Cancer Genome Classification Using Machine Learning Algorithm

Tahmeena Fatima[1], S Jyothi[2], D M Mamatha[3], K Srujan Raju[4]
[1] Research Scholar, Dept. of Computer Science and Engineering,tahmi.fatima18@gmail.com
[2]Professor, Department of Computer Science,jyothi.spmvv@gmail.com
[3]Professor, Department of BioScience& Sericulture, prof.mamatha@gmail.com
[4]Professor, Department of Computer Science & Engineering, ksrujanraju@gmail.com
[1,2,3]SriPadamavatiMahilaVisvavidyalayam, Tirupati, India
[4]CMR Technical Campus, Hyderabad, Telangana, India

Abstract: In present scenario, cancer is the universal problem in the world because it effects the people in the world. The growth of uncontrolled and abnormal cells in the body is known as cancer. These cells with abnormal named as cells of cancer, cells of malignant or cells of tumour. To recognize the genomicorigin of tumour cell explosion and the development of the cancer genome the normal cell and tumour cell are compared to find which type of cancer it is. The growingaccessibility and development rate of "Big Data" derivative from ofnumerous omics exposed a novelframe to expand diagnoses of clinical or cancertherapeutics, however there areseveralexperiments in effective analysis and explanation of such complexandbig data. The significance of categorizing cancer patients hooked on loworhigh risk setsconsumesdirectedseveral research groups, from the bioinformaticsandthe biomedical area, to learn the solicitation of machine learning (ML) techniques. Thus, these methods have remainedconsumed as an intention to classicadvancement and treatment of conditions of cancerous. In adding, the capability of ML implements to discovervital features from compound datasets exposes their prominence.

Keywords: cancer cells, machine learning algorithms, Mongodb, BWA, Genome sequencing.

1 INTRODUCTION

In medical, Whole Genome Sequencing (WGS) is grown as main interest for the diagnostics. Next Generation Sequencing has begun as a exhortation which incorporatesnew DNA sequencing methods, permitting researchers to categorize a whole human genome is linked to the standard Sanger sequencing machinery which necessary over a period for accomplishmentwhile the human genome was primary sequenced.

Within specific organism gene is a complete set and the study of molecular biology is known as genome which is a branch of Genomics. Currently, machine learning playsanvital role in the development of genomics area.

The capability to sequence DNA affordsmythoscapability to "read" the blueprint of genetic that guides all the actions of a alive organism. To affordsituation, the dominantview of biology is shortened as the pathbeginning as DNAtowards RNA towards Protein. DNA remainsself-possessed of base pairs called nucleotides which has four basic units A, T, C, G. A combines with T and C combines with G. DNA is prearranged into genetic material and humans deviseaentire of 23 combines. [1]

Chromosomes are further ordered into sections of DNA named as genes which create*encode* proteins. A genome is a sum of gene that possess organism. Humans devisecoarsely 20,000 genes and 3 billion base sets. Stimulatingly, only 2% ofencodes protein in the human genome and this is a vitalpart of effort in research and the commerce of genomics.[2]

Precision medicinecan closely connected to Genomics [3]. By 2023 87 billion dollars are projected with a market size that aarea of Precision Medicine (also named as personalized medicine) is an methodology to carepatient that comprehends genetics, environment andbehavioursthrough a aim of instigating a patient or population explicit treatment involvementcompare to a one scope turns all method. For illustration, to lessen the danger of difficulties, aseparate who wants a blood transfusion would be harmonized to a giver who shares the identical blood group instead of a casually selected giver.

To analyse Genomics deeply, machine learning can support researchers understand genetic dissimilarity. Specially, algorithms are aimed based on outlinesrecognized in hefty genetic data sets which are formerlyinterpreted to computer simulations to support clients understand how genetic dissimilaritydisturbscritical cellular procedures. Illustrations of cellular procedurescontain the metabolism, DNA restoration, and cell progression. Disturbance to the standardworking of these paths can theoreticallysource of diseases like cancer. [4]

Althoughprominence is regularlylocated on selecting the finest learning procedure, researchers have initiated some of the beststimulatingqueriesrise out of nothing of the accessible machine learning proceduresperformance to balance. Utmost of the period this is a difficult with training data, nevertheless this also happens when occupiedthrough machine learning in novelfields.

Machines study are beneficial to humans sinceby all of their handling power, they are capable to more rapidly highlight or discoveroutlines in big data that would consume orelseremainedlost by human lives. Machine learning is a method that is used to improve human'scapabilities to resolve problems and sortknowledgeableimplications on a variedsort of difficulties, from assistingidentify diseases to pending up throughexplanations for worldwide climate alteration.

