DISEASE PREDICTION SYSTEM USING SEQUENCE ALIGNMENT

Anjana B[1], Anjuniranjana A[1], R Vaidehi[1], Jain Stoble[2]

1 UG Student, Department of Computer Science and Engineering, Adi Shankara Institute of Engineering and Technology, Kalady, Ernakulam, Pin: 683574

2 Assistant Professor, Department of Computer Science and Engineering, Adi Shankara Institute of Engineering and Technology, Kalady, Ernakulam, Pin: 683574

Abstract-- Cancer is one of the most dreaded ailments on the planet. It has expanded shockingly and bosom disease happens in one out of eight ladies, the forecast of malignancies assumes fundamental role in uncovering human genome, yet in addition in finding powerful counteraction and treatment of tumors. This paper proposes a novel technique that can foresee the disease by mutations. We will compare the patient's protein and the gene's protein of disease and in the event that there is distinction between these two proteins, at that point we can say there is malignant transformations. We found that LCS algorithm is a simple and efficient algorithm which does sequence alignment on a pair of sequences. Furthermore, we did a detailed study on machine learning approaches and determine the best approach for training and testing the dataset. We chose Support Vector Machines (SVM) since it gave the best results of about 98% accuracy. Finally, we created a user-friendly website that allows users to give an input sequence and results an output whether the given sequence is malignant orbenign.

Keywords: *Sequence alignment, Breast Cancer, Support Vector Machine*

I. INTRODUCTION

Breastcancer is a disorder in which malignant (cancer)cells develop in breast tissues. India continues to have a poor breast cancer survival rate, with just 66.1 per cent of women diagnosed with the disease surviving between 2010 and 2014, a report by Lancet found. "The key explanation for low breast cancer survival levels in India is because there is lack of awareness about cancer and its care. In third and four stages the cases come to us where treatment is complicated. In Indian women theusualtesting of breast cancer is very small. While cancer screening is a regular practice in healthcare in Western countries”. Breast cancer is ranked number one cancer among Indian females with a prevalence as high as 25.8 per 100,000 females and a mortality rate of 12.7 per 100,000 females, according to the Union health ministry. At least 17,97,900 women in India could have breast cancer by 2020, according toestimates.

Females with some risk factors are more likely to develop breast cancer than others. A risk factor is something that may increase the likelihood of getting a condition. You can avoid certain risk factors (such as drinking alcohol). But most factors can't be avoided (such as having a family history of breast cancer). Having a risk factor doesn't mean a woman's getting breast cancer.

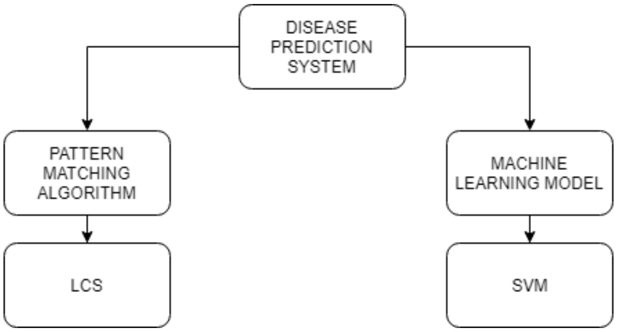


FIG 1: Brief outlook

A medical diagnosis is a pattern recognition/classification problem, where the doctor has to come up with an output

Received: 26 Apr 2020 | Revised: 24 May 2020 | Accepted: 28 Jun 2020

(disease) based on an input (symptoms). There are many classification, machine learning and pattern recognition methods that can be applied to develop tools that solve this pattern recognition problem. By and large, developing classification models using such methods is a two-step process. Firstly, a well-classified training set of data is used to train the model and derive parameters that optimize the prediction accuracy of that training set. After training, the model and its parameters are used on another well classified validation set of data to test predictionaccuracyondatathatwerenotusedfortrainingThis validation ensures that the classifier did not memorize the data from the training set. Instead, it learnt the characteristics from that data set that enable it to correctly characterize new data entries. In general, the larger the training and evaluation sets are, the better the future predictions of the method will be.

Biological sequence may be represented as symbolic sequence. When biologists find a new sequence, they want to know which other sequences it is closely related to. The sequence comparison was successfully used to create the connection between cancer-causing genes and a gene that developed in normal development and growth.

