**CHAPTER 1 INTRODUCTION**

**1. Introduction**

This chapter gives an introduction to the project so that the idea about the overall project is understood well, it also contains the details about the problem statement and the aims and objectives of the given project.

**1.1 Overview**

A major challenge facing healthcare organizations (hospitals, medical centers) is predicting the diseases with greater accuracy and at an early stage.

Here a system is proposed which will predict cancer of patients at its earlier stage by using genomic expression instead of only clinical expressions which will help us to achieve better accuracy. Gene data gives a better advantage since it has the potential ability to indicate cancer at an earlier stage which can be used to train the model more efficiently thus producing overall result more accurately. Different supervised learning algorithms like a highly versatile support vector machine (SVM) algorithm, Naive Bayes theorem, Decision tree and nearest neighbors approach to predict cancer of the patient are being used here. Using these methods, classification of patients will be done to predict whether a patient is suffering from cancer or not.

Over a long period of time, innovation on effective cancer treatment is in progress. Scientists applied different approaches such as screening at an early stage, in order to predict cancer type before the symptoms started to develop. The approach which was used by them was multi omics data viz biological data analysis. With the advancement of new technologies in the field of medicine, vast quantities of cancer data have been collected and are available for medical research. These datasets of new technologies are based on genomic data. However, the accurate prediction of a disease at an early stage is one of the most interesting and challenging tasks for physicians.

**1.2 Dataset used**

Gene Expression profiling is used in the proposed system. It is nothing but genomic data. It is the measurement of activity of ‘n’ number of genes at a single point of time to create a thorough picture of cellular function.

A Laboratory tool called a microarray helps in detecting various gene expressions simultaneously. The microscopic slides that have hundreds of tiny spots printed in specific positions are said to be DNA microarrays. Each spot in microscopic slides is known as DNA sequence or gene. The DNA molecules on such slides acts as probes that helps in detecting gene expression. These molecules are also known as transcriptome or RNA transcripts.

In this microarray analysis procedure the RNA molecules of healthy individual and cancer patient are accumulated at one place. These samples are then converted into DNA samples of complementary version (cDNA). Each sample is labelled with different colors. The two accumulated samples are then combined on the microscopic slides. This process is called as Hybridization. After hybridization process scanning of microarray takes place by which expression of each sample or gene will be found. If the mutation of gene is greater than experimental sample then the spot will turn red otherwise green. If the mutation is equal then it turns yellow. In this way gene expression profile is generated.

**1.3 Methods**

A lot of research is being done on breast cancer. Researchers have developed breast cancer risk models which give probability of cancer occurrence. They make use of Clinical Data. There are few models which provide such risk probability. International breast cancer intervention study model (IBIS), Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm model (BOADICEA), the BRCAPRO model and the Breast Cancer Risk Assessment Tool (BCRAT) also known as the Gail model.

IBIS and BOADICEA models are trained with around 19,000 samples and accuracy obtained by these models were 71% and 70% respectively. Whereas BRCAPRO and BCRAT models underestimated the risk and had an accuracy of about 68% and 60%.

Different methods are used to build such predictive models. Machine learning provides algorithms which can help in building such models. Machine Learning comprises different types of learning like supervised learning, unsupervised learning, semi-supervised learning etc. Supervised learning is used when the datasets consist of labelled output. Unsupervised learning is used when datasets does not have output label with it and semi-supervised learning is used when datasets consists of both labelled and unlabeled values. Here the datasets used to train models have labelled values, so a supervised learning method is used here. Different supervised learning methods are available and such methods are used for prediction. The prediction models built in this study is using Support Vector Machine (SVM), Naïve Bayes, Decision Tree and K-Nearest neighbors (KNN). All these are supervised learning algorithms.

# CHAPTER 2 LITERATURE REVIEW

**2. Literature Review**

A literature review is a text of a scholarly paper, which includes the current knowledge including substantive findings, as well as theoretical and methodological contributions to a particular topic. Literature reviews are secondary sources, and do not report new or original experimental work.

**2.1 Review of Literature**

1. “**Predicting Cancer Prognosis Using Functional Genomics Data Sets**”

Jishnu Das et al. have compared various computational methods that have used different functional genomics datasets.  They identify the molecular patterns that can be used for predicting prognosis of various human cancer tumors. Furthermore, they have outlined the challenges and how such approaches can be useful in solving those [1].

2.  “**Machine learning predicts individual cancer patient responses to therapeutic drugs with high accuracy**”

Cai Huang et al. have designed a software platform which predicts cancer from gene expression profiles. They used SVM based algorithm and for regularization they used Recursive Feature Elimination. Their main finding was that the model works best when it uses all probe-set expression profiles of individual patient tumors. They have achieved more than 75% accuracy [2].

3. “**Machine learning applications in cancer prognosis and prediction**”

Konstantina Kouroua Themis et al. have evaluated all the prominent available ML models. This includes ANNs, BNs, SVMs and DTs. This paper aims to validate the best approaches available so that they can be considered in everyday clinical practice [3].

4. “**Predicting stage-specific cancer related genes and their dynamic modules by integrating multiple datasets**”

Chaima Aouiche et al. have proposed a structure to identify stage specific cancer related genes by integrating multiple datasets. Also they have built a network by taking each sample pathway as vertices and relationships between genes as edges [4].

5. “**Deep Learning Methods for Predicting Disease Status Using Genomic Data**”

Qianfan Wu et al. have studied four articles that predicted cancer using genomic expression. These deep learning methods outperformed existing models such as prediction based on transcript-wise screening and prediction based on principal component analysis [5].

6. “**Dermatologist-level classification of skin cancer with deep neural networks**”

Esteva A et al. used Convolutional Neural Networks to classify skin cancer. They just used skin lesion images and disease labels to train the mode. The model showed great potential [6].

