iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** I **Month of publication:** January 2018

DOI: http://doi.org/10.22214/ijraset.2018.1009

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor :6.887
Volume 6 Issue I, January 2018- Available at www.ijraset.com

Cancer Identification VIA Weighted Entropy Method and Decision Tree Method

[1]Deepa Grover, [2]Dr. Sonica Sondhi, [3] Dr. Nidhika Birla
[1]*Research scholar Ik GPTU, Jalandhar*
[2]*Assistant Professor, Chandigarh group of colleges Landran, Mohali*
[3] *Assistant Professor, Department of Electronics Haryana Engineering College, Jagadhri, Haryana*

Abstract: ***DNA sequence classification is a challenging problem of central importance to genomic research. A number of algorithms have been developed for this problem and implemented as systems that are in widespread use today. In this article, we discuss two systems to predict cancer based on the signal processing of the DNA sequence. In the first method, a system is proposed to predict cancer from DNA sequence through an efficient concept of weighted entropy that takes both entropy and total correlation into consideration. Here, a measure, named as weighted entropy is used which captures the distribution and correlation information of a DNA sequence. In the second method, a decision tree based cancer prediction is proposed, where a split point measure is used namely weighted entropy for pruning the decision tree. For the analysis, we make use of DNA sequences obtained from National Centre for Biotechnology Information (NCBI). Evaluation metrics parameters of sensitivity, specificity and accuracy are found out and compared to the existing literature. Experimental results demonstrate that, this proposed scheme has achieved an accuracy of 90% in classifying DNA sequences.***
Keywords: weighted entropy, DNA sequence, decision tree, classification, cancer prediction

I. INTRODUCTION

The term genome, first introduced in 1920, is defined as the entirety of an organism's genetic information. The whole DNA of a living organism is known as its Genome [5]. For humans, this refers to a full set of 23 chromosomes in gamete cells, consisting of DNA [1]. DNA is a nucleic acid that has two long strands of nucleotides twisted in the form of a double helix and its external backbone is made up of alternating deoxyribose sugar and phosphate molecules. The Adenine, Guanine, Cytosine and Thymine are based on the nitrogenous, which are located in the internal portion of the DNA in duo. Scientifically the DNA and proteins are represented as character strings, in which each character is shown as a letter of the alphabet [11]. Genomics is a very cross-disciplinary field that makes worldview transformations in such various areas as medicine and agriculture. It is accepted that numerous significant scientific and innovative attempts in the 21st century will be identified with the preparing and interpretation of the tremendous information that is at currently exposed from sequencing the genomes of numerous living organisms' forms, including people. The genomic signal is one of the tools for revealing large scale features of DNA [8].Genomic signals carry genomic information to all the processes that take place in an organism [11]. Genomic sign are the measurable occasions starting from DNA succession, mRNA sequence and protein. Based upon current innovation, GSP fundamentally manages separating data from gene expression estimations. The investigation, processing and utilization of genomic signals for increasing biological knowledge constitute the space of GSP. All the instructions expected to direct their action are contained inside the chemical bases of a DNA chain. At the point when a specific instruction gets to be active the comparing gene is said to be turned on or be communicated. Two main objectives of functional genomics are 1) to utilize genomic signals to characterize disease on a molecular level and 2) to screen for genes that focus particular disease and model their activity in such a route, to the point that normal and abnormal performance can be separated [7]. One of the fundamental undertakings in the investigation of genomes is gene recognition. DNA investigation uses techniques, for example, clustering [20], data mining, and gene recognition and gene administrative system demonstrating [25] [26]. These techniques present cutting edge research points and methodologies with the end goal of encouraging coordinated effort in the middle of analysts and bio-informaticians [9].Genomics is getting to be progressively essential in oncology. The last few decades have seen progressively deeper learning of the pathogenesis of cancer and particular variations in the genome of cancer cells. More than 11 million individuals are diagnosed with cancer consistently [1]. Cancer, referred to medically as a harmful neoplasm, is a large group of diseases including unregulated cell development. In cancer, cells partition and develop wildly, structuring malignant tumors, and attack nearby parts of the body [3][2].Cancer is brought about by abnormalities in the genetic material of the transformed cell. Cancer-promoting genetic irregularities might arbitrarily happen

through errors in DNA replication or are inherited and along these lines exhibit in all cells from birth. The heritability of cancer is typically influenced by intricate interactions in the middle of carcinogens and the host's genome. There is a various classification scheme for the different genomic changes which may contribute to the era of cancer cells. A large portion of these progressions are mutations, or changes in the nucleotide sequence of genomic DNA. Small-scale mutations incorporate point mutations, cancellations, and insertions, which may happen in the promoter area of a gene and influence its articulation or may happen in the gene's coding sequence and change the capacity or steadiness of its protein item [10]. Most cancers originate from arbitrary mutations that grow in body cells among one's lifetime - either as an error when cells are experiencing cell division or in response to wounds from ecological agents, for example, introduction to radiation or chemicals. Numerous researchers are utilizing gene microarray and mass spectrography technology to analyze gene expression data for calculation and grouping of cancer. Nanotechnology is likewise used to create precise and sensitive biomedical devices for tumor genome study [4] Profiling irregular genes may help us better classify cancers and select the optimal treatment. Cancer genomics can help anticipate what may happen in cancer cells and figure out what really happens because of that it concentrates on last gene function. This article, summarize the clinical uses of genomic cancer, briefly presenting basic approach in cancer genomics portraying the clinical applications of these technologies.

A. The main contributions of the paper are

1) The paper discusses two cancer prediction methods which use Genomic Signal Processing (GSP).

2) First method makes use of proposed system through a concept of weighted entropy that takes both entropy and total correlation into consideration. In the second method, decision tree based method for cancer prediction is proposed.

3) For the study, we use of DNA accession nos. AF008216.1, AF348525.1, NM_016346.2, NM_005732.3, AF348515.1, NM_012403.1, AF015224.1, AF083883, AF186607.1, AF186613.1, AF007546 and AF065986, of which first 7 DNA sequences are form the cancer cells and the other 5 are from the normal cells.

The rest of the paper organized as follows: the recent research works is analyzed in section 2; The structural profiles of DNA sequence is described in the section 3; the proposed work briefly explained in section 4; In the 5th section, the experimental results along with the comparison result is depicted and the section 6 represents the summary of the paper.

