Modified Gabor Wavelet Transform in Prediction of Cancerous Genes

Lopamudra Das, Anand Kumar, J. K. Das, Sarita Nanda

Abstract*: Cancer is the leading cause of mortality all over the world which in general is the result of some kind of mutation in the genetic sequence. With recent advancements in Digital Signal Processing(DSP) techniques, it has become possible to classify cancerous gene sequences without carrying out extensive biological experiments. In this paper, the Geometric mapping technique along with Modified Gabor wavelet transform (MGWT) has been incorporated to segregate cancerous and non-cancerous gene sequences. This Gabor wavelet based transform technique used in the present research work benefits from the fact that it is independent of the window length which in conjunction with Geometric mapping is used to obtain the spectral components present in the signal accurately and with reduced complexity. This technique has been applied on numerous benchmark datasets and the results obtained prove the performance of the proposed method.

Keywords: Deoxyribonucleic Acid (DNA), Genomic Signal Processing(GSP), Modified Gabor Wavelet Transform (MGWT), Geometric mapping, cancer

I. INTRODUCTION

Cancer is the leading cause for the death of people all around the world, it caused the death of approximately 9.5 million deaths in 2018 as stated in the report[1] presented by the World Health Organization. The genetic abnormality is the main reason for cancer which was presented by D. C. Wertz[2]. The most indispensable task in genomic processing is to locate the exact position of protein coding region also known as exon. Genomic data such as DNA basically is a continuous chain of nucleotides (A, G, T, C) and needs to be converted into numerical form to be processed by digital signal processing techniques. Cristea[3] discussed the conversion of nucleotide sequence into a genomic signal. Several mapping techniques have been discovered to disseminate the difference between protein coding and non-coding regions. Anastassiou[4]explained how protein coding regions (exon) and non-coding regions (intron) can be easily distinguished by the numerical mapping of nucleotides and applying DSP techniques. The exons exhibit a peak at N/3 whereas no such behavior is found in the introns[5]. This behavior is referred to as "periodicity property" or "three base periodicity". L. Das[6] proposed geometric mapping and compared its performance with several other mapping and found this mapping technique to be more efficient with least exon miss and false exons. Vidyanaath [7]explained the role of signal processing concepts in proteomics and genomics.

Cancer arises due to the alteration of nucleotides in the DNA sequence of the gene called mutation. In-depth analysis has been done in [8] that depicts the occurrence of cancer in an individual's genome. Cancer is a dreadful disease and has increased the mortality rate in recent years. It is grounded by scientists that cancer ascends due to the accretion of mutated cells in pivotal gene locations that changes the normal functioning of cell proliferation, metabolization, etc. Early spotting of cancer cell and identification of protein coding region will play a very crucial role in reducing the death due to cancer. Myriad of works has been done using DSP techniques for cancer prediction[9] Qui[10]. used Genomic signal processing in cancer prediction and classification. Dougherty[11] discussed genomic signal processing and how it can be used in diagnosis and therapy.

In the last two decades, there has been significant growth in the implementation of Wavelets in the analysis of DNA sequences and functional genomics data. [12]discussed the implementation of wavelets in Genomic signals and how it overshadows Fourier transform (FT). Wavelet analysis, unlike traditional FT techniques for genomic signal processing, is able to decompose time series into time-frequency space and thus, has gained attention as a potential tool to scrutinize cancer genomic data. Many researchers have implemented Wavelet analysis for several tasks in Genomic processing. Meng[13] used wavelet analysis for cancer identification. The wavelet transform plot was used to pinpoint the location of the mutation in gene sequence[14]. George[15] explains lossless and lossy schemes in DWT and its application in gene segregation. DWT based EIIP technique was implemented in cancer prediction [16]. The proposed techniques discriminate cancerous and healthy gene sequences without any error, also the PSD plots are noise free which is not there in previous works. MATLAB R2015a equipped with a Bioinformatics toolbox has been used in this work to obtain all the plots and outputs.

The rest of the paper is organized as follows: Section II contains the overview of DNA and the cause of cancer. Section III consists of methodology, Section IV contains the proposed algorithm and methods. Section V discusses the simulation and result analysis. Section VI concludes this paper.

Revised Manuscript Received on October 15, 2019
*Correspondence Author
Lopamudra Das, Electronics and Telecommunication Engineering, KIIT University, Bhubaneswar, India.
Anand Kumar, Electronics and Telecommunication Engineering, KIIT University, Bhubaneswar, India.
J. K. Das, Electronics and Telecommunication Engineering, KIIT University, Bhubaneswar, India.
Sarita Nanda*, Electronics and Telecommunication Engineering, KIIT University, Bhubaneswar, India.