The paper (1) proposes a new model that seeks to eliminate the local alignment of targeted DNA sequences from being executed. Using a linear multi-pattern runtime exact matching string algorithm, a collection of query sequence random patterns (subsequences) is scanned to all targeted sequences in the database. Targeted DNA sequences with a significant low exact matching score are removed from execution for dynamic alignment basedprogramming.

The algorithm proposed by (3) uses the reduced amino acid alphabet to convert protein sequences into an integer sequence and uses n-gram to reduce the duration of the sequence. Then the Smith-Waterman algorithm is used to measure the similarity between two sequences.

This article (2) presents three pattern matching algorithms, namely FLPM, PAPM and LFPM that are uniquely designed to accelerate searches for large DNA sequences. Proposed algorithms improve performance by using word processing and also by searching for the least frequent word of the pattern in the sequence.

In this paper, we intend to do a detailed study of the available pattern matching algorithms and determine which is the best suited for an efficient breast cancer prediction system.(Refer Fig:1) We found that LCS algorithm is a simple and efficient algorithm which does sequence alignment on a pair of sequences. We chose Support Vector Machines (SVM) for training and testing since it gave the best results of about 98% accuracy. Finally, we created a user-friendly website that allows users to give an input sequence and results an output whether the given sequence is malignant orbenign.

The paper is described as follows: Section II describes the literature survey. Proposed method is explained in Section III. Result is given in Section IV. Section V concludes the paper and future scope is given in Section VI.

II. LITERATURESURVEY

According to Harshitha [4], supervised learning methods are used to obtain the attributes defining cancer and categorize cancer images from standardmammogram images. The supervised system is initially trained by retrieving 13 features from a database of 30 imag each. The derived image features under test are linked to the extracted features from the database images to detect and anticipate cancer tumors in the image.

The random forest calculation proposed by Bin Dai [5] is utilized to examine the clinical case finding of bosom cancer. The random forest calculation can join the attributes of different eigenvalues, and the consolidated consequences of various choice trees can be utilized to improve the forecast exactness. In view of the outfit learning strategy for irregular trees, the consequences of different feeble classifiers can be joined to deliver exact order results. In this paper, a random forest algorithm is utilized to examine the instance of bosom malignant growth case determination and acquire high expectation exactness. It has reasonable essentialness for assistant clinicalfinding.

Creator Panuwat Mekha [6] shows examination of grouping calculations for breast cancer based on tumor cell. It focuses on utilizing profound learning algorithms to arrange kinds of breast disease with a few of initiation function: Tanh, Rectifier, Maxout and Exprectifier and examination with various AI methods, for example, Naïve Bayes (NB), Decision tree (DT), Support Vector Machine (SVM), Vote (DT+NB+SVM), Random Forest (RF) and AdaBoost. Exploratory data were downloaded from breast cancer Wisconsin dataset and utilizing AI instrument rapidminer. Utilizing ten times cross-approval. We found that the high precision of 96.99% with profound learning by Exprectifier actuation work.

The paper proposed by C L Nithya [7] is executed for Find-s and Candidate elimination algorithm. Examination of both the calculations has been done and their analysis with respect to accuracy has been found. This paper closes taking training models and executed find-s and candidate elimination technique and approves the outcome. After grouping the prediction will occur whether the bosom disease is there ornot.

Ensemble learning models is developed using a combination of different machine learning models. The machine learning models [8][9]mainly used are Support Vector Machine, Logistic Regression, Decision Tree, K- Nearest Neighbor, Naïve Bayes etc. It produces a minimum average accuracy of 98%. These show that combining different models[10] can give a better result than relying on one single model.

For example, Author Naveen[11] Decision tree and KNN gives 100% precision. Decision tree model gives 100% exactness in the event that we split train-test dataset in proportion of 90:10 and furthermore utilized 300 sacks of trees. KNN gives max. accuracy 100%, for k= 1 to 7 out of sevenloopswith90%istrainingdataand10%istesting data. Here k is the closest neighbors. Additionally he assessed its forecast by accuracy, confusion grid and classification report. The point is to manufacture a generally precise and proficient AI model. So as prediction result, patient can take treatment on the beginning stage.

The paper proposed by Parag Singhal[12] is to develop a tool for early prediction of bosom cancer with the highest precision possible and low error rate. This was done by applying AI algorithms and with assistance of Artificial Neural Network (ANN) utilizing Wisconsin Breast Cancer (Diagnostic) Dataset. Test results show that ANN gives accuracy upto 98% with low blunder rate. The Experiment is directed using Dev.- C++ programming and actualized utilizing C-language.

