# **PREDICTION OF BREAST CANCER USING MACHINE LEARNING**

### **A PROJECT REPORT**

#### **Submitted by**

HARSH PANDEY (22BCS17030)

***in partial fulfillment for the award of the degree of***

**BACHELOR OF ENGINEERING**

**IN**

COMPUTER SCIENCE ENGINEERING



**Chandigarh University**

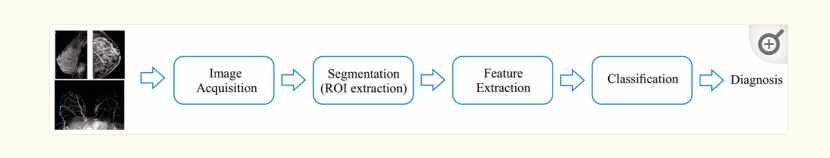
**TABLE OF CONTENT**

CHAPTER 2. LITERATURE REVIEW/BACKGROUND STUDY

* 1. [Timeline of the reported problem 3](#_TOC_250014)
  2. [Existing solutions](#_TOC_250013) 3
  3. [Bibliometric analysis](#_TOC_250012) 3-4
  4. [Review Summary](#_TOC_250011) 5-7
  5. [Problem Definition](#_TOC_250010) 8
  6. [Goals/Objectives](#_TOC_250009) 8
  7. References 9

## **1.1 INTRODUCTION**.

Breast cancer is the second leading cause of death for women (after lung cancer). Invasive breast cancer in women is anticipated to cause 246,660 new cases of diagnosis and 40,450 new cases of mortality in the US in 2016. One type of cancer that begins in the breast is breast cancer. When cells start to proliferate, cancer begins.in control. Breast cancer cells frequently cluster together to create a lump or an x-ray-visible tumour.When cancer cells enter the blood or lymphatic system and are transported to other parts of the body, breast cancer can spread. A Genetic alteration or mutation is one of the causes of breast cancer. There are numerous varieties of breast cancer, and Ductal carcinoma in situ (DCIS) and invasive cancer are typical examples. Some are less frequent, such as phyllodes tumours and angiosarcoma. There are numerous classification systems for the results of breast cancer. Fatigue, headaches, pain and numbness (peripheral neuropathy), bone loss, and osteoporosis are some of the adverse effects of breast cancer.



## **1.2 EXISTING SOLUTIONS**

There are numerous categorization and outcome prediction algorithms for breast cancer. The performance of four classifiers—SVM, Logistic Regression, Random Forest, and kNN—among the most important data mining methods, is compared in the current work. With a portable cancer detection kit or a mammogram during a screening test, it may be medically diagnosed early. As the disease progresses, cancerous breast tissues alter, and this relationship can be established. staging. How far a patient's disease has spread is indicated by the stage of their breast cancer (I-IV). Stages are determined using statistical indications like as tumour size, lymph node metastasis, distant metastases, and so forth. Patients must endure breast cancer surgery, chemotherapy, radiotherapy, and endocrine treatments to stop the cancer from spreading.

**1.3 BIBLIOMETRIC ANALYSIS**

**Background**: Triple-negative breast cancer (TNBC), reported in the early 2000s, is still the most difficult subtype of cancer due to its aggressive behavior, early recurrence, metastatic spread, and non-viable nature. Using machine learning, this study explores the research status and shortcomings of TNBC broadcasts from a macro perspective.  
  
**METHODS**: Searched and retrieved PubMed articles under the heading "Triple Negative Breast Cancer" between January 2005 and 2022. R and Python extract MeSH terms, geographic information, and other details from metadata. Use the Latent Dirichlet Allocation (LDA) algorithm to identify specific research topics.  
The Louvain algorithm creates a subject network and analyzes the relationships between subjects.  
  
**Results**: A total of 16826 documents were identified, with an average annual growth rate of 74.7%. 98 countries and territories worldwide participated in the TNBC study. Molecular pathogenesis and clinical applications are the most studied in TNBC research.  
The published literature mainly focuses on three topics: medical research, scientific research and technology research. The algorithm and citations show that TNBC research is based on technologies that enable TNBC subtyping, new drug development, and clinical trials.  
  
