Journal of Information Systems Engineering and Management
2023, 8(2)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Tumor-Track: A Machine Learning-Based Predictive Framework for Enhanced Breast Cancer Diagnosis

Shilpi Singh[1], Govind Kumar Jha[2*] , Preetish Ranjan[1]
Email Id: shilpi.singh.it@gmail.com, gvnd.jha@gmail.com , pranjan@ptn.amity.edu
[1] Amity School of Engineering & Technology, Amity University Patna, Bihar- 801503, India.
[2] Government Engineering College Munger, Bihar-811202, India
**Corresponding Author: Govind Kumar Jha, ORCID: 0000-0003-2258-1865*

ARTICLE INFO

Received: 15 Mar 2023

Accepted: 25 Apr 2023

ABSTRACT

Breast cancer continues to be a worldwide health challenge with 2.3 million women diagnosed and 670,000 deaths reported in 2022. This paper would like to contribute by creating a model that uses machine learning with high precision and high efficiency in predicting breast cancer. This combines state-of-the-art algorithms in feature and clinical data analysis with a trustworthy tool to aid health professionals in screening and deciding early. The scope of this work is to give the possibility to increase the accuracy in detecting cancer, reduce false positive and offers a scalable solution applicable to mass use for breast cancer detection. The train the system to use Logistic Regression (LR) and K-Nearest Neighbors (KNN) classifiers to perform predictions. The performance is reported using the complete set of features along with a reduced subset of the ten most relevant features. Logistic Regression scores better than K-Nearest Neighbors even when working with the reduced set of the 10 most important features. The new method provides a simple but robust framework, which could help improve clinical diagnosis and patient prognosis.

Keywords: Tumor-Track, Breast Cancer, Machine Learning, Logistic Regression, K-Nearest Neighbor, Classification.

1. Introduction

Contemporary individuals are likely hooked to the internet and lack concern for their physical health. Individuals disregard minor health issues and cease hospital visits, eventually resulting in chronic illnesses. Early diagnosis, particularly in cancer cases, facilitates prompt intervention and treatment, potentially leading to improved patient outcomes and reduced morbidity for individuals and the healthcare system overall. Consequently, it is essential to create effective early detection measures to mitigate the impact of such illnesses on public health.

Conversely, in other nations, the Indian Council of Medical Research(ICMR) has shown a greater prevalence of cancer among Indian women compared to Indian males, with this disparity increasing, as per a recent biannual study. These dismal statistics exclude data from two of India's most populated states, Uttar Pradesh and Bihar. The national cancer incidence rate for 2022 is 100.4 per 100,000 individuals. Breast cancer is a preventable disease, with a diagnosis rate of 105.4 per 100,000 women.

In contrast, 95.6 males per 100,000 have received a diagnosis of lung cancer. The study was conducted at the ICMR's National Centre for Disease Informatics and Research in Bengaluru. The survey indicated that one in every nine Indians is at risk of contracting the illness before the age of 74.

Breast and cervical cancers were the most prevalent malignancies among females, followed by ovarian and uterine corpus tumors. Modeling predicts that breast cancer incidence in India would reach 250,000 by 2030. The expected number of breast cancer cases in the nation is around 182,000. In 2020, the World Health Organization (WHO) reported that 2.3 million women were diagnosed with breast cancer, resulting in 685,000 worldwide fatalities. By the conclusion of 2020, 7.8 million women diagnosed with breast cancer over the preceding five years remained alive, making it the most prevalent disease globally. Cancer is recognized as the aberrant proliferation of human cells that attack healthy cells. Atypical breast cell proliferation often results from unregulated cell division and expansion. Breast cancer is a condition characterized by the formation of malignant cells inside the breast tissues. Breast masses are classified into two categories: non-cancerous (benign) and cancerous (malignant) [2]. Tumors are classified as benign if they are non-threatening and malignant if they pose a threat. In contrast to malignant tumors, most benign tumors do not need treatment. Malignant cells may disseminate damaging enzymes to the adjacent tissues.

Early identification and prognosis of isolated diseases or their development to others may be crucial for effective management and improved patient survival. In machine learning, many researchers start their work by assessing the severity of breast cancer, specifically determining if a tumor is benign or malignant. Two critical components are essential to address such inquiries: the machine's contribution and how machine learning integrates medical data to predict illness severity. Machine learning algorithms enable data-driven decision-making with little human intervention.Machine learning is a subset of artificial intelligence that can acquire knowledge from data, make judgments, identify patterns, and construct analytical models via data analysis. Clinical or medical data include information on human health acquired via standard patient care and clinical trial techniques. This encompasses provider EMR, relevant patient EMR, and patient health records, including patient-related health information.

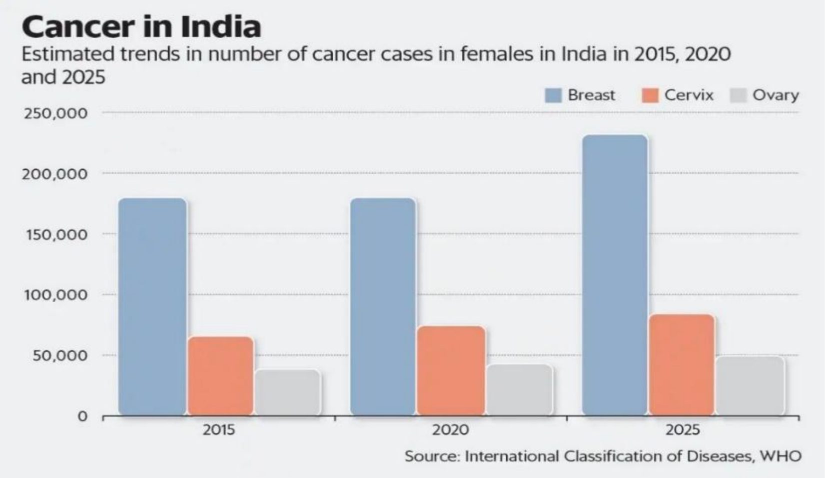


Figure 1. Estimated trends in the number of cancer cases in females in India.

AI can harvest knowledge from health data to digest the knowledge and return precise results to end-users via machine learning [3].The algorithm this method uses recognises patterns of the data, producing its own logic.The AI algorithm aims to identify the association of prevention or treatment with the patient's prognosis [4].

This study aims to construct a powerful prediction model for breast cancer based on machine learning methods. This work seeks to leverage the extensive pools of patients' demographics and clinical and tumor characteristics available by developing decision support systems (DSS) capable of predicting for individuals who might be at elevated risk of developing BC. The workflow consists of several main parts: data gathering, preprocessing, feature reduction, algorithm realization, and performance measurement. An initial data set is obtained that includes all relevant variables for analysis. By carefully preprocessing the data, integrity is preserved and the features are normalised to make a meaningful comparison. Feature importance examination then identifies the most significant predictive variables for breast cancer occurrence. Next, two types of classifiers i.e., logistic regression and k-nearest neighbor (KNN) are used for fitting the models. They were selected due to their ability to deal with linear and non-linear relationships in the data.

Additionally, it is worth noting that the performance of these models are computed on two feature sets which are, 1) All features and 2) A restricted feature set containing the top ten essential features selected through feature selection. This comparative study aims to expose their effects on prediction accuracy and efficiency. This study aims to use powerful machine learning techniques and statistical analysis to improve breast cancer prediction models. In addition, we have designed a user-friendly prototype, TUMOR-TRACK for early prediction of breast cancer. In conclusion, if successful, the findings of this research will provide information that applies to clinical decision-making, improving early detection and curbing the global burden of breast cancer on the individual and the health care system.

The remaining manuscript is structured as follows:

Section 2: Literature Review, which covers previous studies.

Section 3: Technical details of the proposed model, including data investigation and preprocessing, statistical technique modeling, and the overall framework.

Section 4: Application of the proposed model on the dataset and data subsets, along with a balanced comparative analysis of the method and its overall structure.

Section 5: Conclusion.

Section 6: Discussion of future work.

2. Literature Review

Review of Literature The literature review section aims to provide a thorough and critical analysis of the available research on the Breast Cancer Prediction System. A multitude of machine learning (ML) techniques exist for the prediction and diagnosis of breast cancer. The classifiers include Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision Tree, and K-Nearest Neighbours (KNN). Both publications reveal that several researchers have conducted studies on breast cancer using various datasets, including mammography pictures, the SEER dataset, the Wisconsin dataset, and datasets from other hospitals. Consequently, while using these datasets, researchers must identify and pick certain advantageous traits to conduct their studies, thereby advancing the study in this domain.

