Support Vector Machine based Breast Cancer Classification using Next Generation Sequences

Babymol Kurian[1], V.L.Jyothi[2]
{babymolkurian@gmail.com[1], jyothivl15@yahoo.com[2]}

Sathyabama Institute of Science and Technology,Chennai[1],
Department of Computer Science & Applications, GuruShree ShanthiVijai Jain College, chennai[2]

Abstract. Next Generation Sequencing is inevitable for providing better approach for predicting and curing diseases with high success rate in an appreciable timeline. Modern technology such as machine learning support the medical research with high speed and tremendous accuracy from disease prediction to cure. In this paper, the supervised learning model, Support Vector Machine is applied on next generation sequences for the prediction of breast cancer. Ten basic features of DNA sequences such as individual nucleobase average count of A, G, C, T, AT and GC-content, AT/GC composition, G-Quadruplex occurrence, ORF (Open Reading Frame) count and MR (Mutation Rate) are used for framing the feature vector. The feature vectors along with the class value are considered as the dataset for supervised learning. Datasets are prepared to classify (class value) as '0' for normal sequences, '1' for BRCA1 cancer sequences and '2' for BRCA2 cancer sequences. Four different categories of datasets are prepared with 50, 100, 150 and 200 sequences for each class of normal sequence, BRCA1 and BRCA2 cancer sequence. While increasing the dataset size, the outlier, the distribution and scattered features of data were also analysed. The datasets are split into training and testing set with 80:20 ratio for the classification process. SVM model in Python is applied for supervised classification process.

Keywords: Support vector machine, supervised machine learning, breast cancer, multiple classification, next generation sequencing.

1Introduction

Cancer is one of the leading life threatening disease for human beings worldwide. Various types of cancers such as skin, lungs, breast, rectum, stomach, prostate, liver, cervix, oesophagus, bladder, blood and mouth are common in India and worldwide. Major factors for the occurrence of cancer are internal factors such as mutation in genes, hormonal changes, low immunity and external factors such as advanced technological developments, environmental changes, high population and eating habits. Even one among the abovementioned factors alone can cause cancer in human beings. The next generation sequencing as well as technological development in artificial intelligence is playing a vital role in the prediction of cancer.

Machine learning process resumes with sequence collection, both normal and cancer affected sequences from National Center for Biotechnology Information(NCBI) gene repository which is followed by features extraction from sequences. The next process is analysis of the extracted features using Box method for finding the outliers, Histogram for understanding the data distribution and Scatter matrix for finding the relationship among features. Fourth step is vector formation using the extracted features as well as class value to represent the normal class as '0', BRCA1 associated cancer class as '1' and BRCA2 associated cancer as '2'. Following step is to decide the training and testing set with the ratio 80:20, from the dataset. In the sixth step, SVM model is framed for classification process using the training dataset. Next step is testing the framed model with the test dataset. Finally, model generated output for the test data are compared with original/actual output of the test data to measure the model performance. The parameters such as accuracy value and F1 score are extracted from confusion matrix which is derived from generated and actual output of the test dataset.

ICASISET 2020, May 16-17, Chennai, India

DOI 10.4108/eai.16-5-2020.2303953

2 Related Work

ML techniques have been extensively used in intelligent healthcare systems, especially for Breast Cancer (BC) diagnosis and detection over the past few decades. Experience of the physician was used traditionally for accuracy of any kind of diagnosis [1]. Many years of analysis of different patients' symptoms and confirmed diagnoses were the reason for this expertise. Still there is no 100% guarantee for this accuracy. Computing technologies paved the way for a difference to the traditional way of treatment and it is so simple to acquire and store huge medical data preferably the dedicated databases of electronic patient records [2]. The intelligent healthcare system plays a vital role in accurate diagnosis, detection and prediction of diseases.

Breast cancer has a high incidence and mortality rate, which is the most predominant cancer in women. ML techniques are playing a substantial role in diagnosis and prognosis of BC by applying classification techniques. Classification is a predictive modelling problem. A comprehensive elucidation of different classification methods applied to BC would be focused in subsequent sections. Emphasis is on Support Vector Machine (SVM), k-Nearest Neighbour (k-NN), Decision Tree (DT) and Bayesian Hidden Markov Model (HMM) techniques as they are the core methods used in BC diagnosis and prognosis. ML approaches have proved to be more efficient and appropriate than medical statistics for BC datasets as discussed in [3–6].