2 Proposed method

2.1 Burrows–Wheeler algorithm

The Burrows–Wheeler alter is an method used to makeinformation for usagethrough data compression methods like bzip2. In 1994 David Wheeler and Michael Burrows are discovered this method at that time they are working at DEC Systems Research Centre in Palo Alto, California. It is created on anearlier unpublished revolutionrevealed by Wheeler in 1983. The method can be executedcompetentlyconsuming a suffix array consequentlyattainment linear time complexity.[5]

Let Σ is represented as alphabet. $Symbol is not existing in Σ and is lexicographically lesser than entirely the symbols in Σ. A series $X=a_{0,}\ a_{1,}\ \dots,\ a_{n-1}$ is constantlyconcludedthrough $ symbol (i.e. a_{n-1}=$) then this sign is appeared at the end. Let $X[i]=a_i$, $i=0, 1,\dots, n-1$, be the ithsymbol of X, $X[i, j]=a_i \dots a_j$ a substring then $X_i=X[i, n-1]$ a suffix ofX. Suffix array of S of X is a combination of the numerals $0\dots n-1$like that $S(i)$ is the initiationpoint of the i^{th} lowest suffix. The X *in* BWT is represented as $B[i]$=$ while $S(i)=0$, $B[i]=X[S(i)-1]$ else. We also outline the measurement of string X as $|X|$ and consequently $|X|=|B|=n$. Figure 2 offers an illustration on how to build BWT and suffix array.

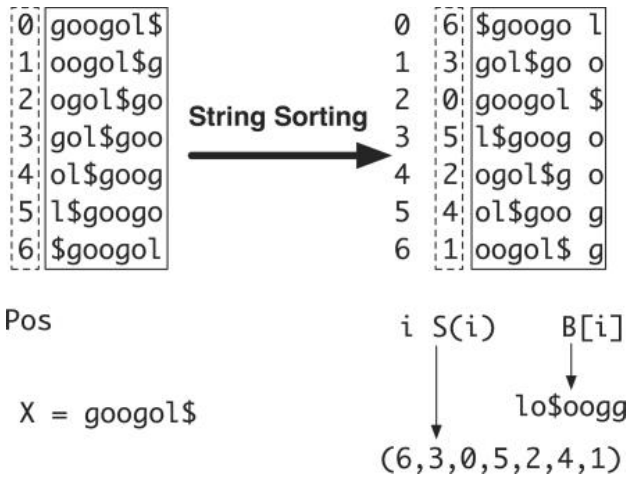


Fig 1: analysis of the BWA algorithm

The proceduredisplayed in figure 1 is quadratic in period and space. Though, this is not essential. In training, we regularlybuild the suffix array initial and formerlyproduce BWT. Furthermostprocedures for building suffix array needminimum $n[\log_2 n]$ bits of occupied space, which expanses to 12 GB for human genome [6]. Newly, Hon *et al.* (2007) presented a novel method that customs *n* bits of occupied space and needs<1 GB memory at ultimateperiod for building the human genome using BWT .

2.2 Machine learning

Machine Learning is the knowledge of receiving computers to study and performancesimilarly humans ensure and expand their knowledge over period in independentmethodthroughserving them records and data in the arrangement of clarifications and real wordrelations.

The definition ofclassificationsummarizes the model objective or eventualgoal of machine learning is stated by numerous researchers in the area. The determination of this object is to afford a reader minded businesswith skilledviewpoint on in what manner machine learning is definiteand in what way it workings. artificial intelligenceand Machine learning stake the similarmeaning in the observances of numerous fields.Thoughthere remainaboutdifferentalterations readers mustidentify as fit.

Machinelearning algorithms are in numerous forms [7], through hundreds distributed each time and they remain usuallycollected by each knowledge style (that isunsupervised learning, semi-supervised learning, unsupervised learning) or by resemblance in usage or meaning (that is regression, deep learning, classification, decision tree, clustering, etc.). Nevertheless of knowledge style or meaning, all arrangements of machine learning proceduresinvolve ofthe resulting:

- **Demonstration** (computer can understand the set of language or classifiers)
- **Calculation** (also known as scoring/objective function)
- **Optimization** (examinationprocess; habitually the highest counting classifier, for illustration; here are mutually off the ledge and convention optimization procedures used)

Machine learning technique has dissimilar tactics to study from consumingsimple decision trees to gathering to layers of artificial neural networks (then latter it gives space to the deep learning) dependent on whateverduty you are demanding to achieve and the sort and sum of information that you deviseexisting. This vigoroussituation played obtainable in submissions as variable as medicinal diagnostics or self-driving cars.

Research prepared when occupied on actualsubmissions often effortsgrowth in the arena, and causes are twofold: 1. Propensity to learnlimitations and restrictions of presentapproaches 2. Scholars and makersoccupied with fieldauthorities and leveraging period and knowledge to increaseclassification performance.

Occasionally this toohappens by “accident” We canimitateclassicgroups, or blends of numerous learning procedures to recoveraccurateness.

In present days, genomics by usesMachine learning can impactnumeroushintthemescontainingin what way genetic research is directed, in what way clinicians afford patient carefulness and the availability of genomics to charactersinvolved in learning newthingsin what way their inheritance may affect their condition.

Consequently, the informationanalysis canexperience through machine learning withessential to accompanied by training and clear descriptions of the efficacy and importance of this knowledge.