Benbrahim et al. [5] observed that the Neuro Network method had superior performance when optimized. Deepika et al. [6] used Naive Bayes and Multi-Layer Perceptron algorithms, concluding that the former exhibited superior accuracy. Mariam et al. [7] suggested Naive Bayes and K-Nearest Neighbors (KNN) for breast cancer classification, with KNN achieving a superior performance of 97.51% compared to Naive Bayes's highest performance of 96.19%. Aruna et al. [8] used Naïve Bayes, SVM, and Decision Tree classifiers on the Wisconsin breast cancer dataset, with SVM achieving the highest accuracy of 96.99%. Chaurasia et al. [9] similarly evaluated Naive Bayes, SVM, Neural Networks, and Decision Trees, with SVM achieving the greatest accuracy of 96.84%. The authors Sakri et al. [10] focused on improving the precision of breast cancer diagnosis in Saudi Arabian women using a Particle Swarm Optimization (PSO) algorithm applied to K-Nearest Neighbors (KNN), Naive Bayes, and Reduced Error Pruning (REP) tree methodologies. In their study, PSO attained an accuracy of 81.3% for Naive Bayes, 80% for REP tree, and 75% for KNNs. Kapil and Rana [11]: They introduced a weight-enhanced decision tree methodology achieving about 99% ideal accuracy based on the WBCD dataset, and consequently 85-90% on an alternative UCI breast cancer dataset. Yue et al. [12] evaluated many machine learning techniques. They established that the Deep Belief Network (DBN) integrated with Artificial Neural Network (ANN) architecture yielded the most excellent accuracy of 99.68%, followed by the Support Vector Machine (SVM) with a two-step clustering approach at 99.1%, and an ensemble method using SVM, Naive Bayes, and J48, which achieved a score of 97.13%.

S. Nayak et al. [13] illustrate many supervised machine learning methods for breast cancer classification using 3D pictures, concluding that SVM provides the best overall performance. B.M. Gayathri et al. [14] conduct a comparative research that emphasizes the Relevance Vector Machine (RVM) for its low computing expense and exceptional efficacy in breast cancer diagnosis, with 97% accuracy despite reducing variables. Asri et al. [15] demonstrate that Support Vector Machine (SVM) outperforms breast cancer prediction and detection, attaining superior precision and a minimal error rate with an accuracy of 97.13%. Youness et al. [16] evaluate machine learning techniques and determine that SVM is the superior classifier, achieving an accuracy of 97.9% in contrast to K-NN, RF, and NB, using a Multilayer Perceptron with 5 layers and 10-fold cross-validation. Latchoumi et al. [17] achieved a classification accuracy of 98.4% by introducing a weighting optimization of the particle swarm (WPSO) grounded on the SSVM. Ahmed et al. [18] provide a diagnostic method for Wisconsin breast cancer (WBCD), with a prediction accuracy of 99.10% by integrating the SVM algorithm with a clustering technique and an effective probabilistic vector support machine.

3. Methodology

The primary aim of our experiment is to ascertain the most efficacious and predictive algorithm for breast cancer detection and then create a user-friendly application. We used machine learning classifiers, namely Logistic Regression and K-Nearest Neighbors (KNN), using the Breast Cancer dataset sourced from Kaggle to do this. Subsequently, we assessed the outcomes to ascertain which model yields superior accuracy. Figure 2 delineates the planned architecture. Our technique starts with data collecting, followed by pre-processing, which encompasses four stages: data cleansing, attribute selection, target role designation, and feature extraction. The processed data is then used to develop machine-learning algorithms that predict breast cancer from novel parameters. To assess the efficacy of these algorithms, we provide them with novel data for which we possess labels. This is often accomplished by dividing the labeled data into two segments via the train_test_split function. We use 75% of the data to construct our machine learning model, known as the training data or training set, and 25% of the data to evaluate the model's performance, termed the test data or test set. Upon evaluating the models, we compare the outcomes to determine the algorithm that yields the maximum accuracy, hence establishing the most predictive algorithm for breast cancer screening.

3.1 Used Dataset

For this work, we collected the Breast Cancer Wisconsin Diagnostic dataset from Kaggle [19], an open-source website providing vast datasets across various domains, including finance, healthcare, social sciences, and more. Users can explore and download these datasets for analysis and experimentation or use in their projects. The datasets range from small and curated to large and complex, catering to different needs and skill levels. In this dataset, there are 569 rows and 31 columns. Out of the 569 values, 357 values have Benign which means they do not have breast cancer, and 212 values have Malignant which means that they have breast cancer.

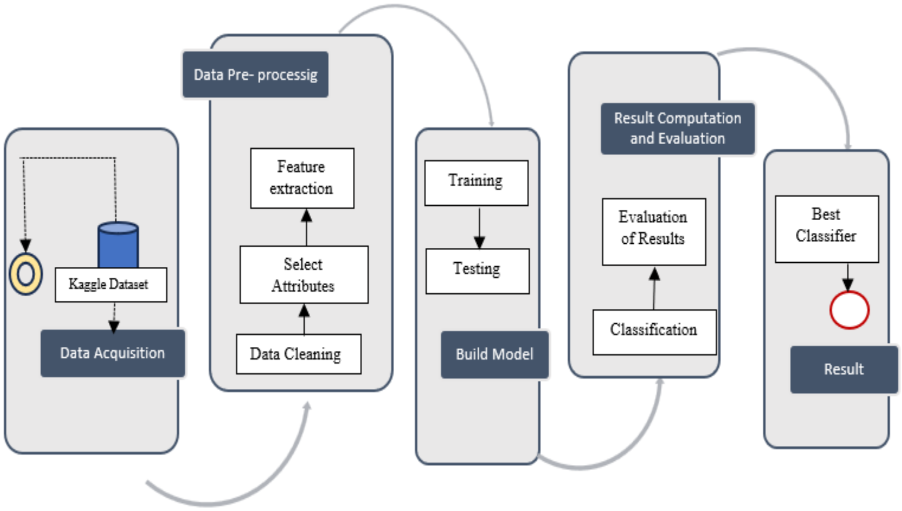


Figure 2. Process Flow Diagram.

The following are some feature names that are present in this dataset:

Worst Perimeter: The most significant distance around the outside of the tumor mass among all measurements, indicating potential tumor size.

Worst Radius: The most significant distance from the center to the edge of the tumor mass among all measurements, suggesting tumor size.

Worst Concave Points: Maximum number of inward curving points on the tumor boundary, indicating irregularities in shape.

Worst Area: The most significant area covered by the tumor mass among all measurements, indicating tumor size.

Mean Concave Points: The average number of concave points across multiple measurements reflects the average irregularity in the tumor boundary.

Concave Points Error: Standard error or variation in the number of concave points among measurements, indicating consistency or variability in tumor shape.

Worst Concavity: The most immense indentation depth in the tumor mass among all measurements, indicating a pronounced concave shape.

Worst Symmetry: The least symmetric measurement of the tumor mass among samples, reflecting irregularity in shape.

Concavity Error: Standard error or variation in the measure of concavity among measurements, providing insight into consistency or variability in tumor shape.

3.2 Experimental Environment

All experiments on the machine learning techniques presented in this work were executed using Jupyter notebook and the Python programming language. Jupyter Notebook offers an interactive platform for developing Python-based data science applications. Previously referred to as IPython notebooks, they provide several characteristics that make them an essential Python machine learning environment component. Scikit-learn, sometimes called sklearn, is an open-source machine learning library can be effectively used within Anaconda's integrated development environments (IDEs). It provides a range of classification, regression, and clustering methods, such as logistic regression, K-nearest neighbors, support vector machines, random forests, k-means, and DBSCAN. It is also designed to integrate effortlessly with the Python numerical and scientific libraries NumPy and SciPy.

3.3 Data Preprocessing

Data preprocessing is essential in developing a predictive system, including cleaning, transforming, and organizing raw data into a format appropriate for analysis and forecasting. The objective is to improve the quality and use of the data for future analytical activities. Initially, we examine the dataset for null values. Subsequently, we delineate the data and use the XGBRFRegressor package from the XGBoost modules to evaluate the significance of the features within the dataset. Figure 3 illustrates the importance of traits in decreasing order.In constructing the model, we designate the goal values as 0 for Malignant and 1 for Benign.To evaluate the accuracy of utilizing all features versus only the salient ones, we identify the ten most significant features (worst perimeter, worst radius, worst concave points, worst area, mean concave points, concave points error, worst concavity, worst symmetry, concavity error, worst texture) and omit the remainder.