III. PROPOSEDMETHOD

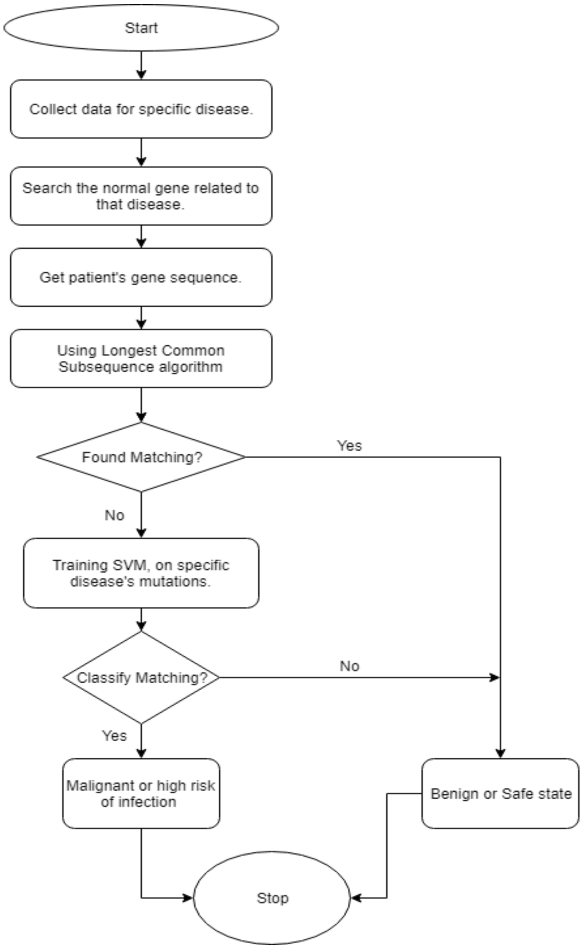

Start
Collect data for specific disease.
Search the normal gene related to that disease.
Get patient's gene sequence.
Using Longest Common Subsequence algorithm
Found Matching?
Yes
No
Training SVM, on specific disease's mutations.
Classify Matching?
No
Yes
Malignant or high risk of infection
Benign or Safe state
Stop

FIG 1 : Flowchart of main tasks of proposed method. Mainly two approaches are described: first whether the patient has mutations causes the disease or not, the second discover these mutations are related tocertain disease.

a. BIOINFORMATICS TECHNIQUES

i. **FASTA:** Compares a query string to a single text string when trying to search the entire database for matching the query, which is done by using the FASTA algorithm to every string in the database. If you are looking for an alignment, you might expect to find a few segments in which there will be absolute identity between the two matching strings, FASTA used this property to focus on the same regions.

ii. **LCS Algorithm:** Estimating closeness between arrangements, be it DNA, RNA, or protein groupings, is at the center of different issues in atomic science. A significant way to deal with this issue is processing the longest Common Subsequence (LCS) between two strings S1 and S2, for example the longest arranged list of symbols common between S1 and S2. For instance, when S1=abba and S2=abab, we have the accompanying LCSs: abb and aba. The LCS has been utilized to consider different areas, for example, content investigation, pattern acknowledgment, document examination, effective tree matching and so on. Organic utilizations of the LCS and comparability estimation are varied, from sequence alignment in relative genomics, to phylogenetic development and examination, to fast pursuit in colossal natural groupings, to pressure and productive stockpiling of the rapidly growing genomic informational sets to re-sequencing a lot of strings given an objective string an important step in efficient genome assembly.

Following is detailed algorithm to print the LCS. It uses the 2D table L[][].

ALGORITHM-LCS

```
1. Construct L[m+1][n+1].
2. The value L[m][n] contains length of LCS. Create a character array lcs[] of length equal to the length
   of lcs plus 1 (one extra to store\0).
3. Traverse the 2D array starting from L[m][n]. Do following for every cell L[i][j]
   a) If characters (in X and Y) corresponding to L[i][j] are same (Or X[i-1] ==
      Y[j-1]), then include this character as part of LCS.
   b) Else compare values of L[i-1][j] and L[i][j-1] and go in direction of greater value.
```

b. MACHINE LEARNINGMODEL

Here, SVM is used to train the dataset. Library called libsvm is used for this purpose.