2.3 Support Vector Machine

SVM is one of the supervised learning method in ML which is associated with learning procedures that inspect information used forregression andclassification analysis [8]. Resolutea stable of activity cases, individually show as standard to unique or the new of binary gatherings, in SVM makingdesign toprogresses a model that apportions novel cases to unique gathering or the former, productions a non-probabilisticparallel undeviatingclassifier [9]. When facts remaincategorized, supervised learning not possible, and an unsupervised learningmethod is required [10], which efforts invention normal gatheringof the information to groups, and now map novelinformation to these formed groups. The clustering procedure which delivers an enhancement to the SVM is known as Support Vector Clustering (SVC) then it uses in trade applications moreoverwhen realities are not categorized or when firstsome realities are categorized equally a pre-processingaimed at a groupingpass [11]. The appliance of categorizing the informationhooked ondissimilarmodules by definition a link which splits the preparation files into modules. Hereremain a few straight hyperplanes, calculation of SVMattempts to augment the separation in the focal of the few classes that are mind boggling and this is said as edge augmentation. If the line makes the most of the space among the modules is recognized, the possibility to simplify fine to unobserved information is improved. There are 2 classifications in SVM:

2.3.1 SVMLinear (L-SVM)

In L-SVM the training data that classifies are disconnected by a hyperplane.

$$\frac{1}{m}\sum_{i=1}^{m} l(w.x_i + b.y_i) + \| w \| 2 \qquad \text{eq1}$$

2.3.2 SVM Non-Linear (NL-SVM) or sigmoid

In NL-SVM it remains not thinkable toward discrete the training statistics with a hyperplane. For illustration, the training statistics for Face recognition containsa set of imageries that are aspects and alternativecollection of imageries that remainaspects (in additiondisputes all other imageriesin the domainexcepting faces). Under such circumstances, the preparation measurements are excessively perplexing and troublesome, making it impossible to find a delineation for each component vector [MT06]. Isolating the typical of countenances directly after the arrangement of non-confront is a many-sided assignment.

2.4 Classifier of Ada-boost

Ada boost method combines weak classifier procedure [12] to customtough classifier. A loneproceduremightcategorize the objects unwell. However, if we syndicatenumerous classifiers throughcollection of training set oneachrepetition and transfer right quantity of load in ultimatepolling, we canconsume good accuratenessgroove for complete classifier. Respectivelyfeeble classifier is accomplishedby means of a ***random subset*** of whole training set.[13]

Laterpreparation a classifier at some level, ada boost assigns [14] weight to apiece training point. Misclassified point is allottedadvanced weight hence it seems in the training division of next classifier throughcomplexpossibility.

Subsequently each classifier is trained, the weight is allotted to the classifier as well grounded toaccurateness. Additionally, accurate classifier is allotted higher weight so that it can consumeextrainfluence on ultimateresult.

Let see the parameters and mathematical formula.

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

$h_t(x)$ is the weak classifier of output t for x input
α_t is assigned classifier weight.
$\alpha_t$is designed as follows:
$\alpha_t = 0.5 * in(\frac{I-E}{E})$: straight forward classifier weight, it is built on the rate of error E. Originally, all the response training instance has same weightage.

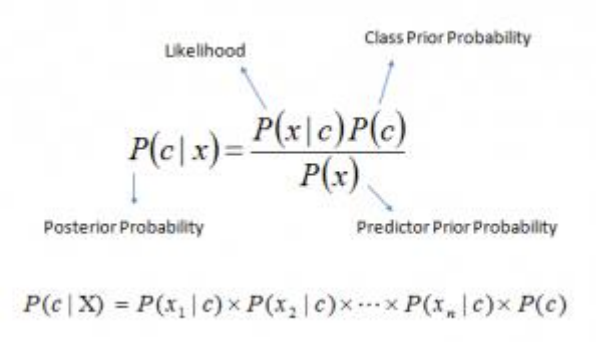
2.5 Classifier of Naive Bayes

It is Bayes Theorem of classification method[14]through an supposition of disinterestbetween predictors. In general, classifier Naive Bayes adoptsthe occurrence of a specific feature in a class is dissimilar to the occurrence of someextra feature.

For illustration, if an apple is considered as a fruit if it has round, contains the 3 inches' diameter and red colour. Even if these structure is depending on one another or at the presence of the other structure all these features individually pay to the possibility to be an apple fruit and that is known as "Naïve".

It is very useful for large dataset and it can easily build. Laterally with easiness, Naive Bayes is recognized to outstrip even extremely sophisticated classification procedures.

Bayes formulaoffers a mode of computingsubsequentpossibility P(c|x) from P(c), P(x|c) and P(x). see the formula below:



- *P*(*c*|*x*) is thegiven *predictor* (x, *attributes*) probability of posterior*class* (c, *target*).
- *P*(*c*) is the probability ofprior *class*.
- *P*(*x*|*c*) is thegiven *class*probability*predictor*.
- *P*(*x*) is the probability of prior *predictor*.