**Conclusions**: This study examines the quantitative nature of TNBC research from a macro perspective and helps inform basic and clinical research for effective diagnosis of TNBC. Medical target research and nanoparticle research are current research. Research on TNBC may be lacking in terms of patient, health economics, and health care.  
TNBC work will require the intervention of new technology.

## **1.4 Some data related to prediction model:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| AUTHOR | DATASET USED | TOOL USED | TECHNIQUE USED | ADVANTAGES |  | ACCURACY | ERROR RATE |
| 1. Wang et al.  [1] | Electronic health records | WEKA | Logistic regression | 5-year survivability prediction using  logistic regression |  | 96.4 % | 0.33 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 2. V Chaurisya  & S Paul [4] | Wisconsin breast cancer | WEKA | Statistical  Feature  Selection | Patient features sorted out from data materials are statistically tested based on the type of individual feature. Then 51 attributes or features are selected out, a feature’s importance score is calculated. XGBoost lgorithm is done by repeating 10-fold cross validation. | 92.3 % | 0.3% |
| 3. Akbugday  [2] | Breast  Cancer Wisconsin dataset | WEKA | KNN and  SVM | Optimal k-Value for a k-NN classifier, g kNN is a lightweight, lazy learning algorithm with very short build times. | KNN-  96.85%  NAÏVE  BAYES -  95.99%  SVM –  96.85% | 0.66% |
| 4. Keles, M.  Kaya, [3] | Wisconsin  Diagnostic  Breast Cancer dataset | Python | SVM vs KNN, decision trees and Naives  bayes | SVMs map the input    vector into a feature  space  of  higher  and  dimensionality  identify  the  hyperplane  that  data  separates  the  two  points  into  classes. The marginal  distance between the  decision  hyperplane  and the instances that  are  closest  to  is  boundary  maximized. | up to 96.91% | 0.33 |
| 5. KELES et  al., (2019) [3] | Wisconsin Dataset | WEKA | RANDOM  FOREST | Each dataset is generated with displacement from the original dataset. Then, trees are developed using a random selection feature, but are not pruned. | 92.2 % |  |
| 6. Chauraisa et. al [4] | UC Irvine machine learning  repository | WEKA | Naive Bayes, J48 Decision  Tree and  Bagging algorithm | Decision tree (C5) is the best predictor on the holdout sample (this prediction accuracy is better than any reported in the literature | 96.5% |  |
| 7. Delen at al.  [5] | Cancer Society | WEKA | ADABOOST | Low in error rate, performing well in the low noise data set. The advantage of this algorithm is that it requires less input parameters and needs little prior knowledge about the weak learner | 97.5 % |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 8. Kavithaa et.  at [6] | Cancer Society | MATL  AB | Ensemble method with Logistic and  Neural  Network | Multiple Learners are combined giving higher accuracy. | 96.3 % |  |
| 9. Sinthia et al.  [7] | Wisconsin  Diagnosis  Breast  Cancer  BCI dataset | CAD  System | Logistic  Regression  and the  Backpropagati  on neural  Network |  | 94.2 % |  |
| [10] Chaurasia  et. at [8] | Wisconsin  Breast  Cancer  (original) datasets | WEKA | SVM | It gives the most optimal hyperplane to distinguish between two classes | 97.13% |  |
| |  |  | | --- | --- | | Khuriwal | | | et.al | [9] |   [11] | Haberman's Survival dataset | WEKA | NAÏVE  BAYES AND  SVM | Helps in marginalizing the hyper-parameters and  differentiating classes. | 74.44% |  |
