IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Predictive Modelling For Cancer Patients Using Machine Learning Techniques

[1]Rohit Pathak [2] Dr. Komal Tahiliani [3] Prof.Nargish Gupta

[1]Research Scholar , Sagar Institute of Science & Technology, Bhopal,India

[2]Associate Professor, Sagar Institute of Science & Technology,Bhopal, India

[3]Assistant Professor, Sagar Institute of Science & Technology,Bhopal, India

Abstract:

Breast cancer develops when abnormal cells accumulate in the breasts. It was the most common cancer in women worldwide in 2020, with an estimated 2.3 million new cases being diagnosed. The aetiology of breast cancer is not yet fully understood; however, it is established that advancing age, familial history of breast cancer, genetic abnormalities, exposure to radiation, and hormonal factors are all recognised as potential risk factors for the development of this disease. The symptoms of breast cancer encompass the presence of a lump or mass in the breast, alterations in breast size or shape, skin dimpling or puckering, nipple discharge or inversion, and breast pain or tenderness. However, not all breast cancers manifest themselves in obvious ways; others are detectable only by mammography or other imaging procedures. Better patient outcomes and lower mortality rates can be achieved through early detection and precise diagnosis. Predicting breast cancer risk, recurrence, and survivability is an area where machine learning algorithms have made significant strides in recent years. This study focuses on utilising machine learning to create precise predictions regarding a range of outcomes related to breast cancer. To begin, a model is constructed to anticipate the probability of acquiring breast cancer before the onset of the disease. This is accomplished through the use of algorithms like Logistic Regression (LR), Decision Trees (DT), and Neural Networks (NN) to examine parameters including age, family history, hormone considerations, and lifestyle factors. After the model has been trained and tested on a sizable dataset of breast cancer patients and healthy individuals, a variety of metrics are used to evaluate the model's performance, including accuracy, sensitivity, specificity, and area under the receiver operating characteristic (AUC-ROC). Second, a model is constructed to foretell the likelihood of a return of breast cancer following initial clearance of the disease. This is accomplished through the use of algorithms like RF, gradient boosting, and deep learning to examine characteristics such tumour size, grade, receptor status, and treatment history. The model's performance is assessed in terms of a number of different outcomes, and it is trained and tested using data from breast cancer patients who have already had treatment and have been followed up on. Finally, a model is created to foretell how breast cancer patients will respond to treatment and whether they will survive. Support vector machines (SVM), Naive Bayes (NB), and K-nearest Neighbours(k-NN) are some of the algorithms used to analyse variables such patient demographics, tumour characteristics, and treatment history to reach this goal. Overall survival, disease-free survival, and progression-free survival are a few of the measures used to assess the model's performance once it has been trained and tested on a dataset of breast cancer patients with known outcomes. The overall goal of developing machine learning models for breast cancer prediction and survivorship is to enable earlier detection, personalised treatment planning, and improved patient outcomes, all of which have the potential to revolutionise breast cancer care.

Keywords: Logistic Regression,Naïve Bayes,Decision trees, Neural networks,Support Vector *Machine*.

INTRODUCTION:

Cancer is a disease that develops when a small number of cells in the body proliferate uncontrollably and metastasize to other parts of the body. Cancer can arise in virtually any of the body's millions of cells. Cell division (also known as cell expansion and multiplication) is a common way for human bodies to replenish their supply of cells. When cells die from natural causes or injury, they are replaced by new ones. By 2023, cancer will have claimed the lives of about 10 million individuals around the globe as shown in figure 1.1. The following were the most common types of newly diagnosed cancer in 2023:

- Breast (2.26 million cases)
- Lung (2.21 million cases)
- Colon and rectum (1.93 million cases)
- Prostate (1.41 million cases)
- Skin (non-melanoma) (1.20 million cases)
- Stomach (1.09 million cases)