Based on the data types and structures, supervised or unsupervised ML algorithm can be chosen. A study of recent research divulges that most of the ML algorithms in the BC diagnosis and prognosis are supervised [7].The concept of Support Vector Machine [8], proposed by Vapnik on the basis of the statistical learning theory [9], has become a major component of ML techniques. The SVM-based supervised machine learning can be applied to predict the effect of mutations on protein stability [10]. The algorithm was also applied for predicting genes in metagenomics fragments [11]. Integrating Rao's score test and SVM supervised machine learning approach, an approach (DriverML) was developed [12]. To find out cancer driver genes, a unique prediction model was proposed for druggable proteins using the Support Vector Machine method [13]. SVM algorithm can be applied to predict cysteines with hyper-reactivity based on local sequence features [14]. The classification accuracy, ROC, F-measure, and computational times of training SVM and SVM ensembles are compared to develop breast cancer prediction models [15].

k-Nearest Neighbour is another core ML technique in classification [16]. k-NN is a lazy non-parametric learning algorithm used for classification as well as regression, which classifies the objects using their "k" nearest neighbours. k-NN only considers the neighbours around the object and not the underlying data distribution. To predict protein structures, machine learning can be applied to sequence alignment generation using K-Nearest Neighbour algorithm [16]. A multiple classification approach for predicting individuals inherited genomic susceptibility was derived using K-NN algorithm [17].

In machine learning, Decision Tree is a predictive model that denotesmapping between object values and object attributes [18,19]. It partitions every possible data outcome recursively into classes. In DT, every non-leaf node specifies a test on a particular attribute, every branch represents a result of that test and every leaf node expresses a classification or decision. The node at the topmost label in the tree corresponds to the best predictor which is called root node. Both numerical and categorical data can be processed using DT. Most prevalent DT methods are Iterative Dichotomiser 3 (ID3) [20], C4.5 [21], C5.0 and Classification and Regression Tree (CART). BC diagnosis and prognosis are carried out using Decision Tree structure, to predict whether a person has malignant or benign BC [20]. In [22], Bayesian hidden Markov model (HMM) with Gaussian Mixture (GM) clustering approach is used to model the DNA copy number change throughout the genome[23][24].

3 Machine Learning Process

The machine learning process comprises of eight steps: sequence collection, feature extraction, feature analysis, vector formation, training/testing set construction, framing the SVM model using the training set, testing the model with testing set and performance analysis to measure the efficiency of the model in cancer classification.

3.1 Sequence Collection

The sequences are collected to form the incremental dataset with 50, 100, 150 and 200 sequences in each class. Collection of sample sequences are shown in **Figure 1**.

Fig. 1. Sample Sequence

3.2Feature Extraction

The next step in the machine learning process is the feature extraction. Eight features such as individual nucleobase average count of A, G, C, T, AT and GC-content, AT/GC composition, G-Quadruplex occurrence are measured based on average occurrence and other two features such as ORF (Open Reading Frame) count and MR (Mutation Rate) are measured for each sequence. The features are evaluated using following equations from (1) to (7).

$$X = \sum G_x / N \quad (1)$$

X ➔ {A, G, C, T}
G_x ➔ The element X in the DNA sequence S
Σ G_x➔ Total occurrence of the element X in sequence S
N ➔ Length of the sequence S

$$AT = (\sum G_A + \sum G_T) /N \quad (2)$$

AT ➔ Average occurrence of A and T in the sequence S
Σ G_A➔ Total occurrence of the genome A in the sequence S
Σ G_T➔ Total occurrence of the genome T in the sequence S
N ➔ Length of the sequence S

$$GC = (\sum G_G + \sum G_C) / N \quad (3)$$

GC➔ Average occurrence of G and C in the sequence S
Σ G_G➔ Total occurrence of the genome G in the sequence S
Σ G_C➔ Total occurrence of the genome Cin the sequence S
N ➔ Length of the sequence S

$$4G = \sum G_{"GGGG"}/N \quad (4)$$

4G ➔ Average occurrence of continuous 4 G's in the sequence S
Σ $G_{"GGGG"}$➔ Total occurrence of continuous 4 G's in the sequence S
N ➔ Length of the sequence S

