International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-4, November 2019

Detection of Huntington's Disease in Human DNA Sequence using Numerical Encoding Method and Machine Learning based Classifier

G. Tamilpavai, C. Vishnuppriya

Abstract— Amino acids are little bio-particles with different properties. The capacity to ascertain the physiochemical properties of proteins is pivotal in many research regions, for example, tranquilize plan, protein displaying and basic bioinformatics. The physiochemical properties of the protein decides its collaboration with different atoms and subsequently its capacity. Foreseeing the physiochemical properties of protein and translating its capacity is of extraordinary significance in the field of medication and life science. The point of this work is to create python based programming with graphical UI for anticipating the physiochemical and antigenic properties of protein. Thus the instrument was named as ASAP-Analysis of protein succession and antigenicity expectation. ASAP predicts the antigenicity of the protein succession from its amino corrosive arrangement, in light of Chou Fasman turns and antigenic file. ASAP computes different physiochemical properties that is required for invitro tests. ASAP utilizes standardization esteems that expansion the affectability of the apparatus.

Keywords: Amino acids, antigenicity, normalization and Protein modeling.

I. INTRODUCTION

Neurodegenerative disorder is a disease, which causes damage in nervous system of human [1]. Neurons in brain and spinal cord are very important for the activation of healthy nervous system. Generally, an affected or damaged neuron cannot be replaced by human body, which would die to do its functionalities. This condition leads to neurodegenerative disorders like HD, Parkinson's disease, Alzheimer's disease, Spinal muscular atrophy etc [2]. Nowadays treatment of these diseases can be possible by replacement of a dead cell with stem cells [3]. But according to neuroscientist this replacement is a hard task and it cannot be work for all diseases or all patients. Hence, medical scientists and other researchers look forward to curing or to give earlier solution for these kinds of diseases.

Manuscript published on November 30, 2019.
* Correspondence Author
G. Tamilpavai*, A Department of Computer Science and Engineering, Government College of Engineering, Tirunelveli, Tamil Nadu, India.(Email: *:* tamilpavai@gcetly.ac.in)
C. Vishnuppriya, Department of Computer Science and Engineering, Government College of Engineering, Tirunelveli,, Tamil Nadu, India.(Email: c.vishnuppriya@gmail.com.)

According to M. Gourie Devi, 2014 [4], people affected by neurodegenerative disorder in India is an average of 2394 per 100000 population. Statistics of neuro degenerative diseases in America is reported by Harvard Neuro Discovery Center as follows, 5 millions are affected by Alzheimer's disease, 1 million affected by Parkinson's disease, 30,000 people affected by HD [5]. HD is a genetic brain disorder, which causes the damage to nerves in brain. Parkinson's disease is a disorder related to movement of human. Alzheimer's disease is a brain disorder which causes memory loss. Spinal muscular atrophy is a disorder related to motor neurons which creates muscle weakness [6].

In 1872, George Huntington provided the description of HD. HD is an autosomal dominant genetic disorder. HD is associated with a gene called Huntingtin (HTT) which is located in 4th chromosome. The number of Cytosine, Adenine and Guanine (CAG) repeats (Trinucleotide repeats) present in HTT gene is used to find whether the human have risk of HD or not. Jerky movements, psychiatric problems are the symptoms of HD. These symptoms of HD are explicitly aware by human only between 30 to 45 years of age. At this stage all HD patients may have children possibly with mutant HTT gene. This mutant HTT is having 50 % chance of pass into next generation. Similarly, HD may occur in all next generation because it is an autosomal dominant disorder [7]. Hence, regular checkup is necessary to all human irrespective of age, to ensure the negative of HD.

L.A. Barboza and N.C. Ghisi (2018) [8] evaluated and reported the current stage of research on Huntington's disease. Authors reported that, more number of researches was going in United States. Out of 92 countries, India took place in 15th position on Huntington's disease research. Due to this, rapid developments are required to save HD patients life in earlier stage.

In this proposed work, computational oriented analysis is focused on human DNA sequence to ensure the positive (presence) and negative (not presence) of HD. The proposed work aims to give the strong prediction of HD, further it helps to neurologist for treating HD patients.

Rest of the paper is organized as follows. Section 2 discuss about related works. Section 3 discuss about methodology used in this proposed work. Section 4 discuss about experimental results. Section 5 discuss about conclusion and future enhancement.