3.4 Data Analysis

During the data analysis phase, we examine the distribution of various features. The distribution plot of each attribute can significantly impact the accuracy of the generated functions. These characteristics must follow a normal distribution, often represented by a Gaussian or bell curve. The Gaussian distribution plot helps identify data skewness, which can be either positive or negative.

The mean is greater than the median in a positively skewed distribution, indicating that the data is skewed towards the lower end. Here, the mean exceeds the median because the median is the middle value, and the mode is the highest value. Conversely, in a negatively skewed distribution, the mean is less than the median, showing that the data skews towards the higher end. In this type of distribution, the mean, median, and mode is negative.

Several key feature distributions are analyzed, such as:

Figure 4: Represents the distribution of target values and shows the accuracy percentage of datasets.

Figure 5: Depicts the mean area distribution.

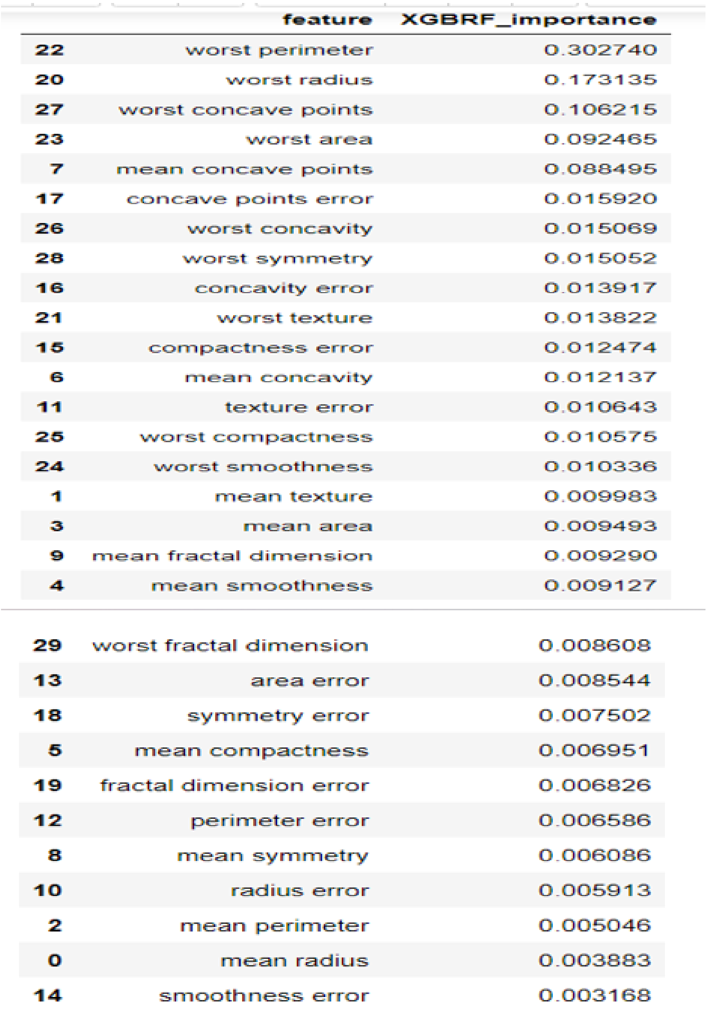
Figure 6: Illustrates the concave points error distribution.

Figure 7: Displays the worst perimeter distribution.

Figure 8: Shows the correlation matrix, highlighting the dependencies between all the features.

3.5 Classification Algorithm

A classification algorithm is a method within machine learning used to sort data into distinct categories or classes based on specific characteristics. These algorithms analyze labeled data to identify patterns and then apply these patterns to classify new, unlabeled data into predefined categories. We have used two classification algorithms: Logistic Regression and K-Nearest Neighbor.



	feature	XGBRF_importance
22	worst perimeter	0.302740
20	worst radius	0.173135
27	worst concave points	0.106215
23	worst area	0.092465
7	mean concave points	0.088495
17	concave points error	0.015920
26	worst concavity	0.015069
28	worst symmetry	0.015052
16	concavity error	0.013917
21	worst texture	0.013822
15	compactness error	0.012474
6	mean concavity	0.012137
11	texture error	0.010643
25	worst compactness	0.010575
24	worst smoothness	0.010336
1	mean texture	0.009983
3	mean area	0.009493
9	mean fractal dimension	0.009290
4	mean smoothness	0.009127
29	worst fractal dimension	0.008608
13	area error	0.008544
18	symmetry error	0.007502
5	mean compactness	0.006951
19	fractal dimension error	0.006826
12	perimeter error	0.006586
8	mean symmetry	0.006086
10	radius error	0.005913
2	mean perimeter	0.005046
0	mean radius	0.003883
14	smoothness error	0.003168