$$C_{ORF} = \sum ORF \quad (5)$$

$$ORF = Start_{Codon} - Stop_{Codon} \quad (6)$$

StartCodon ➔ “ATG”
Stopcodon ➔ {“TAG”, “TAA”, “TGA”}
ORF ➔ The existence of codons between start codon to any one of stop codons in the sequence S
CORF ➔Sum of occurrence of ORFin the entire sequence S

$$\text{Mutation Rate (MR)} = 100 - [(\sum M_S / L_{AS} * 100) + ((\sum I_S + \sum D_S) / L_{AS}) * 100)] \quad (7)$$

$\sum M_S$ ➔ Total number of genomic matches in the aligned sequence with respect to the reference sequence (R) and the sequence (S)
$\sum I_S$ ➔ Total number of genomic insertions in the aligned sequence with respect to the reference sequence (R) and the sequence (S)
$\sum D_S$ ➔ Total number of genomic deletions in the aligned sequence with respect to the reference sequence (R) and the sequence (S)
L_{AS} ➔ Length of the aligned sequence AS.
Note: The sequence alignment between the reference sequence R and the sequence S is done with “Global Alignment Technique”

3.3Feature Analysis

The strength of features are evaluated for classification purpose using three methods, Box Plot graph, Histogram method and Scatter Matrix method.

Box Plot graph. Box plot is a pictorial form of showing the distribution of data or feature values. Box plot is shown in **Figure 2**.

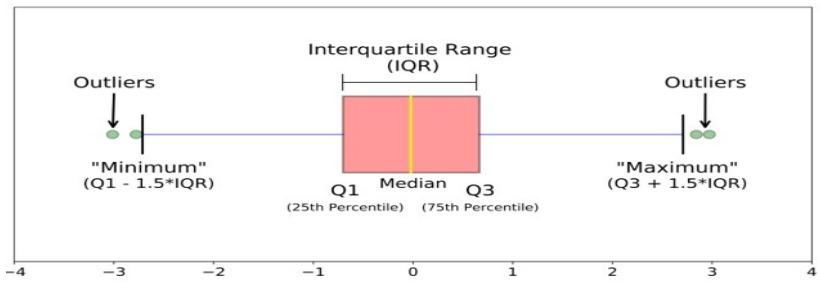

Fig. 2. Box Plot
[Courtesy- https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51]

The box plot is represented by five statistical elements, minimum, first quartile (Q_1), median, third quartile (Q_3) and maximum. The region between Q_1 and Q_3 is referred as Interquartile range (IQR). Data values which lie outside the minimum and maximum range are referred as Outlier. For every feature including the class value, the box plot graph is constructed for all four incremental dataset, shown in **Figure 3**.