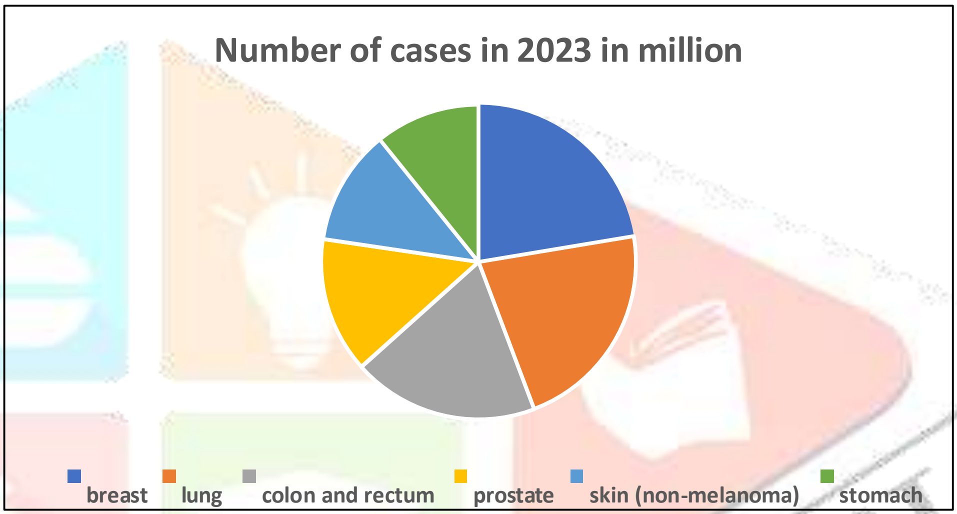


Figure 1: Number of Cases of Different Cancer Types in the Year 2023

Breast cancer is the most common form of cancer in women globally and accounts for a significant portion of cancer-related morbidity and mortality. Over the past few decades, numerous studies have been conducted to improve breast cancer prognosis, treatment, and diagnostics [1-2]. Millions of women around the world are affected by this serious health issue.

Types of Breast Cancer:

1.Non-invasive Breast Cancers

Non-invasive breast cancers are those that are confined to the milk ducts or breast lobules. They have not spread into or penetrated normal breast tissue. Non-invasive malignancies are called carcinomas in situ or pre-cancers.

2.Ductal Carcinoma

The most frequent kind of non-invasive breast cancer is ductal carcinoma in situ (DCIS). Because it has not progressed beyond the milk ducts where it began, this type of breast cancer is said to be non-invasive. Although DCIS itself poses no immediate threat to life, it does raise the likelihood that the patient will eventually develop invasive breast cancer.

3.Lobular Carcinoma

Breast cancer that remains contained within the lobules (the milk-producing glands at the end of the breast ducts) is called lobular carcinoma in situ (LCIS). Since it has not metastasized to other parts of the breast, it is considered to be non-invasive. Although LCIS itself poses no immediate threat to life, having it raises the probability of later developing more serious, invasive breast cancer.

RESEARCH MOTIVATION

Cancer is the primary cause of death in the world, with millions of new cases diagnosed each year and countless lives affected. Despite significant advancements in

cancer research, diagnosis, and treatment, the disease continues to pose immense challenges to patients, healthcare providers, and researchers alike. Early detection, accurate prognosis, and personalized treatment are essential for improving patient outcomes and overall quality of life [9]. The motivation behind this research work lies in addressing these challenges by developing innovative predictive models for cancer susceptibility, recurrence, and survivability, which could ultimately transform cancer care and management. This research is driven by the need to identify individuals at elevated risk of developing cancer prior to the onset of the disease. Early identification can facilitate timely preventive measures, such as lifestyle modifications or increased surveillance, and enable personalized interventions tailored to an individual's unique risk profile. This research aims to integrate multi-omics data, clinical information, and lifestyle factors to create a robust predictive model that could have a substantial impact on cancer prevention and early detection strategies. This work is also motivated by the fact that predicting cancer recurrence, stems from the critical concern of patients who have achieved apparent resolution of their disease but remain at risk for relapse [10]. Accurate predictions of cancer recurrence can significantly impact clinical decision-making and patient management, ultimately improving patient outcomes.