II. RELATED WORKS

A. *Literatures Related to HD Analysis*

Generally neurodegenerative disorder causes very serious problem in brain memory and person's thinking ability [3]. For understanding disease nature, drug targets and signaling pathways, Natasa A Kablar (2019) [9] reviewed the

Parkinson's disease, Alzheimer's disease and HD. Davina J

Hensman Moss et al. (2017) [10] made genome wide association study and identified progression score of HD. Srimanta Pramanik et al. (2000) [11] analyzed CAG and CCG repeats in normal and unrelated HD affected individuals of India. Authors suggested that Indian populations were less affected than the western population. In spite of that they reported, it is necessary to identify the origin of HD mutation.

Johanna Craig (2008) [12] reviewed the complex diseases such as Alzheimer's disease, asthma, Parkinson's disease, etc to find the exact risk factor associated with the corresponding disease. These diseases were caused by genetic abnormality, environmental and other life style factors. Author noted that actually the disease contained samples were rarely available but vast amount of samples are required to perform statistical analysis on the complex disease.

Peggy C. Nopoulos (2016) [13] stated that, sometimes symptoms of HD may not be aware by a person those who affected by HD. This is the severity of HD. Author reported that, the age onset is inversely related to CAG repeat count of HD. This is shown in Table I. Finally author stated that, gene therapy would help for prevention of disease, however still there is a powerful study is required to treat the abnormal condition of HD.

Miao Xu and Zhi-Ying Wu (2015) [14] reviewed the HD patient's epidemiology, clinical characteristics, genotype, phenotype and treatment progression among Asian populations. Babu Srija et al. (2016) [15] reviewed HD based analysis for Indian people. Author reported that incidence of HD is less in India when compared to the western countries. At the same time HD based research also less in India.

Cynthia T.McMurray (2010) [16] reviewed mechanisms of several trinucleotide repeats during human development. Author categorized CAG repeat counts for HD disease as follows, (i) 6 to 29 for normal (ii) 29 to 37 for premutation of HD (iii) 37 and above for HD. This categorization is used in this proposed work.

Table- I: CAG repeat count and age onset details

S. No.	Age onset	CAG repeat count
1	1 to 10 (i.e. childhood HD)	>80
2	11 to 20 (i.e. Adolescent HD)	>60
3	21 to 30 (i.e. Early HD)	>50
4	30 to 55 (i.e. Adult HD)	40 to 49
5	55 to 80 (i.e. Late HD)	36 to 39

B. *Literatures Related to Numerical Encoding Methods*

Tung Hoang et al. (2016) [17] analyzed the evolutionary relationship of genomes using numerical encoding and digital signal processing methods. They used CGR method for numerical and 2D conversion of DNA sequences.

Ning Yu et al. (2018) [18] made survey on various encoding schemes for genomics data. They categorized the encoding schemes based on five perspectives such as biochemical properties, primary structure properties, Cartesian coordinates, binary and information encoding, and graphical representation.

Jorge E. Duarte-Sanchez et al. (2017) [19] proposed the hardware called multifractal processor to analyze the DNA sequences. Multifractal analysis is basically mathematical approach which involves CGR, box counting and linear regression. They concluded that their proposed work was well suited for calculation of fractal dimensions of human genome with less time and low cost.

Several authors used numerical encoding schemes for biological data analysis such as, Chunrui Xu et al. (2016) [20] analyzed the protein sequence using Chaos game and physicochemical properties and Lichao Zhang et al. (2016) [21] predicted the structural class prediction of protein using CGR method. Kang Dai (2007) [22] and Jie Song (2009) [23] analyzed the DNA sequences by their numerical characterization using 2D and 3D graphical representations.

C. *Observation from the Literature Survey*

- Several researchers focus on CGR numerical encoding methods for the analysis of biological data.
- There is a need of strong identification system to detect the HD affected human DNA data from the non HD human data. This identification will helps to the medical practitioners for suggesting the drugs to patients.

Hence, in this proposed work HD based DNA sequence analysis is performed using five various encoding methods such as CGR, ASCII value representation, atomic number representation, molecular mass representation and thermodynamic property representation. Then few machine learning methods are used for classification such as SVM, CT and RF.

III METHODOLOGY

Fig. 1. shows the proposed methodology. Dataset used in this proposed work contains 36 human DNA sequences those are collected from NCBI. In which 25 data are non Huntington data and 11 data are Huntington data. These data are downloaded from NCBI. In addition to this 48 synthetic human DNA sequence data are also constructed and used for process. In which 28 data are non Huntington data and 20 data are Huntington DNA. Table II shows the details of NCBI dataset.