7.  “**ImageNet large scale visual recognition challenge**”

Russakovsky O et al. analyzed the past 5 years of Image classification competition and drew useful patterns and predicted the future development of image classification and its usefulness in disease prediction [7].

8. “**A practical guide to support vector classification**”

Hsu C-W et al. have explained in detail Support Vector Classification and its potential in disease prediction [8].

9. “ **An Overview of Prognostics Markers in Breast Cancer** “

Gu Deshpande et al. all the currently used biomarkers for cancer prediction and concluded that these aren’t enough. He then studied some more biomarkers which can increase the reliability of model if integrated with the existing biomarkers [9].

10. “**A review of feature selection techniques in bioinformatics**”

Sayes Y et al. have performed feature selection techniques by providing basic taxonomy of feature selection, discussing their use, and providing a variety of applications in both common as well as bioinformatics [10].

11. “**Minimum redundancy maximum relevance feature selection approach for temporal gene expression data**”

Radovic M et al. have proposed a temporal minimum redundancy-maximum relevance feature selection approach. The proposed system was able to handle multivariate temporal data without previous data flattening. Redundancy between the gene was computed using a dynamical time wrapping approach [11].

12. “**Highly-accurate metabolomic detection of early-stage ovarian cancer**”

Gaul DA et al. have proposed a system using linear support vector machine. The results which were achieved provided evidence for the importance of lipid and fatty acid metabolism in OC and this can be used for clinical significant diagnostic tests. [12]

13. “**Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines**”

Guan W et al. have developed a system for ovarian cancer in which they developed new approaches for automatic classification of metabolic data. They have used SVM and cross fold validation technique which provided them highly accurate results [13].

14. “**Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell**”.

Hoadley K. et al. have performed an integrative analysis using five genome wide platforms. In this paper methods such as Classification along with Correlation was used inorder to obtain better results [14].

15. “**Computational models for predicting drug responses in cancer research** “

Azuaje F. et al. have developed a model in which matching of tumor characteristics to the most effective therapy available and thus providing the patient with suitable precise medicine [15].

16. “**From molecular mechanisms of leukemia induction to treatment of chronic myelogenous leukemia**”

Salesse S. et al have Performed molecular mechanisms of leukemia induction to treatment of chronic myelogenous leukemia. In this paper they proposed a system with better accuracy [16].

17. “**Database resource of the national genomics data center**”

Wenming Zhao. et al have provided a suite of genomic database resources. With the help of NGDC databases of genomic data a large number of requirements of data was made available publicly for study and research purposes[17].

# CHAPTER 3 METHODOLOGIES AND IMPLEMENTATION

### 3. Methodologies and Implementation

### In this chapter all the methodologies which are used to build this project is presented here. Along with methodologies corresponding implementation of project is given here

### 3.1 Design Details

The aim of proposed methodology is accurate prediction of cancer using genomic data. Cancer is a complex disease and complete causes behind cancer development are not yet fully discovered. Also, cancer treatment causes lots of expenses during treatment, and it increases as tumor grows. So by predicting cancer at an earlier stage, heavy expenses of medication can also be reduced.

The methodology of the proposed model is divided into four phases as shown in fig 1.

### 

**Figure 3.1.1 - Phases of prediction model**

The phases are described below.

i) High Dimensional Input features -

Here microarray gene expression is extracted from online open source repositories [17-18]. The National Center for Biotechnology Information (NCBI) provides access to biomedical and genomic information. The datasets consist of 17,818 genes and 590 samples (including 61 normal tissue samples and 529 breast cancer tissue samples).

ii) Feature Selection/Dimensionality Reduction -

Since there are many genes, the model trained using all such genes may cause overfitting. Also, there are various genes which are not affecting the DNA mutation. To address this issue, major breast cancer causing genes are selected. There are 22 such major cancer causing genes namely BRCA1, BRCA2, ATM, BARD1, BRIP1, CDH1, CHEK2, MRE11A, MSH6, NBN, PALB2, PMS2, PTEN, RAD50, RAD51c, STK11, TP53, CASP8, CTLA4, CYP19A1, FGFR2, LSP1, MAP3K1 [19].

iii) Low Dimensional features -

The dataset having 22 dimensions is preprocessed first. All the field values are numeric values. However, there were many fields where the values were not present, so these values were replaced with mean values.

iv) Prediction Models and Classifiers -

After preprocessing, the dataset is obtained having 530 samples having 22 features (genes). Support Vector Machine algorithm is performed first on weka tool. This tool has various inbuilt machine learning algorithms. It also preprocesses the data and trains the model and plots various graphs. Initially, the dataset was passed to weka tool for model building. Later, SVM was implemented using python 3 on google colab to build the model. Along with SVM, Naive Bayes algorithm based model is also built using python 3.

### 3.2 Algorithms:

### Support Vector machine (SVM)

### The structured support vector machine is a [machine learning](https://en.wikipedia.org/wiki/Machine_learning) algorithm that generalizes the [Support Vector Machine](https://en.wikipedia.org/wiki/Support_vector_machine) (SVM) classifier. Whereas the SVM classifier supports [binary classification](https://en.wikipedia.org/wiki/Binary_classification), [multiclass classification](https://en.wikipedia.org/wiki/Multiclass_classification) and [regression](https://en.wikipedia.org/wiki/Regression_analysis), the structured SVM allows training of a classifier for general [structured output labels](https://en.wikipedia.org/wiki/Structured_learning).