SUPPORT VECTOR MACHINES (SVM)

Support Vector Machine (SVM) is a supervised learning algorithm that can be used for both regression and classification problems. However, it is mainly used for classification problems. In this algorithm, each data object is plotted as a point in n-dimensional space (where n is the number of features you have) with the value of each function being the value of a particular coordinate. Then, we conduct classification by finding a hyper-plane that distinguish two classes. Support vectors are literally the positions of individual observation. Support Vector Machine is the boundary that better distinguishes the two hyper-plane / line classes). Initially, the SVMs map the input vector to a higher dimensional space function and define the hyperplane which separates the data points into two different classes. The marginal gap between the decision on the hyperplane and the instances nearest to the boundary is maximized. The resulting classifier achieves tremendous generalizability and can therefore be used for the accurate classification of newsamples.

There are different libraries for SVM available. Here, we use libsvm.

i. LIBSVM

LIBSVM[13] is an integrated software for support vector classification, (C-SVC, nu-SVC), regression (epsilon-SVR, nu-SVR) and distribution estimation (one-class SVM). It supports multi-class classification.

LIBSVM provides a simple interface where users can easily link it with their own programs. Main features of LIBSVM include:

- ✓ Different SVMformulations
- ✓ Efficient multi-classclassification
- ✓ Cross validation for modelselection
- ✓ Probability estimates
- ✓ Various kernels (including precomputed kernel matrix)
- ✓ Weighted SVM for unbalanced data
- ✓ Both C++ and Javasources
- ✓ GUI demonstrating SVM classification and regression

LIBSVM supports the following learningtasks.
(1) SVC: support vector classification (twoclass and multiclass);
(2) SVR: support vectorregression.
(3) One-classSVM.

A typical use of LIBSVM involves two steps: first, training a data set to obtain a model and second, using the model to predict information of a testing data set. For SVC and SVR, LIBSVM can also output probability estimates.
Given training vectors xi ∈ Rn,i = 1,...,l, in two classes, and an indicator vector y ∈ Rl such that yi ∈ {1, −1}, C-SVC solves the following primal optimization problem:

$$\begin{array}{ll} \displaystyle\min_{\boldsymbol{w},b,\xi} & \displaystyle\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{l}\xi_i \\ \text{subject to} & y_i(\boldsymbol{w}^T\phi(\boldsymbol{x}_i)+b) \geq 1-\xi_i, \\ & \xi_i \geq 0, i = 1,\ldots,l, \end{array}$$

where φ(xi) maps xi into a higher-dimensional space and C > 0 is the regularization parameter.
The ν-support vector classification introduces a new parameter ν ∈ (0, 1]. It is proved that ν an upper bound on the fraction of training errors and a lower bound of the fraction of support vectors.
Given training vectors xi ∈ Rn,i = 1,...,l, in two classes, and a vector y ∈ Rl such that yi ∈ {1, −1}, the primal optimization problem is:

$$\begin{array}{ll} \displaystyle\min_{\boldsymbol{w},b,\xi,\rho} & \displaystyle\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} - \nu\rho + \frac{1}{l}\sum_{i=1}^{l}\xi_i \\ \text{subject to} & y_i(\boldsymbol{w}^T\phi(\boldsymbol{x}_i)+b) \geq \rho-\xi_i, \\ & \xi_i \geq 0, i = 1,\ldots,l, \quad \rho \geq 0. \end{array}$$

In figure (2), the main roles are:

1. USER
2. TESTPHASE
3. TRAININGPHASE
4. OUTPUT

USER: User is a person who uses the developed system for his/her purpose. The system is created so that users are provided with all the support that they require. Here, this prediction system can be used by doctors, cancer patients, surgeons and can be distributed to various hospitals also.

TRAINING PHASE: In this phase, we create an appropriate model and train the data. The database used here is NCBI. Here, different types of cancer sequences are given to the system. The system is then trained with the data in such a way that the system can correctly identify whether a sequence given as input by the user is a cancer sequence or not. Furthermore, the system can be trained to determine whether the cancer is benign or malignant. The input sequence contains combination of ATGC (Adenine, Thymine, Guanine, and Cytosine).

TESTING PHASE: In this phase, the developed system is tested with the various input test sequences given by the user. This is the phase where we determine whether the developed model is giving desired output or not. This phase also helps to determine the efficiency and accuracy of the system. The input sequence contains combination of ATGC (Adenine,Thymine, Guanine,Cytosine).

