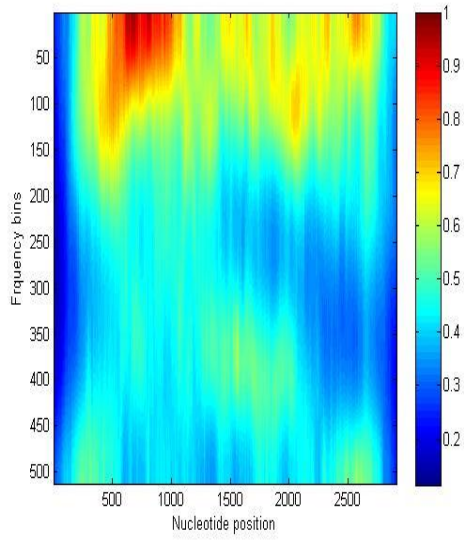


Figure 7: For AAQ08977

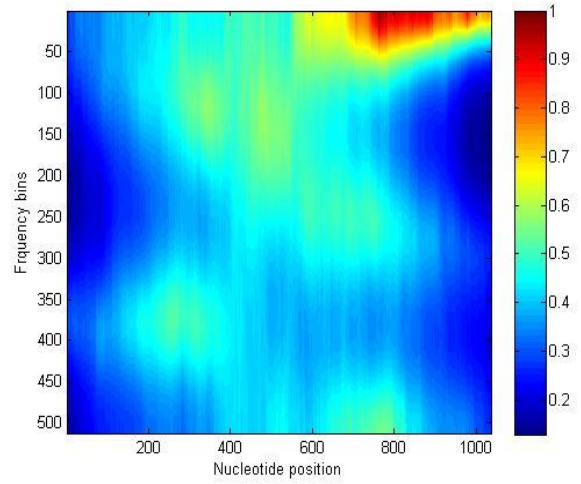


Figure 8: For AAQ08907

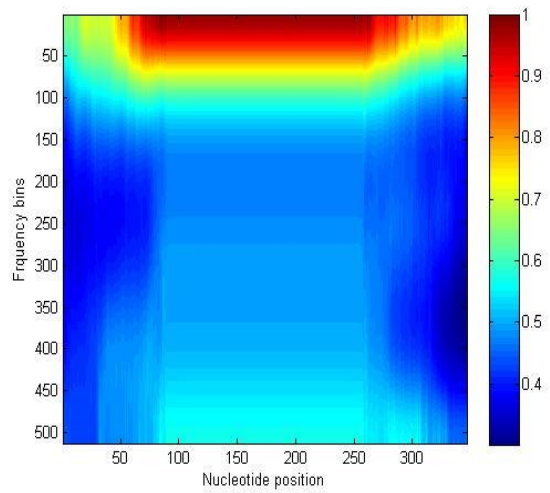
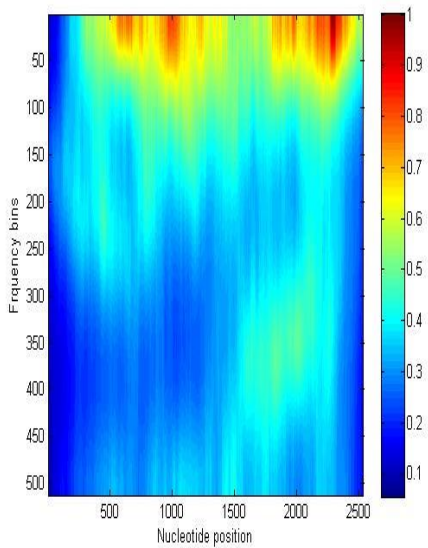