II. OVERVIEW OF DNA

Deoxyribonucleic acid (DNA) molecules contain hereditary information in living beings[17]. A human body is made up of innumerable cells, nearly every cell in a person's body has a similar arrangement of DNA constitutes. DNA molecule has a two-strand double complex helix structure with phosphate and sugar groups in alternate positions as shown in Fig. 1. The four chemical bases, namely adenine (A), guanine (G), cytosine (C), and thymine (T) are assembled to these sugar groups of each DNA strand. A is complementary to T and G is complementary to C, each base in the strands are connected to its complementary nucleotide via Hydrogen bond. In genomic signal processing[4] the protein coding region information is extracted from DNA by a special period-3 property. The nucleotides A, G, C, T in a set of 3 (codons) can be coded into 20 different amino acids in protein coding region. The chain of consecutive codons is responsible for the translation of protein.

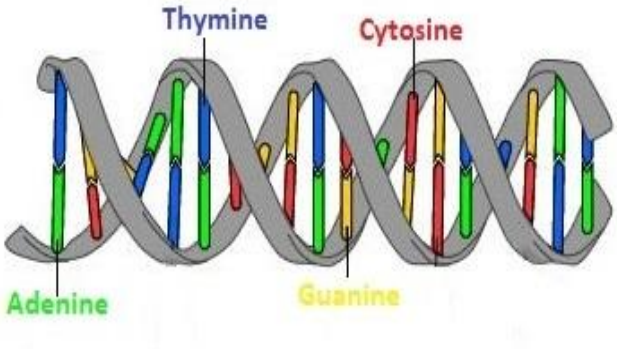

Fig. 1. Double helix structure of DNA

As a cell replicates its DNA before splitting, it makes occasional misprints most commonly a nucleotide is removed or an extra nucleotide is added. This process is termed as mutation and it is the major cause of genetic disorders[18]. Though there are billions of nucleotides in a human body, changes in any single pair may cause dramatic physiological malfunctions. In cancer, cells grow and divide hysterically, forming malicious tumors, and also invade nearby parts of the body. Cancer is caused by internal factors like inherited mutations, immune conditions, misbalance of hormones, and mutation due to metabolism and external constituents like infectious organisms, radiation, tobacco, and chemicals. These internal and external factors may operate together or in sequence to trigger the development of cancer.

Different types of mutations and diseases associated with a certain type of mutations are discussed in [19]. Four types of mutations are usually observed in dominant disorders. They are 1. Substitution, 2. Insertion or Deletion, 3. Translocation, and 4. Copy number alterations. Translocation refers to the process in which a region from one chromosome is aberrantly attached to another chromosome and causes Leukemia. Whereas copy number alteration also considered as duplication occurs when a region of a chromosome is repeatedly copied in certain positions leading to an increase in the increase of dosage in that region and is the reason for the occurrence of some cancers. Substitution and insertion/deletion involves alteration of a nucleotide with other nucleotide and random addition or removal of nucleotides in a gene sequence respectively. The types of mutation discussed above are shown in Fig. 2.

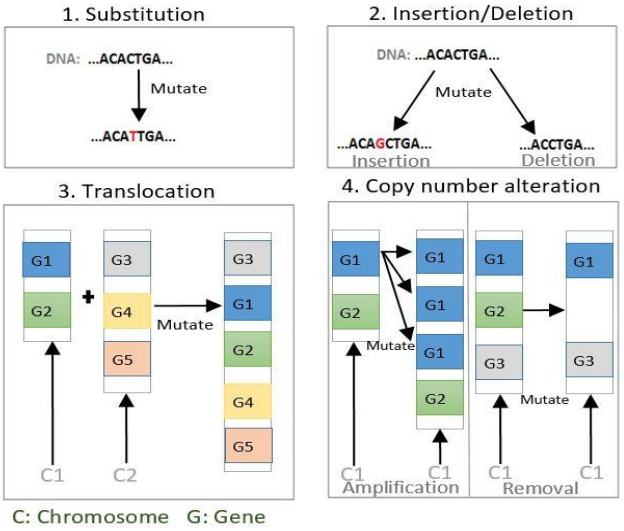

Fig. 2. Classification of mutation

III. METHODOLOGY

In order to accomplish our objective of predicting cancerous gene sequences, the gene sequences undergo certain processes. The steps used in this work are shown in Fig. 3. Firstly, gene sequences are collected from the NCBI website[20]. Before applying a suitable signal processing method, the symbolic representation of the DNA sequence is transformed into a numerical sequence using the geometric mapping technique for the identification of protein coding regions.