A. *Numerical Encoding Methods for DNA sequence*

- *Chaos Game Representation:* Using CGR, genomic sequences can be able to represent into numerical and graphical representation. In this work CGR is used to convert DNA sequences into numerical value. DNA sequences are encoded in a unit square. Four vertices of unit square are encoded with four nucleotides.

Center of the square is the starting point. For the first nucleotide, numerical value can be calculated by half distance between of the starting point to the current first nucleotide coordinate. In this same manner, each nucleotide is converted into numerical form by calculating half distance between the previous point and the current nucleotide coordinate. DNA sequence can be numerically converted by Equation (1), Equation (2) and Equation (3) [17].

$$X_i = 0.5(X_{i-1} + g_{ix}) \quad (1)$$
$$Y_i = 0.5(Y_{i-1} + g_{iy}) \quad (2)$$
$$Z_i = (X_i + Y_i) \quad (3)$$

Where (X_i , Y_i) and (X_{i-1}, Y_{i-1}) are the current and previous plotted points on the coordinate, (g_{ix}, g_{iy}) is the corresponding vertex of the nucleotide and Z_i is the calculated numerical value of the i^{th} nucleotide in the DNA sequence. Where i=1 to n, where n is the length of DNA sequence.

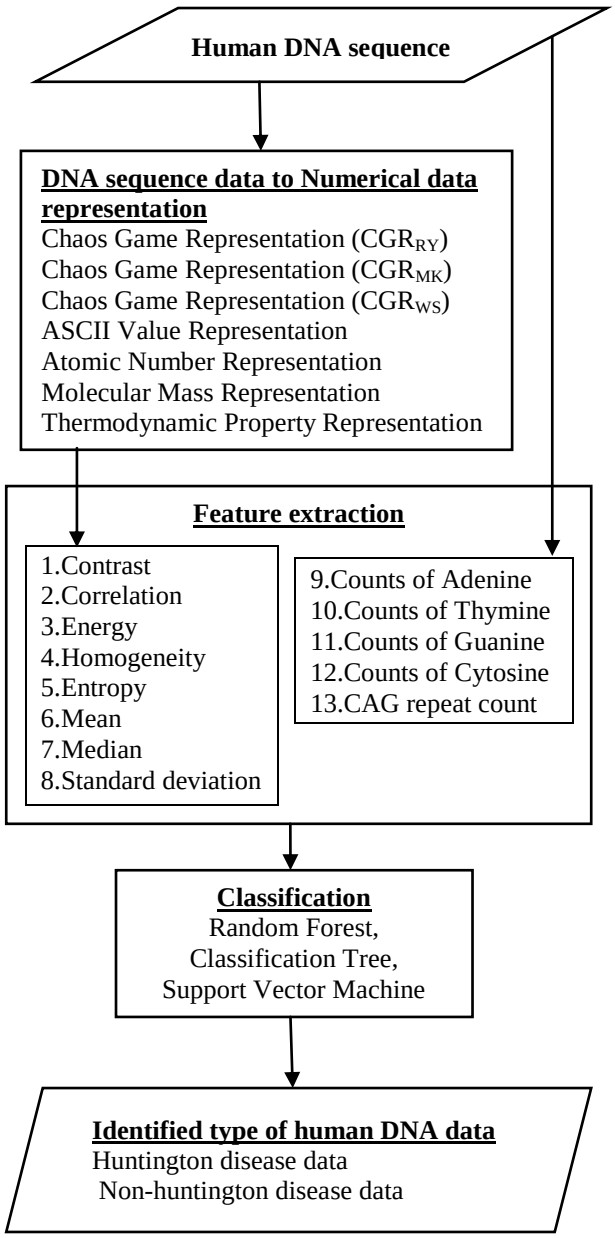

Fig. 1. Proposed methodology

Based on biological properties and three different coordinate assignments, CGR is categorized into three forms and this is shown in Table III.