| [12]Khourdifi  et al. [10] | Wisconin breast cancer dataset | WEKA | Fast  Correlation-  Based Filter with SVM, Random  Forest, Naive Bayes, K-NN and MLP | Attributes are reduced by deleting irrelevant and redundant attributes, which have no meaning in the classification task techniques. | 96.1% | 0.0404 |
| [13]  Mohana,et.  al [11] | WISCONIN  BREAST  CANCER  DATASET | WEKA TOOL | DECISION TREE | Helps in Splitting and  choosing  the  best  attributes | 96.3 % |  |
| [14] Shravya  at al. [12] | UCI  repository | Spyder | SVM | Hyperplane separates two classes which  helps in higher accuracy. | 92.7% |  |
| [15] Wang et. al [13] | Wisconsin  Breast  Cancer  Database  (1991) and  Wisconsin  Diagnostic  Breast  Cancer  (1995) | WEKA | PCA | dimension reduction Wang and Yoon technique, manifests some advantages in terms of prediction accuracy and efficiency. | Eight PCs are chosen based on the scree plot, which  explains  92.6% of total correlation. And ten PCs are selected based on  95%  correlation |  |
| [16]  Bellaachia et.  al [14] | SEER  Public-Use Data. | WEKA | Naïve Bayes | Gives a probabilistic model for classification  Helping in classification. | 96.3 % |  |
| [17] Kim et. al  [15] | The breast cancer survivability  dataset  (1973–2003) from SEER | WEKA | SSL, and SSL-  Co(SemiSupervised  Learning) | SEMI-SUPERVISED  CO-TRAINING  SSL may be a good candidate to use as a predictive model for cancer survivability, particularly when the available dataset for model learning has an abundance of unlabelled patient cases. | 0.84 % |  |
| [18]  Bellaachia et.  al [14] | SEER database. | WEKA | Naïve Bayes,  the back-  propagated neural network, and the C4.5  decision tree | clear and fast classifier [10]. It has been called ‘Naïve’ due to the fact that it assumes mutually independent attributes | 84.5 % -  Naïve Bayes,  86.7% – NN,  81.3% –C4.5 | 0.57 |
| [19] Khuriwal and Mishra  (2018) [16] | Wisconsin  Diagnosis  Breast Cancer dataset. UCI open database | DWT  tool | Ensemble  Voting  Method,  Logistic  Regression | useful for predicting the class a binomial target feature. | 98.50% | 0.99% |
| [20]  AMRANE et  al. (2018) [17] | Breast  Cancer  Dataset  University of  California,  Irvine (UCI) | WEKA | KNN and  NAÏVE  BAYES | KNN classifiers are ranked first in terms of accuracy and duration. | 0.975109-  KNN  0.961932 |  |
| [21] Khuriwal  et. al [16] | Wisconsin  Breast Cancer database. | WEKA | Deep Learning  Neural  Network(conv  olutional neural network) |  | 99.67% | 0.0246 |
| [22] Al-hadidi et al. [18] | General Sample | MATL  AB | Logistic Regression and  Backpropagati  on neural  Network | BPNN is easy to implement and has been used widely for classification purposes. LR needs a hypothesis and a cost function which  optimizes performance. | Greater than  93.7% | Less than  0.07 |
| [23] Kibeom  et. al [19] | Gene  Expression  Dataset  Collection | WEKA | C4.5, Bagging and  ADABOOST  Decision trees | Ensemble Method helps to combine multiple learners. | Single C4.5  – 95.6%,  Bagging  C4.5 –  93.29%,AD ABOOST  C4.5 –  92.62% | Sensitivit  y – 56% and 72% |
| 4] Cruz et. al  [20] | Pubmed  (biomedical literature),  the Science  Citation  Index | MATL  AB | SVM,  NAÏVEBAYES | Helps to form a decision boundary and helps in classification. | 97.3% |  |
| [25] Medjahed  et. al [21] | Wisconsin breast cancer dataset | WEKA | Decision Trees | Helps in splitting | 96.1 % |  |