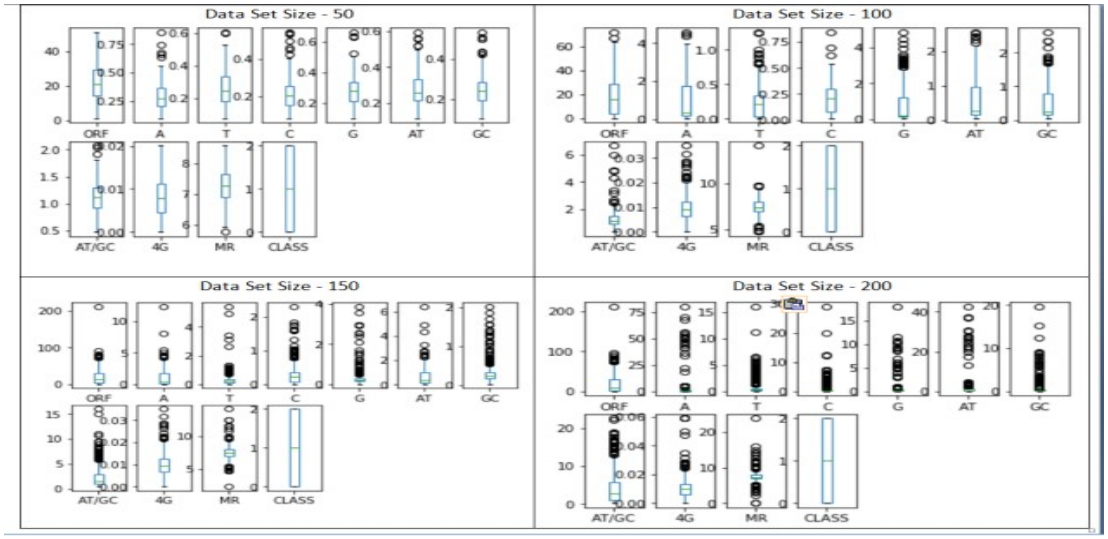

Fig. 3. Features Representation using Box Plot Graph

For the dataset size of 50, outlier datas are less whereas for the dataset size of 200, the outliers are more, due to the variation in the incremental data size. Outlier occurrences in ORF, 4G and MR are comparativelyless. These three features strongly support classification.

Histogram Method. Another visual aid to find the data distribution is Histogram. The X axis in the histogram represents the average occurrence value and the Y axis represents the frequency/count of data at specific intervals. The bar height describes the total number of occurrences of data at specific intervals. The histogram for all features including the class feature are shown in **Figure 4**.

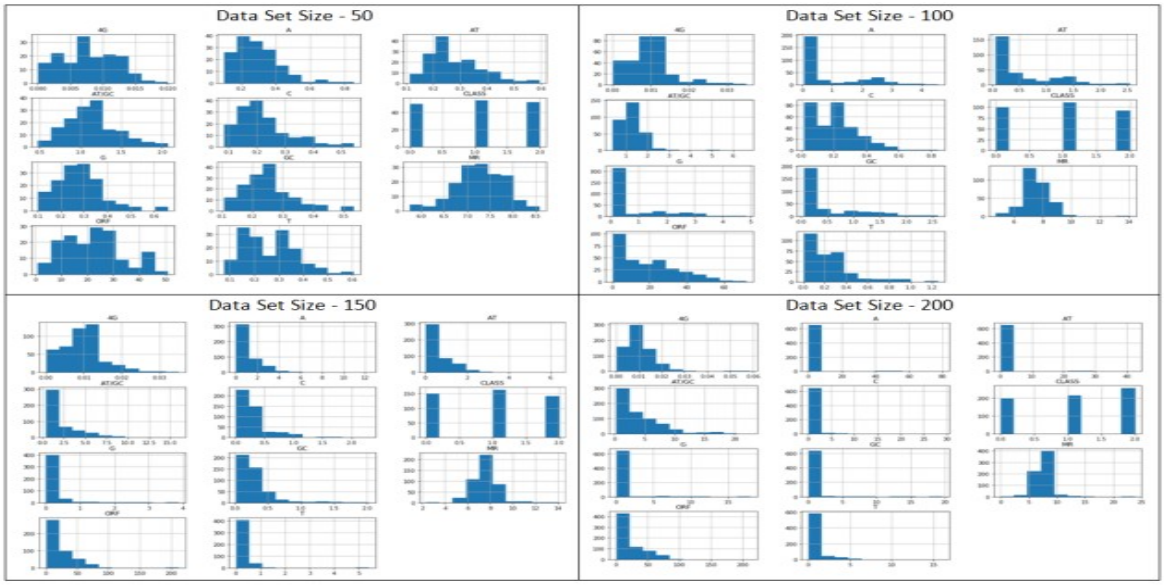

Fig. 4. Histogram Representation for the Features

For the dataset size of 50, the feature bar distribution appears wider than the bar height. This specifies that values are loosely bunched around the mean. Feature bar distribution for the higher dataset size of 200 appears more narrow than the bar height, which designates the feature values are strongly bunched around the mean. This is an indication that the classification accuracy can be increased by increasing the volume of the dataset.

Scatter Matrix. 2D scatter matrix is represented as a squared grid, where the features are specified in rows as well as columns in the same order. A scatter plot chart is generated at each grid cell to show the relation between the intersecting features in the row and column of the grid. All ten features along with the class values

are represented in scatter matrix which is shown in **Figure 5**. The scatter level is wider for the dataset size of 50, which indicates less classification accuracy whereas scatter level is narrow for the dataset size of 200 which indicates more classification accuracy.