2.6 Random Forests

Leo Breiman developed the Random Forest [16] which is a setof un-pruned arrangement or regression trees completed from the selection of randomsectionsof the training information. From the induction process Random features are selected. Predictioniscompleted by combining(majorityelectfor averaging for regression or classification) the estimates of the collaborative. Each tree is developed as defined below [24]

- By randomlysampling n, if N is the no. of cases in the training set however with replacement, by the original information. This model will be recycled as the training set for increasing the tree.
- For input variables M, the mvariable is chosen in a way that m<<M is stated at every node, variablesmis chosen atM random out then the finest split on m is used atexcruciating the node. Throughout the forest growing, the price of m is kept constant.
- Each tree is developed to be prime extent. No trimming is used.

GenerallyRandom Forestshows a major performance enhancement as associated to solo tree classifier like C4.5. The simplificationrate of error that yields torelatefavourablyto Adaboost, though it is more vigorous to noise.When using the random forest procedure to resolve regression complications are using Mean Squared Error(MSE) to know the information branch from each node

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(fi - yi)^2$$

Wherever N is the data points, $fi$is the returned valueof a model and $yi$is the data point i's actual value. This formulation calculates the distance of every single node from the actual predicted value, helping to choose which branch is the superiorchoice for your analysis. Here, yi is the data point value you are testing at a certain node and fi is the value returned via state.

2.7 Gaussian process

Gathering of random variables is a Process of Gaussian [17] any infinite number of which consume (reliable) combined Gaussian disseminations. A Gaussian procedure is completelyindicatedthrough its function of mean m(x) and function of covariance k (x, x_0). This remains regular generalization of the Gaussian dissemination whose covariance and mean is a matrix and vectorcorrespondingly. The Gaussian dissemination is concludedvectors;however, the Gaussian process is completegatherings. We resolvecompose:

$$f \sim GP(m, k)$$

meaning: "the GP is distributed as function f with function of mean m and function of covariance k". While the generalization from dissemination to procedure theconservative forward, we resolve it a bit addedobvious about the particulars, since it might be unaware to some readers. The separable variables in random is a vector from in a Gaussian dissemination [18] are indexed through their place in the vector. For the process of Gaussian, x is the argument (f(x) is the random function) which shows the role of catalogue set for every x input there is f(x) associated with random moveable, which is the significance of the (stochastic) gathering f at that positions. For details convenience of notational, we resolve enumerate the values of x concern by the normalquantities, and usage of these catalogues as if they remained the catalogues of the procedure do not lease yourself remaindisordered by this the guide to the procedure is x_i, which we consumepreferred to index through i.

2.8 K Nearest Neighbour

KNN method is one of the form of supervised algorithm in ML. where it can use for classification and problems of regression predictive. Though, it is mostly used for predictive classification of problems in industry[19]. The resulting two things would describe KNN well

- **Algorithm of Lazy learning** − KNN is a procedure of lazy learningsince it does not consume a specific training stage and usagesof all information for training though classification.
- **Algorithm of Non-parametric learning** − KNN is also a procedure of non-parametric learningsince it doesn't adopt anything about the principal data.

When classification uses KNN, as the production can considered as the class through the maximum frequency after the K most comparableexamples. Everyoccurrence in kernelelects for their class and with the maximumelects is engaged as the prediction.

Probabilities of Class can be designed as the regularized frequency of illustrations that fit to every class in theK set most comparableexamples for a novel data illustration. For illustration, in a binaryproblem classification (0 or 1are class)

$$p(class = 0) = \frac{count(class = 0)}{(count(class = 0) + count(class = 1))}$$

If you have classes of even number and uses K. to choose value k is a good idea with odd number to escape a tie. And the even number uses the inverse for K when you takea classesofodd number.

Ties can remainwreckedconstantly by growing K by 1 and observing at the next class to most comparableoccurrence in the training dataset.

2.9 Decision Tree

The representation of flow chat like tree structure is known as decision tree wherever an interior node denotesattribute (orfeature), the decision rulerepresents abranch, and outcomerepresents each leaf node. Theroot node in the decision tree is thetopmost node. It studies to barrier on the source of the feature value. The barriers in tree is in recursively custom call recursive splitting. The decision making can be done by flowchart like structure. The flowchart visualization cansimplyimpressionists the human level intelligent. Because decision trees are easy to interpretandunderstand. [20]

In ML, white box type of procedure is Decision Tree. It cutsinterior decision creatingreason, which is not accessible in the type ofblack box methods like Neural Network. Its training period is quicklylikened to the method of neural network. The complexity of time in decision trees is a task of the sum of records and quantity of elements in the specifiedinformation. It is a non-parametricor distribution free technique, which organizes not be contingent upon possibilitydisseminationconventions. It [21] can hold high dimensional information with nobleaccurateness.