II. RELATED WORKS

Xingyi Ma et al. [12] had displayed a single nanoplasmonic sensing technology form on restricted surface plasmon resonance for label-free and real-time recognition of exceedingly reliable cancer markers (mutant quality and telomerase) in clinical samples. The sensor particularly recognizes mutant DNA, and can distinguish telomerase from as few as 10 HeLa cells. The methodology could be effortlessly meant recognized other obsessive focuses with high affectability and specificity, and monitor key collaborations between bio- molecules, for example, nucleic acids and proteins amid disease advancement in real time. The system could possibly be further created for on-chip and concurrent investigation of different targets and communications. Seonghwan Lee et al. [13] had presented an early diagnosis stage for prostate cancer; an ultrasensitive electrochemical biosensor taking into account the electro deposition of gold nano particles were planned. The EN2 protein was quantitatively recognized utilizing the electrochemical biosensor, and the computed detection limit was discovered to be 5.62 fM. At last, the specificity and applicability of the biosensor were checked utilizing several proteins and an artificial urine medium. The impedance signals expanded in the instances of EN2, recommending that the system displayed high selectivity to just EN2.P. Filipczuk et al. [14] had reported advanced in computer-aided breast cancer diagnosis taking into account the investigation of cytological images of fine needle biopsies to portray those biopsies as either benign or malignant. As opposed to depending on the exact division of cell nuclei, the nuclei were assessed by circles utilizing the circular Hough transform. The subsequent circles were then sifted to keep only high-quality estimations for further investigation by a support vector machine which characterizes recognized circles as correct or incorrect on the based on the premise of texture features and the percentage of nuclei pixels as per a nuclei mask acquired utilized Otsu's thresholding system. A set of 25 features of the nuclei was utilized as a part of the arrangement of the biopsies by four separate classifiers. The complete analytic system was tried on 737 microscopic images of fine needle biopsies attained from patients and accomplished 98.51% viability. The result analysis showed that a computerized medical diagnosis system taking into account that their system would effective, providing valuable, accurate diagnostic information. Dokyoon Kim et al. [15] had proposed a new graph-based structure that coordinates not only multi-omics information as well as inter-relationship between them for better explaining cancer clinical results. In order to highlight the legitimacy of the proposed system, serous cysta denocarcinoma data from TCGA was received as a pilot undertaking. The proposed model incorporated inter-relationship between distinctive genomic features indicated essentially

enhanced execution contrasted with the model that does not consider inter-relationship when coordinated multi-omics data. For the pair between miRNA and gene expression data, the model incorporated miRNA, for instance, gene expression, and inter-relationship between them with an AUC of 0.8476 (REI) beat the model combining miRNA and gene expression data with an AUC of 0.8404. Comparable results were additionally acquired for different pairs between diverse levels of genomic data. Integration of diverse levels of data and inter-relationship between them could support in removing new biological knowledge by making an integrative determination from numerous pieces of data gathered from various sorts of genomic data, inevitably prompting more powerful screening techniques and alternative therapies that may enhance results.

Seshacharyulu et al. [16] had condensed the part of a several miRNAs that control different oncogenes (KRAS) and tumor suppressor genes (p53, p16, SMAD4, and so forth.) included in PC improvement, their prospective roles as diagnostic and prognostic markers and as a therapeutic targets. Micro RNAs (miRNAs) were short (19–24 nucleotides) non-coding RNA particles embroiled in the regulation of gene representation at post-transcriptional level and play critical parts in different physiological and pathological conditions. Deviant representation of miRNAs had accounted for in a several cancers including PC and was concerned in PC pathogenesis and progression, proposing their utility in diagnosis, prognosis and therapy.

Chih-Wen Lin et al, [17] had implemented a reusable biosensor based on a magnetic graphene oxide (MGO)- modified Au electrode to identify vascular endothelial development component (VEGF) in human plasma for cancer analysis. In that biosensor, Avastin was utilized as the particular bio- recognition element, and MGO was utilized as the carrier for Avastin stacking. The utilization of MGO empowered rapid purification because of its magnetic properties, which prevents the loss of bioactivity. The advantages of the Avastin-MGO- modified biosensor for VEGF location was that it gives a proficient recognition methodology that enhances the identification capacity as well as diminishes the expense and reductions the reaction time by 10-fold, demonstrating its potential as a diagnosis product.

BichenZheng et al. [18] had suggested a feature based cancer diagnose system. To concentrated useful data and diagnosed the tumor, a hybrid of K-means and support vector machine (K-SVM) algorithms was produced. The K- means algorithm was used to perceive the concealed patterns of the benign and malignant tumors independently. The membership of every tumor to those patterns was ascertained and treated as a new feature in the training model. At that point, a support vector machine (SVM) was utilized to get the new classifier to separate the approaching tumors. Based on 10-fold cross acceptance, the proposed methodology enhanced the precision to 97.38%, when tested on the Wisconsin Diagnostic Breast Cancer (WDBC) data set from the University of California – Irvine machine learning archive. Six unique tumor features were extricated from the 32 unique features for the training phase. The results not just delineate the ability of the proposed approach on breast cancer diagnosis, additionally show time savings amid the training phase. Doctors could also profit from the mined abstract tumor features by better understanding the properties of diverse sorts of tumor.

George et al. [19] had developed a completely automated strategy for cell nuclei identification and segmentation in breast cytological images. The areas of the cell nuclei in the image were identified with circular Hough transform. The elimination of false-positive (FP) discoveries (uproarious circles and platelets) was attained to utilizing Otsu's thresholding strategy and fuzzy c-means clustering method. The segmentation of the nuclei boundaries was fulfilled with the application of the marker-controlled watershed transform. Next, a shrewd breast cancer classification framework was created. Four classification models were utilized, specifically, multilayer perceptron utilizing back- propagation algorithm, probabilistic neural network (PNN), learning vector quantization, and support vector machine (SVM). The classification results were attained utilizing tenfold cross validation. The execution of the systems was compared at in view of resulted error rate, correct rate, affectability, and specificity.

III. STRUCTURAL PROFILES OF DNA SEQUENCE

Deoxyribo Nucleic Acid (DNA) is the significant chemical present in the nucleus of all cells. It makes up chromosomes which is responsible for passing on the genetic information from parent cell to offspring during reproduction. The most important function of DNA is to provide instructions for protein synthesis. After the sensational discovery of double helix structure of DNA by Watson & Crick, researchers from all fields have focused their attention in this particular field of biology keeping in view the vast information content and functional importance of DNA. A DNA is a double helix structure consisting of two complementary strands of sugar-phosphate group with bases attached to it.A DNA sequence is made up of nucleotides, which can be distinguished by the four bases: Adenine (A), Thymine (T), Cytosine (C) and Guanine (G). Thus, a DNA sequence can be formally viewed as a symbol string, consisting of the four alphabet characters (A, C, G, and T).Figure 1 and 2 shows DNA and A, C, G, and T structure description.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor :6.887
Volume 6 Issue I, January 2018- Available at www.ijraset.com