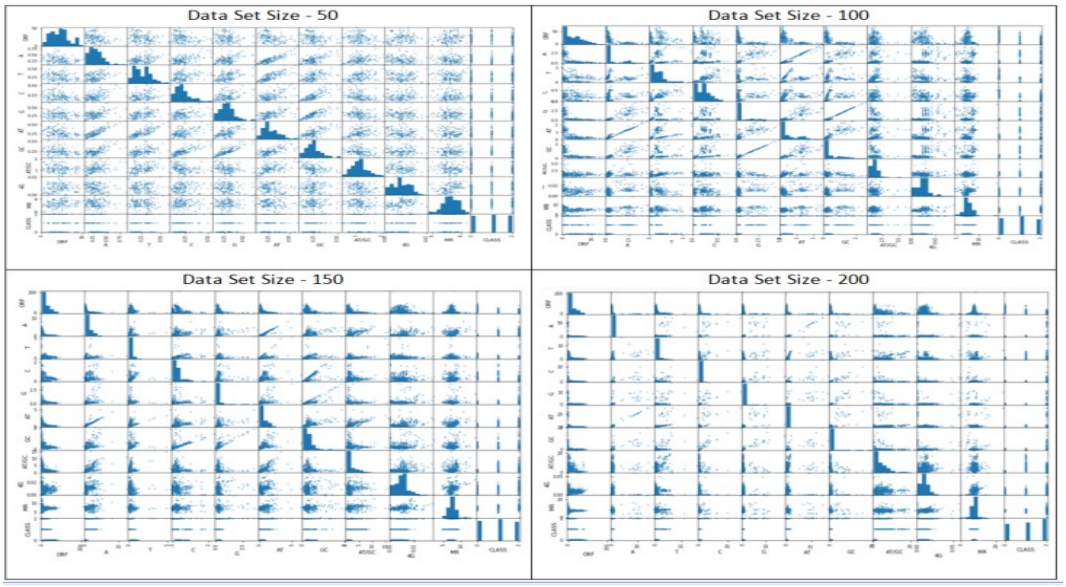

Fig. 5. Feature Representation using Scatter Matrix

Vector Formation. All the ten features along with the class value represent the vector value. The class value is the target fixed for the relevant vector value. The sample is generated with one vector for normal sequence and another vector for BRCA1 sequence, as shown in Table 1.

Table 1. Vector representation

ORF	A	T	C	G	AT	GC	AT/GC	4G	MR	CLASS
14	0.305	0.1378	0.2354	0.2426	0.2214	0.239	0.92636	0.0102	7.61523	0
20	0.345	0.3332	0.19218	0.2977	0.3400	0.2449	1.3883	0.0068	6.8250	1

Training and Testing Dataset Construction. From the dataset, the training and testing dataset are constructed with the ratio of 80:20 which is shown in equation (8). Python method "train_test_split( )" is used for training and testing dataset formation. The set {x_train, y_train} represent the training dataset. The x_train variable describes the training vector. The y_train variable has class values of x_train training vectors. The set {x_validation, y_validation} represent the testing dataset. x_validation has a collection of feature vectors which are mutually exclusive with x_train dataset. The class values for the x_validation is stored in y_validation, which is used to measure the performance of the model for completing the testing process.

(x_train,x_validation,y_train,y_validation = train_test_split | x,y,test_size=0.20,random_state=1) (8)

SVM Model Formation using Training Dataset. Support Vector Machine classifier is applied for the classification. Three classes are derived for the dataset. In Python, sklearn.svm module is used for performing the supervised learning. The SVC method is imported for constructing the Support Vector Machine model. Python script for SVM model formation is shown below. In the script, the import command loads all the necessary libraries, classes and methods related to SVC. The model is constructed using SVC() method. Thereafter the model is fine-tuned with the training dataset, x_train, having feature vectors and y_train havingclass values related to the feature vectors.

```
from sklearn.svm import SVC
model = SVC(gamma = 'auto')
model.fit(x_train,y_train)
```

Testing the SVM Model. The fine-tuned SVM model is ready for the classification process. The model has to be tested with set of data which are not in the training set. For testing, the predict() method is used with x_validation (testing feature vectors from testing dataset) as the parameter, shown below. The x_validation feature vectors are passed on to the SVM framed model and their classes are predicted and stored into the 'Prediction' variable.

```
Prediction = model.predict(x_validation)
```

Performance Evaluation. The performance of the SVM model is evaluated using confusion matrix. The confusion matrix is constructed by comparing the original class values of the testing set, 'y_validation' and the model generated class values for the testing set, 'Prediction'. From the confusion matrix, precision, recall, F1score and support values for each class (0,1 and 2) are extracted. Finally, the overall accuracy of the model is also extracted. Training, testing and performance measures are carried out for all four types of incremental datasets of sizes 50, 100, 150 and 200 in each class.Overall system architecture is shown in **Figure6**.