They are 1) CGR- Purine and Pyrimidine i.e. CGR_{RY}
2) CGR- Amino and Keto group i.e. CGR_{MK}
3) CGR-Weak and Strong hydrogen bond i.e. CGR_{WS}

Table- II: NCBI Dataset details

S.No.	Accession number of NCBI data (FASTA format)	Length of data (number of base pairs)
1	BD227048.1	327
2	BD227049.1	331
3	BD227050.1	470
4	BD227051.1	565
5	BD227052.1	233
6	BD227053.1	578
7	BD227054.1	390
8	BD227055.1	547
9	BD227056.1	436
10	BD227057.1	469
11	BD227058.1	359
12	BD227059.1	209
13	BD227060.1	485
14	BD227061.1	468
15	BD227062.1	393
16	BD227063.1	421
17	BD227064.1	498
18	BD227065.1	427
19	BD227066.1	367
20	BD227067.1	502
21	HV308698.1	1543
22	HV308707.1	1096
23	L37198.1	1006
24	L37199.1	1356
25	M92292.1	611
26	HV308701.1	1735
27	AH003045.2	20909
28	AK290544.1 (CDS portion)	1749
29	BC172756.1 (CDS portion)	9435
30	L20431.1 (CDS portion)	1749
31	NG_009378.1 (CDS portion)	9435
32	NM_002111.8 (CDS portion)	9435
33	AY157849.1	4140
34	HV308700.1	1900
35	KJ535072.1	1778
36	L34020.1	4105

Table- III: CGR Coordinate assignments [18]

S.No.	CGR type	Coordinates
1	CGR_{RY}	A(0,0) , T(1,0), C(0,1), G(1,1)
2	CGR_{MK}	A(0,0) , T(1,0), G(0,1), C(1,1)
3	CGR_{WS}	A(0,0) , G(1,0), C(0,1), T(1,1)

- *ASCII Value Representation*: ASCII values of characters are used for the numerical conversion of DNA data. ASCII values of nucleotides are A=65, C=67, T=84, G=71.
- *Atomic Number Representation*: Based on the atomic number, the nucleotide is converted into numerical value [18]. Atomic number for the nucleotides are A= 70, T=66, C=58, G=78.
- *Molecular Mass Representation:* Molecular mass of the nucleotides are A=134, T=125, C=110, G=150. [18]
- *Thermodynamic Property Representation:* Thermodynamic property values for nucleotide interactions are AA=9.1, AC=6.5, AG=7.8, AT=8.6, TA=6.0, TC=5.6, TG=5.8, TT=9.1, CA=5.8, CC=11.0, CG=11.9, CT=7.8, GA=5.6, GC=11.1, GG=11.0, GT=6.5. [18]

B. Feature Extraction from Numerical Values of DNA Sequence

- *Statistical features*: Statistical information of DNA sequence can be calculated using mean, median, standard deviation and Gray Level Co-occurrence Matrix (GLCM) features. GLCM features are mainly used for image related processing. In this proposed work GLCM is applied to DNA sequences [17]. GLCM features include contrast, correlation, energy, homogeneity, entropy.

Contrast gives the depth of texture i.e. how the variance is between the neighboring objects, correlation gives the randomness, energy gives the uniformity and homogeneity gives the local changes between neighboring objects and entropy gives the statistical randomness.

C. Feature Extraction from DNA Sequence

- *Nucleotide counts:* Counts of Adenine, Guanine, Thymine and Cytosine in DNA sequences are accounted.
- *CAG repeat count*: It is the codon which is related to Huntington disease. Counts of CAG repeats in DNA sequences are calculated and used.

D. Classification of DNA Sequences

Three kinds of classifiers such as Support Vector Machine, Classification Tree and Random Forest Tree are used in this work to detect the DNA sequences into Huntington disease data and non Huntington disease data.

- *Support Vector Machine*: It is a supervised learning system which uses linear separability for classification task. It solves the problem by searching an optimal hyperplane. Hyperplane with the largest margin is called as Maximum Marginal Hyperplane. Separating hyperplane is given by Equation (4). [24]

$$(W*X)+b=0 \quad (4)$$

where W- weight vector, X – input vector and b- bias.