### ii. Decision Tree

A decision tree is a [decision support](https://en.wikipedia.org/wiki/Decision_support_system) tool that uses a [tree-like](https://en.wikipedia.org/wiki/Tree_(graph_theory)) [model](https://en.wikipedia.org/wiki/Causal_model) of decisions and their possible consequences, including [chance](https://en.wikipedia.org/wiki/Probability) event outcomes, resource costs, and [utility](https://en.wikipedia.org/wiki/Utility). It is one way to display an [algorithm](https://en.wikipedia.org/wiki/Algorithm) that only contains conditional control statements.

Decision trees are commonly used in [operations research](https://en.wikipedia.org/wiki/Operations_research), specifically in [decision analysis](https://en.wikipedia.org/wiki/Decision_analysis), to help identify a strategy most likely to reach a [goal](https://en.wikipedia.org/wiki/Goal), but are also a popular tool in [machine learning](https://en.wikipedia.org/wiki/Decision_tree_learning)

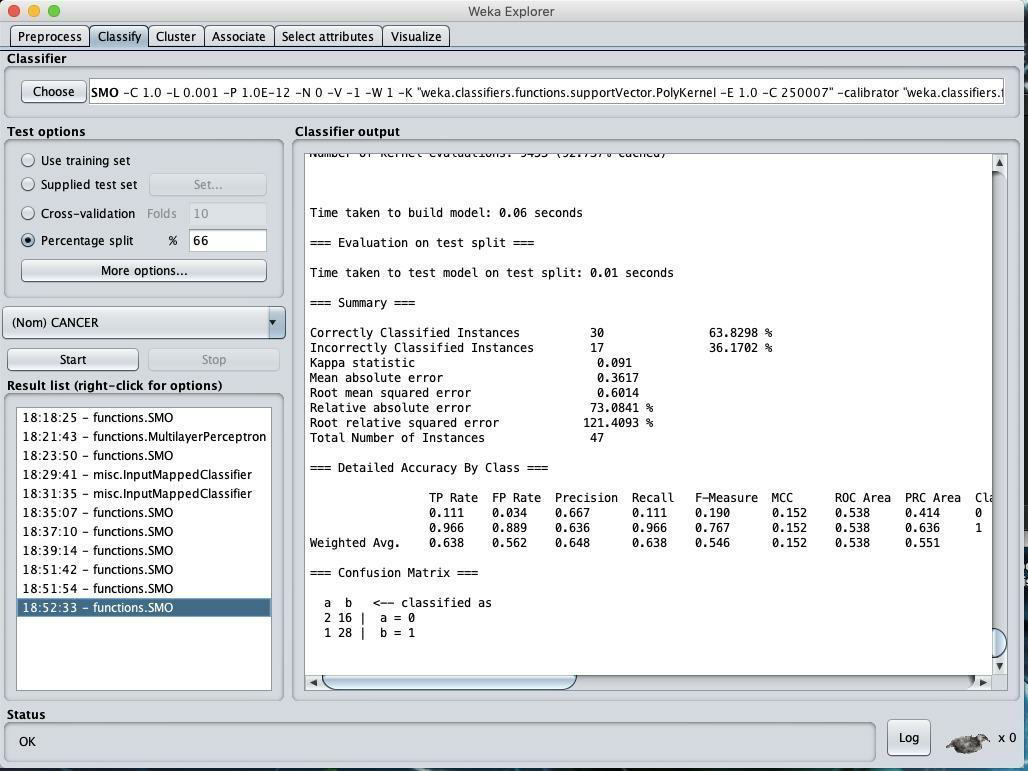
iii. Naïve Bayes

In [machine learning](https://en.wikipedia.org/wiki/Machine_learning), naïve Bayes classifiers are a family of simple "[probabilistic classifiers](https://en.wikipedia.org/wiki/Probabilistic_classification)" based on applying [Bayes' theorem](https://en.wikipedia.org/wiki/Bayes%27_theorem) with strong (naïve) [independence](https://en.wikipedia.org/wiki/Statistical_independence) assumptions between the features. They are among the simplest [Bayesian network](https://en.wikipedia.org/wiki/Bayesian_network) models.

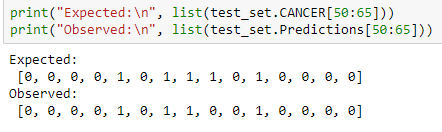
### iv. K – Nearest neighbors

### *k*-NN is a type of [instance-based learning](https://en.wikipedia.org/wiki/Instance-based_learning), or [lazy learning](https://en.wikipedia.org/wiki/Lazy_learning), where the function is only approximated locally and all computation is deferred until function evaluation.