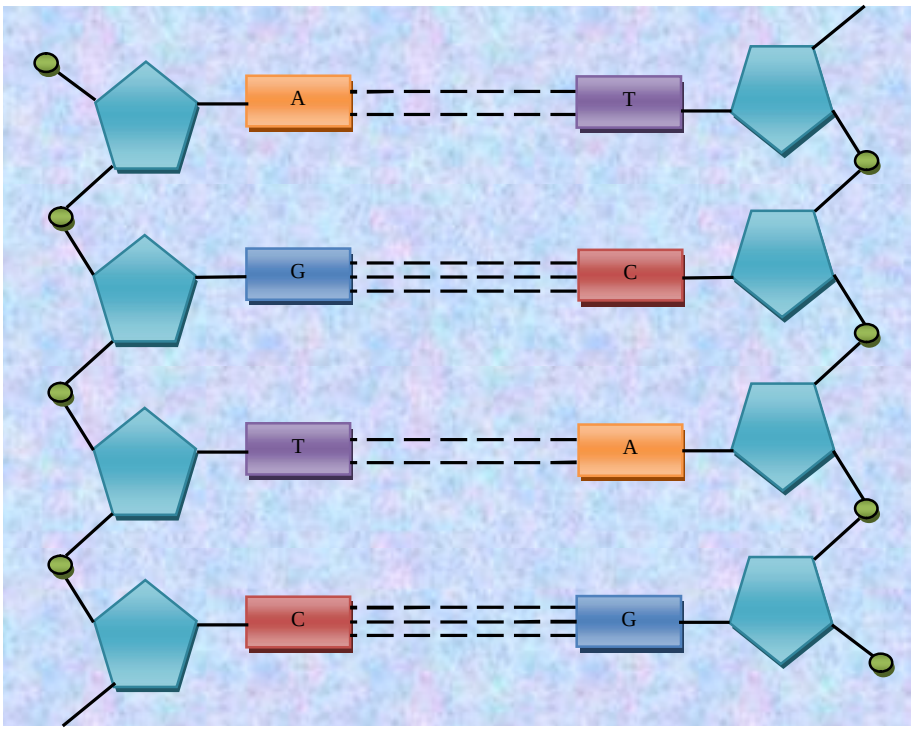


Figure 1: Structure for DNA

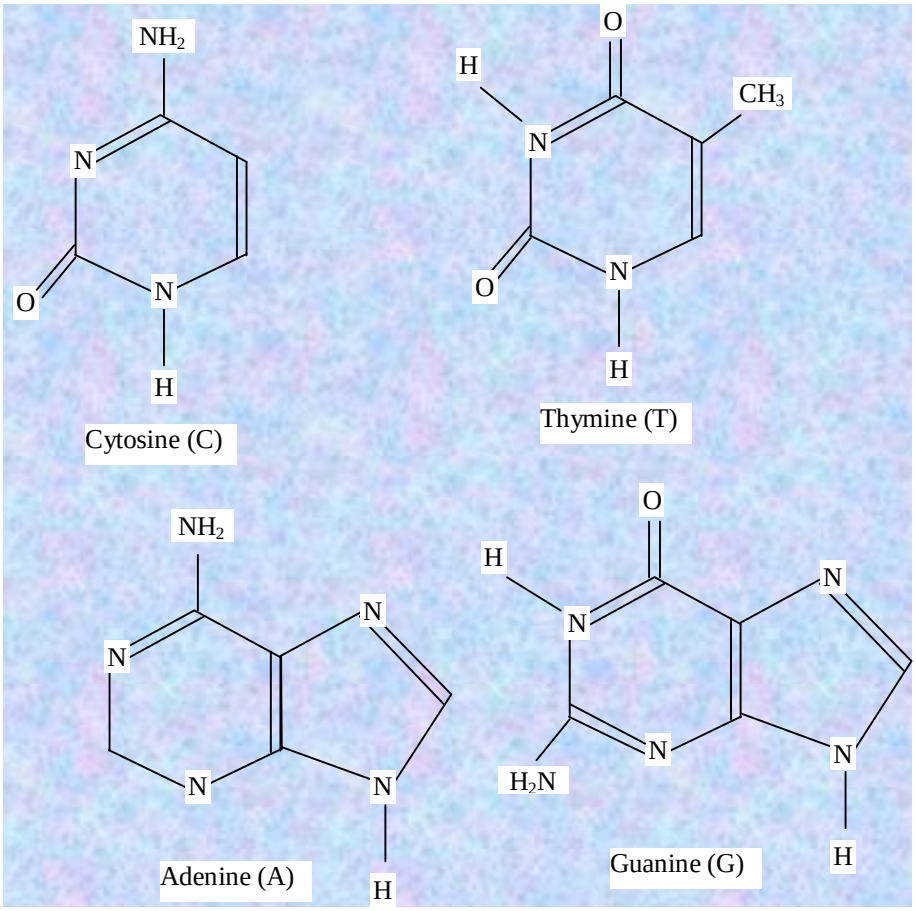


Figure 2**:** Structure for Adenine (A), Thymine (T), Cytosine (C) and Guanine (G)

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor :6.887
Volume 6 Issue I, January 2018- Available at www.ijraset.com

IV. METHODS FOR CANCER PREDICTION FROM DNA SEQUENCE

In this section, we describe the proposed cancer prediction from DNA sequence using two methods. In the first one, the weighted entropy [20] is found out directly for the DNA sequence in order to predict for cancer. In the second method, we develop a decision tree based method to predict cancer prediction.

A. Weighted Entropy Based Prediction

The first method to predict the cancer cell from the DNA sequence is based on the weighted entropy value. In this method, a system is proposed via anefficient concept of weightedentropy [20] that takes both entropy and total correlation into consideration.Here, there is no need for transformation and the calculation happens in the time domain, which is to find the entropy of the DNA sequence and predict the cancer based on the weighted entropy value.

Initially, the unique groups having a length l are found out from the input DNA sequence $D[n]$ having length k. The DNA sequence consists of four letters where A stands for Adenine, T stands for Thymine, G stands for Guanine and C stands for Cytosine.

Let the number of unique groups be represented by N and in our case l is taken as 4. Consider the input DNA sequence example as shown in equation (1)

$$D[n] = ATTCGATTCTCATTC \quad (1)$$

So the first task is to find the unique groups having length l, which is taken as 4 in our case. The unique groups for the example in equation (1) are given in equation (2):

$$U = \{ATTC, TTCG, TCGA, CGAT, GATT, TTCB, TCTC, CTCA, TCAT, CATT\} \quad (2)$$

So the number of unique terms N obtained for the example is 10.

To weight the entropy of each U_i (where $0 < i \le N$), we propose to employ a reverse sigmoid function of the entropy, as follows:

$$w(a_i) \rightarrow 2\left(1 - \frac{1}{1 + \exp(-H_x(a_i))}\right) \quad (3)$$

This reverse sigmoid is a decreasing function ranging between (0, 2).