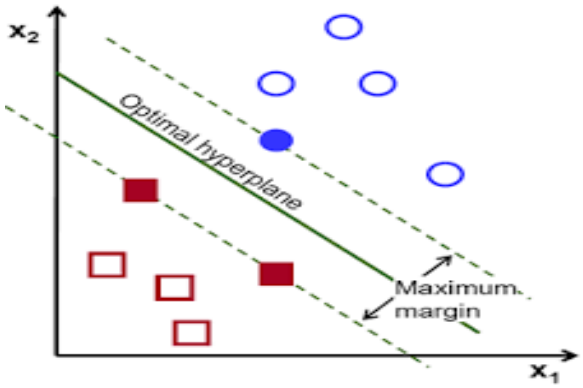

Fig. 2. Support Vector Machine Classifier – Optimum hyperplane

- *Classification Tree*: It follows supervised learning strategy. It can be also called as decision tree or Classification and Regression Tree (CART). Tree contains root, internal and leaf nodes. Attributes (feature values) are represented in root and internal nodes. Leaf node contains class labels. It classifies non-linear data as well [25].
- *Random Forest*: It is a supervised learning classifier. It is a collection of classification tree predictors. Classifier builds a set of classification trees by the given set of class labelled data. From the training data, each tree is developed randomly with the best attributes. Finally, the classification is done based on the majority vote against the test sample from the developed trees in the forest. [26]

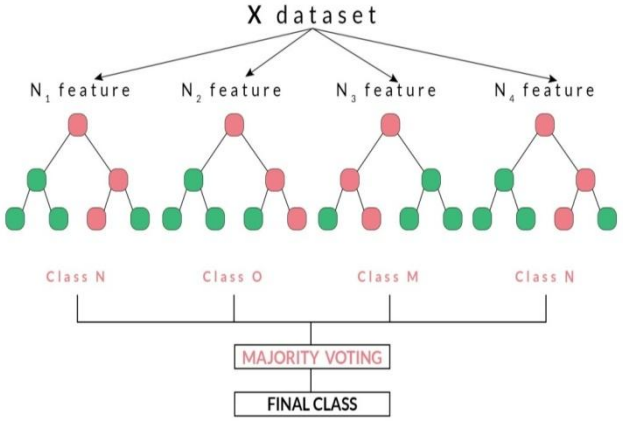

Fig. 3. Random Forest Tree Classifier

IV. EXPERIMENTAL RESULTS AND DISCUSSION

This section discusses the implementation results of this work. Matlab 2013b and orange 2.7 (data mining software) is used for implementation.

A. Results of Encoding Methods of DNA Sequence

Table IV and Table V shows the results of encoding methods for DNA sequence data. Header, portion of sequence information of DNA data is shown. In Table IV, the portion of converted numerical value of sequence data is shown for encoding methods such as CGR_{RY}, CGR_{MK} and CGR_{WS} representation respectively.

Table V shows the portion of converted numerical value of sequence data for encoding methods such as ASCII value representation, atomic number representation, molecular mass representation and thermodynamic property representation respectively. Nucleotide interaction for thermodynamic property is represented as NI in TABLE V.

Table- IV: CGR_{RY}, CGR_{MK} and CGR_{WS} representation -Portion of numerical form of DNA sequence data

Header: M92292.1 Homo sapiens Huntington's disease			
Portion of DNA Sequence: GGATCC			
Sequence	Numerical value		
	CGR_{RY}	CGR_{MK}	CGR_{WS}
G	1.5000	1.0000	1.0000
G	1.7500	1.0000	1.0000
A	0.8750	0.5000	0.5000
T	0.9375	0.7500	1.2500
C	0.9688	1.3750	1.1250
C	0.9844	1.6875	1.0625

Table- V: ASCII, atomic number, molecular mass, thermodynamic property representation -Portion of numerical form of DNA sequence data

Header: M92292.1 Homo sapiens Huntington's disease					
Portion of DNA Sequence: GGATCC					
Sequence	Numerical value				
	ASCII	Atomic number	Molecular mass	Thermo dynamic property	
G	71	78	150	NI	value
G	71	78	150	GG	11.0
A	65	70	134	GA	5.6
T	84	66	125	AT	8.6
C	67	58	110	TC	5.6
C	67	58	110	CC	11.0

B. Results of Feature Extraction

Totally 17 features are extracted from DNA data. GLCM features are such as contrast, homogeneity, energy and correlation for two offset values [1, 0], [0, 0]. In offset first value represents row distance and second value represents column distance. Numerical form of DNA data has the single column with multiple rows representation. So that this two offset values are used. Here [1, 0] takes the one row distance and no column distance between the point of interest and its neighbor for the feature calculation. Likewise [0, 0] calculates the feature value with no distance between the row and column. Hence, totally GLCM results eight feature values. Then entropy, mean, median, standard deviation is extracted. It gives 4 feature values. Next counts of nucleotides such as Adenine, Cytosine, Guanine and Thymine are extracted (i.e. 4 feature values). Afterwards CAG repeat counts are extracted (i.e. 1 feature value).