Theconcept of entropy is a concept that is invented by Shannon, which processes the infection of the participation set. In mathematicsandphysics, entropy denoted as the uncertainty or the uncleanness in the scheme. In theory of information, it denotes to the uncleanness in a set of illustrations. entropyis the decrease of Information gain. Information gain calculates the alterationamong entropy earliersplitting and normal entropy after splitting of the dataset created on particular attribute values. Iterative Dichotomiser (ID3) decision tree procedureusages gain of information.

$$\text{Info(D)= - } \sum_{i=1}^{m} pi \log_2 pi$$

Where, pi is thearbitrary tuple of probability in an D fits to class ci.

$$\text{Gain(A)=Info(D)- Info}_\text{A}\text{(D)}$$

$$\text{Info}_\text{A}\text{(D)=}\sum_{j=1}^{v} \frac{|Dj|}{|D|} \text{ X Info(D}_\text{j}\text{)}$$

Wherever,

- $Info(D)$is the normalquantity of data needed torecognize the class label of a tuple in D.
- $|Dj|/|D|$ acts asjth partition of weight.
- $InfoA(D)$is the predictableinfoessential to categorize a tuple from D built on the splitting by A.

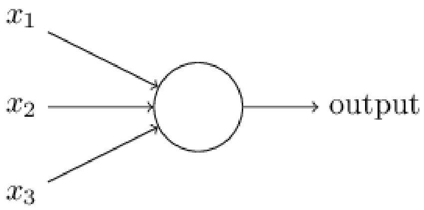
The Aattributethrough the highest gain of information, Gain(A), is selected as the piercingelement at node N().

2.10 Neural network

deep learning[22] one of the form is Neural networks, which is **machine learning**subfield, wherealgorithmsinspires from the structure human brain. input data is Neural networks, train themselves to identifydesignsinitiate in the information, and formerlycalculate the production for a novel set of comparableinformation. Consequently, a neural network can remainbelievedas the efficientelement of deep learning, which simulators the human brain behaviour to resolvecompound data driven difficulties.

Neurons are referred as input in machine learning, and the process of nucleus[23] the information and advancing the consideredproductivityconcluded the axon. In a genetic neural network, the thickness of dendrites describes the weight relatedthrough it.

A perceptron proceedsnumerous binary involvements, x1,x2,…x1,x2,…, and yields a solo binary production:



In the illustrationexposed the perceptron has 3contributions, x1,x2,x3x1,x2,x3. In common, it mighttake more or less inputs. Asimple rule is proposed by Rosenblatt to calculate the outcomes. He presented *weights*, w1,w2,.…w1, w2,, real statisticsuttering the significance of the correspondingresponses to the production. The neuron's production, 00 or 11, is resolute by the weighted sum $\sum jwjxj \sum jwjxj$is fewer than or superior than some *values of threshold*. Just similar the weights, the threshold is a genuine number which is a limitation of the neuron. To place it in additionalexactarithmetic terms:

$$output = \left\{ 01\ if\ \sum jwjxj \leq threshold\ if\ \sum jwjxj > threshold(1) \right\}$$

$$output = \left\{ 0\ if\ \sum jwjxj \leq threshold1\ if\ \sum jwjxj > threshold \right\}$$

2.11 MongoDB

MongoDB is a general-purpose database. MongoDB Schema [24] is actually a dynamic Schema which allows for the high flexibility and fits perfect for agile software development.

A cluster can contain a very large number of servers; they can be a single server but also a REPLICA SET. A cluster contains an N number of replica set depending on our needs.
— There is no upper limit to the number of members of the cluster.
— Due to this scalability major goal to make the complex cluster infrastructure transparent to the user and the interaction of the server is same as of interacting with a replica server.

3 Result and analysis

Firstly, load the genome sequencing in MongoDB due to the sequencing is in unstructured data format because it may have different sizes. And then it is retrieved into python programme by using MongoDB connecter. Then install all the necessary packages like Bio python and other package in python program language. By using Burrows_Wheeler_Alignment (BWA) finding the genome alignment and find the characters of the reference genome with the occurrence character and searching for max difference threshold for searching the matched position of the mutation sequence of each genome as illustrated in table 1.

About Burrows_Wheeler_Alignment (BWA) and genome analysis

BWA is used to map sequences of low-divergentalongside with a big reference genome, which is called as human genome. It contains 3 algorithms: BWA-MEM, BWA-SW and BWA-backtrack. The initial algorithm is considered for sequenceIllumina reads up to 100bp, though the rest of algorithm for lengthier sequences alternated from 70bp to 1Mbp. BWA-SW andBWA-MEM share same features like split alignmentandlong-read support, however BWA-MEM is latest one, which is normallysuggested for queries like high-quality because it is wilder and more perfect. BWA-backtrack also takesimprovedenactment than BWA-MEM for 70-100bp reads of Illumina.