Subsequently for every U_i (where $0 < i \le N$) the probability is found out. The probability is found out using the equation (4):

$$P(a_i) = \frac{Number\ of\ times\ a_i\ appears\ in\ D[n]}{C_i} \quad (4)$$

Where, a_i is the attribute, C_i is the number of comparisons made (in the sliding window style).

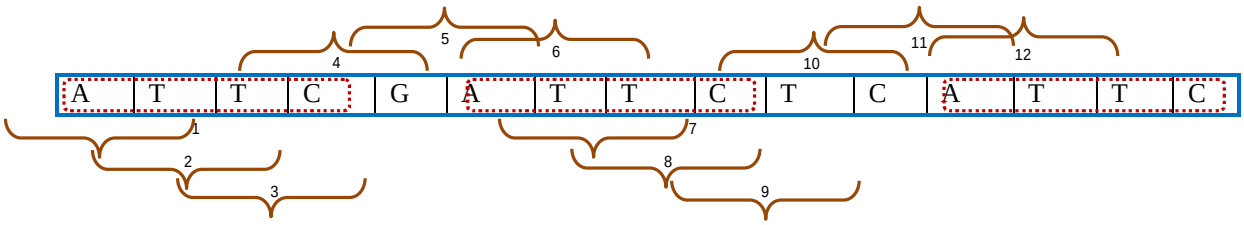


Figure 3:Example showing the probability calculation

For the example in equation (4), probability of the first unique term will be as in equation (5):

$$P(TTCG) = \frac{Number\ of\ times\ TTCG\ appears\ in\ D[n]}{C_{TTCG}} \quad (5)$$

The number of times the term appears and the number of comparisons are clear from the figure 3. The method involves finding the entropy value $E(X)$ for the input DNA sequence. The entropy is a measure of the uncertainty associated with a random variable and is given by the formula in equation (6):

$$H_x(a_i) \rightarrow -\sum_{j=1}^{N} P(a_i)\log_2 P(a_i) \quad (6)$$

Using equation (3) and (6), we derive the weighted entropy $W_x(A)$ which is the sum of the weighted entropy on each attribute of the DNA sequence.

$$W_x(A) = \sum_{i=1}^{m} w(a_i)H_x(a_i) \quad (7)$$

$W_x(A)$ stands for the weighted entropy value of x^{th} input sequence. Similarly, for every sequence the weighted entropy is calculated and is used for prediction of cancer by fixing a threshold T_e. When the entropy value is above the threshold, cancer is predicted else it's a normal cell.

$$when \quad \begin{matrix} E_i > T_e \text{ , Cancer cell is predicted} \\ E_i < T_e \text{ , Normal cell is predicted} \end{matrix} \quad (8)$$

Figure 4 shows the flow diagram of weighted entropy based prediction.

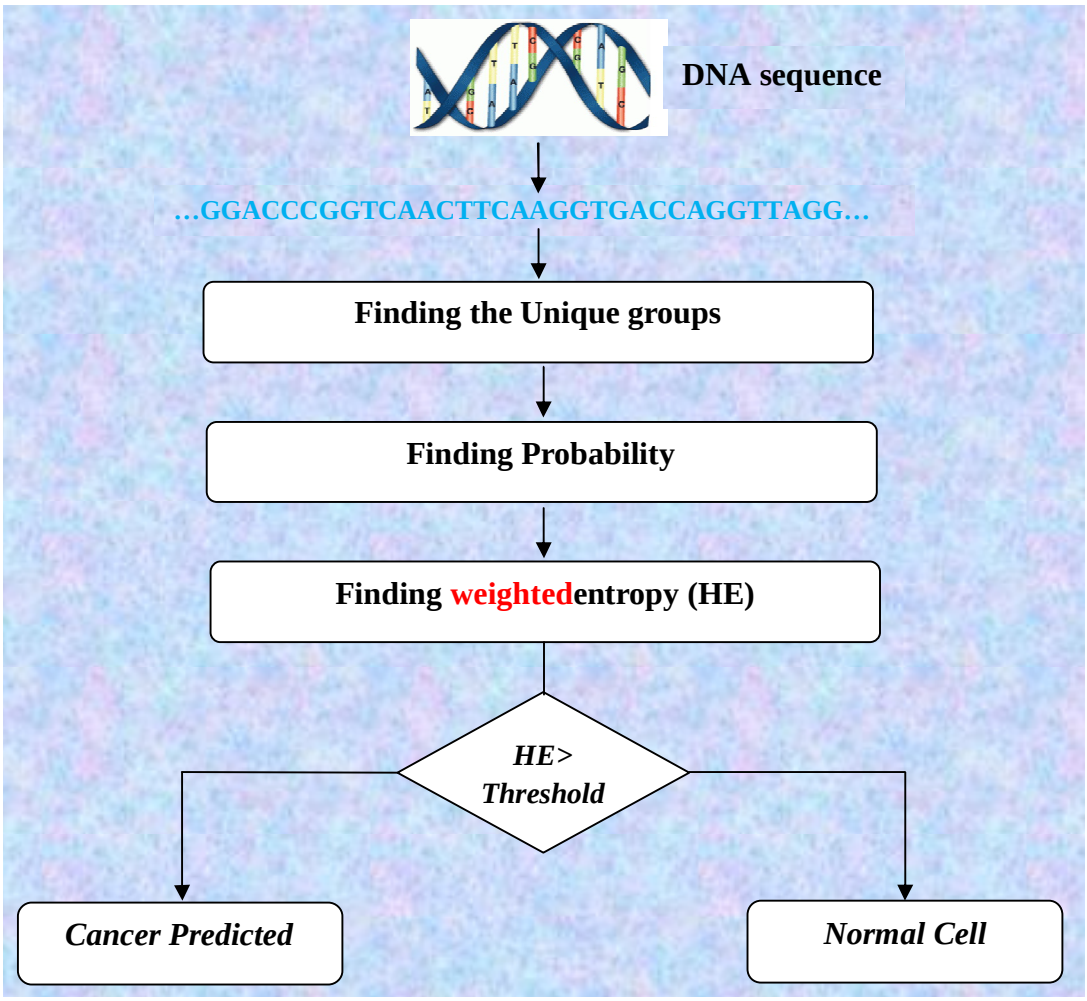


Figure 4: Flow diagram of weighted entropy based prediction

B. Decision Tree and Weighted Entropy Based Prediction

In this section, we develop a decision tree based cancer prediction on DNA sequences. Each node R of a decision tree is associated with an attribute A_{jn} and a split point $s_n \in dom(A_{jn})$. An internal node has two children, which are labelled "left' and "right' respectively. Each leaf node r in the decision tree is associated with a probability distribution P_r over C. For each $c \in C$, $P_r(c)$ gives a probability reflecting how likely a tuple assigned to leaf node r would have a class label of c.