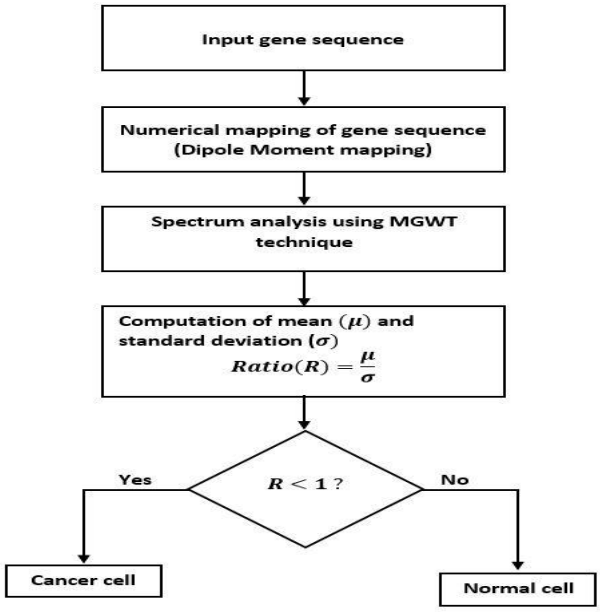

Fig. 3. Flow of the proposed method

To implement signal processing technique to the gene sequence it is first converted into a numerical form from the string of characters. Several mapping techniques have been used to map the gene sequences like Voss mapping, binary mapping, EIIP mapping, etc. and many more are discussed in [21]. In this work we have used mapping technique based on the physiochemical property i.e. trigonometric mapping [6]. It suppresses false exons with

very less miss rate along with discriminating factor. The values associated with the nucleotides are given in Table- I.

X[n] = ACCGTTA…. (1)

On assigning the values associated with the nucleotides in (1):

Table-I: Values of nucleotides in Trigonometric Mapping (TM)

Nucleotide	TM value
A	- cos(θ) + j*sin(θ)
G	cos(θ) + j*sin(θ)
T	cos(θ) - j*sin(θ)
C	- cos(θ) - j*sin(θ)

Secondly, MGWT is applied to the numerical sequence (2) to obtain the output. The method is capable of handling both small coding regions with small scale sequences as well as large coding regions with long scale sequences. It doesn't have overhead computational complexity like other signal processing techniques along with de-noising the signal.

IV. ALGORITHMS AND METHODS

A multiscale transform of a signal u can be defined as

$$U(a,b) = \int u(x)\psi(x,b,a)dx \quad (2)$$

Where, ψ represents the analyzing function, $a > 0$ is the scale parameter, and b is the position (or time) parameter.

Different functions may be used to adopt this transform. Short-time Fourier transform (STFT) in particular is mostly used along with a Gaussian window. It is also known as Gabor transform and can be written as:

$$\psi_{GWT}(x,b,a) = e^{-\frac{\left(\frac{x-b}{a}\right)^2}{2}} e^{j\omega_o\left(\frac{x-b}{a}\right)} \quad (3)$$

where ω_o is the basic frequency of ψ_{GWT}.

The normal Gabor and Gabor-wavelet transform cannot be used directly in the analysis of genomic sequences as it analyses signals at different frequencies. Some modifications were done to analyze the signal [22]. Modified Gabor-wavelet transform is given as follows:

$$\psi_{MGWT}(x,b,a) = e^{\frac{-(x-b)^2}{2a^2}} e^{j\omega_0(x-b)} \quad (4)$$

$$U(b,a) = \int u(x)\, e^{\frac{-(x-b)^2}{2a^2}} e^{j\omega_0(x-b)} \quad (5)$$

In this work two methods are used to distinguish cancer and normal cells.

A. Digital Spectral Analysis

All the gene sequences used in this paper are obtained from the NCBI website [20]. To make the work more robust we have implemented this in two different types of gene sequences i.e. a normal homo sapiens gene sequence and in cancer affected gene sequence. After performing trigonometry mapping to all the gene sequences. The normal homo sapiens gene sequence is manually mutated as discussed in section II. Wavelet transform is applied to all the gene sequences i.e. normal and mutated to observe distinct patterns. Wavelet transform has the advantage of visualizing the alteration in DNA sequence at different levels occurred due to mutation. The Scalogram plot obtained on performing wavelet transform greatly located the place of mutation and so can be used to distinguish a normal and mutated gene sequence.

B. Power Spectral Density (PSD)

Several gene sequences of causing different types of cancer disease, as well as normal homo sapiens gene sequences, are obtained from the NCBI website. Similar steps are followed in this case and lastly PSD plot is obtained after applying Modified Gabor-wavelet transform. Modified Gabor-wavelet gives noise free PSD plot with a considerable difference between exon and intron. Several evaluation parameters have been used to show a distinction between normal and cancer cells.