Figure 9: For AF338650

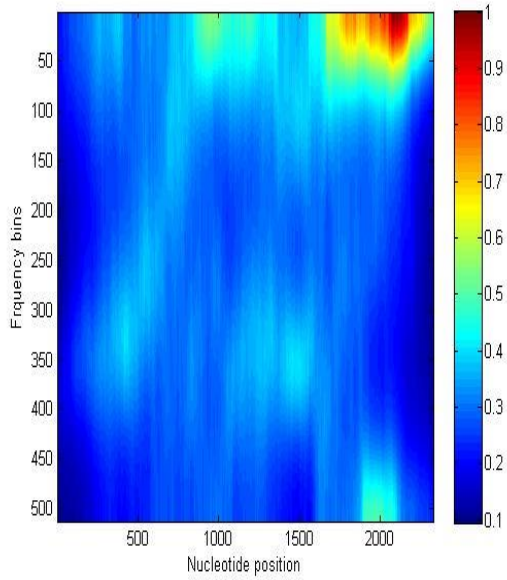


Figure 11: For AF338370

Figure 10: For AAQ08944

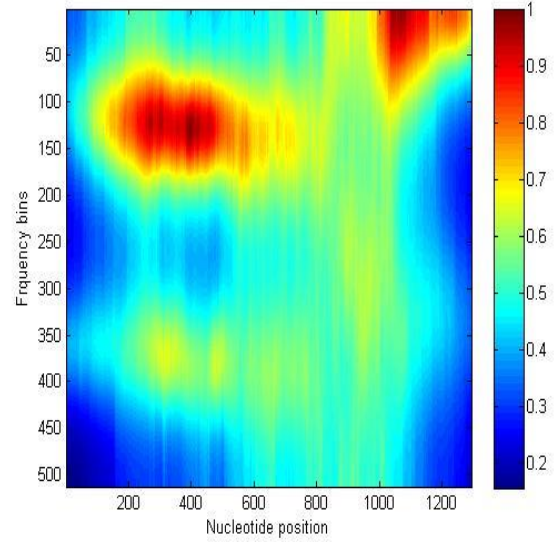


Figure 12: For NM\_001127182

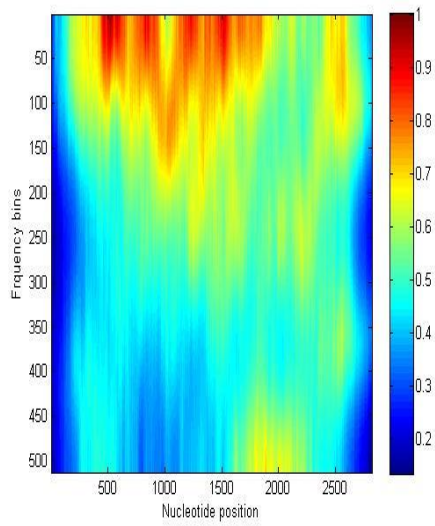


Figure 13: For AAQ0878

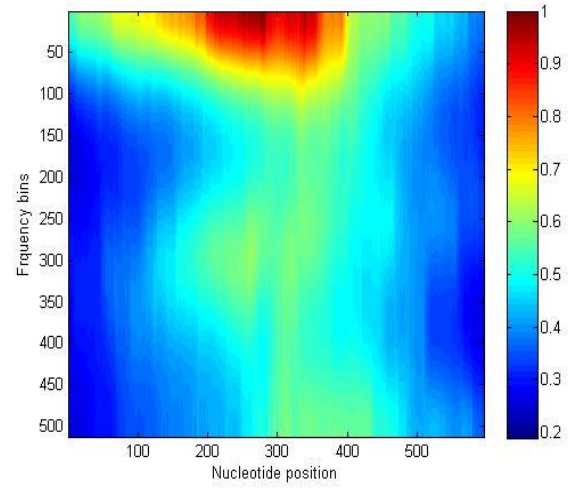


Figure 14: For AAQ08976

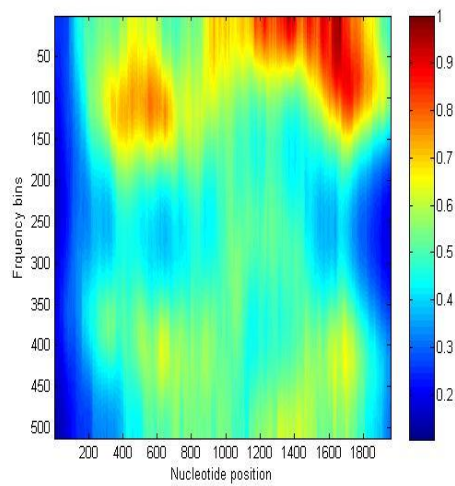


Figure 15: For NM\_001127182

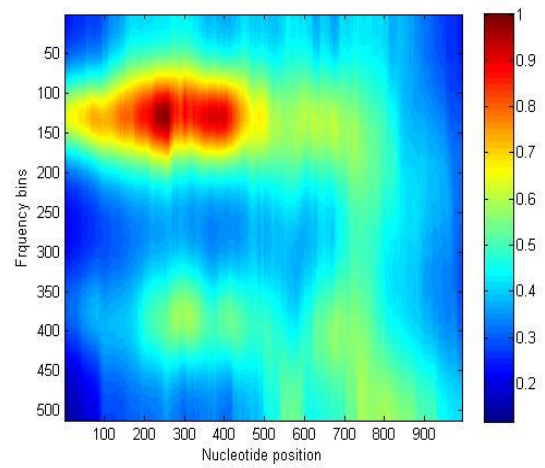


Figure 16: For NM\_001127182

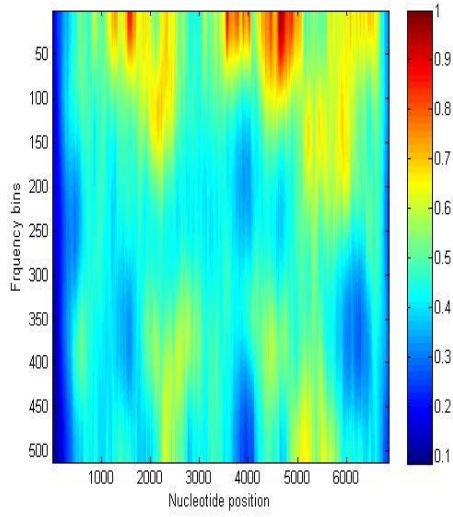


Figure 17: For NM\_001127185

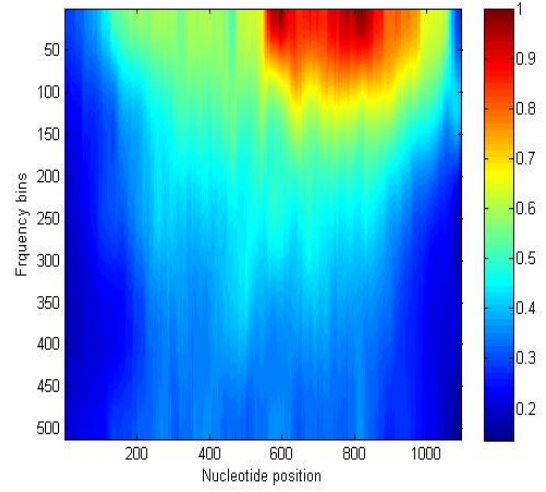


Figure 18: NM\_001127199

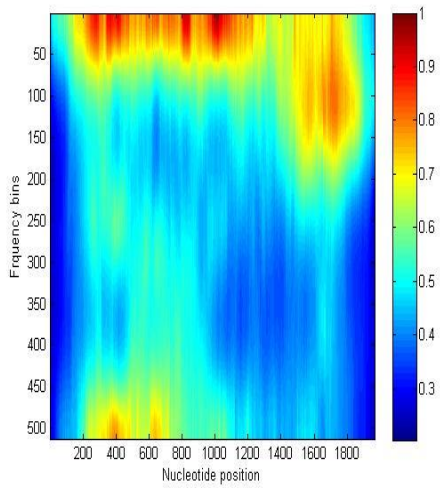


Figure 19: For NM\_0011271689

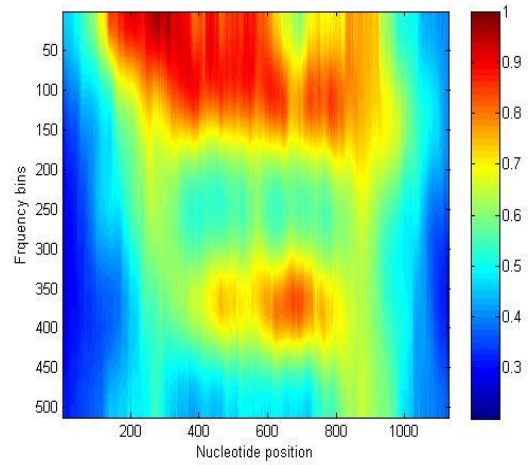
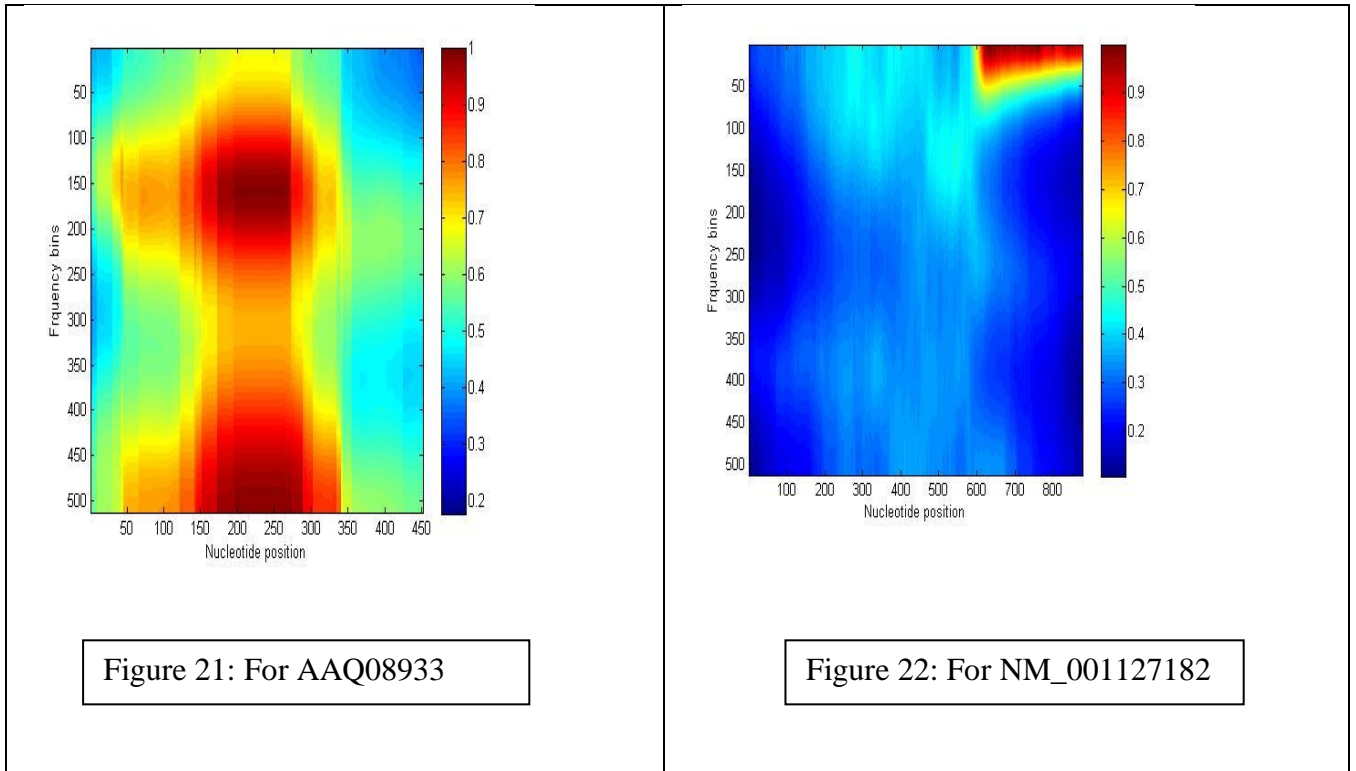