Accordingly, at first, the decision tree algorithm checks whether all the tuples possess same class label, then all the tuples will be grouped into one leaf node. If the classes are different, a split function is utilized to put the tuples into left of right of the leaf node. The split function s_n uses the mean or median value of the corresponding tuple to plot it into either left or right of the leaf nodes. We can consider the mean or median as the decision parameter x, then,

$$if\ (x > s)$$
$$put\ it\ left\ node$$
$$else$$
$$put\ \ it\ right\ node$$

Now, we calculate the probability of the nodes in left and right with respect to the class labels. The probability can be subjected as $P_r(c)$ and the tuples with higher probability regarding a class label then the tuple is assigned to that class. Thus we obtain the probability range of all the tuples corresponding to that particular input. We set an average probability of classes $P_r(c)_{avg}$ for each class c_i in set C. This will be used to classify an unknown DNA sequence data.

Our algorithm works as follows: Starting from the region of the entire DNA sequence, a decision tree is constructed to recursively find the best binary split point on the residues in this DNA sequence region. At each iteration, we utilize the weighted entropy based measure that is efficiently used in decision tree algorithm to evaluate all possible split points and find the best split those results in minimized weighted entropy [20].

$$W_x(A) = \sum_{X=L,R} \frac{|X|}{S}\left[\sum_{i=1}^{m} w(a_i) H_x(a_i)\right] \quad (9)$$

Where,

$$H_x(a_i) \rightarrow -\sum_{j=1}^{N} P(a_i)\log_2 P(a_i) \quad (10)$$

$$w(a_i) \rightarrow 2\left(1 - \frac{1}{1+\exp(-H_x(a_i))}\right) \quad (11)$$

Each iteration determines a best binary split with minimum weighted entropy and splits the current region into two sub-regions. This is the same as generating two decision tree nodes that split the data from the parent node.

In the testing phase of the decision tree, an unknown test tuple is given to the trained decision tree. The test tuple will be in the similar format like a training tuple but with the class label field empty or unknown. The test tuple t_{test} is given to the decision tree algorithm and the split function plot it into either left of right. Then correspond probability is compared with the probability of test tuple. As per the obtained probability value, the test tuple is plotted to the corresponding class. If a single class label is desired as the result, we select the class label with the highest probability as the final answer.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor :6.887
Volume 6 Issue I, January 2018- Available at www.ijraset.com

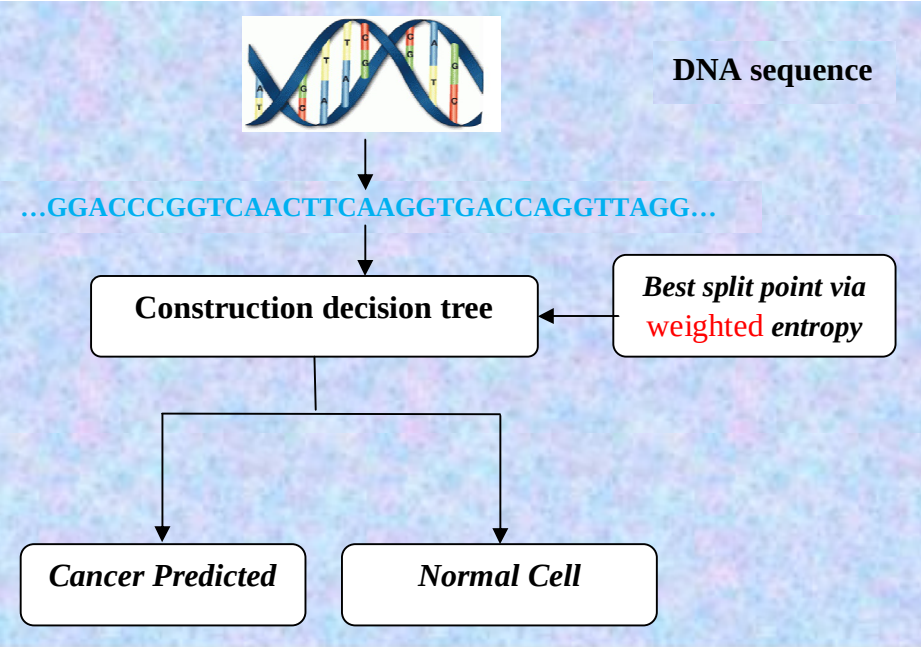


Figure 5: Flow diagram of decision tree based prediction

V. RESULT AND DISCUSSION

We have offered the results of our suggested methodology and have examined their presentation in this part. The proposed cancer diagnosis is implemented in the MAT LAB program and the cancer diagnosis is experimented with the Gene bank. For implementing the proposed technique we have used Mat lab version (7.12) and the technique is done in windows machine having Intel Core i5 processor with speed 1.6 GHz and 4 GB RAM.

A. Evaluation metrics

The evaluation of proposed cancer diagnosis using genomic signals are carried out using the following metrics as suggested by below equations,

1) *Sensitivity:* The Sensitivity is the proportion of true positives that are correctly identified by a diagnostic test. It shows how good the test is at detecting a disease. This relation can be expressed as,

$$S_t = \frac{T_p}{T_p + F_n} \quad (12)$$

2) *Specificity:* The specificity can be evaluated by taking the relation of number of true negatives to the combined true negative and the false positive. It suggests how good the test is at identifying normal (negative) condition. The specificity can be expressed as,

$$S_p = \frac{T_n}{T_n + F_p} \quad (13)$$

3) *Accuracy:* The Accuracy is the proportion of true results, either true positive or true negative, in a population. It measures the degree of veracity of a diagnostic test on a condition.

The accuracy can be described by the following equation

$$A = \frac{T_p + T_n}{T_p + F_p + F_n + T_n} \quad (14)$$

Where,

T_p → True positive

T_n →True negative

F_p → False positive

F_n →False negative

B. Experimental Result

The basic idea of our research is to diagnosis the cancer signal using Holo entropy with decision tree. Here, the cancer diagnosis method is based on Holo entropy. The figure 6 shows the sequence with cancer signal and figure 7 shows the sequence with non-cancer signal. For evaluation of methods, we make use 12 DNA sequence having accession numbers AF008216.1, AF348525.1, NM_016346.2, NM_005732.3, AF348515.1, NM_012403.1, AF015224.1, AF083883, AF186607.1, AF186613.1, AF007546 and AF065986 [14]. Of these DNA sequences, first 7 are form the cancer cells and the other 6 are from the normal cells.