LITERATURE REVIEW

For this research work, we have studied more than hundred research papers. Clinical data is used to train a machine learning model to make a diagnosis of metastatic breast cancer (MBC). Using a variety of Python modules for text mining, data processing, and machine learning, the authors [13] created a non-invasive classification method for diagnosing cancer metastases. Using EMR and blood profile data, they implemented their approaches, with a decision tree classifier attaining 83% accuracy and an AUC of 0.87. The authors postulate that this strategy may aid in the identification of high-risk MBC patients, leading to better survival rates. Researchers [14] presents a literature review on the use of machine learning and deep learning techniques to the diagnosis of breast cancer. The authors accessed breast cancer data sets and mammography images from the state of Wisconsin. Finding the best breast cancer diagnostic model was the key focus of the study. The research [15] investigates the feasibility of using machine learning methods for breast cancer forecasting. On a breast cancer dataset, the authors employed multiple machine learning classification methods, including Naive Bayes, Logistic regression, Support Vector Machine, K-Nearest Neighbour, and Decision Tree, as well as ensemble methods, such as Random Forest, Adaboost, and XGBoost. The best accuracy (97%) was obtained using both the decision tree and the XGBoost classifier, with the

XGBoost classifier achieving the best AUC (0.999). Several machine learning techniques, such as C 5.0, Naive Bayes, logistic regression, random forest, ctree, KNN, K-Mean, GBM, adaBoost, and a decision tree model, are explored in [16] for their potential in early breast cancer detection. The Wisconsin and SEER datasets were used for training and testing these models, and their efficacy was measured using accuracy, precision, recall, and the F1 measure. Using information like tumour size, disease stage, and expected survival time, the researchers determined whether or not the tumour was still alive. Their efforts help educate the public about breast cancer and reduce anxiety about tumours. The Neighbour Component Analysis (NCA) method is used in conjunction with several machine learning techniques for breast cancer prediction [17], including logistic regression, decision tree, random forest, and K-nearest neighbour. A high level of accuracy, around 98.5%, was reached by an automated system employing the KNN with NCA method, showing its potential usefulness for early disease identification and subsequently enhancing patient survival chances. The authors in [18] examined the efficacy of machine learning techniques like Support Vector Machine, Decision Tree, Multi-layer Perceptron, and Naive Bayes in classifying breast cancer patient data into various classes using actual patient data from HealthCare Global Enterprises Ltd

(HCG) hospitals. Death, progression, recurrence, and metastasis were used as main class variables for model training and evaluation. The findings of this study highlight the promise of machine learning in predicting key parameters and facilitating early breast cancer diagnosis. This meta-analysis looked at studies published between 1997 and 2014 that used machine learning to make predictions about breast cancer recurrence. [19] shed attention on the difficulty of collecting adequate data on breast cancer recurrence and the ongoing debate over which factors are most predictive. Below mentioned table 1 presents summary of some most relevant papers.

Table 1: ***Summary of research papers with key findings***

Author(s)	Methods	Dataset Used	Performance Metrics	Findings
Botlagunta et al. (2023) [13]	Text mining, Data processing, Machine Learning, Decision Tree classifier, Flask	Electronic Medical Records (EMR)	Accuracy: 83%, AUC: 0.87, ROC: Not specified	MBC patients had considerably fewer monocytes than healthy controls. Removing outliers improved the ML model. Decision Tree classifier with 83% accuracy and 0.87 AUC. Blood profile-based ML algorithms could prioritise MBC critical care patients.
Kajala, Aditi, and V. K. Jain (2020) [14]	Machine Learning algorithms, Deep Learning algorithms, Hybrid Machine Learning approaches	WBCD, Mammography imaging datasets	NA	The study tested ML methods using Wisconsin breast cancer registries and mammography image datasets to find the best breast cancer diagnosis model.
Nemade & Fegade (2023) [15]	NAB, LR, SVM, KNN, DT, RF,Ada_Boost, XG_Boost	Breast cancer dataset	Accuracy: 97% (DT & XGBoost), AUC: 0.999 (XGBoost)	Decision Tree and XGBoost classifiers had the highest accuracy (97%) and AUC (0.999). ML helps breast cancer prediction.
Singh et al. (2023) [16]	C 5.0, NAB, LR,RF, ctree, KNN, K_Means, GBM, Ada_Boost, DT	WBCD, SEER datasets	Accuracy, Precision, Recall, F1 measure (Not specified)	Classifying breast cancer tumours using ML models and comparing Wisconsin and SEER datasets. To educate the public about breast cancer and ease tumour fears, tumour size and other factors were used to predict cancerousness.
Ravale & Bendale (2023) [17]	LR, DT, RF, K-NN with (NCA)	Not specified	Accuracy: 98.5%	KNN with NCA has the maximum accuracy of 98.5%, making it a reliable, effective, and fast breast