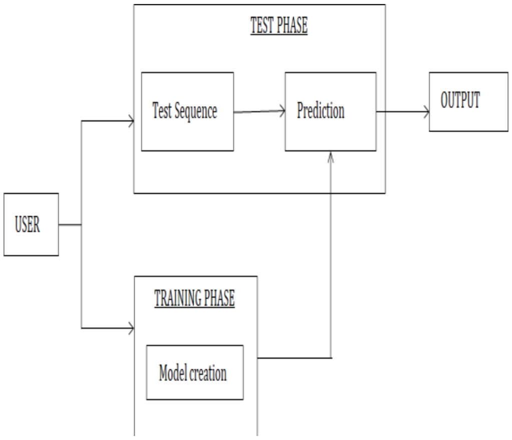


FIG 2: Level 1 diagram of model

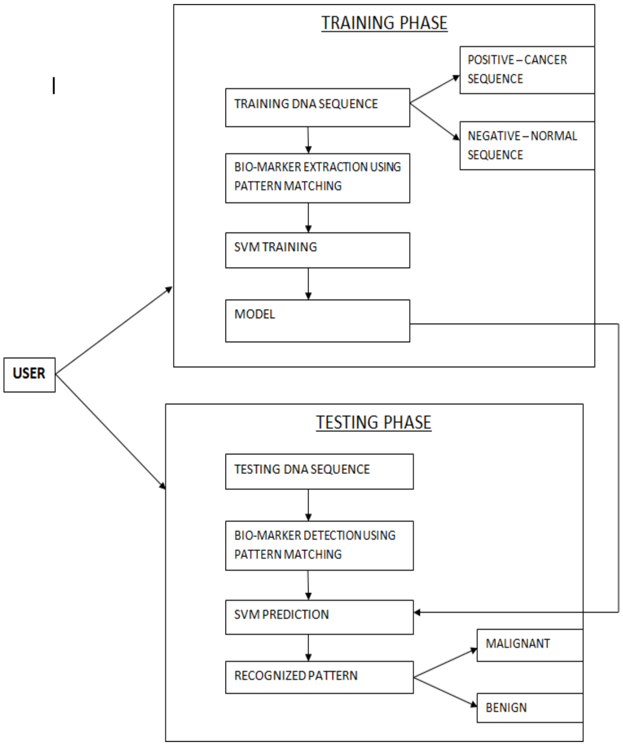

TRAINING PHASE
POSITIVE – CANCER SEQUENCE
TRAINING DNA SEQUENCE
NEGATIVE – NORMAL SEQUENCE
BIO-MARKER EXTRACTION USING PATTERN MATCHING
SVM TRAINING
MODEL
USER
TESTING PHASE
TESTING DNA SEQUENCE
BIO-MARKER DETECTION USING PATTERN MATCHING
SVM PREDICTION
MALIGNANT
RECOGNIZED PATTERN
BENIGN

FIG 3: Level 2

Figure (3) illustrates the training and testing phase of the proposed system.

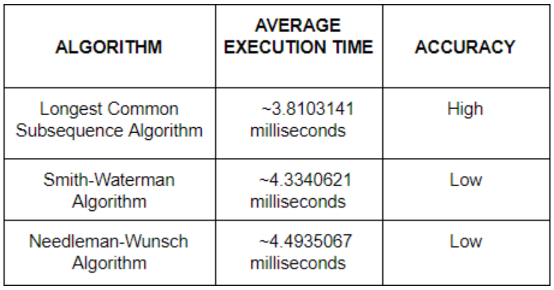
The classification model which gives one of the two outputs:

- POSITIVE (1) : This means that the given input sequence is a cancersequence.
- NEGATIVE (0) : This means that the giveninput is a normal or non-cancersequence.

Here, the model created in training phase is feeded to the SVM prediction in testing phase. We detect the similarities in two sequences and produce an output which indicates the percentage of similarity. Depending upon the value obtained, we can say whether the cancer is malignant orbenign.