V. SIMULATION AND RESULT ANALYSIS

The method used in this paper to detect cancerous gene sequences are useful as it efficiently discriminates cancerous cells from normal cells and does not involve any biological experiments. Several normal and cancerous gene sequences are collected from the NCBI website as given in Table- II and Table- III respectively along with its exon locations.

Two different approaches have been implemented to differentiate mutated and healthy gene sequences. In the Method-I Scalogram plot of both healthy and mutated gene sequences is done. A normal homo sapiens gene sequence having accession number AF007546.1 whose Scalogram plot is in Fig. 4(a) and on performing manual mutation to the same gene sequence. A difference can be observed in the position of mutation in the Scalogram plot as shown in Fig. 4(b) and Fig. 4(c). Same procedures were repeated with a cancerous gene sequence and the presence of mutation can be clearly seen the Scalogram plot as shown in Fig. 5. All the mutated spots are marked in the respective figures.

Whereas, in Method-II, the MGWT technique is used to all the geometrically mapped gene sequences to obtain the PSD plot. The PSD plot shows a considerable difference in both healthy and cancerous gene sequences. Spikes can be observed in the case of cancerous gene sequences in Fig. 6 and Fig. 7 which is absent in the case of normal gene sequences as shown in Fig. 8 and Fig. 9. Several evaluation parameters have been used to test its efficacy. The gene sequences carrying cancer have ***R*** <1 as well as ***CV*** (%) is >100. Which is because there is high degree of randomness in case of cancerous genes which leads to higher value for standard deviation compared to mean of a gene sequence. In case of normal gene sequence there is very less degree of randomness and so has less value of standard deviation than mean in the gene sequence. Several normal and cancer affected cells have been tested in this work and the values obtained after analysis are written in Table- IV and Table-V for Normal and Cancer cells respectively. Graphs of variation of ***R*** and ***CV*** of the gene sets have been plotted in Fig. 10 and Fig. 11 respectively.

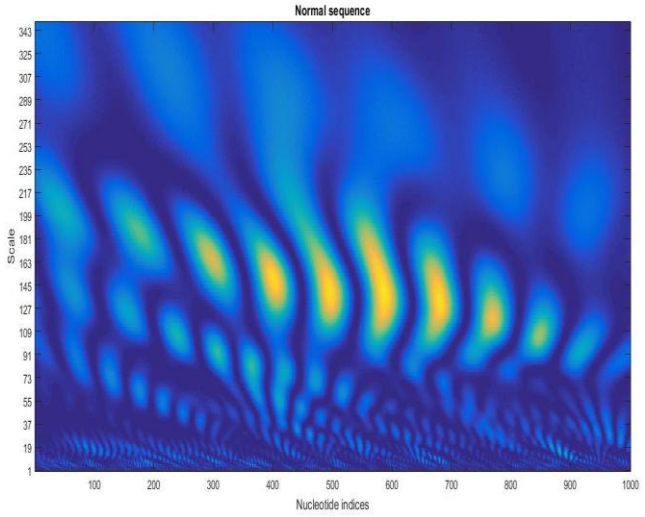

Fig. 4(a). Scalogram plot of sequence AF007546.1 (Normal)

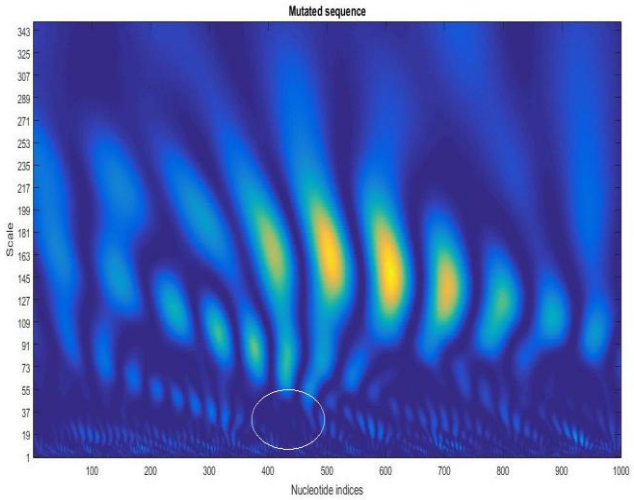

Fig. 4(b). Scalogram plot of sequence AF007546.1 (Mutated)