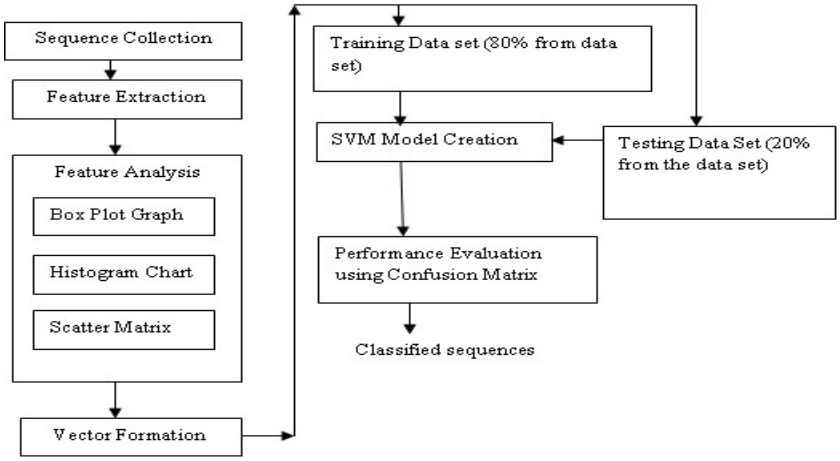

Fig. 6. Architecture Diagram for Classification Process

4 Results and Discussion

Supervised machine learning process is done with Support Vector Machine method. The dataset is constructed with various sizes approximately from 50 in each class to 200 which is shown in Table 2. For dataset size of 50, total volume of entire three classes were 156 sequences, for size of 100, 303 sequences, for 150 size, 454 sequences and for 200 size, 667 sequences were considered for the process. Four types of datasets were represented as {125, 242, 363, 533} for the training set and {32, 61, 91, 134} for the testing set. SVM model is constructed with the training set for making multiple classification as '0' for normal sequences, '1' for BRCA1 sequences and '2' for BRCA2 sequences.

Table 2. Dataset Description

Defined Data size of each class	Dataset Size (for 3 classes)	Dataset Assigned by System	
		Training set (80%)	Testing Set (20%)
50	156	125	32
100	303	242	61
150	454	363	91
200	667	533	134

The constructed SVM model is tested with the testing dataset. The system generated classes for testing data are stored as 'Prediction'. The original classes for testing set are stored as 'y_validation'. For data size of 50, the test set size is 32. The outcome of system generated (P) test values are compared with the original testing values (O), shown in **Figure 7**. The red colour indicates the wrongly predicting classes. Black colour specifies the correctly predicting classes. For class '0', correctly identified ('0' with black in both rows), are 4, wrongly identified as value '1' are 3 and wrongly identified as '2' are 3. Class 0 is specified as'C1' and the data identified as '0','1' and '2' are listed in the first row of the confusion matrix for the dataset size of 50 in Table 3.

```
O-> [2 1 1 0 0 0 0 2 1 2 2 1 1 1 1 0 2 0 0 0 1 2 2 1 1 1 1 0 2 0 2 2 ]
P-> [1 1 1 2 0 2 1 1 1 1 1 0 1 2 1 1 1 0 0 0 1 2 2 1 2 1 2 1 0 2 0 0 ]
```

Fig. 7. Classification Comparison of Original vs System Generated Test Dataset

In the class 0, total dataset taken for the data size 50 is 10. Of 10, only 4 is identified correctly. For the same class value with the data size of 100, total dataset taken is 18 and 15 are identified correctly. For the class 0, 23 of 25 are classified correctly in the category of data size of 150. In the dataset size of 200, for class 0, 36 of 38 are classified correctly. It is observed from the confusion matrixthat,more the volume of the dataset, better the result.