Table VI and Table VII shows the extracted GLCM features of two offsets for sample of eight DNA sequences by CGR_{RY}. Table VI corresponds to offset [1,0] and Table VII corresponds to offset [0,0]. Table VIII shows the extracted features of entropy, mean, median and standard deviation. Nucleotide counts i.e. bases_cnt.A, bases_cnt.C, bases_cnt.G, bases_cnt.T and CAG repeat counts are shown in Table IX.

Table- VI: GLCM features for offset [1,0] – Contrast, Homogeneity, Energy, Correlation

Contrast1	Homogeneity1	Energy1	Correlation1
3.009202	0.711846334	0.330366	0.372013698
2.533333	0.765616883	0.422075	0.413310876
3	0.710726977	0.327772	0.412368761
2.787234	0.724886018	0.34992	0.387664444
2.143868	0.7924227	0.375007	0.788337467
2.840961	0.709668873	0.326727	0.285261759
3.364533	0.677105101	0.27796	0.393794538
2.8627	0.708591451	0.325593	0.283156762

Table- VII: GLCM features for offset [0,0] – Contrast, Homogeneity, Energy, Correlation

Contrast2	Homogeneity2	Energy2	Correlation2
0	1	0.504793	1
0	1	0.590438	1
0	1	0.491127	1
0	1	0.53431	1
0	1	0.498752	1
0	1	0.511845	1
0	1	0.454238	1
0	1	0.510414	1

Table- VIII: Features – Entropy, Mean, Median, Standard deviation

Entropy	Mean	Median	Standard deviation
3.572707	1.023288	1.023643	0.365582
3.244022	1.109138	1.063225	0.403309
3.93948	1.048574	1.006122	0.402414
3.961312	1.040752	1.02747	0.367584
3.592673	1.010897	1	0.517099
4.07619	1.068737	1.062213	0.379217
4.491738	1.015657	1.024131	0.395965
4.081055	1.067593	1.061819	0.37948

Table IX: Features –Nucleotide counts (A,C,G,T) and CAG repeat count

bases_cnt.A	bases_cnt.C	bases_cnt.G	bases_cnt.T	CAG_count
69	103	77	78	10
66	99	103	63	7
103	120	126	121	13
113	165	136	151	18
3201	3626	3627	3847	457
368	525	488	368	68
2282	2530	2430	2193	350
369	525	487	368	68

C. Classification performance analysis

Extracted features are given to classifiers SVM, CT and RF. Three classifiers are performs well for the undertaken DNA sequences. Table X shows the training and testing split

up of DNA data set which is used for performance analysis. In Table X, NNHD represents NCBI Non Huntington Data, SNHD represents Synthetic Non Huntington Data, NHD represents NCBI Huntington data and SHD represents Synthetic Huntington Data.

Table- X: DNA data- Training and Testing split up

	Class 1 (Non HD)	Class 2 (HD)
Training samples (70 % data)	38 (NNHD :17, SNHD :21)	22 (NHD:8, SHD:14)
Testing samples (30 % data)	15 (NNHD :8, SNHD :7)	9 (NHD:3, SHD:6)

Table XI shows the performance values of the proposed system by CGR_{RY}, CGR_{MK} and CGR_{WS} representation encoding methods and Table XII shows performance values of ASCII value representation, atomic number, molecular mass and thermodynamic property representation encoding methods respectively. CT yields 100% accuracy for all encoding methods and it outperforms than the RF and SVM for HD identification.

Accuracy obtained for numerical encoding methods by RF, SVM and CT is graphically represented in Fig. 4. The proposed work gives better performance for balanced and imbalanced dataset also.