Figure 20: For AAQ0821



## 4.2 Concentration Measure Calculated for Experimental Data

Concentration measure calculated for cancerous and non cancerous genes has been calculated and summarized in table 4.2

**Table 4.2:** Concentration measure for cancerous genes

GENES SEQUENCE	CM VALUES
AAQ08976	5.1814e-07
AB489153	8.5738e-07
AF045584	1.3460e-06
AF304370	8.6114e-07
AF338650	2.4214e-07
AF455138	3.4517e-07
AF540953	1.2462e-06
NM_001127182	1.1431e-06
NM_001178078	6.7928e-06
NM_001272046	1.0650e-06
NM_001276342	1.2778e-06

AY572973	1.2045e-06
----------	------------

**Table 4.3:** Concentration measure for non-cancerous genes:

GENES SEQUENCE	CM VALUES
AF335477	2.1371e-06
AJ459782	4.2844e-06
AJ459784	2.9709e-06
AJ512346	2.5521e-06
BT006816	7.5519e-06
NM_012278	1.34816e-06
NM_001276364	9.7095e-07
NM_017436	8.1871e-07
NM_018713	4.3529e-06
NM_024533	2.6720e-06
NM_030754	4.7844e-06
NM_001276342	4.8655e-06

After checking the concentration measure for both healthy and non-healthy genes sequences, we estimate the value of threshold:

$$\text{Threshold} = 1.3462 \times 10^{-6}$$

### 4.3 Performance analysis.

To analyse the performance of the proposed hypothesis, following parameters have been computed. True positive (TP), True Negative (TN), False positive (FP), False Negative (FN), Sensitivity, Specificity and Accuracy.

- FP = 1
- FN = 2
- TP = 11
- TN = 10
- Sn =  $(TP) / (TP + FN) = 0.84$
- Sp =  $(TN) / (TN + FP) = 0.90$
- Accuracy =  $(TP + TN) / (TP + FP + FN + TN) = 0.875$  or 87.5%.