Table 3. Confusion Matrix

Confusion Matrix

Dataset Size 50

Class	C1	C2	C3
C1	4	3	3
C2	1	8	3
C3	3	5	2

Dataset Size 100

Class	C1	C2	C3
C1	15	1	2
C2	7	13	4
C3	8	6	5

Dataset Size150

Class	C1	C2	C3
C1	23	2	0
C2	7	22	9
C3	6	6	16

Dataset Size 200

Class	C1	C2	C3
C1	36	1	1
C2	3	25	13
C3	0	4	51

From the confusion matrix, other statistical factors such as precision, recall, F1 score and support are generated to measure the model performance, shown in **Figure 8**. The first column represents class values with values 0.0, 1.0 and 2.0 in the report. The recall values are nothing but the accuracy values which are shown for each class. The overall accuracy is measured in the last row as average/total. F1 score is directly proportional to recall values. Support column describes the data taken under each class for testing process. The last row value for support column project the total data taken for the classification process under each category of dataset.

System Generated Classification Report

Data set Size 50

	precision	recall	f1-score	support
0.0	0.50	0.40	0.44	10
1.0	0.50	0.67	0.57	12
2.0	0.25	0.20	0.22	10
avg / total	0.42	0.44	0.42	32

Data set Size 100

	precision	recall	f1-score	support
0.0	0.50	0.83	0.62	18
1.0	0.65	0.54	0.59	24
2.0	0.45	0.26	0.33	19
avg / total	0.54	0.54	0.52	61

Data set Size 150

	precision	recall	f1-score	support
0.0	0.64	0.92	0.75	25
1.0	0.73	0.58	0.65	38
2.0	0.64	0.57	0.60	28
avg / total	0.68	0.67	0.66	91

Data set Size 200

	precision	recall	f1-score	support
0.0	0.92	0.95	0.94	38
1.0	0.83	0.61	0.70	41
2.0	0.78	0.93	0.85	55
avg / total	0.84	0.84	0.83	134

Fig. 8. Classification Report

The accuracy values are represented in the graph shown in **Figure 9**. The trend line in the graph is projected up as the dataset size increases. This chart clearly depicts that, the classification process can be optimal for large dataset.

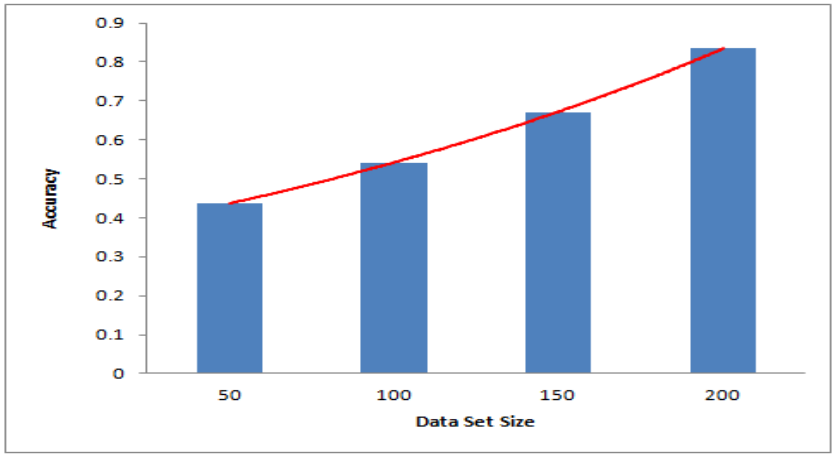

Fig. 9. Classification Accuracy

5 Conclusion

The need for prediction and cure of cancer is eminently important in the reduction of mortality and prevalence worldwide. In this paper, the DNA sequences were collected for three different categories(classes), such as normal human sequences, BRCA1 and BRCA2 cancer sequences. One of the supervised machine learning models, Support Vector Machine model was constructed to classify sequences. The model was framed and tested with four incremental datasets of size 50, 100, 150 and 200 approximately in each class. The datasets were split as training and testing set with ratio of 80:20. Confusion matrix was generated for allfour incremental datasets. From the confusion matrix, some statistical factors such as precision, recall and F1score were generated. The recall value projects the accuracy value. The accuracy was more as the dataset sizeincreased from 50 to 200. To conclude, the classification accuracy with a total of 667 datasets(with a miimum of 200 datasets in each class) could reach83% approximately.