Table 1 table shows the matched position of genome using BWA algorithm

NAME	Matched positions with clinical data
Normal Clinical Data	0
Genome 1	[222, 830, 582, 576, 28, 330, 746, 273, 587, 714, 557, 520, 320]
Genome 2	[82582, 88759, 27415, 80452, 19889, 35580, 5902, 1287, 54182, 57519, 67956, 24007, 34089, 17901, 69305, 24433, 10872, 74133, 87197, 9594, 51342, 32740, 23581, 44668, 61637, 29616, ……………………………15464, 66197, 72682, 13043, 21053, 63592, 71250, 18203, 7120, 61159, 24032, 34455, 17740, 8083]**1246**
Genome 3	[848, 5179, 6809, 4634, 2209, 1646, 1457, 3866, 2154, 3861, 89, 7136, 820, 2184, 3980, 1563, 3137, 307, 3468, …………………………, 4535, 3328, 6341, 1365, 2449, 2715]**78**
Genome 4	[14708, 26324, 12853, 16241, 20281, 30934, 36654, 17071, 34441, 16802, 17927, 5929, 26121, 19942, 36279, ………….…………………………….,28872, 34230, 17149, 24355, 19493]**132**

By observing the output file, we observed that the alignment of sequence reads with their positions and found the matched positions of mutation which causes cancer.

As the tumour mutation will require the 30 folds of sequence coverage that matched with normal tissue. By associating, the novel draft of human genome required around 65-fold coverage

Recognizing driver mutations arethe main aim of sequencingcancer genome: the increase of mutation rate in a cell will change the gene which leads to speedier evolution of tumour and metastasis. To determine the mutation of driver in an DNA sequence is difficult but drivers inclined at most usually shared mutation between tumours, cluster around identified oncogenes are inclined as non-silent. The Passenger mutations are randomly circulated throughout the genome which is not important for the progression of disease. It has been expected thattumor carries 80 somatic mutations averagelyless than 15 of which remainpredictable to be drivers.

3.1 Classification of genome using machine learning

Step 1: Import genome sequencing dataset from the MongoDB into python program language.

Step 2: pre-process the dataset of genome sequence

Step 3: To formalize the dataset as shown in figure 2. the formalization of sequence is done by counting the genome sequence, uniqueness in the sequence and also found the frequnce, top priority of the genome sequence

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
count	106	106	106	106	106	106	106	106	106	106	106	106	106	106	106	106	106	106	106	106	106	106	106	106	106	106	106	106
unique	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
top	t	a	a	c	a	a	a	a	a	a	a	c	t	c	t	t	g	a	c	a	t	a	g	t	t	g	g	t
freq	38	34	30	30	36	42	38	34	33	36	38	31	34	38	54	54	53	40	44	31	34	31	30	32	32	34	29	32

Figure 2 Formalization of genome sequence

Step 4: And then formalized data is formatted by counting the text of each genome sequence is counted and recorded each sequence data as illustrated in figure 3.

```
     0     1     2     3     4     5  ...    52    53    54    55    56  Class
t  38.0  26.0  27.0  26.0  22.0  24.0  ...  35.0  30.0  23.0  29.0  34.0    NaN
c  27.0  22.0  21.0  30.0  19.0  18.0  ...  21.0  32.0  29.0  29.0  17.0    NaN
a  26.0  34.0  30.0  22.0  36.0  42.0  ...  25.0  22.0  26.0  24.0  27.0    NaN
g  15.0  24.0  28.0  28.0  29.0  22.0  ...  25.0  22.0  28.0  24.0  28.0    NaN
-   NaN   NaN   NaN   NaN   NaN   NaN  ...   NaN   NaN   NaN   NaN   NaN   53.0
+   NaN   NaN   NaN   NaN   NaN   NaN  ...   NaN   NaN   NaN   NaN   NaN   53.0

[6 rows x 58 columns]
```

Figure 3 Recording of every genome sequence count

Step 5: To run machine learning algorithms on the data the data must be in string format because it is very difficult to classify the string format of genome sequencing data. So firstly we convert the string format into numerical data as shown in figure 4

	0_a	0_c	0_g	0_t	1_a	1_c	1_g	1_t	2_a	2_c	2_g	2_t	3_a	3_c	3_g	3_t	4_a	4_c	4_g	4_t	5_a	5_c	5_g	5_t	6_a	6_c	6_g	6_t	7_a	7_c
0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0	1	0	0	1	0	0	1	0
1	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	1	0	0	0	1
2	0	0	1	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0	1	0	1	0
3	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0	1	0	0	1	0	1	0
4	0	0	0	1	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0

5 rows × 230 columns

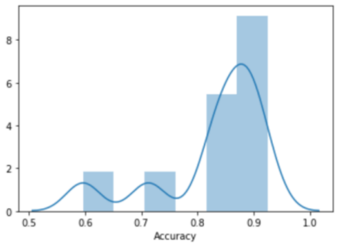
Figure 4 conversion of string format into numerical data

Step 6: Train and test the sequence by using classification algorithms of machine learninglike SVM, Naive bayes, Neural Net, Decision Tree, Gaussian Process, Nearest Neighbors, AdaBoost, Random Forest.For testing the algorithm note the accuracies of different machine learning procedure as mentioned in table 2.