## **1.5 Problem definition**

Breast cancer is the most common type of cancer in women. When cancers are found early, they can often be cured. There are some devices that detects the breast cancer but many times they lead to false positives, which results in patients undergoing painful, expensive surgeries that were not even necessary. These types of cancer are called Benign which do not require surgeries and we can reduce these unnecessary surgeries by using Machine Learning. We take a classified dataset of the previous breast cancer patients and train the model to predict whether the cancer is benign or malignant. These classifications will help doctors to do surgeries only when the cancer is malignant, thus reducing the unnecessary surgeries for woman.

## **1.6 Prediction table outcomes:**

This model can be used for compiling as well as prediction of the data of benign and malignant cancer cells and their patients.The improvement achieved with this algorithm in accurate classification of women with and without breast cancer.Doctors can easily find the record of a particular patient with the tumour using this model. And further it can predict on the basis of past data that the new patient has or have the cancer tumour of which category whether it is benign or malignant (classification and prediction).

**REFERENCES**

[1]. Bansal N, Maurya A, Kumar T, Singh M, Bansal S. Cost performance of QoS Driven task scheduling in cloud computing. In: Procedia Computer Science. 2015.

[2]. Bala S, I. S, P. R. A Pheromone Based Model for Ant Based Clustering. Int J Adv Comput Sci Appl. 2012;

[3]. Sharda V, Agarwal RP. Analysis of Graphene Nanoribbon (GNR) interconnects with multi-gate device technology for VLSI applications. In: 2015 IEEE UP Section Conference on Electrical Computer and Electronics, UPCON 2015. 2016.

[4]. Kumar S, Kumar K, Pandey AK. Dynamic Channel Allocation in Mobile Multimedia Networks Using Error Back Propagation and Hopfield Neural Network (EBP-HOP). In: Procedia Computer Science. 2016.

[5]. Jain S, Gupta R, Dwivedi RK. Generating patterns from pizza ontology using protégé and weka tool. In: Proceedings of the 2018 International Conference on System Modeling and Advancement in Research Trends, SMART 2018. 2018.

[6]. Gupta N, Kumar Agarwal A. Object identification using super sonic sensor: Arduino object radar. In: Proceedings of the 2018 International Conference on System Modeling and Advancement in Research Trends, SMART 2018. 2018.

[7]. Gupta PK, Gupta S. Generation of green electricity with pedal generator. In: Proceedings of the 2018 International Conference on System Modeling and Advancement in Research Trends, SMART 2018. 2018.

[8]. Xu J, Yang G, Chen Z, Wang Q. A survey on the privacy-preserving data aggregation in wireless sensor networks. China Communications. 2015.

[9]. Verma KG, Kaushik BK, Singh R. Propagation Delay Variation due to Process Induced Threshold Voltage Variation. In: Communications in Computer and Information Science. 2010.

[10]. Bhardwaj S, Singhal N, Gupta N. Adaptive neurofuzzy system for brain tumor. In: Proceedings of the International Conference on Innovative Applications of Computational Intelligence on Power, Energy and Controls with Their Impact on Humanity, CIPECH 2014. 2014.

[11]. Mathiyalagan G, Devaraj D. A machine learning classification approach based glioma brain tumor detection. Int J Imaging Syst Technol. 2021;

[12]. Verma S, Biswas R, Singh JB. Extension of superblock technique to hyperblock using predicate hierarchy graph. In: Communications in Computer and Information Science. 2010.

[13]. Kishore N, Singh S. Torque ripples control and speed regulation of Permanent magnet Brushless dc Motor Drive using Artificial Neural Network. In: 2014 Recent Advances in Engineering and Computational Sciences, RAECS 2014. 2014.

[14]. Bakker EM. Image and video retrieval : Second International Conference, CIVR 2003, Urbana-Champaign, IL, USA, July 24-25 2003 : proceedings. Lecture notes in computer science. 2003.

[15]. Sowah R, Ampadu KO, Ofoli A, Koumadi K, Mills GA, Nortey J. Design and implementation of a fire detection and control system for automobiles using fuzzy logic. In: IEEE Industry Application Society, 52nd Annual Meeting: IAS 2016. 2016.

[16] International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN: 2347-5552, Volume-9, Issue-6, November 2021

[17] Murat Dener et al. / Procedia - Social and Behavioral Sciences 195 (2015) 1846 – 1850

[18] Z. Liu\*, G. Ferrier\*\*, X. Bao\*\*, X. Zeng\*\*, Q. Yu\*\*, A. Kim\* \*Fire Risk Management Program, National Research Council Canada, Ottawa, Ontario, Canada, K1A

[19] Akmalbek Abdusalomov 1 , Nodirbek Baratov 1 , Alpamis Kutlimuratov 2 and Taeg Keun Whangbo 3,\*(2021).

[20] Majid Bahrepour, Nirvana Meratnia, Paul Havinga Pervasive Systems Group, University of Twente(2007).

[**https://breast-cancer-research.biomedcentral.com/**](https://breast-cancer-research.biomedcentral.com/)

[**https://www.cancerresearchuk.org/about-cancer/breast-cancer**](https://www.cancerresearchuk.org/about-cancer/breast-cancer)

[**https://www.cancer.gov/types/breast/research**](https://www.cancer.gov/types/breast/research)