References

1. Meesad P, Yen GG. Combined numerical and linguistic knowledge representation and its application to medical diagnosis. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans. 2003 Aug 11;33(2):206-22.
2. Pavlopoulos SA, Delopoulos AN. Designing and implementing the transition to a fully digital hospital. IEEE Transactions on information technology in biomedicine. 1999 Mar;3(1):6-19.
3. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine. 2005 Jun 1;34(2):113-27.
4. Funahashi KI, Nakamura Y. Approximation of dynamical systems by continuous time recurrent neural networks. Neural networks. 1993 Jan 1;6(6):801-6.
5. Razi MA, Athappilly K. A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. Expert Systems with Applications. 2005 Jul 1;29(1):65-74.
6. Subasi A, Ercelebi E. Classification of EEG signals using neural network and logistic regression. Computer methods and programs in biomedicine. 2005 May 1;78(2):87-99.
7. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer informatics. 2006 Jan;2:117693510600200030.
8. Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995 Sep 1;20(3):273-97.
9. Vapnik VN. " An Overview of Statistical Learning Theory", IEEE transactions on neural networks, vol. 10,№ 5.
10. Pandurangan AP, Blundell TL. Prediction of impacts of mutations on protein structure and interactions: SDM, a statistical approach, and mCSM, using machine learning. Protein Science. 2020 Jan;29(1):247-57.
11. Al-Ajlan A, El Allali A. The Effect of Machine Learning Algorithms on Metagenomics Gene Prediction. InProceedings of the 2018 5th International Conference on Bioinformatics Research and Applications 2018 Dec 27 (pp. 16-21).
12. Han Y, Yang J, Qian X, Cheng WC, Liu SH, Hua X, Zhou L, Yang Y, Wu Q, Liu P, Lu Y. DriverML: a machine learning algorithm for identifying driver genes in cancer sequencing studies. Nucleic acids research. 2019 May 7;47(8):e45-.
13. Kandoi G, Acencio ML, Lemke N. Prediction of druggable proteins using machine learning and systems biology: a mini-review. Frontiers in Physiology. 2015 Dec 8;6:366.
14. Wang H, Chen X, Li C, Liu Y, Yang F, Wang C. Sequence-based prediction of cysteine reactivity using machine

learning. Biochemistry. 2018 Jan 30;57(4):451-60.
15. Huang MW, Chen CW, Lin WC, Ke SW, Tsai CF. SVM and SVM ensembles in breast cancer prediction. PloS one. 2017 Jan 6;12(1):e0161501.
16. Moreno-Seco F, Micó L, Oncina J. A modification of the LAESA algorithm for approximated k-NN classification. Pattern Recognition Letters. 2003 Jan 1;24(1-3):47-53.
17. Kim BJ, Kim SH. Prediction of inherited genomic susceptibility to 20 common cancer types by a supervised machine-learning method. Proceedings of the National Academy of Sciences. 2018 Feb 6;115(6):1322-7.
18. De Mántaras RL. A distance-based attribute selection measure for decision tree induction. Machine learning. 1991 Jan 1;6(1):81-92.
19. Mingers J. An empirical comparison of selection measures for decision-tree induction. Machine learning. 1989 Mar 1;3(4):319-42.
20. Quinlan JR. Induction of decision trees. Machine learning. 1986 Mar 1;1(1):81-106.
21. Quinlan J. C4. 5: programs for machine learning. Elsevier; 2014 Jun 28.
22. Manogaran G, Vijayakumar V, Varatharajan R, Kumar PM, Sundarasekar R, Hsu CH. Machine learning based big data processing framework for cancer diagnosis using hidden Markov model and GM clustering. Wireless personal communications. 2018 Oct 1;102(3):2099-116.
23. Vinod Jagannath Kadam, Shivaji rao Manik rao Jadhav, K.Vijayakumar, “Breast Cancer Diagnosis Using Feature Ensemble Learning Based on Stacked Sparse Auto encoders and Soft max Regression”, Image & Signal Processing, springer, june 2019.
24. K. Vijayakumar , K. Pradeep Mohan Kumar ,Daniel Jesline, “Implementation of Software Agents and Advanced AoA for Disease Data Analysis”, journal of medical systems, Part of Springer Nature 2019.