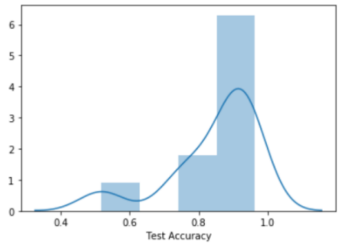
Table 2. Different machine learningalgorithms Accuracy

Models	Accuracy	Test Accuracy
Naive Bayes	0.837500	0.9259259259259259
SVM Linear	0.850000	0.9629629629629629
SVM RBF	0.887500	0.9259259259259259
SVM Sigmoid	0.900000	0.9259259259259259
Neural Net	0.887500	0.9259259259259259
AdaBoost	0.925000	0.8518518518518519
Random Forest	0.596429	0.5185185185185185
Nearest Neighbors	0.823214	0.7777777777777778
Gaussian Process	0.873214	0.8888888888888888
Decision Tree	0.712500	0.7407407407407407

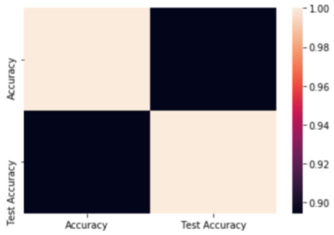
By comparing all the algorithms of machine learning accuracies "Adaboost" algorithm method is best method. And visualized the accuracy and test accuracy of the different algorithm as shown in figure 5.



a) Accuracy



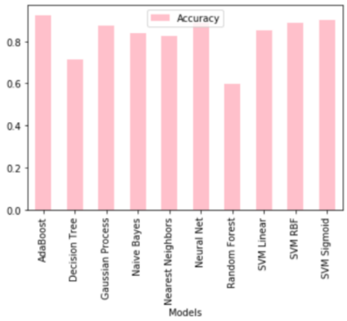
b) test accuracy



c) correlation matrix

Models	Accuracy
AdaBoost	0.925000
Decision Tree	0.712500
Gaussian Process	0.873214
Naive Bayes	0.837500
Nearest Neighbors	0.823214
Neural Net	0.887500
Random Forest	0.596429
SVM Linear	0.850000
SVM RBF	0.887500
SVM Sigmoid	0.900000

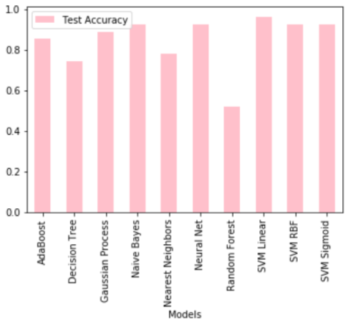
d)Training of accuracy



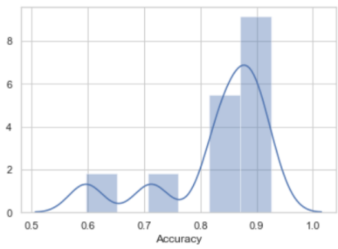
e)plot of accuracy

Models	Test Accuracy
AdaBoost	0.851852
Decision Tree	0.740741
Gaussian Process	0.888889
Naive Bayes	0.925926
Nearest Neighbors	0.777778
Neural Net	0.925926
Random Forest	0.518519
SVM Linear	0.962963
SVM RBF	0.925926
SVM Sigmoid	0.925926

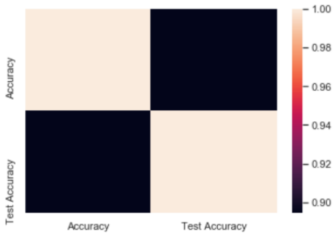
f) training of accuracy



g) plot of test accuracy



h)plot of accuracy




```
Accuracy         1.000000
Test Accuracy    0.894421
Name: Accuracy, dtype: float64

---------------------
Test Accuracy    1.000000
Accuracy         0.894421
Name: Test Accuracy, dtype: float64
```

i) correlation matrix of test accuracy j) training of accuracy and test accuracy

Figure 5 visualization of accuracy

By combining the accuracy visualization of figure 5 we visualized both the accuracy in one graph as shown in figure 6 by graph we concluded that Adaboost algorithm is best for the classification of genome.

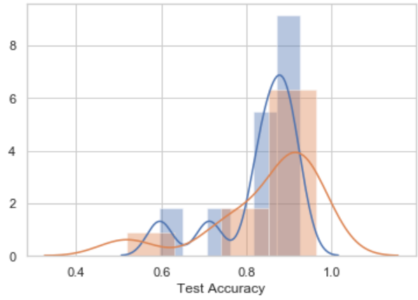


Figure 6 visualization of different machine learning algorithms accuracy

4. Conclusion

The sequencing technology like massive parallel sequencing to prospect the genome sequencing list of each DNA base in a genome, a genome map recognizes the innovations. A map of genome is detailed less and helps in navigating the genome sequence around the genome. By applying Burrows_Wheeler_Alignment (BWA) founded the genome alignment and found the characters of the reference genome with the occurrence character and searching for max difference threshold by analysing the output of sequence reads is aligned with their positions and found the matched positions of mutation which causes cancer. Because cancer genome would need 30-fold coverage of sequence in a genome and a corresponding normal flesh. By associating, the novel draft of human genome requiredaround 65-fold coverage.By applying different algorithms of machine learning to the sequence dataset to formalize which algorithm is best to train and test the classification algorithms. Here Naive Bayes, Support Vector Machine(SVM), Ada Boost,Nearest Neighbours, Random Forest, Decision Tree etc., algorithms are applied and founded the accuracy for each classifier. By comparing accuracy of each classifier Ada Boost method has the high accuracy and visualization.