## 4.4 Pseudo Code

Classification of cancerous and non-cancerous genes using spectrogram and concentration measure

```
// Read the file from the folder
```

```
//giving the location to access the genes  
a=char(x); // initializing a as character
```

```
[r2 c2]=size(a); //giving the size
```

```
//after that
```

We have converted the DNA characters in to numerical form

Defining the length as L

```
for i=1:D_length-1 //applying for loop
```

```
code=c(i+1); // i+1 to code
```

```
end
```

```
// then to remove the dc component from the signal by
```

```
Sd1 // insilising sd1
```

```
I5=mean(I5); //then taking mean
```

```
plot(sd1); //plotting on graph
```

```
l_win=512 //window length
```

```
w(n)=rectwin(l_win); //then doing zero Padding
```

```
// calculated concentration measure
```

```
// select the fixed threshold based on the experimental analysis by Threshold=1.3462*10(-6);  
6);
```

```
Threshold=1.3462*10(-6);
```

//applying it to cancerous and non cancerous in if else case

CM > Threshold //non-cancerous

Else it is a cancer gene

## **CHAPTER 5** **CONCLUSION**

In this project a hypothesis has been proposed, which is based on the digital signal processing based spectral features. The concentration measures of the spectrogram of the DNA sequences of genes have been considered as spectral feature. The spectrograms of the DNA sequences of genes have been generated using short time Fourier transform. Based on the values of the concentration measures of the cancerous and non cancerous gene, an appropriate value of the threshold has been selected manually. The performance of the proposed hypothesis has been applied on the 12 cancerous and 12 non cancerous genes and it has found that the accuracy of the proposed hypothesis is 87.5%.

In future the proposed hypothesis can tested on large data set and its performance can be compared with the other exist methods.

## REFERENCES

- [1] Franco-Zorrilla JM, López-Vidriero I, Carrasco JL, Godoy M, Vera P, Solano R. DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proceedings of the National Academy of Sciences*. 2014 Feb
- [2] Peterson RT, Shaw SY, Peterson TA, Milan DJ, Zhong TP, Schreiber SL, MacRae CA, Fishman MC. Chemical suppression of a genetic mutation in a zebrafish model of aortic coarctation. *Nature biotechnology*. 2004 May;22(5):595-9.
- [3] Bostwick DG, Cooner WH, Denis L, Jones GW, Scardino PT, Murphy GP. The association of benign prostatic hyperplasia and cancer of the prostate. *Cancer*. 1992 Jul 1;70(S1):291-301.
- [4] Kubota H, Matsuse K, Nakano T. DSP-based speed adaptive flux observer of induction motor. *IEEE transactions on industry applications*. 1993 Mar;29(2):344-8.
- [5] Anastassiou D. Genomic signal processing. *IEEE signal processing magazine*. 2001 Jul;18(4):8-20.
- [6] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. *Nucleic acids research*. 2001 Jan 1;29(1):308-11.
- [7] Church GM, Gilbert W. Genomic sequencing. *Proceedings of the National Academy of Sciences*. 1984 Apr 1;81(7):1991-5.
- [8] Shendure J, Ji H. Next-generation DNA sequencing. *Nature biotechnology*. 2008 Oct;26(10):1135-45.
- [9] Venter JC, Smith HO, Hood L. A new strategy for genome sequencing. *Nature*. 1996 May;381(6581):364-6.
- [10] Chakraborty S, Gupta V. Dwt based cancer identification using EIIP. In 2016 second international conference on computational intelligence & communication technology (CICT) 2016 Feb 12 (pp. 718-723). IEEE.
- [11] Bracewell RN, Bracewell RN. *The Fourier transform and its applications*. New York: McGraw-Hill; 1986 Feb.
- [12] (2006) Guo SB, Lyu MR, Lok TM. Quality choice dependent on common data for the classification of multi-class malignancy.
- [13] 2013 (John SM) order of IJCER for enormous microarray directed learning.
- [14] 2004 L. Bharadwaj method of storing DNA
- [15] M. Akhtar, 2007, On DNA numerical representations on Genomic Signal Processing
- [16] Norman E (2015) About cancer

[17] 1997 Gurney K neural network introduction

[18] 2014 [Hariprasad SANovel cancer detection approach. Mysore, India.

[19] Galleani L, (2010) The base entropy planning spectrum of a DNA grouping.

[20] Grosse I, Herzel H, Buldyrev SV, Stanley HE (2000) Species indubiousness of shared data in coding and noncoding DNA.