Fig. 3 Features importance in decreasing order

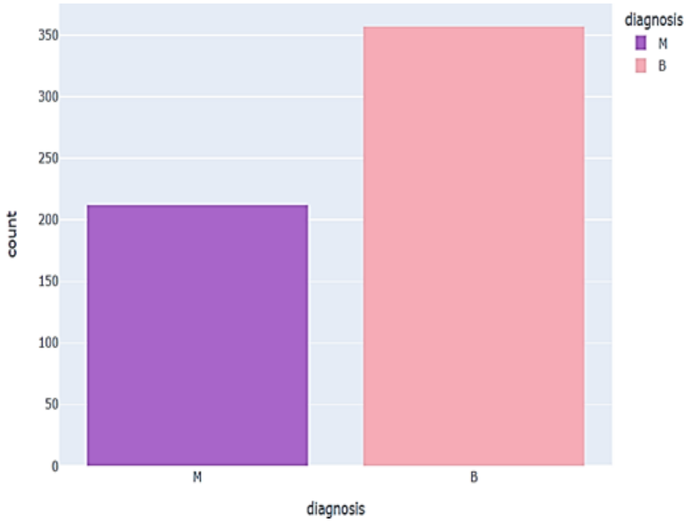


Fig 4. Target value Distribution

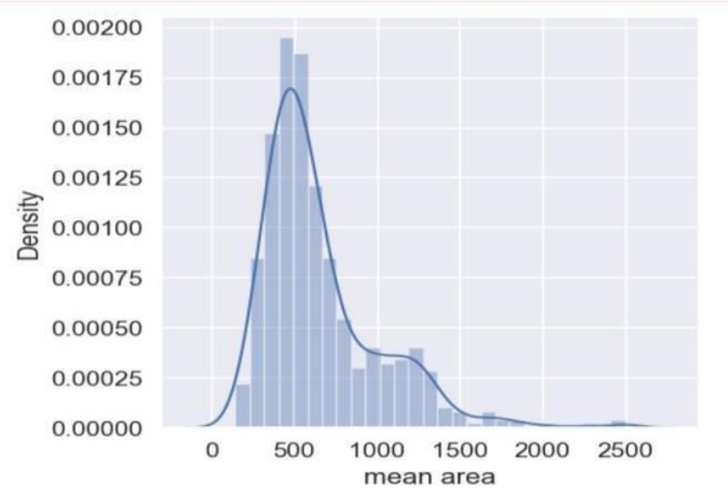


Fig 5. Mean area Distribution

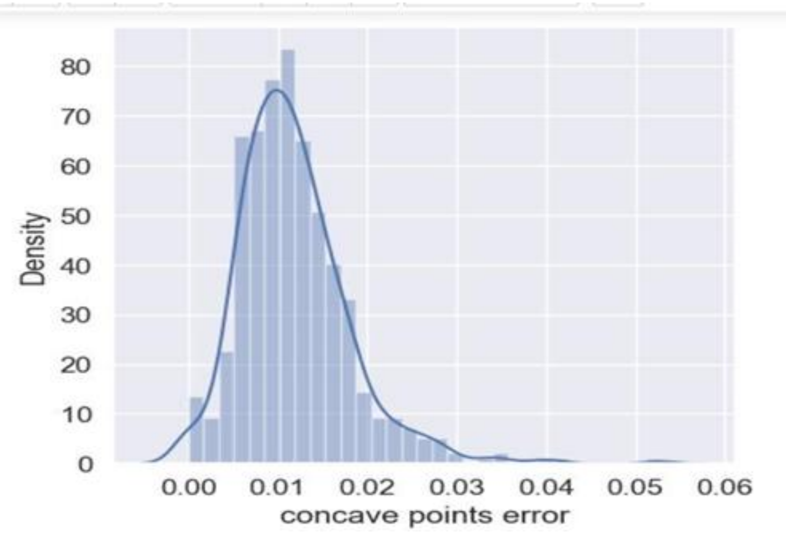


Fig 6. Concave points error Distribution

Research Article

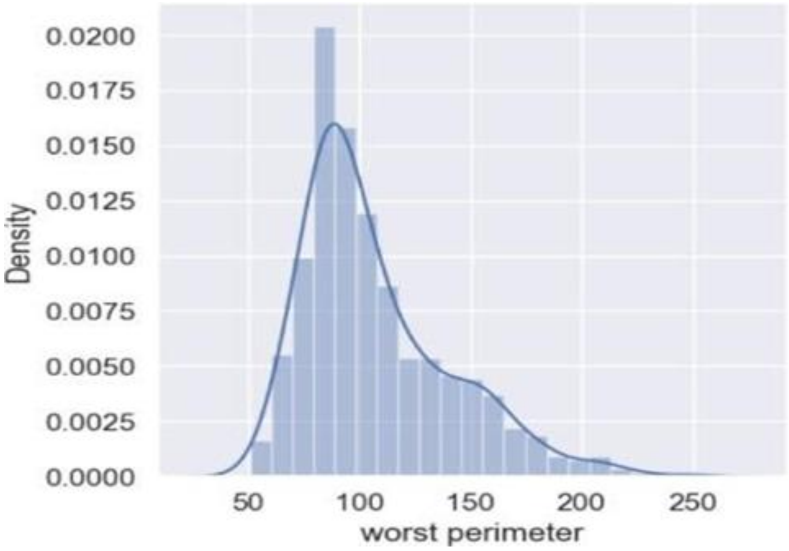


Fig 7. Worst perimeter Distribution

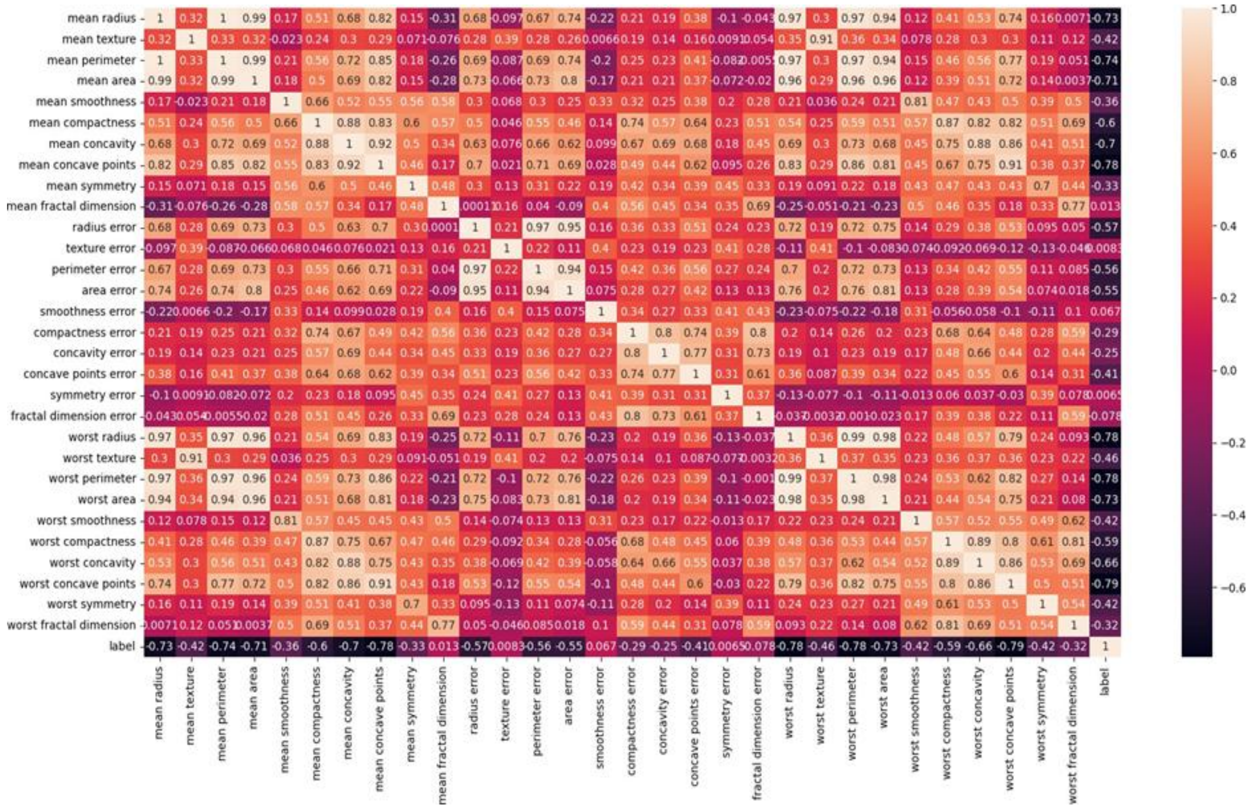


Fig 8. Correlation matrix

Logistic Regression is a statistical technique used for binary classification problems, whereby the outcome variable is categorical and has just two potential outcomes (e.g., yes/no, true/false). It delineates the association between a dependent variable and one or more independent factors by evaluating the likelihood of the dependent variable being categorized into a certain class. Notwithstanding its designation, logistic regression is a classification method, not a regression technique. The method involves applying a logistic function to the observed data, facilitating the prediction of probabilities and the classification of new data points into one of two groups according to a defined threshold. Figure 9 depicts the flowchart of the comprehensive logistic model.

K-Nearest Neighbor- KNN (K-Nearest Neighbors) is a flexible algorithm employed widely in various machine learning classification and regression problems. The Model also maintains the training dataset, in which each instance includes features and a class label or target value. When a KNN prediction is made for the new data point, the KNN algorithm compares (using distance theorems such as Euclidean, Manhattan or Minkowski distance) the new data point with its training dataset neighbors. It transfers the label of its 'K' most similar neighbors to this new data point. It then finds the 'k' closest data points of the new data points and estimates the outcome variable for this new data point. For classification, it makes a prediction that is the majority class among the closest neighbors while for regression it depths the mean (or weighted mean) of the target values. Important details related to the K-NN search are the length of a circular neighborhood 'k' and the type of distance function, which significantly influence the behavior of the algorithm. Furthermore, it allows for weighting schemes where nearer neighbors contribute more to predictions. The overall KNN model flowchart is described in Figure 10.

The following is the formula for the distance calculation.

$Euclidean\ Distance\ =$

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \qquad (1)$$

$Manhattan\ Distance\ =$

$$d(x,y) == \sum_{i=1}^{n}|x_i - y_i| \qquad (2)$$

$Minkowski\ Distance\ =$

$$d(x,y) == \left(\sum_{i=1}^{n}|x_i - y_i|^p\right)^{1/p} \qquad (3)$$

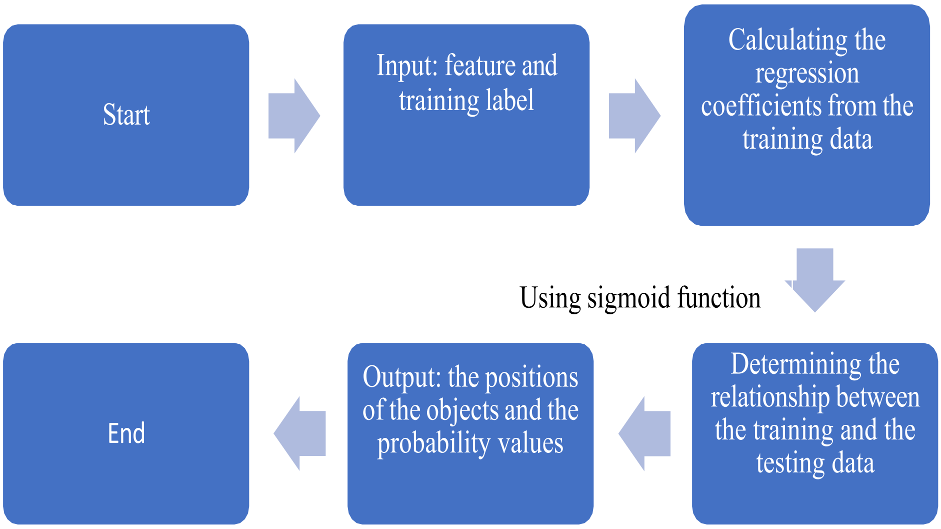

Start
Input: feature and training label
Calculating the regression coefficients from the training data
Using sigmoid function
Determining the relationship between the training and the testing data
Output: the positions of the objects and the probability values
End