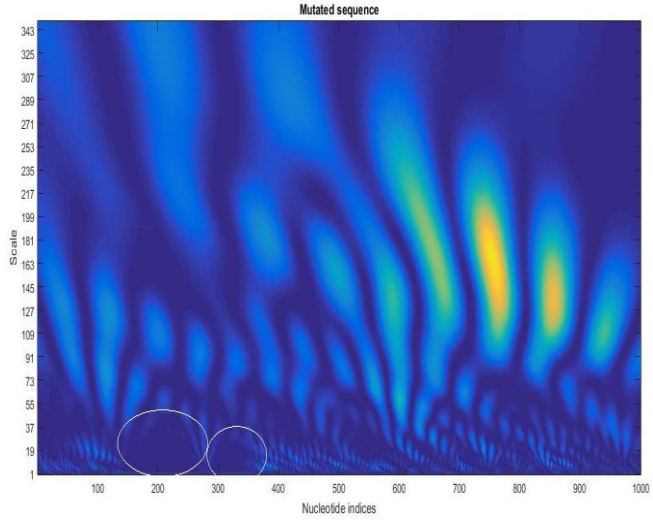

Fig. 4(c). Scalogram plot of sequence AF007546.1 (Mutated)

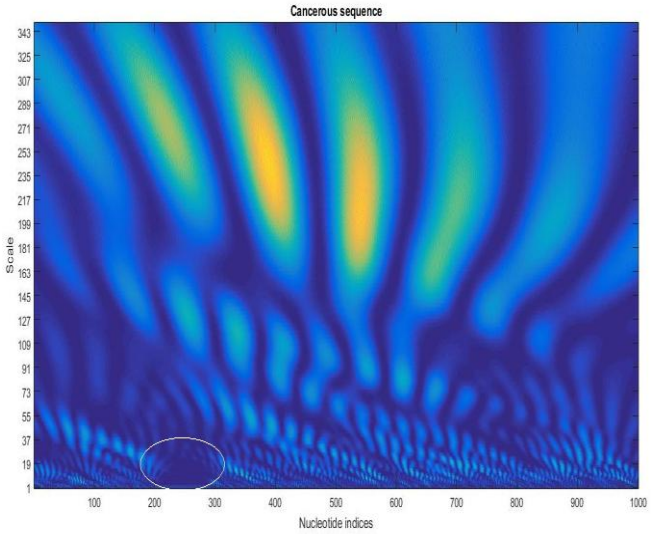

Fig. 5. Scalogram plot of cancerous sequence AF012108.1 (Breast cancer)

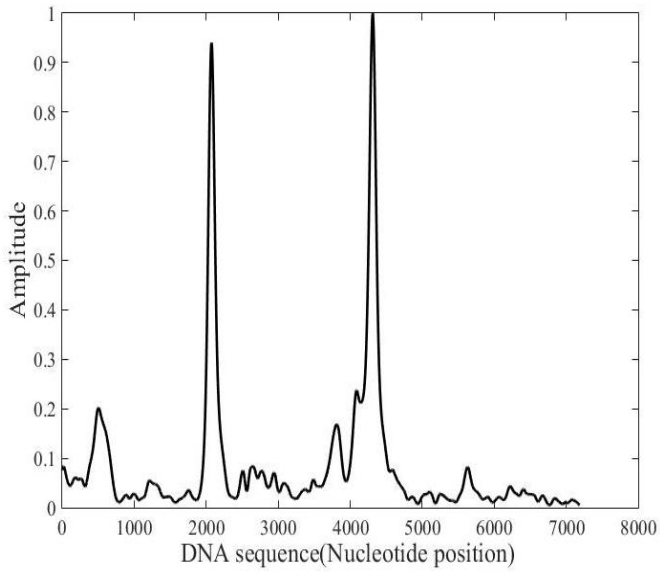

Fig. 6. PSD plot of sequence AF012108.1 (Breast cancer)

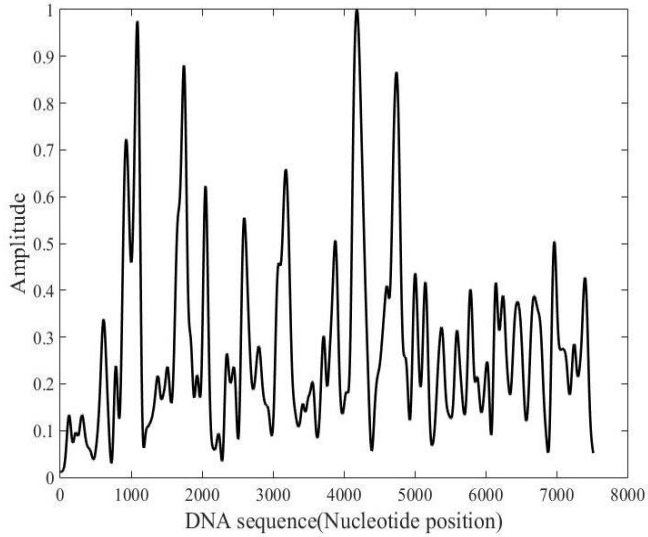

Fig. 7. PSD plot of sequence AF284036.1 (Prostate cancer)