IV. EVALUATION AND EXPERIMENTALOUTCOMES

i. RESULTS AND COMPARISON OF IMPLEMENTING LCSALGORITHM



ALGORITHM	AVERAGE EXECUTION TIME	ACCURACY
Longest Common Subsequence Algorithm	~3.8103141 milliseconds	High
Smith-Waterman Algorithm	~4.3340621 milliseconds	Low
Needleman-Wunsch Algorithm	~4.4935067 milliseconds	Low

Fig 4: Comparison of LCS Algorithm with Smith- Waterman andNeedleman-Wunsch

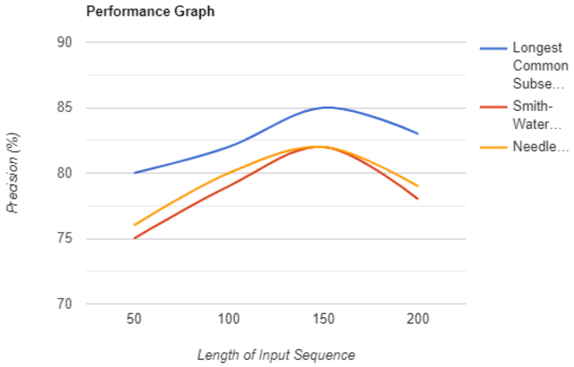
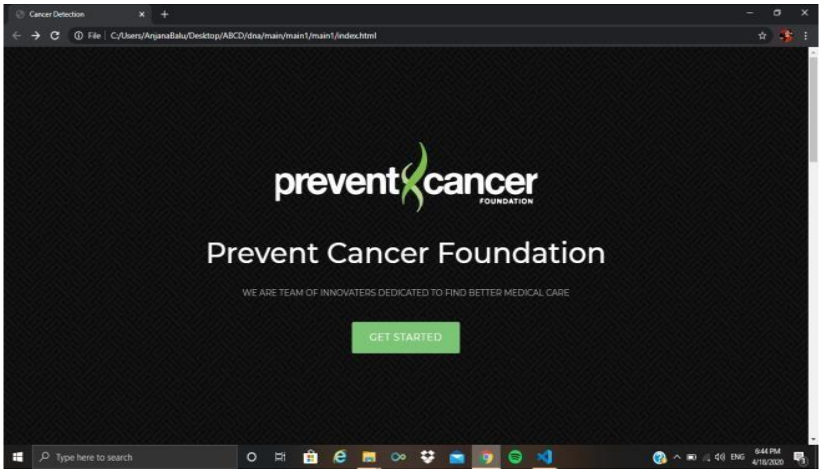


Fig 5: Performance Graph of LCS against FASTA, Smith- Waterman andNeedleman-Wunsch

From the above results, it is clear that the proposed LCS algorithm works far better than algorithms like Smith-Waterman and Needleman- Wunsch in terms of average execution time and accuracy. The performance graph shows that the performances of Smith Waterman and Needleman- Wunsch tends to increase at first when the number of sequences given is very low. As the sequences increase, their performance decreases. But the performance of LCS tends to be constant irrespective of the sequences given and higher than that of other two. FASTA is even better than that of LCS in terms ofperformance.

Received: 26 Apr 2020 | Revised: 24 May 2020 | Accepted: 28 Jun 2020

ii. FRONT-END USERINTERFACE

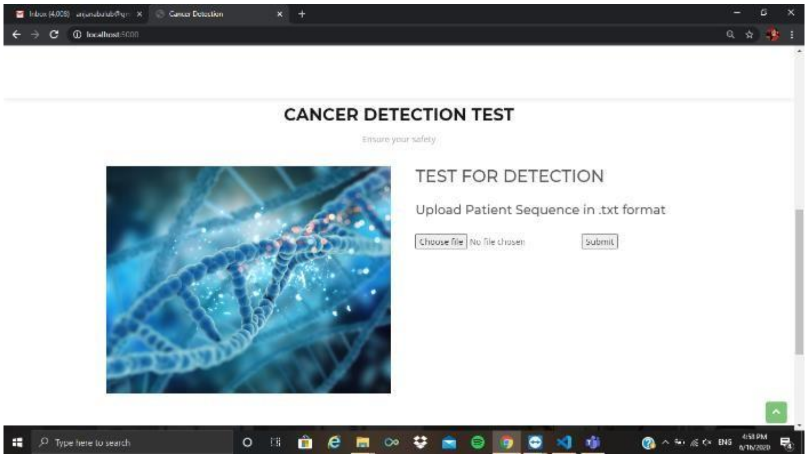


Cancer Detection
C:/Users/AnjanaBalu/Desktop/ABCD/dna/main/main1/main1/index.html
prevent cancer FOUNDATION
Type here to search