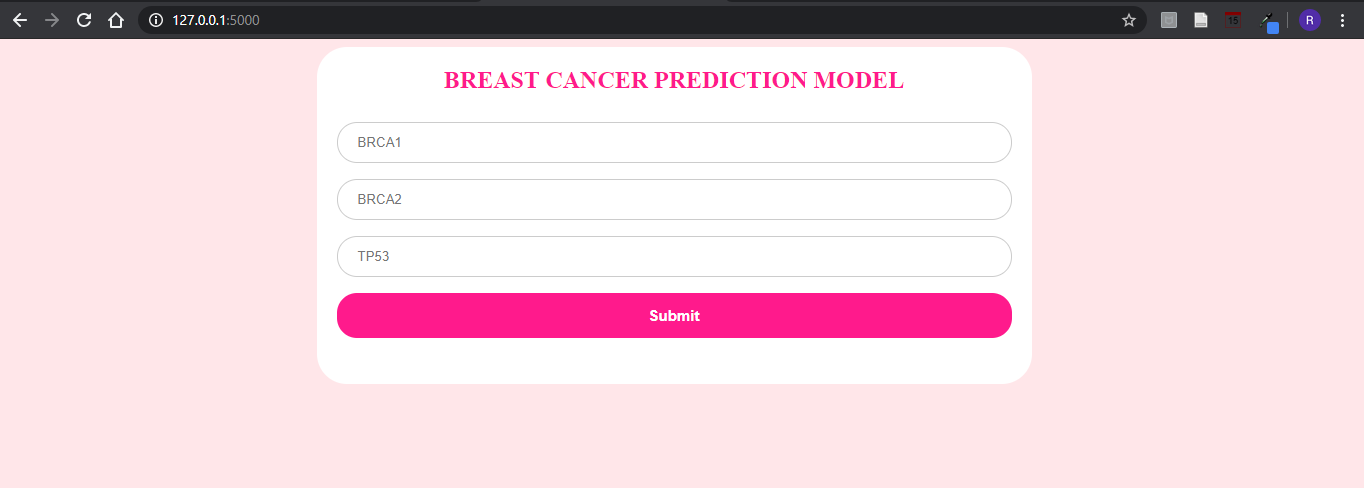
### 3.3 Implementation



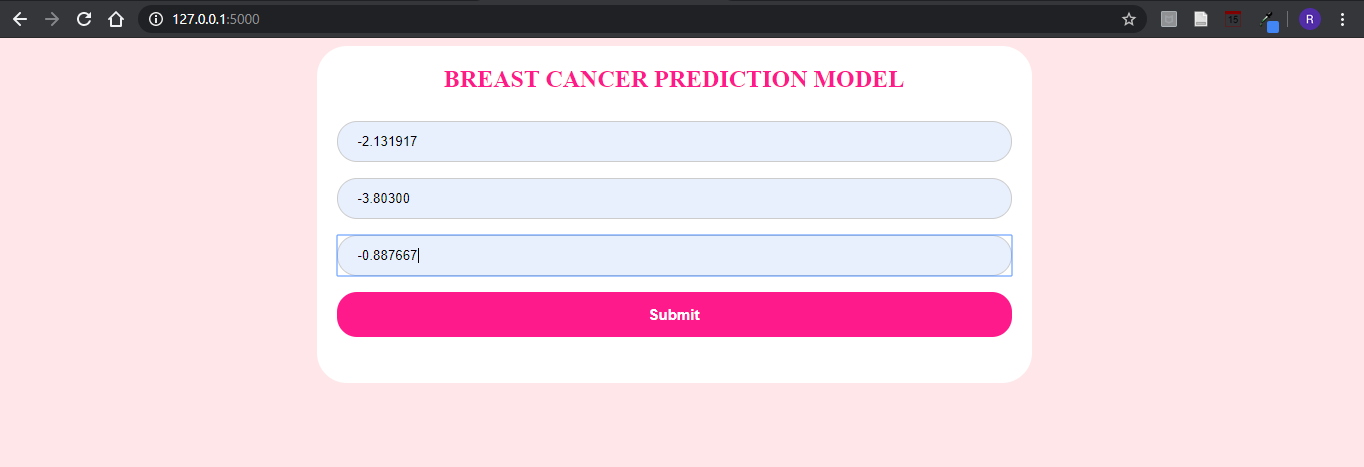
**Figure 3.3.1 – SVM model trained on weka**



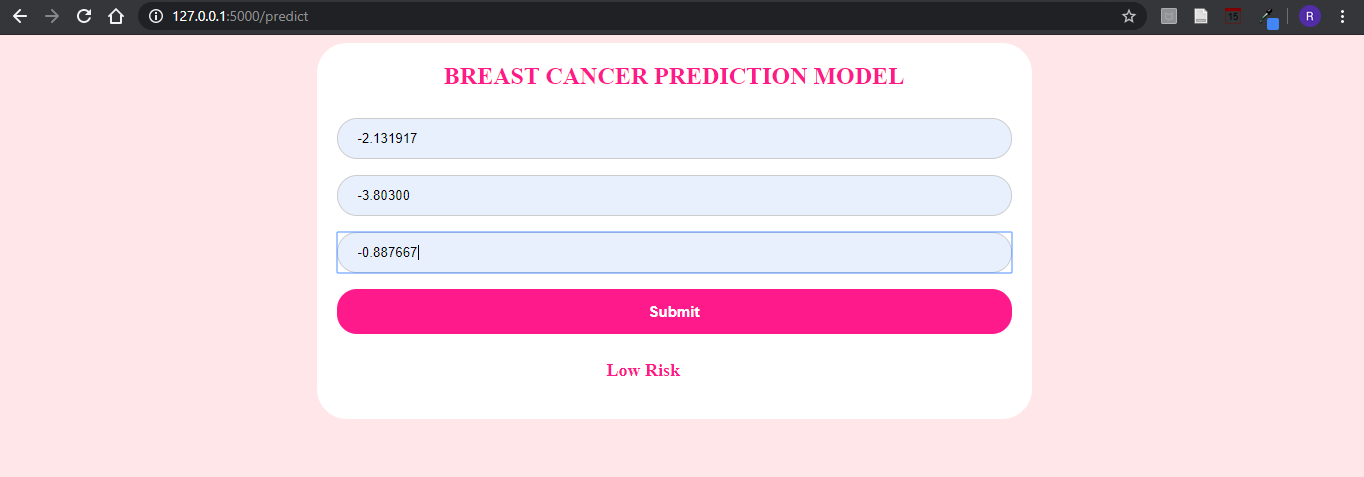
**Figure 3.3.2 - Expected and Observed results for Cancer in SVM Model**