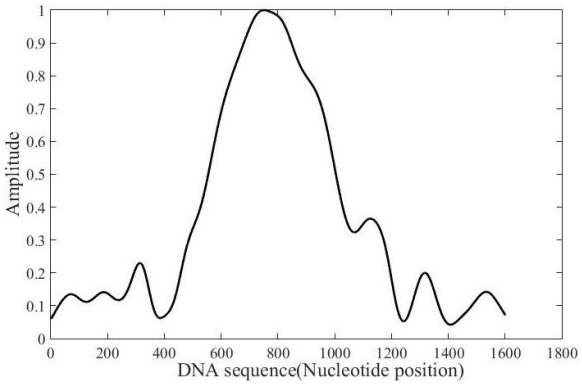

Fig. 8. PSD plot of sequence AF007189.1 (Healthy)

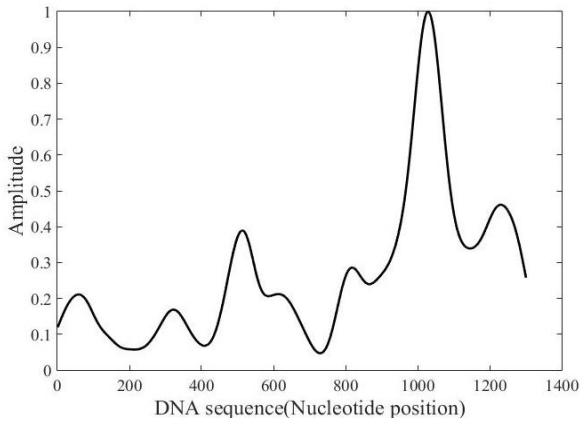

Fig. 9. PSD plot of sequence AF186607.1 (Healthy)

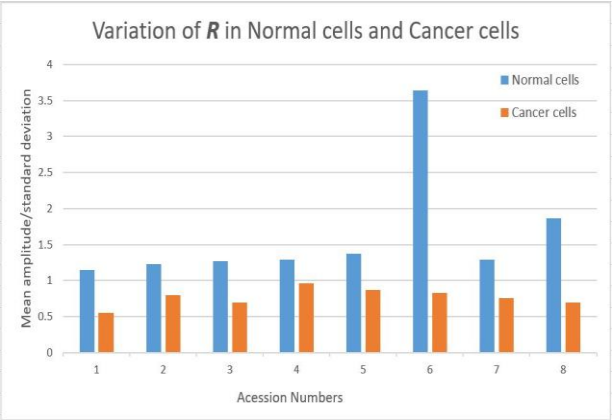

Fig. 10. Bar plot of Ratio (*R*)

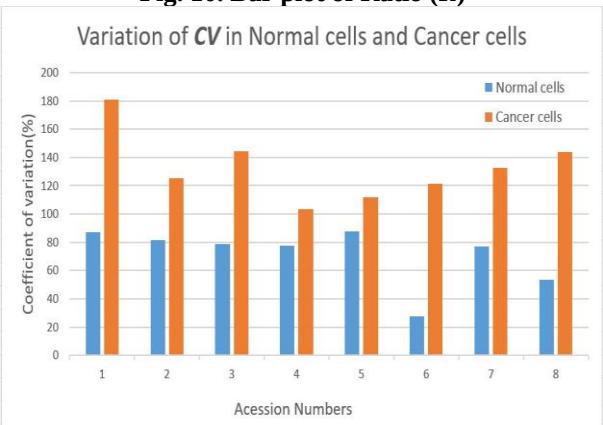

Fig. 11. Bar plot of coefficient of variation *(CV)*

Table- II: Normal gene sequences with their accession number

Accession number	Gene name	Protein name	Exon position(s)
AF007546.1	Homo sapiens beta-globin gene		1475:1603
AF083883.1	Homo sapiens mutant beta-globin gene	HBB	27:118,24:.471, 1322:1450
AF186613.1	Homo sapiens haplotype beta-globin gene	HBB	987:1078,1209:1300
AF186607.1	Homo sapiens haplotype A11a beta-globin gene	HBB	988:1079,1210:1301
AF007189.1	Homo sapiens claudin 3 gene	CLDN3	477:1139
AF186616.1	Homo sapiens haplotype C17 beta-globin gene	HBB	992:1083,1214:1305
AF007190.1	Homo sapiens SIB 227C intestinal mucin mRNA	MUC3	1:1539
AF348448.1	Homo sapiens beta-globin gene	HBB	139:230,361:444