				cancer diagnosis algorithm.
Shastri et al. (2018) [18]	SVM, DT, Multi-layer Perceptron, NAB	HCG Hospital patient records	Sensitivity , Specificity , Accuracy	Machine learning methods classified patients' outcomes by mortality, progression, recurrence, and metastasis. The model predicts breast cancer early detection.
Abreu et al. (2016) [19]	Machine Learning Techniques (Not specified)	Local and open source databases (1997-2014)	NA	Researchers found no way to predict breast cancer recurrence. Combining machine learning approaches and creating breast cancer recurrence predictors may improve results.

RESEARCH METHODOLOGY

Machine learning (ML) enables high accuracy breast cancer prediction, enabling early detection and tailored treatment options. The Breast Cancer data set is used in these investigations. These datasets include a variety of clinical indicators, such as insulin, glucose, resistin, adiponectin, and leptin, as well as age and the obesity index. (MCP1). A variety of machine learning (ML) techniques were used to analyse the data and predict the occurrence of breast cancer, including LR, k-nearest neighbour, support vector machines, DT, RFs, Nave Bayes [57,58,59].

Machine learning algorithms have the potential to assist healthcare professionals in offering patients improved treatment options and making better- informed decisions. Both investigations evaluated the performance of several ML algorithms, and the results revealed that these algorithms are capable of making precise predictions about breast cancer. Two research studies were undertaken to assess and authenticate the efficacy of various machine learning models in forecasting the onset of breast cancer in women. The Breast Cancer dataset was utilised in the initial study to compare the performance of SVC, ETC, KNN, LR, and RF among other classification methods.

The Breast Cancer dataset has several characteristics, including texture, radius, perimeter, area, fractal dimension, compactness, concavity, symmetry, and smoothness. These characteristics were applied to compartmentalise and classify the many subtypes of breast cancer. Upon evaluating the precision, recall, and accuracy measures of various machine learning algorithms, it was determined that SVM exhibited the highest efficacy in predicting breast cancer. The study demonstrated how ML techniques can significantly reduce inaccuracy and save time by effectively categorising and compartmentalising different measurements of data.
In this research work we have used Wisconsin Breast cancer data set .Table 2 describes the Wisconsin breast cancer dataset in detail.

Table 2: Description of WBCD Datasets

Feature	**Wisconsin Breast Cancer Diagnostic Dataset (WBCD)**
Number of instances	569
Number of attributes/features	32 (30 input features, 1 patient ID, 1 diagnosis label)
Patient demographics	-
Laboratory measurements	-
Cell nucleus features	Texture, Radius, Perimeter, Area, Compactness, Concavity, Concave points, Smoothness, Symmetry, Fractal dimension
Classification labels	M: Malignant, B: Benign
Data collection methods	Digitized FNA images, Cytomorphology Lab, University of Wisconsin-Madison
Data source and accessibility	UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets /Breast+Cancer+Wisconsin+ (Diagnostic)

MACHINE LEARNING METHODS FOR BREAST CANCER DISCERNMENT ASSESSMENMT

Logistic Regression (LR)

Logistic regression is a supervised learning method that has been adopted by machine learning from the statistical domain [65,66]. As a binary classification algorithm, Logistic Regression (LR) assigns each data point to one of several possible groups. In LR, in contrast to Linear Regression, the outcome or target variable is a scalar.

LR is commonly used in breast cancer diagnosis to predict tumor type (benign or malignant). It calculates the probability of a tumor being malignant, providing a linear decision boundary in the feature space. Unlike linear regression, which predicts continuous values, logistic regression is a binary classification algorithm that estimates the probability of an event occurring (in this case, the presence or absence of breast cancer). It is an easily interpretable model that can be evaluated using metrics such as precision, recall, accuracy, and f1-score. It‘s performance can be improved by using regularization, feature selection, and engineering techniques.

Linear Support Vector Machine

The Support Vector Machine (SVM) is a supervised machine learning methodology that is widely employed in the realm of pattern recognition and classification quandaries, particularly when the dataset comprises precisely two classes. Support Vector Machines (SVMs) are employed to determine the optimal hyperplane that can effectively separate the different classes [68]. The classifier utilizes an input pattern, referred to as a feature vector, to ascertain its classification. The algorithm is capable of effectively categorizing data that can be separated linearly, however, it may encounter difficulty when presented with feature vectors that are not linearly separable. The utilization of the kernel trick has been employed to address this issue, as noted in reference. Support Vector Machines (SVM) employ kernel techniques to transform input data into higher dimensional space and offer a rapid training algorithm.

This technique is utilized for the purpose of pattern classification and regression analysis. The efficacy of a support vector machine (SVM) classifier is contingent upon the selection of an appropriate kernel function. Distinct kernel functions are employed for diverse classification tasks. The implementation of Support Vector Machines (SVM) was carried out in this project through utilization of the SVC class available in the sci-kit-learn library. Support Vector Machines (SVM) can pose challenges in terms of memory usage and may require intricate interpretation and tuning.

PERFORMANCE PARAMETERS OF MACHINE LEARNING ALGORITHM

Accuracy: In the context of classification problems, accuracy of a model is measured as the percentage of right predictions relative to the total number of predictions across all classes. When the classes of the target variable are roughly distributed throughout the dataset, then that is a suitable measure of accuracy. There are two types of cases in this data collection; benign and malignant cases account for 60% and 40% of the total, respectively.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

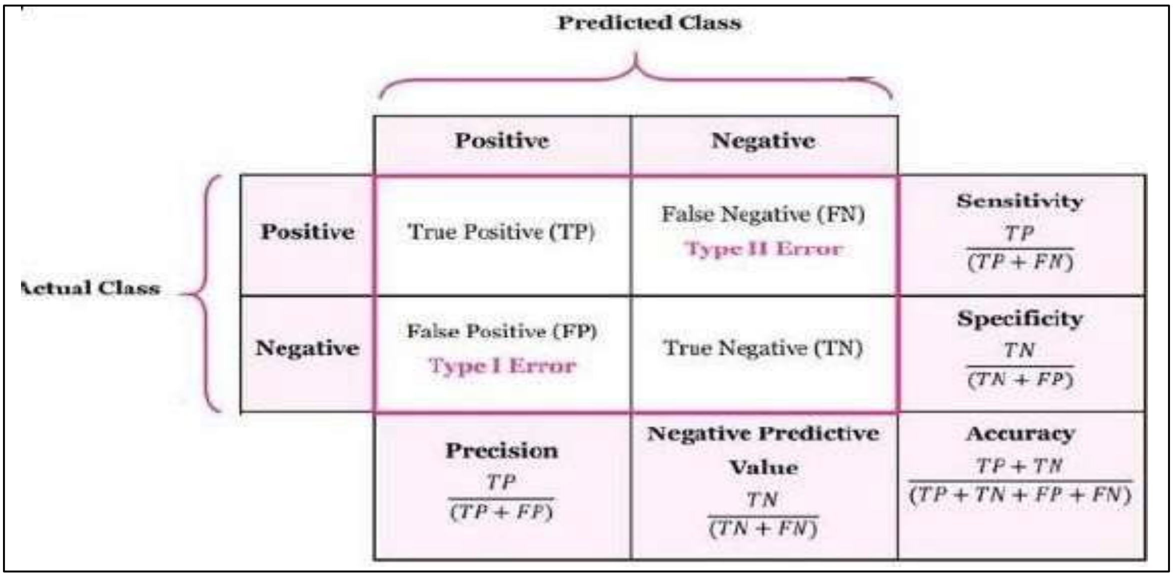
Sensitivity : Sensitivity is used to measure how many patients with malignant tumours were correctly diagnosed by the algorithm, as a percentage of all patients with tumours. People with malignant tumours are categorised as either true positives (TP) or false negatives (FN) in the current study, with TP referring to those people for whom the model correctly predicted the presence of malignant tumours. Therefore, it is crucial to enhance sensitivity to approach 100% in order to prioritise the decrease of False Negatives.

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

Precision: Precision in a binary classification task is defined as the fraction of correct classifications divided by the sum of correct and incorrect classifications.

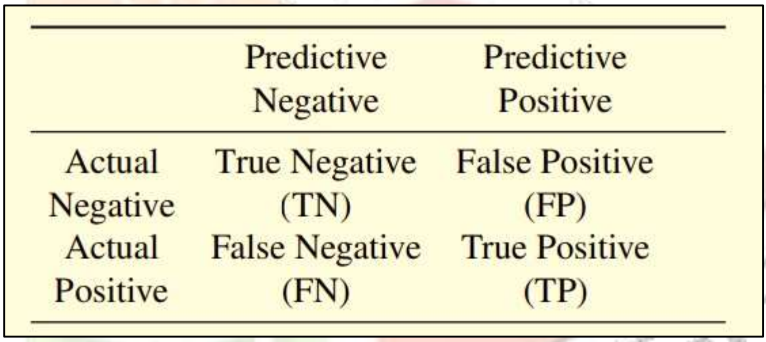
Precision is a statistical metric that quantifies the ratio of true positive cases,

i.e. patients who have been correctly diagnosed with a malignant tumour, to the total number of patients diagnosed with a malignant tumour. The individuals who are forecasted to have a malignant tumour, including both true positives (TP) and false positives (FP), are compared to the actual population of individuals with malignant tumours, also consisting of true positives (TP).



		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

Confusion Matrix: The Confusion matrix is a highly intuitive and straightforward metric utilised for determining the accuracy and correctness of a given model. The present study employs logistic regression for the purpose of addressing classification problems characterised by multiple classes of output, rendering it an ideal analytical technique for the current investigation. The utilisation of a table layout, also known as a matrix layout, facilitates the visualisation of algorithmic performance



	Predictive Negative	Predictive Positive
Actual Negative	True Negative (TN)	False Positive (FP)
Actual Positive	False Negative (FN)	True Positive (TP)

COMPARISON OF RESULTS

Results of classification using support vector classifier on different kernels(Linear,Poly,Rbf) over the Wisconsin Breast Cancer Diagnostic dataset were assessed. The test results and predictions represented as a confusion matrix for SVM classifier, the SVM on linear model exhibit the highest levels of precision, specificity, and accuracy. Findings demonstrate that SVM with Feature Selection (SelectKBest) correctly classify tumors as benign or malignant.

Algorithms Comparison Chart:

SVM	linear	poly	rbf
Model Training with PCA	Accuracy Score: 0.9707602339181286 F1 Score: 0.9612403100775193 Recall Score: 0.9841269841269841 Precision Score: 0.9393939393939394	Accuracy Score: 0.9122807017543859 F1 Score: 0.8648648648648648 Recall Score: 0.7619047619047619 Precision Score: 1.0	Accuracy Score: 0.9707602339181286 F1 Score: 0.9606299212598425 Recall Score: 0.9682539682539683 Precision Score: 0.953125

Model Training with Feature Selection (SelectKBest)	Accuracy Score: 0.9824561403508771 F1 Score: 0.976 Recall Score: 0.9682539682539683 Precision Score: 0.9838709677419355	Accuracy Score: 0.8366666666666667 F1 Score: 0.8361204013377926 Recall Score: 0.8116883116883117 Precision Score: 0.8620689655172413	Accuracy Score: 0.9298245614035088 F1 Score: 0.8947368421052632 Recall Score: 0.8095238095238095 Precision Score: 1.0

CONCLUSION:

In conclusion, the study compared SVM on various models(Linear,poly and rbf), for breast cancer classification using the Wisconsin Breast Cancer Dataset . For the Wisconsin breast cancer dataset (WBCD), with SVM showing the most promising results. The dataset is trained separately using both PCA and SelectKBest and model training with feature selection –selectKBest gives better accuracy.The analysis of heat map helped clarify connections between variables. The most crucial features identified were glucose, age, resisting, BMI, and insulin, which highlighted the significance of obesity/metabolic dysregulation in breast cancer prediction.

References:

1. D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: The next generation," Cell, vol. 144, no. 5, pp. 646–674, 2020.

2. S. Katuwal, P. Jousilahti, and E. Pukkala, "Causes of death among women with breast cancer: A follow-up study of 50,481 women with breast cancer in Finland," International Journal of Cancer, vol. 149, no. 4, pp. 839–845, 2021.

3. Tawam Hospital | Medical News. (N.D.). Retrieved November 19, 2014, from http://www.tawamhospital.ae/english/news/print.aspx?Newsid=367

4. S. F. Khorshid and A. M. Abdulazeez, "Breast cancer diagnosis based on k- nearest neighbors: a review," Palarch's Journal of Archaeology of Egypt/Egyptology, vol. 18, no. 4, pp. 1927–1951, 2021.

5. M. Karabatak, ―A new classifier for breast cancer detection based on naive bayesian,‖ measurement, vol. 72, pp. 32– 36, 2022

6. M. Kumar, S. K. Khatri, and m. Mohammadian, ―breast cancer identification and prognosis with machine learning techniques-an elucidative review,‖ j. Interdiscip. Math., vol. 23, no. 2, pp. 503–521, 2020,

7. R. Kohavi, "Glossary of terms," Special Issue on Applications of Machine Learning and the Knowledge Discovery Process, vol. 30, no. 271, pp. 127– 132, 2021.

8. A. L. Samuel, "Some studies in machine learning using the game of checkers," IBM Journal of Research and Development, vol. 3, no. 3, pp. 210–229, 1959.

9. O. L. Mangasarian, W. N. Street, and W. H. Wolberg, "Breast cancer diagnosis and prognosis via linear programming," Operations Research, vol. 43, no. 4, pp. 570–577, 2000.

10. S. Belciug, A.-B. Salem, F. Gorunescu, and M. Gorunescu, "Clustering-based approach for detecting breast cancer recurrence," in 2010 10th International Conference on Intelligent Systems Design and Applications, pp. 533–538, Nov. 2019.

11. D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: A comparison of three data mining methods," Artificial Intelligence in Medicine, vol. 34, no. 2, pp. 113–127, Jun. 2022.

12. World Cancer Research Fund (WCRF). (N.D.). Breast cancer statistics. Retrieved from https://www.wcrf.org/cancer-trends/breast-cancer-statistics/

13. M. Botlagunta, M. D. Botlagunta, M. B. Myneni, D. Lakshmi, A. Nayyar, J. S. Gullapalli, & M. A. Shah, "Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms," Scientific Reports, 13(1), 485, 2023.

14. A. Kajala and V. K. Jain, "Diagnosis of breast cancer using machine learning algorithms - a review," in 2020 International Conference on Emerging Trends in Communication, Control and Computing (ICoNc3), pp. 1-5, 2020, IEEE.

15. V. Nemade and V. Fegade, "Machine learning techniques for breast cancer prediction," Procedia Computer Science, vol. 218, pp. 1314-1320, 2023.

16. D. Singh, R. Nigam, R. Mittal, and M. Nunia, "Information retrieval using machine learning from breast cancer diagnosis," Multimedia Tools and Applications, 82(6), 8581-8602, 2023.

17. U. Ravale and Y. Bendale, "Breast cancer prediction using different machine learning algorithms," in Sentiment Analysis and Deep Learning: Proceedings of ICSADL 2022, pp. 493-502, Singapore: Springer Nature Singapore, 2023.

18. S. S. Shastri, P. C. Nair, D. Gupta, R. C. Nayar, R. Rao, and A. Ram, "Breast cancer diagnosis and prognosis using machine learning techniques," in Intelligent Systems Technologies and Applications, pp. 327-344, Springer International Publishing, 2021.