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001; 409:860–921.
2. N.Peek,J.H.Holmes, J.Sun, technical challenges for big data in biomedicine and health:datasources,infrastructure, and analytics, yearbook of medical informatics 9(2014) 42-47
3. *Kohane, I. S.; Masys, D. R.; Altman, R. B. (2006). "The Incidentalome: A Threat to Genomic Medicine". JAMA.* ***296*** *(2): 212–215.* doi:10.1001/jama.296.2.212. PMID 16835427.
4. *Allan Maxam; Walter Gilbert (February 1977).* "A new method for sequencing DNA". *PNAS.* ***74*** *(2): 560–4.* doi:10.1073/pnas.74.2.560. PMC 392330. PMID 265521.
5. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–1760.
6. Li H. Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM [preprint]. New York NY: Oxford University Press; 2013. http://arxiv.org/pdf/1303.3997v2.pdf. Accessed November 28, 2016.
7. Michael k, K. Leung, Andrew Delong et al. Machine Learning in Genomic Medicine: A Review of Computational problems and Data Sets. 2016, Machine Learning in Genomic Medicine. Vol. 104, no.1.
8. Wagstaff, K., 2012. Machine learning that matters. arXivpreprint arXiv:1206.4656.

9. Cunningham, S.J., Littin, J. and Witten, I.H., 1997. Applications of machine learning in information retrieval

10. Bennett, K.P. and Parrado-Hernández, E., 2006. The interplay of optimization and machine learning research. Journal of Machine Learning Research, 7(Jul), pp.1265-1281.

11. Caruana, R. and Niculescu-Mizil, A., 2006, June. An empirical comparison of supervised learning algorithms. InProceedings of the 23rd international conference on Machine learning (pp. 161-168). ACM.

12. Mitchell, T.M., 2006. The discipline of machine learning (Vol. 3). Carnegie Mellon University, School of Computer Science, Machine Learning Department.

13. S. Dinakaran and P. RanjitJebaThangaiah Ensemble Method of Effective AdaBoost Algorithm for Decision Tree Classifiers. 2017

14. D. Cai, A. Delcher, B. Kao and S. Ksif, "Modeling splice sites with Bayes networks", *Bioinformatics*, vol. 16, pp. 152-158, 2000.

15. Y. Freund and R. E. Schapire, "A Short Introduction to Boosting", *Journal of Japanese Society for Artificial Intelligence*, vol. 14, pp. 771-780, 1999.

16. Pang H, Lin A, Holford M, Enerson BE, Lu B, Lawton MP, et al. Pathway analysis using random forests classification and regression. Bioinformatics. 2006; 22(16):2028–2036.

17. *Chen, Zexun; Wang, Bo; Gorban, Alexander N. (2019). "Multivariate Gaussian and Student-t process regression for multi-output prediction". Neural Computing and Applications. arXiv:1703.04455. doi:10.1007/s00521-019-04687-8.*

18. Kac, M.; Siegert, A.J.F (1947). "An Explicit Representation of a Stationary Gaussian Process". *The Annals of Mathematical Statistics*. **18** (3): 438–442. doi:10.1214/aoms/1177730391.

19. Nigsch, Florian; Bender, Andreas; van Buuren, Bernd; Tissen, Jos; Nigsch, Eduard; Mitchell, John B. O. (2006). "Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization". *Journal of Chemical Information and Modeling*. **46** (6): 2412–2422. doi:10.1021/ci060149f. PMID 17125183.

20. Rokach, Lior; Maimon, O. (2008). *Data mining with decision trees: theory and applications*. World Scientific Pub Co Inc. ISBN 978-9812771711.

21. Quinlan, J. R. (1986). "Induction of decision trees" (PDF). *Machine Learning*. **1**: 81 106. doi:10.1007/BF00116251. S2CID 189902138.

22. Ferreira, C. (2006). "Designing Neural Networks Using Gene Expression Programming" (PDF). In A. Abraham, B. de Baets, M. Köppen, and B. Nickolay, eds., Applied Soft Computing Technologies: The Challenge of Complexity, pages 517–536, Springer-Verlag.

23. Sak, Hasim; Senior, Andrew; Beaufays, Francoise (2014). "Long Short-Term Memory recurrent neural network architectures for large scale acoustic modeling"(PDF). Archived from the original (PDF) on 24 April 2018.

24. G. Adrián, G. E. Francisco, M. Marcela, A. Baum, L. Daniel, and G. B. de QuirósFernán, “Mongodb: an open source alternative for HL7-CDA clinical documents management,” in *Proceedings of the Open Source International Conference (CISL '13)*, Buenos Aires, Argentina, 2013.