Table- XI: Performance values of RF, SVM and CT for CGR_{RY}, CGR_{MK} and CGR_{WS} encoding methods

	Classifier	Accuracy	Sensitivity	Specificity	Precision	Recall
CGR_{RY}	RF, CT	100	100	100	100	100
	SVM	83	100	68	84	100
CGR_{MK}	RF,CT	100	100	100	100	100
	SVM	93	100	81	90	100
CGR_{WS}	RF, CT	100	100	100	100	100
	SVM	90	94	81	90	94

Table- XII: Performance values of RF, SVM and CT for ASCII value, atomic number, molecular mass, thermodynamic property encoding methods

	Classifier	Accuracy	Sensitivity	Specificity	Precisin	Recall
ASCII value	RF	91	100	77	88	100
	SVM	98	100	95	97	100
	CT	100	100	100	100	100
Atomic number	RF, SVM, CT	100	100	100	100	100
Molecular mass	RF, SVM, CT	100	100	100	100	100
Thermo dynamic property	RF	98	100	95	97	100
	SVM, CT	100	100	100	100	100

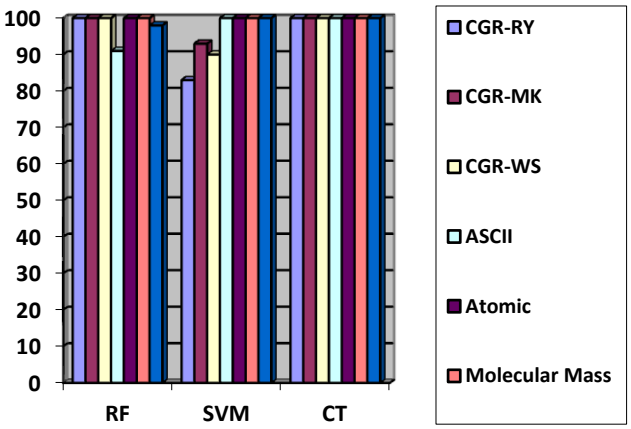

Fig. 4. Accuracy of the proposed system by RF, SVM and CT

V. CONCLUSION

The proposed system, works well for the detection of DNA data into Huntington data and non Huntington DNA data. Constructed synthetic data are seems to be non biased one for the classification accuracy. CGR_{RY}, CGR_{MK} and CGR_{WS} encoding methods give mostly similar kind of numerical values. There is only minimal difference in the converted numerical values. According to feature extraction, CGR_{RY}, CGR_{MK}, CGR_{WS} methods gives the appropriate feature values than the other encoding methods those used. So these three methods are suitable for conversion of DNA data into numerical one. Nucleotide count and CAG repeat count features plays the major role for detecting the DNA data into HD affected one or free from HD. From the performance analysis, CT classifier is concluded as the best suitable one for this proposed work.

VI. FUTURE ENHANCEMENT

This work can be extended in future using Deep Neural Network (DNN) algorithm for comparing the performance results of machine learning based classifier..

REFERENCES

1. https://www.news-medical.net/health/What-is-Neurodegeneration.aspx# Last Updated: Aug 23, 2018, SalleyRoberston [Accessed on 05.08.2019]
2. https://www.neurodegenerationresearch.eu/about/what/ [Accessed on 05.08.2019]
3. https://kids.frontiersin.org/article/10.3389/frym.2018.00070. What are Neurodegenerative Diseases and How Do They Affect the Brain? Published: December 12, 2018. Authors : Tary Berman, Armin Bayati [Accessed 03.09.2019]
4. M. Gourie –Devi, "Epidemiology of neurological disorders in India: Review of background, prevalence and incidence of epilepsy, stroke, Parkinson's disease and tremors", Neurology India, 2014, Vol.62, issue 6.
5. https://neurodiscovery.harvard.edu/challenge [Accessed 03.09.2019]
6. https://medlineplus.gov/ [Accessed 03.09.2019]
7. Heidi Chial, "Huntington's disease: The Discovery of the Huntingtin Gene", Nature Education, 2008, 1(1):71
8. L.A. Barboza and N.C. Ghisi, "Evaluating the current state of the art Huntington disease research: a scientometric analysis", Brazilian Journal of Medical and Biological Research 2018, 51 (3).