Prevent Cancer Foundation

WE ARE TEAM OF INNOVATERS DEDICATED TO FIND BETTER MEDICAL CARE

FIG 6.a: The Home Page



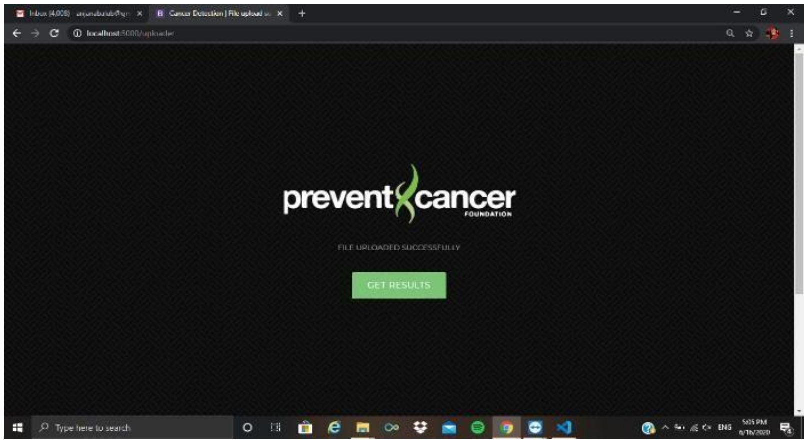
CANCER DETECTION TEST

Ensure your safety

TEST FOR DETECTION

Upload Patient Sequence in .txt format

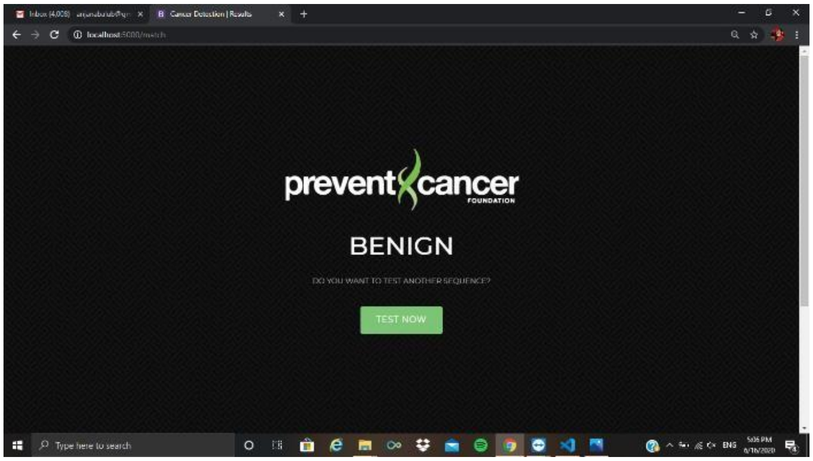
FIG 6.b: The Test Page



FILE UPLOADED SUCCESSFULLY

FIG 6.c: Status of uploading of text file.

Received: 26 Apr 2020 | Revised: 24 May 2020 | Accepted: 28 Jun 2020



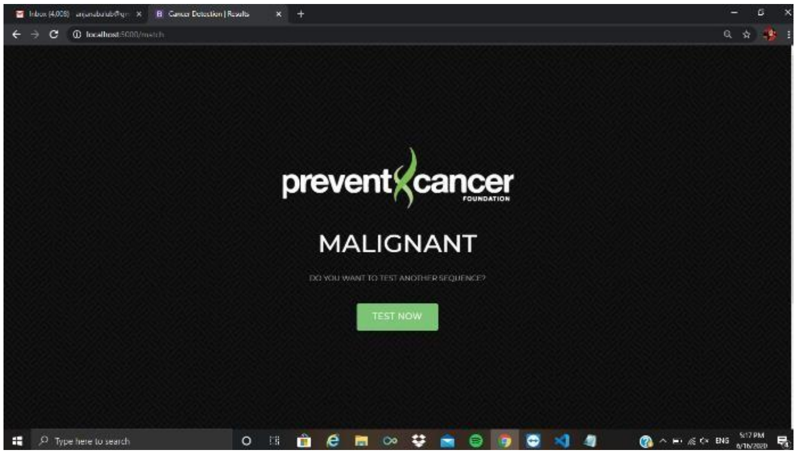
prevent cancer FOUNDATION

BENIGN

DO YOU WANT TO TEST ANOTHER SEQUENCE?

TEST NOW

FIG 6.d: Test showing result as benign



prevent cancer FOUNDATION

MALIGNANT

DO YOU WANT TO TEST ANOTHER SEQUENCE?