Fig 9. Flowchart For Logistic Regression

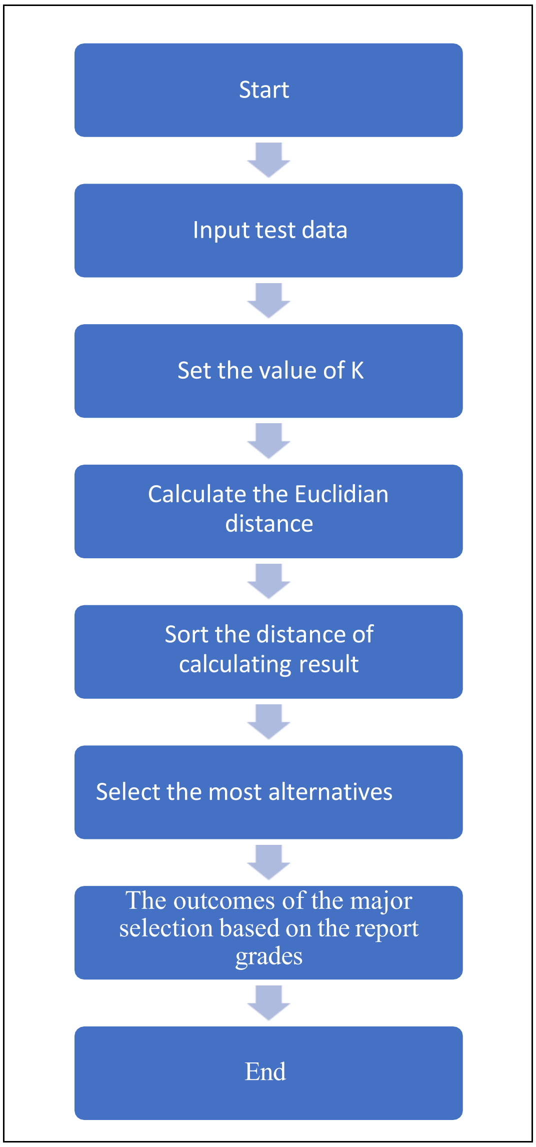

Start
Input test data
Set the value of K
Calculate the Euclidian distance
Sort the distance of calculating result
Select the most alternatives
The outcomes of the major selection based on the report grades
End

Fig 10. Flowchart For K-Nearest Neighbors

3.6 Check the accuracy

Accuracy in machine learning refers to the percentage of correctly classified instances from the total cases in a dataset. This fundamental metric assesses the performance of classification models. Higher accuracy indicates better predictive capabilities, making it a critical factor in model selection as it allows for comparing different models or their variations, aiding in decision-making. Additionally, high accuracy boosts user confidence in the model's predictions, potentially reducing costs by minimizing errors. However, accuracy may not always be the most suitable metric, especially in imbalanced class distributions or varying error costs. In such scenarios, alternative metrics like precision, recall, F1-score, or AUC-ROC provide a more comprehensive evaluation. We evaluated both models' accuracy using all significant features and only the top ten important features in this context, as mentioned in Table 1.

Table 1. Comparison table for accuracy

Classification Algorithm	Accuracy with all features	Accuracy with an important feature
Logistic Regression	0.92	0.93
K-Nearest Neighbors	0.91	0.89

4. Results and Discussion

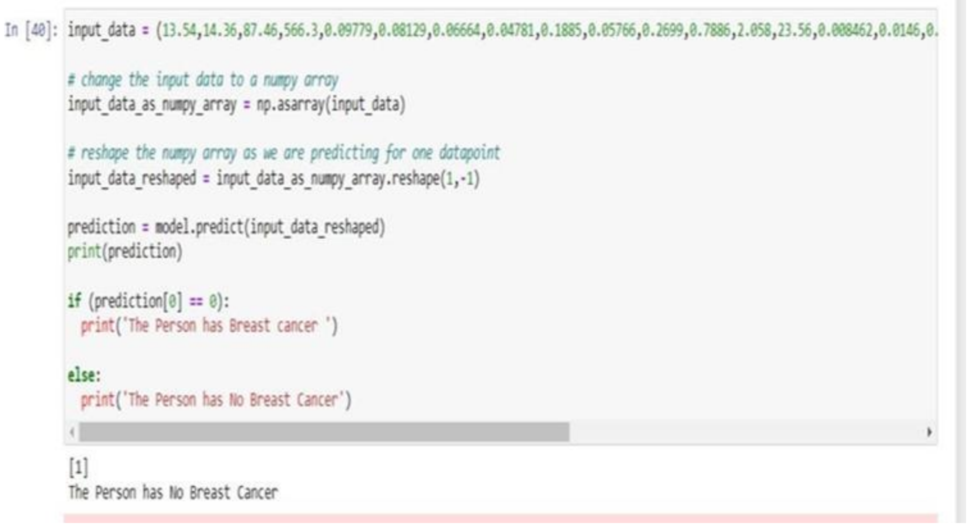
The task was executed on an Intel Core i7 CPU operating at 2.30GHz, with 16 GB of RAM and 500 GB of external storage. All tests on the classifiers detailed in this paper were executed using libraries from the Anaconda machine-learning environment. We divided the data into 70% for training and 30% for testing in the experimental investigations.

Jupyter was used to develop a suite of machine learning algorithms for data preprocessing, classification, regression, clustering, and association rule mining. These algortithms were used to address a range of real-world issues.

The project's outcomes are as follows:

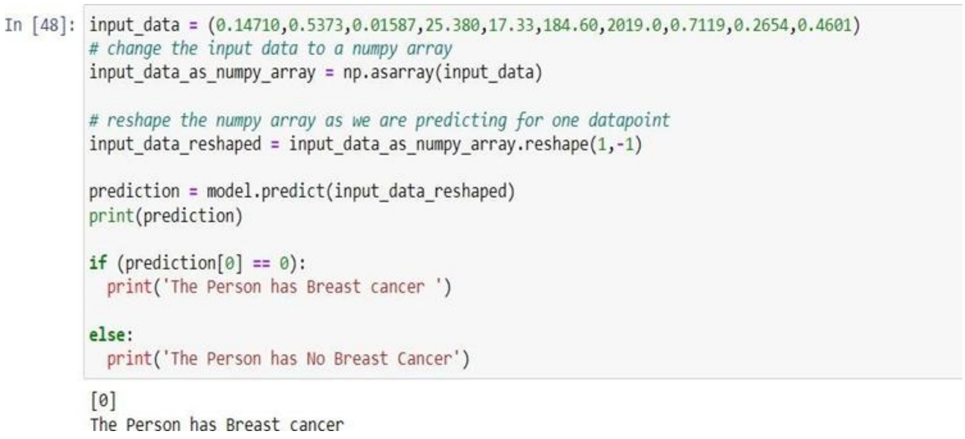
Employing logistic regression with all variables, we attained an accuracy of 0.9298. The accuracy improved to 0.9385 while using just the essential characteristics. The accuracy of KNN using all characteristics was 0.9122, however it declined to 0.8947 while utilizing just the significant features. User feedback was also included in the study. Figure 11 displays the result including all characteristics, whereas Figure 12 presents the output with the 10 most significant features. We evaluated the accuracy of both methods using ROC curves: Figure 13 illustrates the accuracy graph of the logistic regression model including all characteristics, whereas Figure 14 presents the graph with significant features. Figure 15 illustrates the accuracy graph of the KNN model using all characteristics, whereas Figure 16 presents the graph employing substantial features.