9. Natasa A Kablar, "Signaling Pathways Perspective of Neurodegenerative Diseases: Parkinson's, Alzheimerr's and Huntington's Disease", ACTA scientific Microbiology, 2019, vol.2, issue 4.
10. Davina J Hensman et al., "Identification of genetic variants associated with Huntington's disease progression: a genome- wide association study", Lancet Neurol 2017.
11. Srimanta Pramanik et al., "Analysis of CAG and CCG repeats in Huntintin gene among HD patients and normal populations of India", European Journal of Human Genetics, 2000, vol.8, pp.678 – 682.
12. Johanna Craig, "Complex diseases: Research and applications", Nature Education, 2008, 1(1):184.
13. Peggy C. Nopoulos, "Huntington disease: a single-gene degenerative disorder of the striatum", Dialogues in Clinical Neuroscience – 2016, Vol.18. No.1.
14. Miao Xu, Zhi-Ying Wu, "Huntington Disease in Asia", Chinese medical journal, 2015, vol.28, issue 13.
15. Babu Srija et al., "Huntington's disease: An Indian update on genetics and widespread", International Journal of Research and Development in Pharmacy and Life Sciences", 2016, vol.6, No.1, pp.2274-2483.
16. Cynthia T.McMurray, "Mechanisms of trinucleotide repeat instability during human development", Nat Rev Genet, 2010, 11(11), 786-799.
17. Tung Hoang, Changchuan Yin and Stephen S.T.Yau, "Numerical encoding of DNA sequences by Chaos game representation with application in similarity comparison", Elsevier –Genomics, 2016.
18. Ning Yu, Zhihua Li and Zeng Yu, "Survey on encoding schemes for Genomic Data Representation and Feature Learning – From Signal Processing to Machine Learning", Big data mining and analytics, 2018, Vol.1, issue 3, pp.191-210.
19. Jorge E. Duarte-Sanchez, Jaime Velasco-Medina and Pedro A. Moreno, "Hardware Accelerator for the Multifractal Analysis of DNA sequences", IEEE –ACM Transactions on Computational Biology and Bioinformatics, 2017.
20. Chunrui Xu, Dandan Sun, Shenghui Liu and Yusen Zhang, " Protein Sequence Analysis by Incorporating Modified Chaos Game and Physicochemical Properties into Chou's General Pseudo Amino Acid Composition", Elsevier – Journal of Theoretical Biology, 2016.
21. Lichao Zhang, Liang Kong, Xiadong Han and Jinfeng Lv, "Structural class prediction using novel feature extraction method from chaos game representation of predicted secondary structure", Elsevier – Journal of Theoretical Biology, 2016, vol.400, pp.1-10.
22. Kang Dai, "A Novel Two-Dimension Graphical Representation of DNA Sequences and its Numerical Characterization", IEEE conference, 2007.
23. Jie Song, "A new 3-D graphical representation of DNA sequences and their numerical characterization", IEEE conference – International Conference on Computer Science & Education, 2009.
24. http://www.robots.ox.ac.uk/~az/lectures/ml/lect2.pdf -Support Vector Machine lecture notes. [Accessed on 27.08.2019]
25. https://www.geeksforgeeks.org/decision-tree-introduction-example/ - Decision Tree classifier document. [Accessed on 06.09.2019]
26. https://docs.biolab.si//3/visualprogramming/widgets/model/randomforest.html -Random Forest Tree classifier document.[Accessed on 27.08.2019].

AUTHORS PROFILE

Dr.G.Tamilpavai, she completed her B.E in Computer Science and Engineering from Thiagarajar College of Engineering, Madurai, Tamil Nadu, India. She did her P.G in Government College of Engineering, Tirunelveli, Tamil Nadu, India. She Completed her Ph.D. at Anna University, Chennai, Tamil Nadu, India. Her area of interest includes medical image processing, remote sensing, bio informatics and operating systems. She is working as Associate Professor (CAS) and Head in Department of Computer Science and Engineering at Government College of Engineering, Tirunelveli. She has 20 years of teaching experience. She is recognized guide in Anna University, Chennai, Tamil Nadu, India. She has 15 publications in international journals especially in biomedical image processing and bio informatics. She has published many papers in National and International conferences. She has life membership in ISTE, IE and BMESI. She received fund from SERB, Department of Science and Technology, Government of India for the project entitled "Detecting Defective DNA motifs to find genetic disease sequence in a Human DNA using SOM". SERB sanctioned Rs.14,78,000 for the project (project duration 2017- 2020).

C.Vishnuppriya, she completed her B.Tech. degree in Information Technology in 2014 and M.E. degree in Computer Science and Engineering in 2016 from Anna University Regional Campus- Tirunelveli Region, Tirunelveli, Tamil Nadu, India. Her research interests include image processing and bioinformatics. She has one publication in national journal for Siddha medicine related image processing and two publications in international journal for bio informatics. She has published 3 papers in international conferences. She is working as Senior Research Fellow in Department of Computer Science and Engineering at Government College of Engineering, Tirunelveli, Tamil Nadu, India.