**Figure 3.3.3– UI for model 2**



**Figure 3.3.4 – Input values for genes**



**Figure 3.3.5 – Result displayed based on model trained**





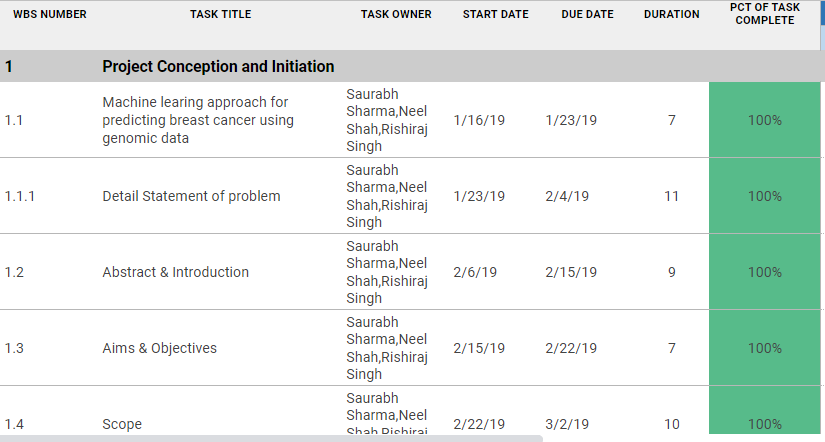
**Figure 3.3.6 – Dataset with same values as given to** **UI**

# CHAPTER 4 PROJECT ANALYSIS

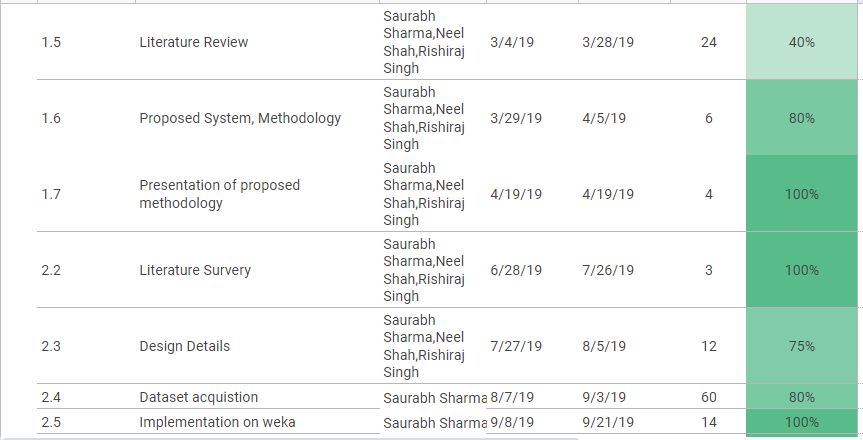
### 4. Project Analysis

### This chapter should give detail design of the project. It includes Block diagram of proposed system and UML diagrams (Use case diagram, Data flow diagram, Sequence diagram etc.) as applicable to the project.

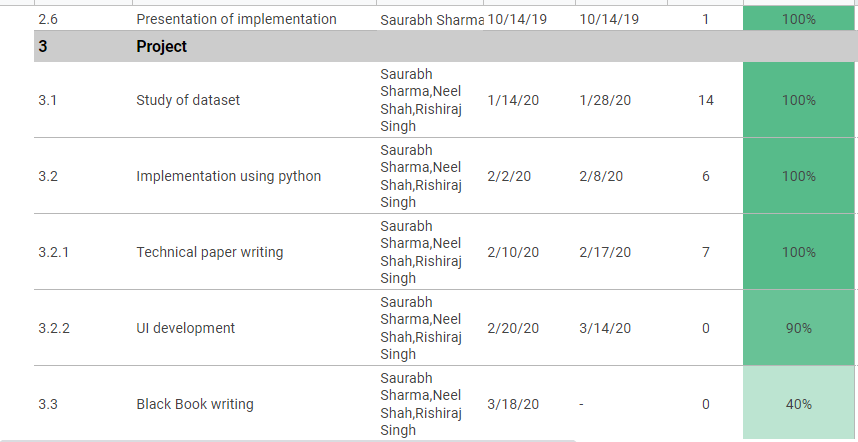
### 4.1 Project TimeLine



**Figure 4.1 Project Timeline 1**



**Figure 4.2 Project Timeline 2**



**Figure 4.3 Project Timeline 3**

### 4.2 Task Distribution

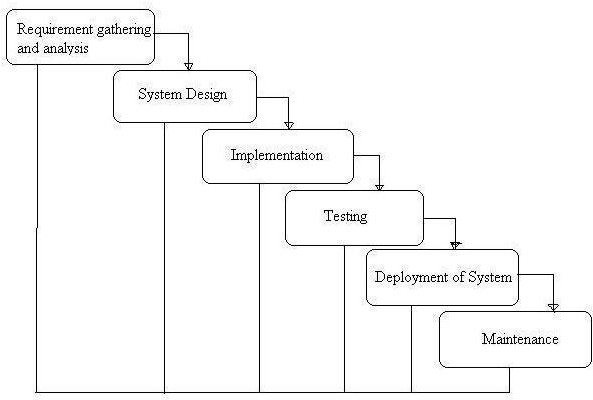
**Table 4.2 Task Distribution**

|  |  |  |
| --- | --- | --- |
| **TASK LIST** | **ASSIGNED TO** | **STATUS** |
| Defining Project | Saurabh Sharma | Complete |
| Neel Shah |
| Rishiraj Singh |
| Literature Review | Saurabh Sharma | Complete |
| Neel Shah |
| Rishiraj Singh |
| Survey Paper | Saurabh Sharma | Complete |
| Neel Shah |
| Rishiraj Singh |
| Project Plan | Saurabh Sharma | Complete |
| Neel Shah |
| Rishiraj Singh |
| Project Analysis | Saurabh Sharma | Complete |
| Neel Shah |
| Rishiraj Singh |
| Input Page Design | Saurabh Sharma | Complete |
| Neel Shah |
| Rishiraj Singh |
| Documentation of Synopsis | Saurabh Sharma | Complete |
| Neel Shah |
| Rishiraj Singh |
| Dataset formatting | Saurabh Sharma | Complete |
| Neel Shah |
| Rishiraj Singh |
| Implementation | Saurabh Sharma | Complete |
| Neel Shah |
| Rishiraj Singh |
| Testing | Saurabh Sharma | Complete |

|  |  |  |
| --- | --- | --- |
|  | Neel Shah |  |
| Rishiraj Singh |
| Final Report | Saurabh Sharma | Complete |
| Neel Shah |
| Rishiraj Singh |
| Final Presentation | Saurabh Sharma | Complete |
| Neel Shah |
| Rishiraj Singh |