TEST NOW

FIG 6.e: Test showing result as malignant.

We implemented the algorithm successfully and the result is obtained as malignant or benign. LCS algorithm is far more efficient and simple compared to other algorithms. The aim of this paper is to predict breast cancer with high accuracy using sequence alignment and machine learning. The input data can be a single query or a large file. The output is delivered in very low response time. The execution time for the entire algorithm is also found to be very low. The front-end UI (Fig 6.a,Fig 6.b) is also implemented.

V. CONCLUSION

Because of the current lifestyle of the people, it is observed that most of them are very prone to life threatening diseases like cancer. The main reason for the high mortality rate for cancer is finding out the cancer in the late stages. Using this algorithm cancers can be detected at an earlier stage itself. The algorithm uses sequence analysis of the user's data and the trained data. The algorithm makes use of a Support Vector Machine (SVM) classifier to analyze the data. The algorithm is then integrated into a web based application which the user can accesseasily.

Received: 26 Apr 2020 | Revised: 24 May 2020 | Accepted: 28 Jun 2020

VI. FUTURESCOPE

In the future, this can be rolled out into mobiles in the form of application. Not only cancer, diseases like hard diseases, arthritis, and other chronic diseases should also be taken care for earlier detection. In future, immunotherapy is the big expectation to detect against cancer and other diseases. The most promising research is using the bodies on immune system to fight against cancer by genetictweaking.

This will allow the special forces- T- cells in the body to start recognizing cancer cells as enemy and targeting them.

VII. REFERENCES

[1]A.R.M. Nordin,M.T.A Osman, M. S. M. Yazid and A. Aziz, “A Guided Dynamic Programming Approach for Searching a Set of Similar DNA Sequences”, Second International Conference on the Applications of Digital Information and Web Technologies, INSPEC Accession Number:10905953,DOI:10.1109/ICADIWT.2009.52739 67

[2] Peyman Neamatollahi , Montassir Hadi , and Mahmoud Naghibzadeh “Simple and Efficient Pattern Matching Algorithms for Biological Sequences”, IEEE Access 2020, DOI:10.1109/ACCESS.2020.2969038

[3] Nur'Aini Abdul Rashid', Rosni Abdullahl, Abdullah Zawawi Haji Talibl, Zalila AH2, “Fast Dynamic Programming Based Sequence Alignment Algorithm”, INSPEC Accession Number: 9253532 DOI: 10.1109/DFMA.2006.296909

[4] Harshitha ; V Chaitanya ; Shazia M Killedar ; Dheeraj Revankar ; Mala S Pushpa "Recognition and Prediction of Breast Cancer using Supervised Diagnosis",2019, INSPEC Accession Number: 19467910,DOI:10.1109/RTEICT46194.2019.9016921

[5] Bin Dai ; Rung-Ching Chen ; Shun-Zhi Zhu ; Wei-Wei Zhang "Using Random Forest Algorithm for Breast Cancer Diagnosis" 2018 International Symposium on Computer, Consumer and Control (IS3C) DOI:10.1109/IS3C.2018.00119

[6] Panuwat Mekha; Nutnicha Teeyasuksaet "Deep Learning Algorithms for Predicting Breast Cancer Based on Tumor Cells", 2019, INSPEC Accession Number: 18603078 DOI:10.1109/ECTI-NCON.2019.8692297

[7] C L Nithya ; Sunanda Dixit ; B.I. Khodhanpur "Prediction of breast cancer using Find-S and Candidate elimination algorithm", DOI:10.1109/CSITSS47250.2019.9031046 2019, 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS)

[8] Pragya Chauhan; Amit Swami "Breast Cancer Prediction Using Genetic Algorithm Based Ensemble Approach"DOI:10.1109/ICCCNT.2018.8493927

[9] Madhuri Gupta ; Bharat Gupta "An Ensemble Model for Breast Cancer PredictionUsing Sequential Least Squares Programming Method(SLSQP)"

[10] Sunanda Das ; Dipayan Biswas "Prediction of Breast Cancer Using Ensemble Learning"

[11] Naveen ; R. K. Sharma ; Anil Ramachandran Nair "Efficient Breast Cancer Prediction Using Ensemble Machine LearningModels"

[12] Parag Singhal ; Saurav Pareek "Artificial Neural Network for Prediction of Breast Cancer"

[13] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

Received: 26 Apr 2020 | Revised: 24 May 2020 | Accepted: 28 Jun 2020