Table- III: Cancerous gene sequences with their accession number

Accession number	Gene name	Disease	Exon position(s)
AF012108.1	AIB1	Breast cancer	201:4463
NM_007294.3	BRCA1	Pancreatic cancer	233:5824
AF284036.1	COPEB	Prostate cancer	1692:1793, 4494:5067, 6480:6603, 7120:7171
NM_012403.1	ANP32C	Lung cancer	1:705
AB035807.1	DRCC1	Colon cancer	25:1563
NM_000142.4	FGFR3	Bladder cancer	257:2677
NM_005732.4	RAD50	Colorectal cancer	349:4287
NM_001145155.2	NR2F2	Prostate cancer	265:1110

Table -IV: Analysis of normal cells

S. no	Accession number	Mean (μ)	Standard deviation(σ)	Ratio $(R) = \frac{\mu}{\sigma}$	Coefficient of variation % $(CV) = \frac{\sigma}{\mu}$
1	AF007546.1	0.2587	0.2256	1.1467	87.20
2	AF083883.1	0.3131	0.2550	1.2278	81.44
3	AF186613.1	0.2725	0.2139	1.2739	78.49
4	AF186607.1	0.2714	0.2104	1.2899	77.52
5	AF007189.1	0.3577	0.3144	1.3772	87.89
6	AF186616.1	0.6709	0.1841	3.6442	27.44
7	AF007190.1	0.2724	0.2104	1.2946	77.23
8	AF348448.1	0.4374	0.2341	1.8684	53.52

Table-V: Analysis of cancerous cells

S. no	Accession number	Mean (μ)	Standard deviation (σ)	Ratio $(R) = \frac{\mu}{\sigma}$	Coefficient of variation % $(CV) = \frac{\sigma}{\mu}$
1	AF012108.1	0.0804	0.1456	0.5521	181.09
2	NM_007294.3	0.1737	0.2181	0.7964	125.56
3	AF284036.1	0.1914	0.2766	0.6919	144.51
4	NM_012403.1	0.3156	0.3273	0.9642	103.70
5	AB035807.1	0.1861	0.2086	0.8705	112.09
6	NM_00142.4	0.2857	0.3468	0.8238	121.38
7	NM_005732.4	0.2348	0.3116	0.7535	132.70
8	NM_001145155.2	0.1683	0.2419	0.6957	143.73

VI. CONCLUSION

In this paper modified Gabor wavelet transform along with geometric mapping has been implemented for analyzing DNA sequences for predicting cancerous genes. The geometric mapping reduces the computational complexity to a great extent when compared to other mapping techniques. On compared to typical Fourier based spectral technique Modified Gabor wavelet technique is noise resistant and is adaptive to the length of the gene sequence. All the methods mentioned distinctly detect normal and cancer cells and several evaluating parameters have been used to mathematically buttress the spectrally obtained result. This work is one more step towards the early detection of cancer as it solely depends on DNA and shows any changes in the gene sequence without performing any biological experiments and without using any costly equipment.