We used confusion matrices to assess the efficacy of the categorization systems. These matrices provide a comprehensive analysis of true positives, true negatives, false positives, and false negatives, facilitating the enhancement of model accuracy and the identification of mistake kinds. Figure 17 illustrates the confusion matrix for the logistic regression model using all data, whereas Figure 18 presents the matrix with significant features. Figures 19 and 20 illustrate the confusion matrices for the KNN model using all features and significant features, respectively.



```
In [40]: input_data = (13.54,14.36,87.46,566.3,0.09779,0.08129,0.06664,0.04781,0.1885,0.05766,0.2699,0.7886,2.058,23.56,0.008462,0.0146,0.

         # change the input data to a numpy array
         input_data_as_numpy_array = np.asarray(input_data)

         # reshape the numpy array as we are predicting for one datapoint
         input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

         prediction = model.predict(input_data_reshaped)
         print(prediction)

         if (prediction[0] == 0):
           print('The Person has Breast cancer ')

         else:
           print('The Person has No Breast Cancer')

         [1]
         The Person has No Breast Cancer
```

Fig 11. Output with all features



```
In [48]: input_data = (0.14710,0.5373,0.01587,25.380,17.33,184.60,2019.0,0.7119,0.2654,0.4601)
         # change the input data to a numpy array
         input_data_as_numpy_array = np.asarray(input_data)

         # reshape the numpy array as we are predicting for one datapoint
         input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

         prediction = model.predict(input_data_reshaped)
         print(prediction)

         if (prediction[0] == 0):
           print('The Person has Breast cancer ')

         else:
           print('The Person has No Breast Cancer')

         [0]
         The Person has Breast cancer
```