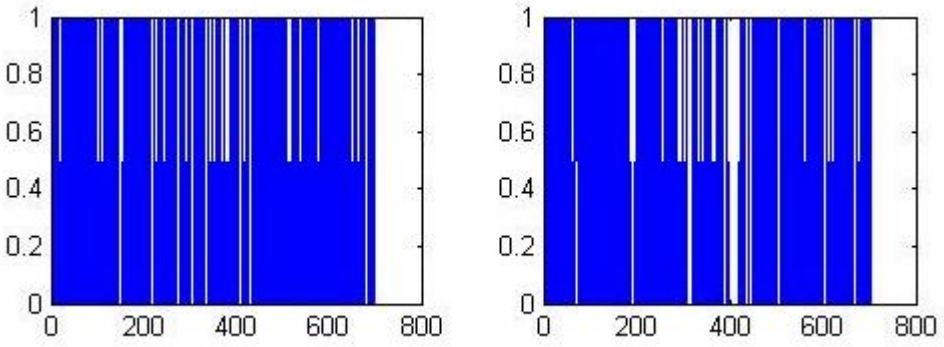


Figure 6: Sequence with cancer signal

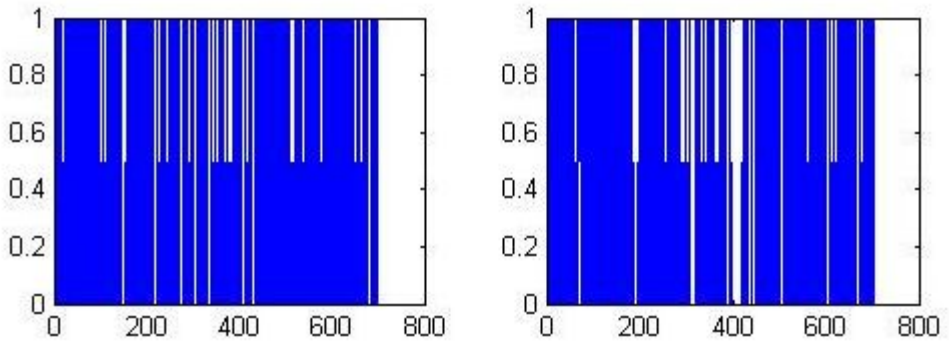


Figure 7: Sequence with non-cancer signal

C. Comparative analysis of Proposed Approach

The diagnostic performance of a proposed system was assessed using the receiver operating characteristic (ROC) curve. The performance of cancer signal diagnosis is often measured in terms of accuracy, sensitivity and specificity which are most significant performance parameters. In signal classification first we pre-process the signal which signal is used for further processing. After that we calculate the signal based on the proposed approach. The performance is made up of two process training and testing process. Depending on the training and texting process we calculate the accuracy and that accuracy value shows the efficiency of our proposed work. In training and testing process we use the k-fold cross validation process for improving the accuracy of the system. The k-fold cross validation is a common technique for estimating the performance of a classifier. The performance comparison of proposed work is explained in the figure 8 to 10.

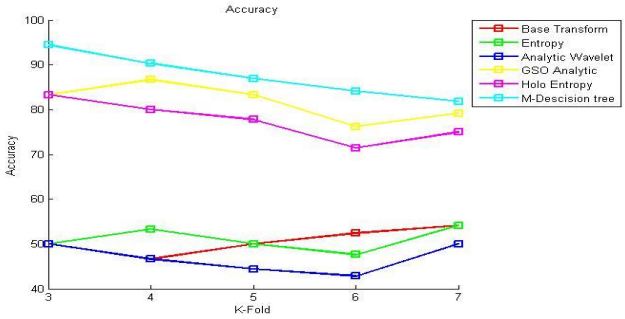


Figure 8: Performance comparison of the accuracy plot

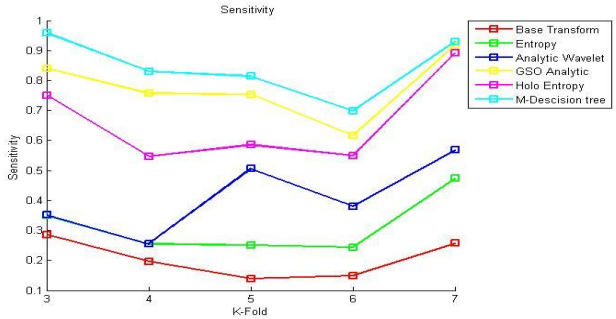


Figure 9: performance comparison of the sensitivity plot

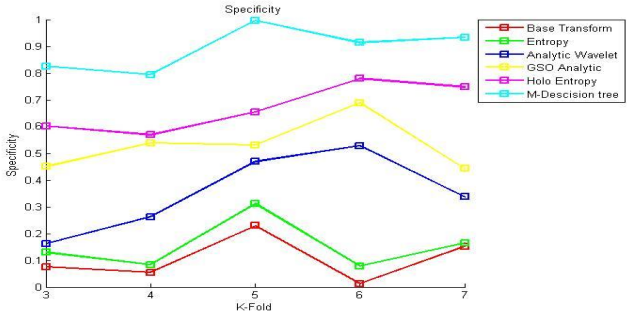


Figure 10: performance comparison of the specificity plot

The methods base transform, Entropy, Analytic wavelet, GSO analytic and Holo entropy are the best among existing schemes for cancer diagnosis. Furthermore, they characterize local details of the signal, feature analysis representation. Therefore, we have chosen to compare the performance of our proposed algorithm against that of these ones. From the figure 8 to 10, one can observe that the Holo entropy, GSO analytic and our proposed one yield the best performances followed by base transform, Entropy and Analytic wavelet. This is because of that these methods very well describe the feature of signal. Compare to these three methods our proposed method (Holo Entropy with Modified Decision Tree) is slightly better than other methods. When analyzing figure 8, our proposed approach achieves the maximum accuracy of 95% which is better than all the other approaches. In the same figure, Base transform obtained the accuracy of 52%, accuracy of entropy approach 53%, accuracy of 50% for analytic wavelet, accuracy of 85% for GSO entropy and accuracy of 82% for Holo entropy. In figure 9, we obtain the maximum sensitivity of 98% which value is better than the existing works. In figure10 shows the performance comparison of specificity plot. Here, when we take the 50% of signal for training and testing we obtain the maximum specificity value of 99%.