REFERENCES

1. International Agency for Research on Cancer. Globocan.: "Http://Gco.Iarc.Fr/Today." *Http://Gco.Iarc.Fr/Today*. vol. 876, pp. 2018–2019, 2018.
2. D.C. Wertz, J.C. Fletcher, and K. Berg: "Review of Ethical Issues in Medical Genetics. Report of Consultants to WHO." *World Heal. Organ. Hum. Genet. Program. www.who.int*. pp. 110, 2003.
3. P.D. Cristea: "Conversion of nucleotides sequences into genomic signals." *J. Cell. Mol. Med.* vol. 6, no. 2, pp. 279–303, 2002.
4. 4. D. Anastassiou: "Genomic signal processing." *IEEE Signal Process. Mag.* vol. 18, no. July, pp. 8–20, 2001.
5. J. Sanchez and I. Lopez-Villasenor: "A simple model to explain three-base periodicity in coding DNA." *FEBS Lett.* vol. 580, no. 27, pp. 6413–6422, 2006.
6. L. Das, S. Nanda, and J.K. Das: "A novel DNA mapping scheme for improved exon prediction using digital filters." *Proc. - 2017 2nd Int. Conf. Man Mach. Interfacing, MAMI 2017*. vol. 2018-March, no. 1, pp. 1–6, 2018.
7. P.P. Vaidyanathan and B.-J. Yoon: "The role of signal-processing concepts in genomics and proteomics." *J. Franklin Inst.* vol. 341, pp. 111–135, 2004.
8. M.R. Stratton, P.J. Campbell, and P.A. Futreal: "The cancer genome." *Nature*. vol. 458, no. 7239, pp. 719–724, 2009.
9. G.N. Satapathi, P. Srihari, A. Jyothi, and S. Lavanya: "Prediction of cancer cell using DSP techniques." *Int. Conf. Commun. Signal Process. ICCSP 2013 - Proc.* pp. 149–153, 2013.
10. Peng Qiu, Z.J. Wang, and K.J.R. Liu: "Genomic processing for cancer classification and prediction - Abroad review of the recent advances in model-based genomoric and proteomic signal processing for cancer detection." *IEEE Signal Process. Mag.* vol. 24, no. 1, pp. 100–110, 2007.
11. E.R. Dougherty and A. Datta: "Genomic signal processing: Diagnosis and therapy." *IEEE Signal Process. Mag.* vol. 22, no. 1, pp. 107–112, 2005.
12. W. Trust and G. Campus: "Biology : State of Art and Perspectives." *Genome*. vol. 19, no. 1, pp. 2–9, 2003.
13. T. Meng, A.T. Soliman, M.L. Shyu, Y. Yang, S.C. Chen, S.S. Iyengar, J.S. Yordy, and P. Iyengar: "Wavelet analysis in current cancer genome research: A survey." *IEEE/ACM Trans. Comput. Biol. Bioinforma.* vol. 10, no. 6, pp. 1442–1459, 2013.
14. T.T. Gayathri and S.A. Christe: "Wavelet Analysis in Prediction and Identification of Cancerous Genes." *Int. J. Sci. Eng. Res.* vol. 8, no. 3, pp. 720–727, 2017.
15. T.P. George and T. Thomas: "Discrete wavelet transform de-noising in eukaryotic gene splicing." *BMC Bioinformatics*. vol. 11, no. SUPPLL.1, pp. 1–8, 2010.
16. S. Chakraborty and V. Gupta: "DWT based cancer identification using EIIP." *Proc. - 2016 2nd Int. Conf. Comput. Intell. Commun. Technol. CICT 2016*. pp. 718–723, 2016.
17. H. Rolston: "What is a gene? from molecules to metaphysics." *Theor Med Bioeth*. vol. 27, no. 6, pp. 471–472, 2006.
18. 18. S. Clancy: "Genetic Mutation | Learn Science at Scitable," https://www.nature.com/scitable/topicpage/genetic-mutation-441.
19. N. Mahdieh and B. Rabbani: "An overview of mutation detection methods in genetic disorders." *Iran. J. Pediatr.* vol. 23, no. 4, pp. 375–88, 2013.
20. "National Center for Biotechnology Information," https://www.ncbi.nlm.nih.gov/.
21. M. Ahmad, L.T. Jung, and A.A. Bhuiyan: "From DNA to protein: Why genetic code context of nucleotides for DNA signal processing? A review." *Biomed. Signal Process. Control*. vol. 34, pp. 44–63, 2017.
22. J. Mena-Chalco, H. Carrer, Y. Zana, and R.M. Cesar: "Identification of protein coding regions using the modified Gabor-wavelet transform." *IEEE/ACM Trans. Comput. Biol. Bioinforma.* vol. 5, no. 2, pp. 198–206, 2008.

AUTHORS PROFILE

Lopamudra Das received a B.Tech degree in Electronics and Telecommunication engineering from the Utkal University, India and then received M.Tech. degree in communication system engineering from the University College of Engineering, Burla, India, in 2005. She has nearly 14 years of experience in teaching different organizations and currently she is pursuing a Ph.D. degree with KIIT University, Bhubaneswar, India. Her current research interests include Genomic signal processing, Image processing, and embedded system

Anand Kumar is currently pursuing a B.Tech degree in Electronics and Telecommunication engineering at Kalinga Institute of Industrial Technology, Bhubaneswar, India. His other fields of interest are digital signal processing, Wireless communication and Networking, Wireless sensor networks and application of power quality estimation.

Jitendra Kumar Das received the M.Tech. degree in Electronics and communication engineering from NIT, Rourkela, India, in 2004. He has completed a Ph.D. degree with NIT, Rourkela, India. He is currently an Associate Professor with the School of Electronics Engineering at KIIT University, Bhubaneswar, India. His current research interests include digital signal processing and embedded system.

Sarita Nanda received the M.Tech. degree in communication system engineering from the University College of Engineering, Burla, India, in 2005. She has completed the Ph.D. degree from Sambalpur University, Sambalpur, India. Currently, she is an Associate Professor with the School of Electronics Engineering at KIIT University, Bhubaneswar, India. Her current research interests include digital signal processing, Mixed signal VLSI, Application to power quality and embedded system.