Fig 12. Output with ten important features

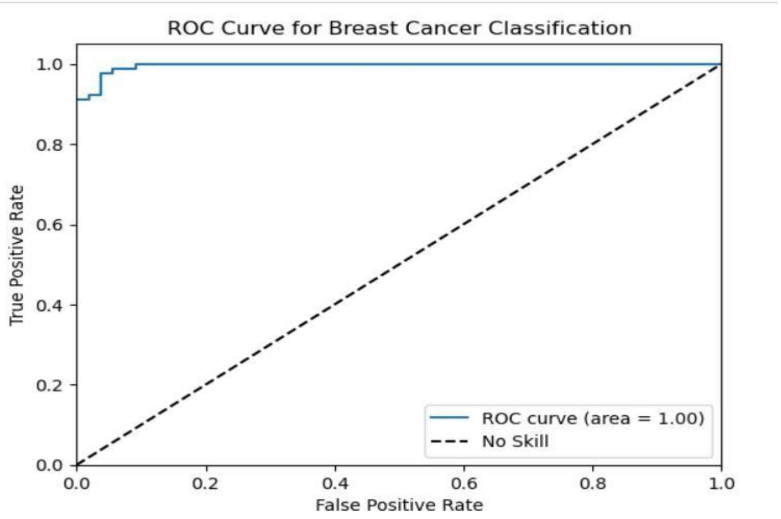


Fig 13. Accuracy Graph of LR model with all features

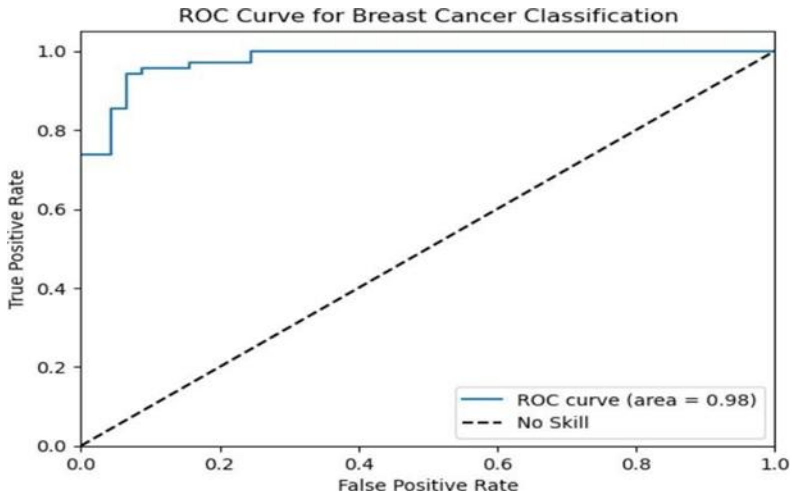


Fig 14. Accuracy Graph of LR model with ten important features

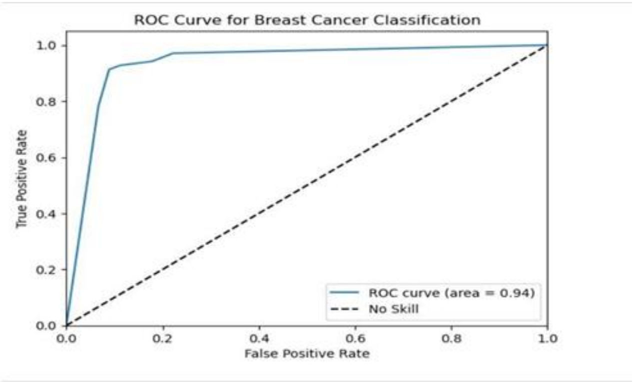


Fig 15. Accuracy Graph of KNN model with all features

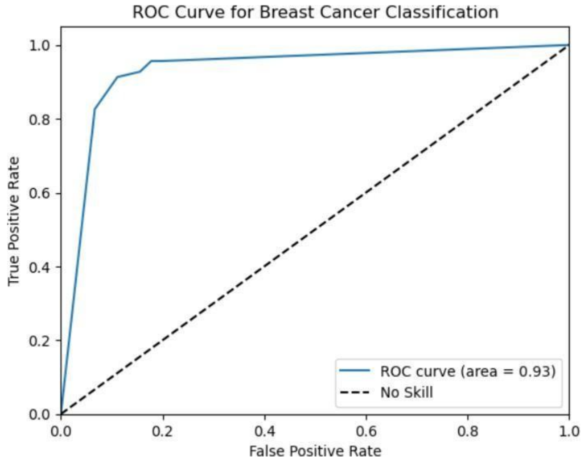


Fig 16. Accuracy Graph of KNN model with ten important features

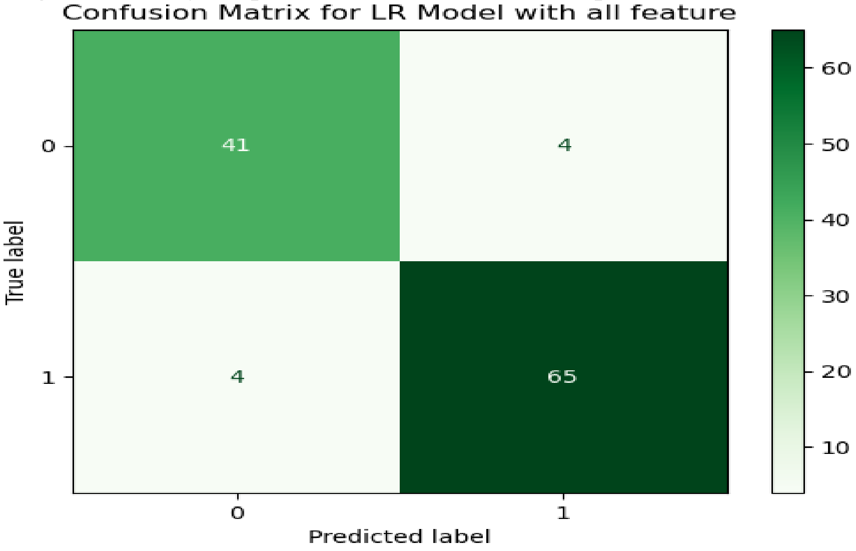


Fig 17. Confusion matrix for LR model with all features

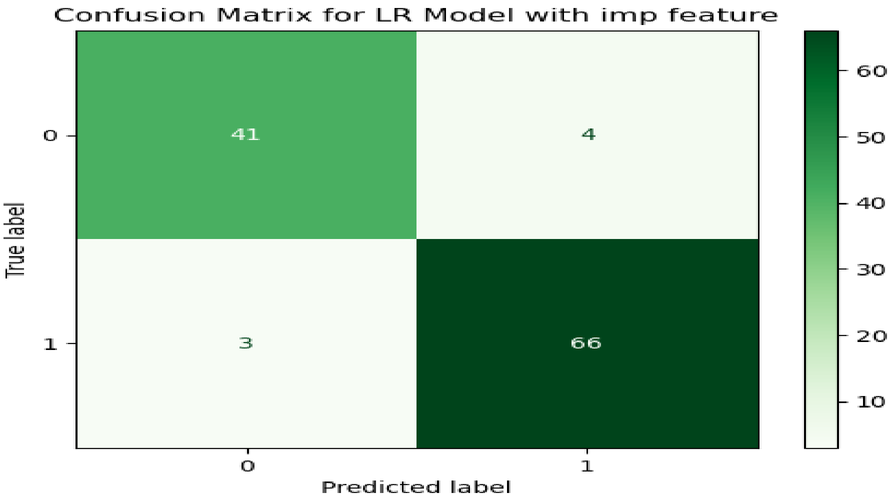


Fig 18. Confusion matrix for LR model with important features

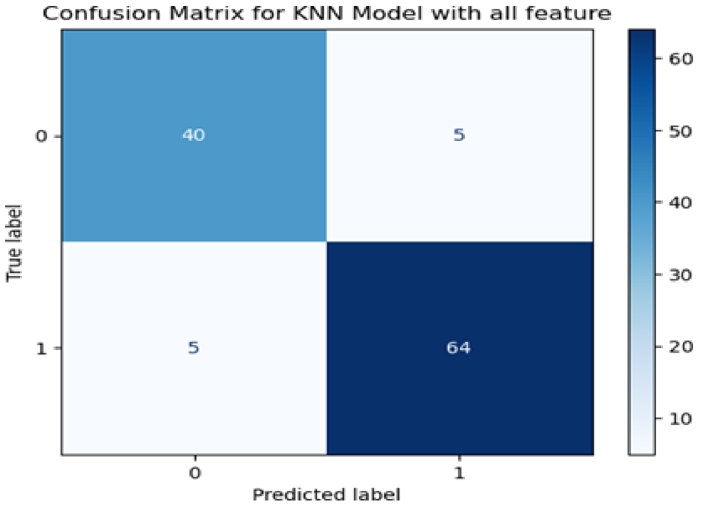


Fig 19. Confusion matrix for KNN model with all features

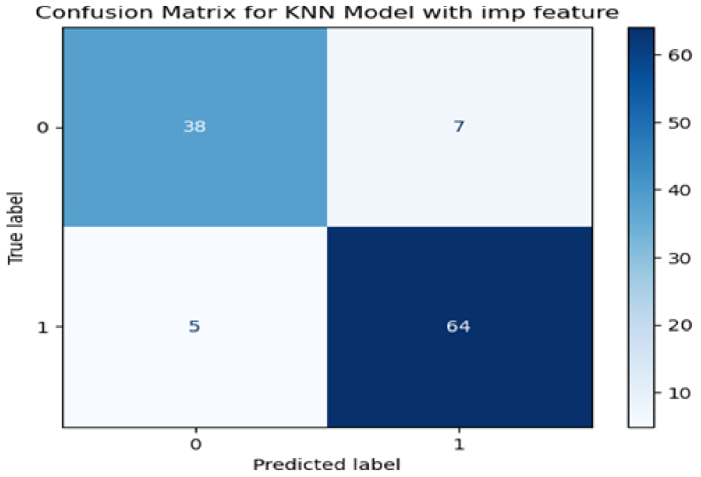


Fig 20. Confusion matrix for the KNN model with important features

1.1 Implementation of a user-friendly TUMOR-TRACK Interface

For ease of use, we have integrated a machine learning model trained in clinical data to assist users, including medical professionals and patients, in assessing the likelihood of breast cancer based on given input features. The tool offers a user-friendly interface for data input and instant prediction while maintaining high data security and patient privacy. The system integrates a frontend application designed for ease of use by healthcare professionals and patients, coupled with a robust backend powered by trained ML models.

Figure 21 shows the prediction page of the ML-based framework. After entering the values of essential features, the user gets the result page (Figure 22), where the user can clearly see the result whether a person has breast cancer or not. The result is an accessible and reliable tool that bridges advanced ML technologies with practical healthcare applications, empowering users to make informed decisions.

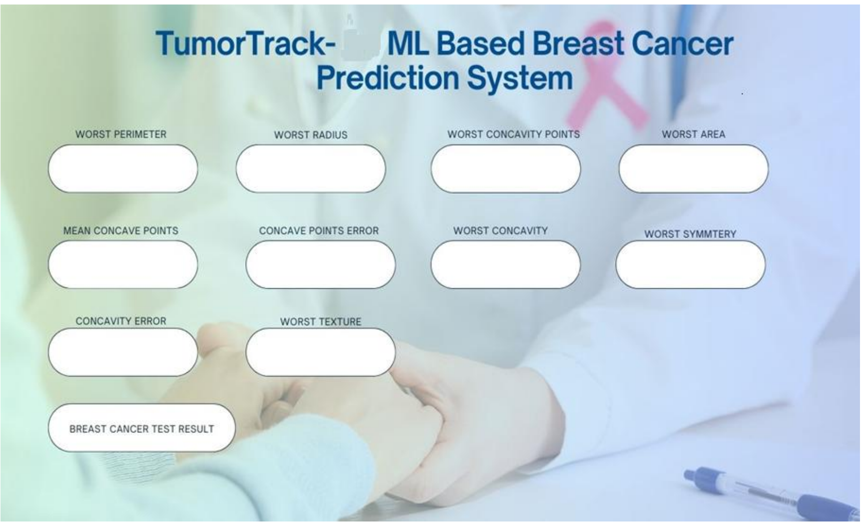
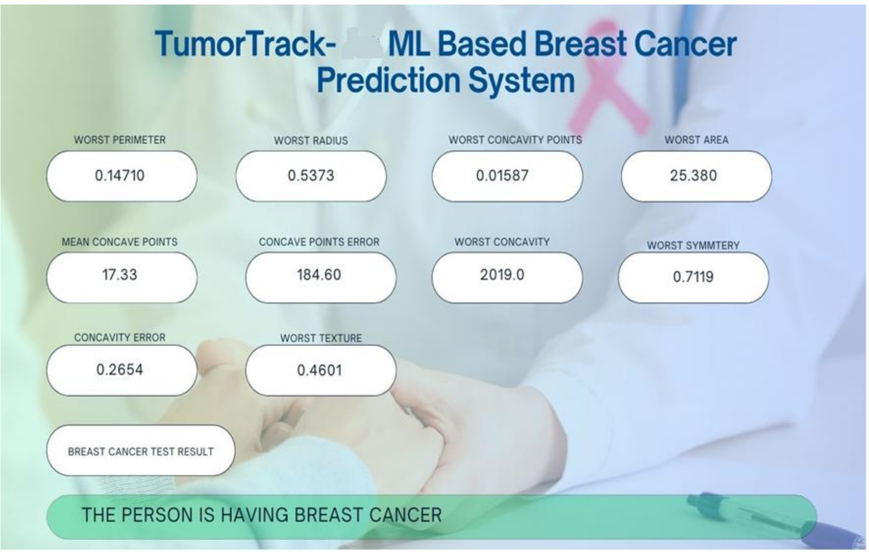


Fig 21. Prediction Page of the Tumor-Track


TumorTrack- ML Based Breast Cancer Prediction System
BREAST CANCER TEST RESULT

WORST PERIMETER: 0.14710
WORST RADIUS: 0.5373
WORST CONCAVITY POINTS: 0.01587
WORST AREA: 25.380
MEAN CONCAVE POINTS: 17.33
CONCAVE POINTS ERROR: 184.60
WORST CONCAVITY: 2019.0
WORST SYMMTERY: 0.7119
CONCAVITY ERROR: 0.2654
WORST TEXTURE: 0.4601

THE PERSON IS HAVING BREAST CANCER

Fig 22. Result Page of the Tumor-Track

2. Ethical Implications of Tumor-Track

Different critical questions appear repeatedly while analyzing the ethical challenges posed by machine learning-based breast cancer prediction with an academic research perspective. These issues encompass algorithmic bias, data privacy, security and privacy, and over-dependence on technology. The present review is based on the findings of several research papers to discuss some of these ethical considerations and possible approaches.

Algorithmic Bias and Fairness As the literature suggests, a key concern for policymakers is the risk of bias among machine learning algorithms. Obermeyer et al. [20] show that biases can occur when algorithms are trained on non-representative data. When the training data are unbalanced or the diversity of the population is not well taken into account, the es- timator may have poor performance for the underrepresented populations and it produces biased predictions. This bias can further marginalize pre-existing health inequities as observed by Adamson and Smith [21] possibly leading to inequitable healthcare access and treatment for minority communities. To overcome this, we raise the significance of diverse and representative data sets and frequent audits of algorithms to detect and correct biases.

