

**Machine Learning approach for Predicting Breast Cancer using Genomic Data**

Submitted in partial fulfillment of the requirements of the degree of

Bachelors of Engineering in Information Technology By

**SAURABH SHARMA**

**NEEL SHAH**

**RISHIRAJ SINGH**

Supervisor

**MS. REENA LOKARE**



**DEPARTMENT OF INFORMATION TECHNOLOGY**

**K.J. SOMAIYA INSTITUTE OF ENGINEERING AND INFORMATION TECHNOLOGY, AYURVIHAR, SION, MUMBAI-22**

**2020**

 

**CERTIFICATE**

This is to certify that the project entitled “**MACHINE LEARNING APPROACH FOR PREDICTING BREAST CANCER USING GENOMIC DATA**” is a Bonafede work of students Saurabh Sharma, Neel Shah, Rishiraj Singh submitted to University of Mumbai in partial fulfillment of the requirement for the award of the degree of Bachelor of Engineering in Information Technology.

MS. Reena Lokare Project Guide

Department of Information Technology

|  |  |
| --- | --- |
|  |  |
| Dr. Radhika Kotecha | Dr. Suresh Ukarande |
| Head of Department | Principal |

Place: Sion, Mumbai-400022

Date:

**Project Report Approval for B.E.**

This project report entitled “**Machine Learning approach for Predicting Breast cancer using Genomic data**” by

**Saurabh Sharma (32)**

**Neel Shah (31)**

**Rishiraj Singh (33)**

is approved for the degree of Bachelor of Engineering in Information Technology.

Examiners: 1.

2.

Date:

Place: Sion, Mumbai-400022.

**Declaration**

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Saurabh Sharma

Neel Shah

Rishiraj Singh

Date:

Place: Sion, Mumbai-400022.

**ACKNOWLEDGEMENT**

We would like to convey our sincere thanks to everyone who guided us throughout this project endeavor, before we present our project entitled “**Machine Learning approach for Predicting Breast Cancer using Genomic Data** ”.

We wish to express our sincere thanks to **Dr. Suresh Ukarande**, our Principal, for providing us with all the necessary facilities for the research.

We place on record, our sincere thank you to **Dr. Radhika Kotecha**, Head of Department, for the continuous encouragement.

We are also grateful to **Ms. Reena Lokare**, Project Supervisor. We are extremely thankful and indebted to her for sharing expertise, and sincere and valuable guidance and encouragement extended to us.

We take this opportunity to express gratitude to all the Department faculty members for their help and support. We also thank our parents for unceasing encouragement, support and attention.

Saurabh Sharma

Neel Shah

Rishiraj Singh

Date:

**ABSTRACT**

Cancer prediction at an early stage is very crucial as the patient can then prepare for dealing with it. There are several Machine Learning models that help in predicting cancer by identifying samples of independent persons at high risk, facilitating the design and planning of cancer trials. These models use biomarkers like age, menopause, tumor-size, inv-nodes, breast, breast-quad dimensions to predict breast cancer. However, these models had major drawbacks of late prediction as well as low accuracy. So here presenting the system which uses gene expression profiles (genomic data) to predict breast cancer at an early stage. This model is built using different machine learning algorithms like a highly versatile support vector machine (SVM), Naive Bayes theorem, Decision tree and nearest neighbors’ approach to predict breast cancer using gene expression profiles.

## TABLE OF CONTENT

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **TITLE** | **Page No.** |
|  | List of Abbreviations | iii |
|  | List of Figures | iv |
|  | List of Tables | v |
| 1 | INTRODUCTION |  |
| 1.1 | Overview | 3 |
| 1.2 | Dataset used | 4 |
| 1.3 | Methods | 5 |
| 2 | LITERATURE REVIEW |  |
| 2.1 | Review of Literature | 6 |
| 3 | METHODOLOGIES AND IMPLEMENTATION | |
| 3.1 | Design Details | 10 |
| 3.2 | Algorithms | 12 |
| 3.3 | Implementation | 13 |
| 4 | PROJECT ANALYSIS |  |
| 4.1 | Project TimeLine | 17 |
| 4.2 | Task Distribution | 19 |
| 4.3 | Development Methodology | 20 |
| 5 | SYSTEM REQUIREMENTS |  |
| 5.1 | Hardware Requirements | 22 |
| 5.2 | Software Requirements | 22 |
| 6 | TESTING |  |
| 6.1 | Test Approach | 20 |
| 6.2 | Test Cases | 21 |
| 7 | RESULTS ANALYSIS |  |
| 7.1 | Evaluation Parameters | 27 |
| 7.2 | Result | 28 |
| 8 | CONCLUSION |  |
| 8.1 | Conclusion | 33 |

|  |  |  |
| --- | --- | --- |
| 8.2 | Future Scope | 33 |
|  | BIBLOGRAPHY | 34 |
|  | PUBLICATION and CERTIFICATES |  |

### List of Abbreviations

SVM – Support Vector Machine

BRCA1 – Breast cancer 1

BRCA2- Breast cancer 2

TP53-  Tumor Protein p53

PALB2 - Chr. Partner and localizer of BRCA2

CHEK2- Checkpoint kinase 2

CDH1- Cadherin-1

SDLC - Software Development Life Cycle

### List of Figures

|  |  |  |
| --- | --- | --- |
| **Figure no.** | **Title** | **Page no.** |
| 3.1.1 | Phases of prediction model | 10 |
| 3.3.1 | SVM model trained on weka | 13 |
| 3.3.2 | Expected and Observed results for Cancer in SVM Model | 13 |
| 3.3.3 | UI for model 2 | 14 |
| 3.3.4 | Input values for genes | 14 |
| 3.3.5 | Result displayed based on model trained | 15 |
| 3.3.6 | Dataset with same values as given to UI | 15 |
| 4.1 | Project Timeline 1 | 17 |
| 4.2 | Project Timeline 2 | 17 |
| 4.3 | Project Timeline 2 | 18 |
| 4.4 | Waterfall model | 20 |
| 7.2.1 | Scatter plotfor brca1, brca2 | 28 |
| 7.2.2 | Scatter plot for brca2, tp53 | 29 |
| 7.2.3 | Scatter plot for tp53 and brca1 | 29 |
| 7.2.4 | Line chart for 50 rows | 30 |
| 7.2.5 | Histogram for brca1 | 30 |
| 7.2.6 | Histogram for brca2 | 31 |
| 7.2.7 | Histogram for tp53 | 31 |

**List of Tables**

|  |  |  |
| --- | --- | --- |
| **Table no.** | **Title** | **Page no.** |
| 4.2  6.1 | Task Distribution  Test Cases for model 1 | 19  25 |
| 6.2 | Test Cases for model 2 | 25 |
| 7.2.1 | Comparison of performance of Machine learning algorithms for model 1 | 28 |