### 4.3 Development Methodology

This section describes the project as per the various stages of the Software Development life cycle. The model of software development life cycle used in this project is the waterfall method. The Waterfall Method is comprised of a series of very definite phases, as shown below in figure 4.7, each one run intended to be started sequentially only after the last has been completed, with one or more tangible deliverables produced at the end of each phase of the waterfall method of SDLC. Essentially, it starts with a heavy, documented, requirements planning phase that outlines all the requirements for the project, followed by sequential phases of design, coding, test-casing, optional documentation, verification (alpha-testing), validation (beta-testing), and finally deployment/release.



**Figure 4.4 Waterfall Model**

# CHAPTER 5 SYSTEM REQUIREMENTS

### 5. System Requirements

### The motto of this chapter is to identify the platform needed to run the proposed system. Team will

### Study the hardware as well as software requirements that will help in order to develop the system.

### 5.1 Hardware Requirements

**Processor:** Intel ® Core(TM) i3-7100U

**Main Memory (RAM):** 8 GB **Cache Memory:** 8 MB

**Monitor:** 13.3” Color Monitor

**Keyboard:** 108 keys

**Mouse:** Optical Mouse

**Hard Disk:** 32GB or more

**System Requirements:** 64-bit OS, x64-based processor

### Software Requirements

**Front End/Language:** html, bootstrap

**Back End/Database:** Python3, Flask

**Platform:** Google Colab

**Operating System:** Windows 7/Windows 8/ Windows 10

# CHAPTER 6 TESTING

### 6. Testing

### This chapter gives information about the test results

### 6.1 Test Approach

Software testing is an investigation conducted to provide stakeholders with information about the quality of the product or service under test. Software testing also provides an objective, independent view of the software to allow the business to appreciate and understand the risks of software implementation.

#### Black box testing

In black box testing we test the system at random for some random functionalities and depending on the output that we get we come to the conclusion that whether the system we have built is right or wrong. Internal system design is not considered in this type of testing. Tests are based on requirements and functionality. The number of modules and number of java files required for each module is checked.

#### White box testing

This testing is based on knowledge of the internal logic of an application code. Also known as Glass box Testing. Internal software and code working should be known for this type of testing. Tests are based on coverage of code statements, branches, paths, conditions. All the modules are tested for their logic whether it functions properly or not. Code is checked by inserting different inputs to check its functionality.

#### Unit testing

Testing of individual software components or modules. Each module was runs separately to check the output. Unit testing focuses first on the modules, independently of one another, to locate errors. This enables the tester to detect errors in coding and logical errors that is contained within that module alone. Those resulting from the interaction between modules are initially avoided. Here we test each module individually and integrate the overall system. Unit testing focuses verification efforts even in the smallest unit of software design in each module.

#### Integration testing

Integration testing is the testing process in software testing to verify that when two or more modules are interact and produced result satisfies with its original functional requirement or not. Integrated testing will start after completion of unit testing.

#### User Acceptance Testing

User acceptance testing of the system is the key factor for the success of any system. A system under consideration is tested for user acceptance by constantly keeping in touch with the prospective system at the time of development and making change whenever required. This is done with regard to the input screen design and output screen design. Here we will test whether the proposed system is having well defined UI so that the citizens can interface the application more easily.

#### Functional Testing

Functional testing is a technique in which all the functionalities of the program are tested to check whether all the functions that were proposed during the planning phase is full filled. This is also to check that if all the functions proposed are working properly. This is further done in two phases One before the integration to see if all the unit components work properly. Second to see if they still work properly after they have been integrated to check if some functional compatibility issues arise.

### Test Cases

A test case is a specification of the inputs, execution conditions, testing procedure, and expected results that define a single test to be executed to achieve a [software testing](https://en.wikipedia.org/wiki/Software_testing) objective. In this project, our test cases are listed below in the table.

**Fig 6.1- Test case for model 1**

|  |  |  |  |
| --- | --- | --- | --- |
| **SR NO** | **INPUTS** | **EXPECTED OUTPUT** | **OBSERVED OUTPUT** |
| 1. | Dataset row number = 51 | 0 | 0 |
| 2. | Dataset row number = 657 | 0 | 1 |
| 3. | Dataset row number = 709 | 1 | 1 |
| 4. | Dataset row number = 719 | 1 | 0 |
| 5. | Missing values | Error | Error |

**Fig 6.2 – Test case for model 2**

|  |  |  |  |
| --- | --- | --- | --- |
| **SR NO** | **INPUTS** | **EXPECTED OUTPUT** | **OBSERVED OUTPUT** |
| 1. | Dataset row no = 1 | 1 | 0 |
| 2. | Dataset row no = 4 | 0 | 0 |
| 3. | Dataset row no = 16 | 1 | 1 |
| 4. | Dataset row no = 34 | 0 | 1 |
| 5. | Missing values | Error | Error |

# CHAPTER 7 RESULT ANALYSIS

### 7. Result Analysis

### In this chapter the obtained result are analyzed and comparison between different algorithms are done on the basis of few parameters and data is visualized.