5.1 Privacy and Security of Patient Data: Privacy and security of patient data were also ethical issues mentioned in the studies. Machine learning models rely on large datasets, which may include personal health data. The risk that data could be hacked or that the information could be misused is an important concern as noted by Kaushal et al. [22]. Patient privacy and data confidentiality must be guaranteed. Ways to mitigate these concerns are through strong data encryption, access controls, and clear data governance policies. Furthermore, explicit consent to use patient data for research is necessary to preserve trust and meet ethical obligations.

5.2 Over-reliance on machine learning predictions: Another ethical concern is the risk of over-reliance on machine learning predictions. As with other machine learning research in medical diagnostic images [23], Topol's research suggests that although machines can improve diagnostic accuracy, they should not substitute human judgment. If we are relying too much on risk scores rather than thinking about the history of an individual patient, it could be a missing point, or missed diagnosis, or a delayed diagnosis. This is best addressed by incorporating diagnostic tools as decision- support interfaces, rather than as independent diagnostic tools. In this way, machine learning is designed to enhance, rather than substitute for, the knowledge of healthcare workers.

5.3 Ethical Frameworks and Governance: The development and application of ethical frameworks and governance mechanisms are essential to addressing these challenges. Research by Floridi et al. [24] recommended ethical guidelines to ensure transparency, accountability, and inclusion in the development and deployment of machine learning systems. Frequent ethical audits and fusing multidisciplinary teams, composed of ethicists, clinicians, and data scientists, ensure that machine learning in breast cancer prediction is ethically sound and just for all patients.

In summary, machine learning could massively enhance the ability to predict breast cancer and thereby improve patient outcomes, however, consideration of the ethical issues involved is crucial. By addressing algorithmic bias, protecting patient data, avoiding dependence of “black box” technology, and integrating strong ethical frameworks, researchers and clinicians can leverage the power of machine learning while ensuring fairness, privacy, and trust in healthcare.

5. Conclusion

Breast cancer is still a major global health problem with millions of deaths per year and it is one of the leading causes of death in the world. Therapeutic decisions today are based on several established prognostic and predictive factors. The proposed model provides a useful tool for the doctor, makes disease prediction convenient, and could relieve clinicians' workload. As enabling earlier and more accurate diagnosis, the model also benefits patients by reducing delays and associated health care costs. According to the experimental result, the Logistic Regression (LR) model achieved higher accuracy than the K-Nearest Neighbors (KNN) model, hence it is the better choice for predicting breast cancer. This model improves accuracy and relevance by seeking subsets of the most significant features, because there are 30 features to check.

To further improve breast cancer prediction, future efforts should focus on integrating multi-omics data, refining feature selection techniques, and exploring ensemble learning methods. Incorporating longitudinal data analysis can bolster model robustness and clinical applicability. Additionally, ensuring patient engagement, addressing ethical considerations, and integrating decision support tools will be crucial for the responsible and effective implementation of predictive models in clinical practice.

Compliance with Ethical Standards

Conflict of interest - The authors have no conflicts of interest to declare relevant to this article's content.

Ethical approval - This article contains no studies with human participants or animals performed by the authors.

Declaration of Competing Interest :

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The authors have not received any funding/support from any organization.

REFRENCES

[1] https://www.who.int/news-room/fact-sheets/detail/breast-cancer
[2] https://www.nationalbreastcancer.org/about-breast-cancer/, 2019.
[3] Luca M, Kleinberg J, Mullainathan S. Algorithms need managers, too. Brighton: Chapman & Hall Ltd; 2016.
[4] Coiera E. Guide to medical informatics, the Internet and telemedicine. London: Chapman & Hall Ltd; 1997.
[5] Benbrahim, H., Hachimi, H., & Amine, A. (2020). Comparative study of machine learning algorithms using the breast cancer dataset. In Advanced Intelligent Systems for Sustainable Development (AI2SD'2019) Volume 2-Advanced Intelligent Systems for Sustainable Development Applied to Agriculture and Health (pp. 83-91). Springer International Publishing.
[6] Verma, D., & Mishra, N. (2017, September). Comparative analysis of breast cancer and hypothyroid dataset using data mining classification techniques. In 2017 IEEE international conference on power, control, signals and instrumentation engineering (ICPCSI) (pp. 1624-1626). IEEE.
[7] Amrane, M., Oukid, S., Gagaoua, I., & Ensari, T. (2018, April). Breast cancer classification using machine learning. In 2018 electric electronics, computer science, biomedical engineerings' meeting (EBBT) (pp. 1-4). IEEE.
[8] Aruna, S., Rajagopalan, S. P., & Nandakishore, L. V. (2011). Knowledge based analysis of various statistical tools in detecting breast cancer. Computer Science & Information Technology, 2(2011), 37-45.
[9] Chaurasia, D. V., & Pal, S. (2014). Data mining techniques: to predict and resolve breast cancer survivability. International Journal of Computer Science and Mobile Computing IJCSMC, 3(1), 10-22.
[10] Sakri, S. B., Rashid, N. B. A., & Zain, Z. M. (2018). Particle swarm optimization feature selection for breast cancer recurrence prediction. IEEE Access, 6, 29637-29647.
[11] Juneja, K., & Rana, C. (2020). An improved weighted decision tree approach for breast cancer prediction. International Journal of Information Technology, 12(3), 797-804.
[12] Yue, W., Wang, Z., Chen, H., Payne, A., & Liu, X. (2018). Machine learning with applications in breast cancer diagnosis and prognosis. Designs, 2(2), 13.
[13] S. Nayak and D. Gope, "Comparison of supervised learning algorithms for RF-based breast cancer detection," 2017 Computing and Electromagnetics International Workshop (CEM), Barcelona, 2017, pp.
[14] B.M. Gayathri and C. P. Sumathi, "Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer," 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai, 2016, pp. 1-5.
[15] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, 'Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis', Procedia Computer Science, vol. 83, pp. 1064–1069, 2016, doi: 10.1016/j.procs.2016.04.224.
[16] Y. khoudfi and M. Bahaj, Applying Best Machine Learning Algorithms for Breast Cancer Prediction and

Classification, 978-1-5386- 4225-2/18/$31.00 ©2018 IEEE.

[17] L. Latchoumi, T. P., & Parthiban, "Abnormality detection using weighed particle swarm optimization and smooth support vector machine," Biomed. Res., vol. 28, no. 11, pp. 4749–4751, 2017.

[18] A. H. Osman, "An Enhanced Breast Cancer Diagnosis Scheme based on Two-Step-SVM Technique," Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 4, pp. 158–165, 2017

[19] https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data.

[20] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

[21] Adamson, A. S., & Smith, A. (2018). Machine learning and health care disparities in dermatology. JAMA Dermatology, 154(11), 1247-1248.

[22] Kaushal, A., Altman, R., & Langlotz, C. P. (2020). Health care AI systems are biased: Here's how to make algorithms fairer. Harvard Business Review.

[23] Topol, E. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56.

[24] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28, 689-707.

Authors' Profiles

Dr. Shilpi Singh is an Assistant Professor in the Department of Computer Science and Engineering at Amity School of Engineering and Technology, Amity University Patna, India. She earned her Ph.D. in Information Technology from Babasaheb Bhimrao Ambedkar University, Lucknow, Uttar Pradesh, India. Her primary research interests lie in software engineering, artificial intelligence and image processing. She has published several research papers in reputed journals and conferences.

Dr. Govind Kumar Jha works as an Assistant Professor and Head of the Department of Computer Science and Engineering at Government Engineering College Munger (Bihar) . He received his M.Tech. and Ph.D from Dr. APJ Abdul Kalam Technical University, Lucknow, India. He has over 15 years of teaching and administrative experience with reputed universities/institutes. He has published various research papers in national & international journals and conferences. His research areas are Recommender Systems and Machine Learning. He is a Lifetime Member of the Computer Society of India.

Dr. Preetish Ranjan is an Assistant Professor in the Computer Science and Engg. Department at Amity University Patna. He received his Ph.D from IIIT Allahabad. His research area is the implementation of data mining in social network analysis, call data record analysis, recommender systems, and VAPT in network infrastructure. He has published several papers in Scopus and SCI-indexed journals.