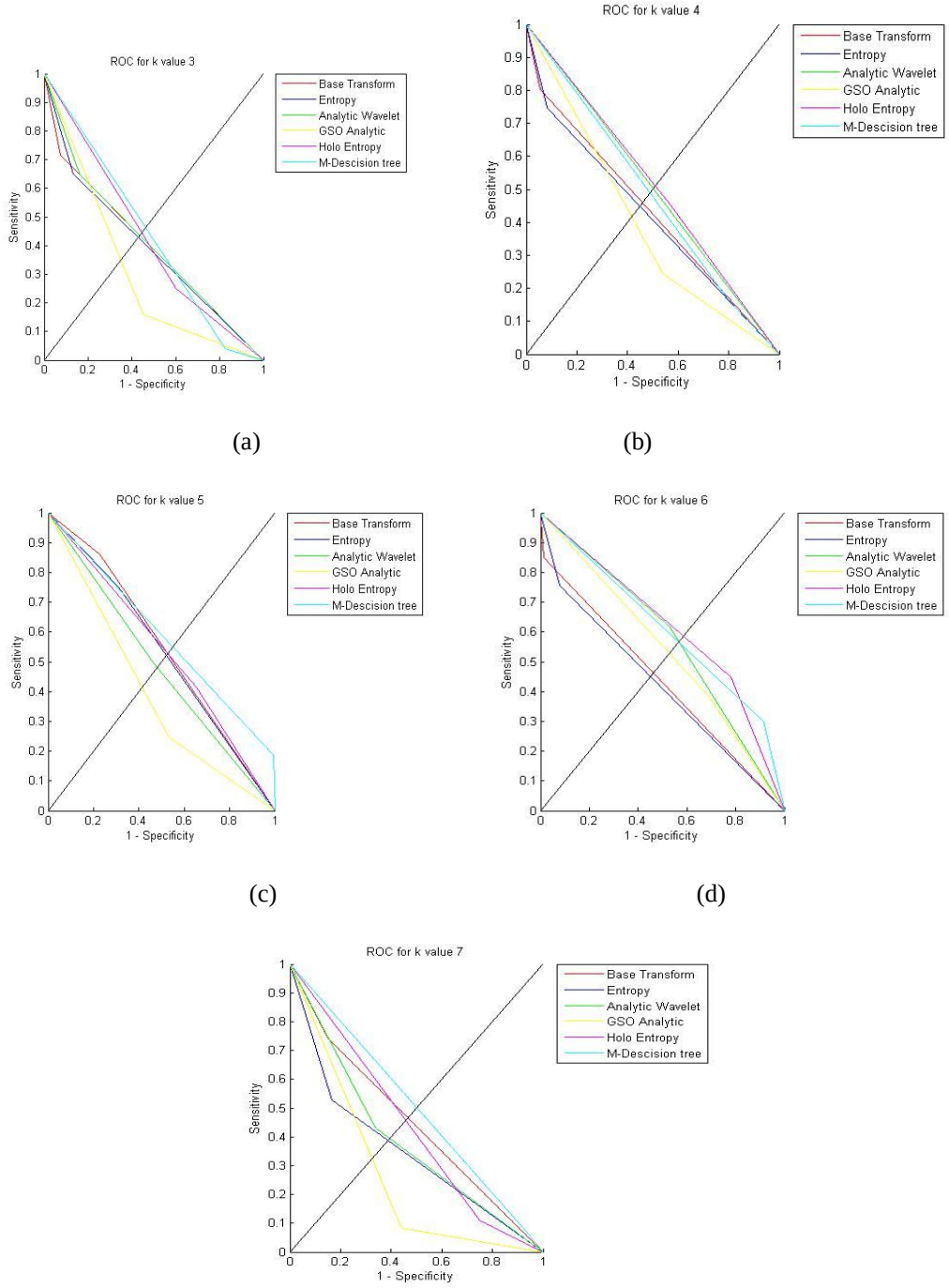


Figure 11: k-fold validation performance of the ROC curve

The ROC curve of this test is a straight diagonal from the lower left- hand corner to the upper right-hand corner of the graph. The region under this 'curve' is 0.5 (50% of total area). The ROC contains the optimal operating point (i.e., TPF=1 and FPF=0), corresponding to the upper left-hand corner of the ROC graph. The area under this ROC curve is 1.0 (100% of the total area). Consider the figure 11(a), the cut off value shows that sensitivity of 0.49 and specificity of (1-specificity) 0.51 for the proposed approach. But in case of entropy and analytical wavelet approaches are obtain the similar sensitivity and specificity value for K=3. However, in figure 11(b), when using k=4 Holo entropy and analytical wavelet are slightly better than the proposed approach. In figure 11(c), our proposed approach Holo entropy with modified decision tree achieves the maximum sensitivity and specificity

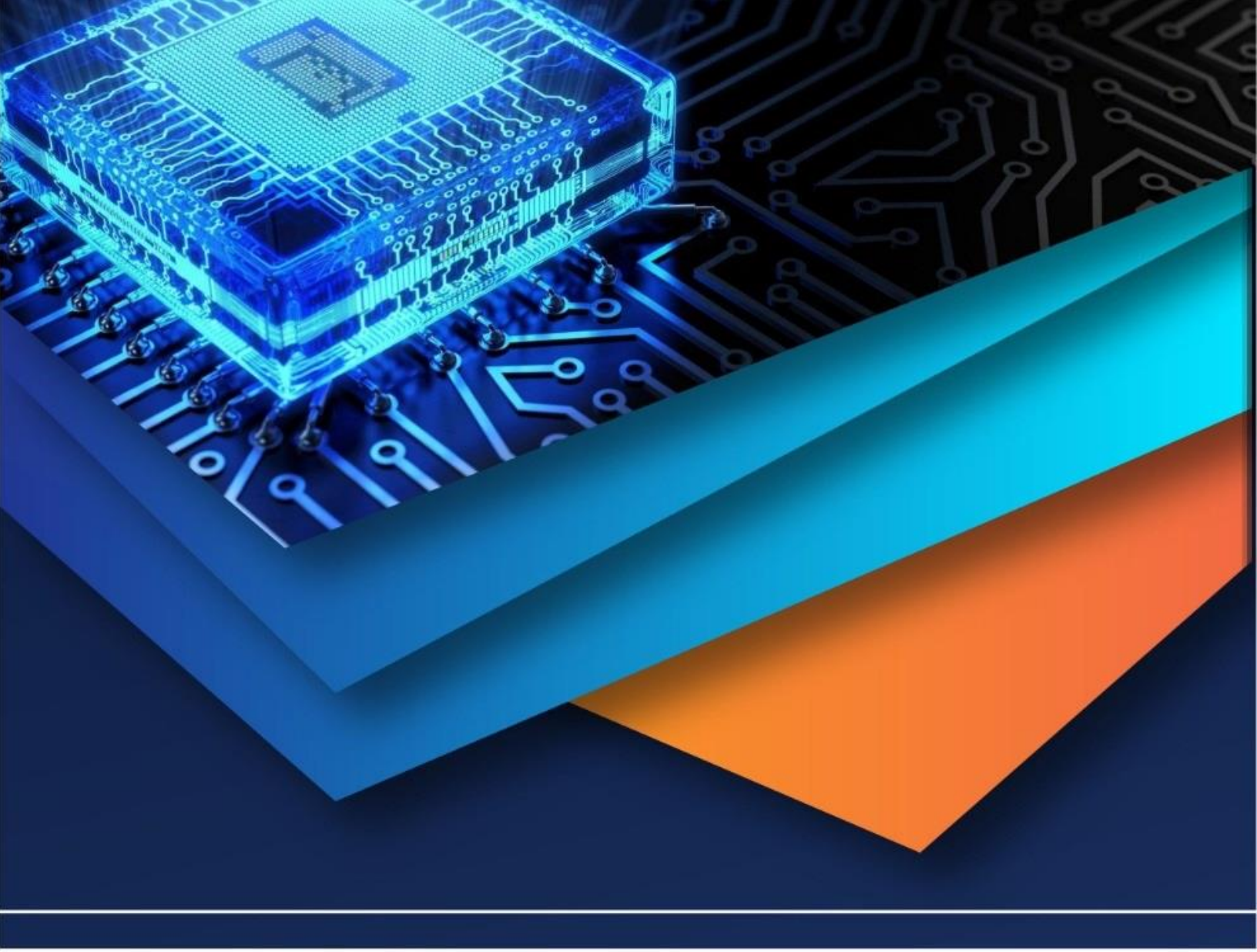
value. In the same way, figure 11(e) shows the performance of the K- fold validation using K=7. Here also our proposed approach obtains the good result compare to the existing works.