### 7.1 Evaluation Parameters

1. Accuracy

The accuracy of a [machine learning](https://deepai.org/machine-learning-glossary-and-terms/machine-learning) classification algorithm is one way to measure how often the algorithm [classifies](https://deepai.org/machine-learning-glossary-and-terms/classifier) a data point correctly. Accuracy is the number of correctly predicted data points out of all the data points.

2. Precision

 Precision, or the positive predictive value, refers to the fraction of relevant instances among the total retrieved instances.

Precision = TP / (TP + FP)

3. Recall

 Recall, also known as sensitivity, refers to the fraction of relevant instances retrieved over the total amount of relevant instances.

Recall = TP / (TP + FN)

4. F1 Score

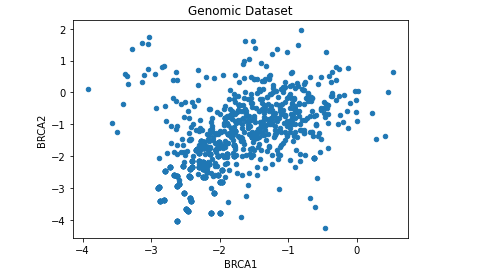
The F score, also called the F1 score or F measure, is a measure of a test’s accuracy. The F score is defined as the weighted [harmonic mean](https://deepai.org/machine-learning-glossary-and-terms/harmonic-mean) of the test’s [precision and recall](https://deepai.org/machine-learning-glossary-and-terms/precision-and-recall). F1 Score is calculated as,



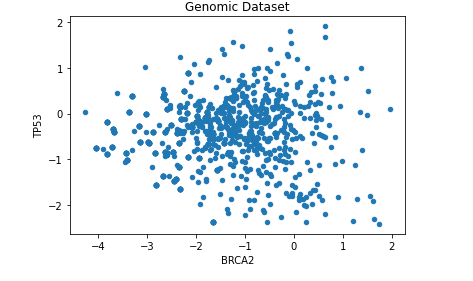
### 7.2 Result

**Table 7.2.1  Comparison of performance of Machine learning algorithms for model 1**

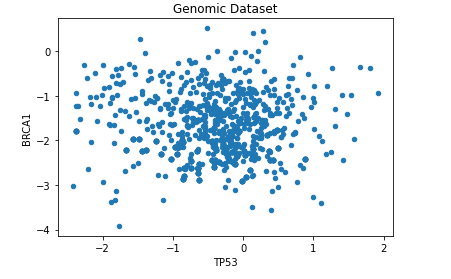
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sr no** | **Algorithm used** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
| **1** | **SVM** | **0.9768** | **0.99** | **0.96** | **0.97** |
| **2** | **Naïve Bayes** | **0.9259** | **0.94** | **0.91** | **0.92** |
| **3** | **Decision Tree** | **0.9898** | **0.96** | **0.95** | **0.96** |
| **4** | **KNN** | **0.9305** | **1.0** | **0.86** | **0.92** |



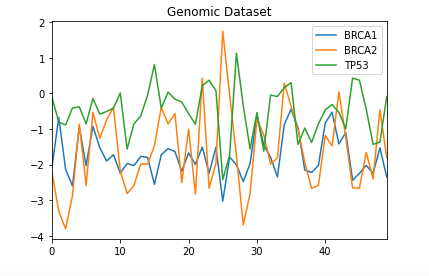
**Figure 7.2.1-  Scatter plotfor brca1, brca2**



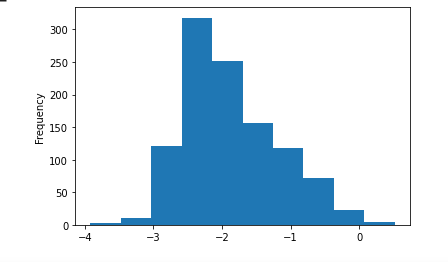
**Figure 7.2.2- Scatter plot for brca2, tp53**



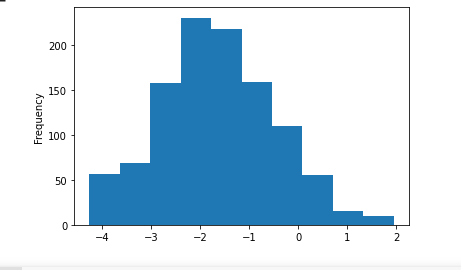
**Figure 7.2.3- Scatter plot for tp53 and brca1**



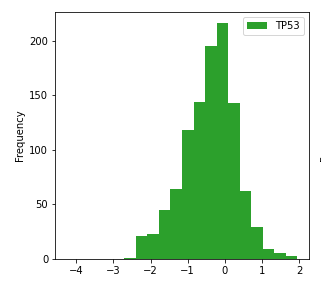
**Figure 7.2.4- Line chart for 50 rows**



**Figure 7.2.5- Histogram for brca1**



**Figure 7.2.6- Histogram for brca2**



**Figure 7.2.7- Histogram for tp53**

# CHAPTER 8 CONCLUSION

### Conclusion

From the above study, it is clear that the cancer prognosis is possible in most cases using machine learning on high dimensional genomic data. Conventional cancer prediction models don't accurately predict cancer at an early stage. By using genomic data this void can be filled as it helps in early prediction.  The microarray gene expression represents the mutation of genes, so if such genes are mutated then chances of tumour growing increases and eventually causing cancer. Thus, due to such microarray gene expression early prediction of cancer is feasible.

**8.2 Future Scope**

In this Application, four machine learning models for prediction of cancer were implemented. However, this is a partial system. For early prediction of cancer, more dimensions of the individual sample may be required. These dimensions can be the lifestyle of the individual, hereditary etc. Acquisition of such dimensional datasets and combining it with gene expression will be the future task and based on such datasets machine learning models can be built.

**BIBLOGRAPHY**

#### Journal Paper

1. Jishnu Das, Kaitlyn M Gayvert, and Haiyuan Yu “Predicting Cancer Prognosis Using Functional Genomics Data Sets” Published online 2014 Nov 2. doi: 10.4137/CIN.S14064 PMCID: PMC4218897 PMID: 25392695

1. Cai Huang, Evan A. Clayton, Lilya V. Matyunina, L. DeEtte McDonald, Benedict B. Benigno,FredrikVannberg, and John F. McDonald, “ Machine learning predicts individual cancer patient responses to therapeutic drugs with high accuracy” Published online 2018 Nov 6. doi: 10.1038/s41598-018-34753-5
2. Konstantina Kouroua Themis, P .Exarchosab Konstantinos,  P. Exarchosa Michalis V .Karamouzisc Dimitrios, I .Fotiadisab “ Machine learning applications in cancer prognosis and prediction ” Published online  [doi](https://doi).org/10.1016/j.csbj.2014.11.005 15 November 2014.
3. Chaima Aouiche, Bolin Chen, and Xuequn Shang “Predicting stage-specific cancer related genes and their dynamic modules by integrating multiple datasets” BMC Bioinformatics. 2019; 20(Suppl 7): 194. Published online 2019 May 1. doi: [10.1186/s12859-019-2740-6](https://dx.doi.org/10.1186%2Fs12859-019-2740-6) PMCID: PMC6509867 PMID: [31074385](https://www.ncbi.nlm.nih.gov/pubmed/31074385)
4. Qianfan Wu, Adel Boueiz,and Weiliang Qiu “ Deep Learning Methods for Predicting Disease Status Using Genomic Data” Published online 2018 Dec 11 PMCID: PMC6530791 NIHMSID: NIHMS1024586 PMID: 31131151
5. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542: 115–118. doi: [10.1038/nature21056](https://doi.org/10.1038/nature21056) [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/28117445)]
6. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. Int J Comput Vision. 2015;115: 211–252.
7. Hsu C-W, Chang C-C, Lin C-J. A practical guide to support vector classification, Technical Report Department of Computer Science and Information Engineering, National Taiwan University, Taipei 106, Taiwan, 2003
8. Gu Deshpande and Ramji Rai An Overview of Prognostics Markers in Breast Cancer [Med J Armed Forces India](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5531823/). 1999 Apr; 55(2): 129–132. Published online 2017 Jun 26. doi: [10.1016/S0377-1237(17)30268-X](https://dx.doi.org/10.1016%2FS0377-1237(17)30268-X) PMCID: PMC5531823 PMID: [28775603](https://www.ncbi.nlm.nih.gov/pubmed/28775603)
9. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23: 2507–2517. doi: [10.1093/bioinformatics/btm344](https://doi.org/10.1093/bioinformatics/btm344) [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/17720704)]
10. Radovic M, Ghalwash M, Filipovic N, Obradovic Z. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. BMC Bioinformatics. 2017;18: 9 doi: [10.1186/s12859-016-1423-9](https://doi.org/10.1186/s12859-016-1423-9) [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5209828/?report=reader)][[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/28049413)]
11. Gaul DA, Mezencev R, Long TQ, Jones CM, Benigno BB, Gray A, et al. Highly-accurate metabolomic detection of early-stage ovarian cancer. Sci Reports. 2015;5: 16351. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4647115/?report=reader)] [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/26573008)]
12. Guan W, Zhou M, Hampton CY, Benigno BB, Walker LD, Gray A, et al. Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines. BMC Bioinformatics. 2009;10: 259–274. doi: [10.1186/1471-2105-10-259](https://doi.org/10.1186/1471-2105-10-259) [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2741455/?report=reader)] [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/19698113)]
13. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S. et al.Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 2014;158: 929–44. doi: [10.1016/j.cell.2014.06.049](https://doi.org/10.1016/j.cell.2014.06.049)[[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4152462/?report=reader)] [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/25109877)]
14. Azuaje F. Computational models for predicting drug responses in cancer research. Brief Bioinform. 2016; pii: bbw065 (Epub ahead of print). [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5862310/?report=reader)][[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/27444372)]
15. Salesse S, Verfaillie CM. BCR/ABL: from molecular mechanisms of leukemia induction to treatment of chronic myelogenous leukemia. Oncogene. 2002;21: 8547–59. doi: [10.1038/sj.onc.1206082](https://doi.org/10.1038/sj.onc.1206082) [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/12476301)]
16. Wenming Zhao, Yiming Bao, Shunmin He, Guoqing Zhang et al.(2020) "Database resource of the national genomics data center"
17. Xie, Haozhe; Li, Jie; Jatkoe, Tim; Hatzis, Christos (2017), “Gene Expression Profiles of Breast Cancer”, Mendeley Data, v1
18. National Center for Biotechnology Information. Accessed on: Feb 13, 2020. Available: [https://www.ncbi.nlm.nih.gov/guide/genes-expression](https://www.ncbi.nlm.nih.gov/guide/genes-expression/)

Breastcancer.org. Accesses on:  Feb 13, 2020. Available: <https://www.breastcancer.org/risk/factors/genetics>.

#### Websites

Breastcancer.org. Accesses on:  Feb 13, 2020. Available: <https://www.breastcancer.org/risk/factors/genetics>

## PUBLICATIONS & CERTIFICATES

* + 1. “Abstractive text summarization using artificial intelligence”, 2nd International Conference on Advances in Science & Technology (ICAST 2019) SSRN, Elsevier - Abstract id - 3370795.
    2. Participated and won the National Level Project Competition KJSIEIT – INTECH ’19