VI. CONCLUSION

In this paper, we discussed two systems to predict cancer based on the signal processing of the DNA sequence. A system is proposed in the first method to predict cancer from DNA sequence, through an efficient concept of weighted entropy that takes both entropy and total correlation into consideration. In the second method, a decision tree based cancer prediction was proposed, where a split point measure was used namely weighted entropy for pruning the decision tree. For the analysis, we make use of DNA sequences obtained from National Center for Biotechnology Information (NCBI). Evaluation metrics parameters of sensitivity, specificity and accuracy were found out and compared to the existing literature. From the results obtained, we can infer that all our methods have attained better evaluation metrics parameter values than the existing approach. Both the weighted entropy based method and the decision tree based method yielded the best results with an accuracy of 91.6%. Weighted entropy method gave 66.6% accuracy when compared to 58.3% for the existing literature.

REFERENCES

[1] Anant Narayan Bhat, RohitMathur,AbdullahFarooque, AmitVerma and Dwarakanath, “cancer bio markers –current perspcectives”, Indian Journal of Medical Research, pp. 129-149, 200
[2] Cancer Research Institute, UK www.cancer.org/index
[3] Hariprasad, SaneeshCleatus, Anjali Chitaranjan,Ashwini Data and Monisha Ganesh., “Novel approach on cancer detection “, Proceedings of Asar International Conference, 14th may-2014, mysore, india,
[4] Barman (Mandal), M.Roy, S.Biswas and Saha, ” Prediction of Cancer Cell Using Digital Signal Processing”, Interational Journal of Engineering, Vol.1, pp. 91-95, 201
[5] Swapnoneel Roy, MinhazurRahman, and Ashok Kumar Thakur, "Sorting Primitives and Genome Rearrangement in Bioinformatics: A Unified Perspective", World Academy of Science, Engineering and Technology, Vol.3, No.38, 2008
[6] Rames, Prabhu Shankar, Mallikarjunaswamy and Puttaiah, “Genomic Signal Processing (GSP) Of Rheumatic Arthritis (RA) Using Different Indicator Sequences”, International Journal of Computer Science and Mobile Computing, Vol. 2, No. 5, pp.332 – 337, 201
[7] AkankshaKulshreshtha,.Karthika and ArchanaTiwari, “Genome Signal Analysis: An Approach For Predicting Hivi Variability Based on Phase Analysis of Protease Gene”, International journal of pharma and bio sciences, Vol. 4, No. 4, pp. 519-535, 2011
[8] Sathish Kumar and Duraipandian,” An Effective Identification of Species from DNA Sequence: A Classification Technique by Integrating DM and ANN”, International Journal of Advanced Computer Science and Applications, Vol. 3, No.8, 2012
[9] PengQiu, Wang Z.Jane, and K.J. Ray Liu, “Genomic Processing of Cancer Classification and Prediction”, IEEE Signal Processing Magazine, 2007
[10] Jeremy Volkening and Stephen Spatz "Purification of DNA from the cell-associated herpesvirusMarek’s disease virus for 454 pyrosequencing using micrococcal nuclease digestion and polyethylene glycol precipitation", Journal of Virological Methods, Vol. 157, pp. 55– 61. 2009.
[11] Xingyi Ma, Phuoc Long Truong, Nguyen Hung Anha and Sang Jun Sima, “Single gold nanoplasmonic sensor for clinical cancer diagnosis based on specific interaction between nucleic acids and protein”, Biosensors and Bioelectronics, Vol. 67, pp. 59–65, May 2015,
[12] Seonghwan Lee, Hunho Jo, Jin Her, Ho Yong Lee and Changill Ban, “Ultrasensitive electrochemical detection of engrailed-2 based on homeodomain-specific DNA probe recognition for the diagnosis of prostate cancer”, Biosensors and Bioelectronics, Vol. 66, No.15, pp.32–38, 2015
[13] P. Filipczuk, T. Fevens, A. Krzyzak and R. Monczak, “Computer-Aided Breast Cancer Diagnosis Based on the Analysis of Cytological Images of Fine Needle Biopsies”, IEEE Transaction on Medical Imaging, Vol.32, No. 12, pp. 2169 - 2178, 2013
[14] Dokyoon Kim, Hyunjung Shin, Kyung-Ah Sohn, AnuragVerma, Marylyn D. Ritchie and Ju Han Kim, “Incorporating inter-relationships between different levels of genomic data into cancer clinical outcome prediction”, Methods, Vol. 67, No. 3, pp. 344–353, 2014
[15] Seshacharyulu, DhanyaHaridas, SeemaChugh and Surinder K. Batra, “Clinical implications of miRNAs in the pathogenesis, diagnosis and therapy of pancreatic cancer”, Advanced Drug Delivery Reviews, Vol. 81, pp. 16–33, 2015
[16] Chih-Wen Lin, Kuo-Chen Wei, Shih-sheng Liao, Chiung-Yin Huang, Chia-Liang Sun, Pei-Jung Wu, Yu-Jen Lu, Hung-Wei Yang and Chen-Chi M. Ma, “A reusable magnetic graphene oxide-modified biosensor for vascular endothelial growth factor detection in cancer diagnosis”, Biosensors and Bioelectronics, Vol. 67, pp. 431–437, 201
[17] BichenZheng, Sang Won Yoon and Sarah S. Lam, “Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms”, Expert Systems with Applications, Vol. 41, No. 4, pp. 1476–1482, 2014
[18] Y.M .George, H.H. Zayed, M.I. Roushdy and B.M. Elbagoury, “Remote Computer-Aided Breast Cancer Detection and Diagnosis System Based on Cytological Images”, IEEE Transaction on Systems, Vol.8 ,No.3, pp. 949 – 964, 2013
[19] ShuWu,Shengrui Wang, "Information-Theoretic Outlier Detection for Large-Scale Categorical Data," Knowledge and Data Engineering, Vol.25, no.3,589 - 602,2013.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY