Data and Metadata. 2024; 3:.346
doi: 10.56294/dm2024.346

ORIGINAL

Prediction and Diagnosis of Breast Cancer using Machine Learning Techniques

Predicción y diagnóstico del cáncer de mama mediante técnicas de aprendizaje automático

Gufran Ahmad Ansari[1], Salliah Shafi Bhat[2], Mohd Dilshad Ansari[3], Sultan Ahmad[4,5], Hikmat A. M. Abdeljaber[6,7]

[1]School of Computer Science, Dr. Vishwanath Karad MIT World Peace University. Pune 411038, India
[2]B. S. Abdur Rahman Crescent Institute of Science and Technology Chennai-48. India.
[3]Department of Computer Science and Engineering, SRM University. Delhi-NCR, India.
[4]Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University. Alkharj 11942, Saudi Arabia.
[5]University Center for Research and Development (UCRD), Department of Computer Science and Engineering, Chandigarh University. Gharuan, Mohali 140413, Punjab, India.
[6]Department of Computer Science, Faculty of Information Technology, Applied Science Private University. Amman, Jordan.
[7]MEU Research Unit, Middle East University. Amman 11831, Jordan.

Cite as: Ansari GA, Shafi Bhat S, Dilshad Ansari M, Ahmad S, A. M. Abdeljaber H. Prediction and Diagnosis of Breast Cancer using Machine Learning Techniques. Data and Metadata. 2024; 3:.346. https://doi.org/10.56294/dm2024.346

Submitted: 13-02-2024 **Revised:** 23-05-2024 **Accepted:** 04-10-2024 **Published:** 05-10-2024

Editor: Adrian Alejandro Vitón Castillo

Corresponding Author: Sultan Ahmad

ABSTRACT

Introduction: one of the most common types of cancer and a significant contributor to the high death rates among women is breast cancer. It usually occurs in women. It is crucial to acquire a diagnosis early in order to kill cancer from becoming worse.
Objective: the traditional diagnosing procedure takes more time. A fast and useful option can apply Machine Learning Technique (MLT) to identify illnesses. However new technology creates a variety of high-dimensional data kinds particularly when it comes to health or cancer data.
Method: data classification techniques like Machine Learning are efficient. Particularly in the medical field where such techniques are often utilised to make decisions via diagnosis and analysis. Using Wisconsin Breast Cancer Dataset, the proposed research was carried out (WBCD). Some of these issues may be solved using the feature selection approach.
Results: this research analyses the classification accuracy of different MLT: Logistic Regression, Support Vector Machine, and K-Nearest Neighbour. According to experiment results, SVM has the best accuracy of all algorithms, at 97,12 %.
Conclusion: the mentioned prediction models are based on several supervised MLT. Tenfold cross validation is applied. Additionally, author also proposed a Flow chart of breast Cancer using MLT.

Keywords: Breast Cancer; Machine Learning Algorithms; Logistic Regression; Support Vector Machine; K-Nearest Neighbour; Diagnosis.

RESUMEN

Introducción: uno de los tipos de cáncer más frecuentes y que contribuye de forma significativa a las elevadas tasas de mortalidad entre las mujeres es el cáncer de mama. Suele aparecer en mujeres. Es crucial adquirir un diagnóstico precoz para evitar que el cáncer se agrave.
Objetivo: el procedimiento tradicional de diagnóstico requiere más tiempo. Una opción rápida y útil es aplicar la técnica de aprendizaje automático (MLT) para identificar enfermedades. Sin embargo, las nuevas tecnologías crean una gran variedad de tipos de datos de alta dimensión, especialmente cuando se trata de datos sobre salud o cáncer.

Método: las técnicas de clasificación de datos como Machine Learning son eficientes. Especialmente en el campo de la medicina, donde estas técnicas se utilizan a menudo para tomar decisiones a través del diagnóstico y el análisis. La investigación propuesta se llevó a cabo utilizando el conjunto de datos de cáncer de mama de Wisconsin (WBCD). Algunos de estos problemas pueden resolverse utilizando el enfoque de selección de características.
Resultados: esta investigación analiza la precisión de clasificación de diferentes MLT: Regresión Logística, Máquina de Vectores de Soporte y K-Nearest Neighbour. Según los resultados del experimento, SVM tiene la mejor precisión de todos los algoritmos, con un 97,12 %.
Conclusiones: los modelos de predicción mencionados se basan en varios MLT supervisados. Se aplica una validación cruzada de diez veces. Además, el autor también propuso un diagrama de flujo del cáncer de mama utilizando MLT.

Palabras clave: Cáncer de Mama; Algoritmos de Aprendizaje Automático; Regresión Logística; Máquina de Vectores Soporte; K-Nearest Neighbour; Diagnóstico.

INTRODUCTION

According to estimates cancer is the sixth most common cause of death in the world. A prevalent kind of cancer among women is breast cancer. There are numerous instances because of factors like hereditary, lifestyles and dietary behaviour.[1] About 27 % of all cancers in women in India are breast cancers which are widespread. Women's cancer cases in India are the third highest worldwide.[2] According to statistics breast cancer will affect 1 in 28 women throughout the course of their lives. obesity, menopause hormone replacement medication, breast cancer in the family, inactivity and long-term electromagnetic radiation exposure getting their first period at an early age, not having children at all, having children later in life, and other variables[3] are among those that contribute to breast cancer in women. Every day 2000 new cases of cancer in women are discovered 1 200 of which are found in late stages. In comparison to early-stage diagnosis, late detection[4] lowers the survival rate by 3 to 17 times and increases expenses by 1,5 to 2 times. Breast cancer has a death rate that is 1,6-1,7 times greater. India has the highest worldwide breast cancer death rate in 2021. Breast cancer early detection may significantly improve Patients life expectancies. A breast cancer diagnosis may benefit greatly from the use of MLT. The majority of these cancers are identified when the illness is still limited due to the most recent, most effective, and most advanced diagnostic methods. Machine learning methods are becoming ever more useful in healthcare analysis. The analysis of the patient clinical data and the doctor expertise are undoubtedly the most important aspects of diagnosis. The majority of potential medical faults may be prevented by adopting categorization. For example the use of MLT in medical research is growing quickly due to its great performance in predicting outcomes, lowering medical expenses, increasing patient health, enhancing the value and quality of healthcare, and providing the main decision to save lives.[5,6] There are several classification and outcome prediction algorithms for breast cancer. Diabetes mellitus, heart disease, and hypertension are a few examples of health-related factors that MLT has been successful in identifying and predicting.[7] Various data with high dimensions are produced by technical advances, especially medical data on breast cancer. There are nine characteristics according to the WBCD. Data quality issues including biased-unrepresentative data, noise, and missing or duplicate data may be impacted by high dimensional data. Additionally, it could be difficult to obtain data from them. Pre-processing data such as feature selection, is required to get around issue. More features may be produced by the use of feature selection that will lower the model training process processing costs. A small feature subset in the field of medical diagnosis results in lower test and diagnostic costs, hence the usage of few features is essential.[8]

Breast cancer is diagnosed using imaging methods such as radiology, ultrasound, and MRI.[9] All of these strategies describe the breast cell depicted in the images using various characteristics, such as cell size, cell texture etc. These results allow us to forecast whether a person will develop breast cancer or not.

The main motivation of this research is to detect breast cancer in its early stages this research's goal was to compare and contrast several machine learning algorithms.

KNN, SVM and LR technique is utilised in this work to predict breast cancer. The framework model for machine learning is created to predict breast cancer in women. The dataset is used to display the important variables involved in predicting the risk of breast cancer. Finally, an evaluation and a visualisation of the model's performance are created. It calculates, summarizes and displays the model's performance. This research is unique and better than other studies in the same field since it uses the KNN, SVM and LR. Better results are produced by the SVM algorithm (97,12 %) by utilizing output of the previous various stages. This Research Paper is divided into five segmentations one introduction the concept of breast cancer discusses how it affects women throughout the world and establishes the foundation for the topic of this research. The prior study presented in Section Two includes a Literature review of manuscripts and their classification and categorization in order to shortlist the papers that are most closely related to and relevant to the topic at mind. The processes for choosing the dataset and the parameters for computing the findings are presented in Section three of the study Proposed Flow chart of Breast Cancer in Women. Section Four presents Experimental Result and Discussion and finally, Section five includes conclusion and Future work in order to wrap up research.

Literature Review

This research examines a dataset from the UCI Machine Learning Repository called Wisconsin Breast Cancer (WBC).[10] The majority of the breast cancer research that have been published in the literature on medical data analysis have high classification accuracy. Suggested LS-SVM classification system for diagnosing breast cancer.[11] They were able to get a classification accuracy of 89,53 % using 10-fold cross validation. By using the support vector classification method and the best predictive factors, a new approach for detecting breast cancer was able to obtain a classification accuracy of 87,02 % without the usage of cross-validation. The breast cancer recurrence dataset was processed to feature selection.[11] The classifier in use is a rapid decision tree learner, Naive Bayes, and KNN. According to all tested machine learning models, using particle swarm optimization could enhance the efficacy of classification.[12] Many breast cancer classification algorithms were developed, and the accuracy of many of them was tested using data from the Wisconsin breast cancer database. The optimal learning vector approach, for example, achieved 84,7 % performance, while the big LVQ method scored 86,13 % accuracy, the highest in the literature. Utilised WBCD Dataset to assess how well supervised and unstructured breast cancer classification algorithms performed. KNN be used in cancer classification to evaluate false positive rates effectiveness. In general, naive Bayesian classifiers are used to forecast biological, chemical, and physiological features. In order to generate prognostics or classification models for cancer, NBC is occasionally combined with other classifiers, such as decision trees. To identify the features, the data were divided into training and testing sections at a ratio of 70:30 using scaling and principal component analysis. They proposed that the ensemble voting technique is appropriate as a breast cancer prediction model. Several models were evaluated and trained using feature selection techniques. Researchers have found that only four models ensemble voting classifier, logistic regression, support vector machine, and AdaBoost had an average accuracy of around 90 % among all those used for the prediction.[13] According to their analysis the suggested model performed well in terms of accuracy, recall tests, ROC-AUC (area under the receiver operating characteristic curve-receiver operating characteristic curve), F1-measure and amount of computation.[14] Introduced a hybrid method in their research for diagnosing hepatitis illness using a raw set and an extreme machine learning algorithm. The dataset for hepatitis came from the UCI repository. Rough set theory was used to create 20 reducing with three to seven attributes. The majority of published research has examined the accuracy of classifiers, i.e., in comparison to false positives and false negatives the percentage of true positives (TPs) and true negatives (TNs) (FNs). So far it is equally important to assess performance in terms of false positives (precision), false negatives (recall) and F-measures. Score as failing to recognise a condition could have negative patient outcomes[15] developed a CAD based on the morphological features of breast ultrasonography that use ML. The method extracts features from the ROI using a hybrid model. Instead of one feature, there are seven moments, FD, and HOG.A total of 250 ultrasounds have been used, with 100 benign and 150 malignant tumours. The ANN, which is used to categories ultrasound images, had an accuracy rate of 87,1 % for malignant lesions and 82,4 % for benign lesions. SVM, KNN, DT, RF and ADA Boost are contrasted with the Turgut Machine Learning technique. The random forest method was found to be the most effective (89 %)[16] Used four different breast cancer datasets to compare SVM and ANN. The researchers demonstrated that SVM outperformed ANN in terms of performance and outcome. Researched on extracting features such as variance, range, and compactness. They analysed the performance with SVM classification. Their results revealed the highest variance (95 %), as well as the highest compactness (86 %). To determine best classifier in breast cancer datasets assess performance criteria of supervised learning classifiers including Naïve Bayes, SVM, RBF kernel, RBF neural networks, Decision trees (J48) and basic CART. Outcome of experiment demonstrates that the SVM-RBF kernel outperforms other classifiers, scoring an accuracy of 90,84 % in the Wisconsin Breast Cancer (original) datasets[17] examine how an ensemble of machine learning approaches can predict a breast cancer patient's survival time. developed a model with better performance than traditional machine learning techniques, which usually need the practitioner to possess domain expertise of input data in order to select best latent representation. SVM may be considered as a useful method for predicting breast cancer based on its outcomes. Patients with breast cancer can be identified and have their survival rates improved using a variety of diagnostic (X-ray, Bone, CT, MRI, PET scans) and laboratory testing but metastatic spread of the disease which is a major factor in poor survival is not. An early breast cancer screening strategy based on machine learning was suggested as an outcome of this research.

Proposed Flow Chart of Breast Cancer in Women

Classification model with increased accuracy is proposed for predict breast cancer Consider the problem statement from the introduction section. The methodology for this research is shown in figure 1. Collecting of the Breast Cancer Dataset is the first step. The normalising procedure was then used for each dataset. To assess the effectiveness of feature selection for classification we implement 10 cross-validations. Model for a supervised machine learning-based flowchart for the detection of breast cancer. Supervised MLT a machine trains using labelled data, such as an input where the desired output is known and then performs classification using model data. Following the analysis of the training data the algorithm develops a learning algorithm that, given sufficient practise, can accurately anticipate the outcome for the current data. The information in our research is categorised as either positive or adverse. LR (Logistic Regression), SVM (Support Vector Machine) and KNN are a few of the methods used in supervised machine learning to forecast the results (K-Nearest Neighbor). The framework consists of the following significant phases:

- Step 1 Data Set.
- Step 2 Data Pre-processing.
- Step 3 Feature Selection.
- Step 4 Machine Learning Techniques.
- Step 5 Measure Accuracy.

Components are described in full below along with the steps performed against each component.

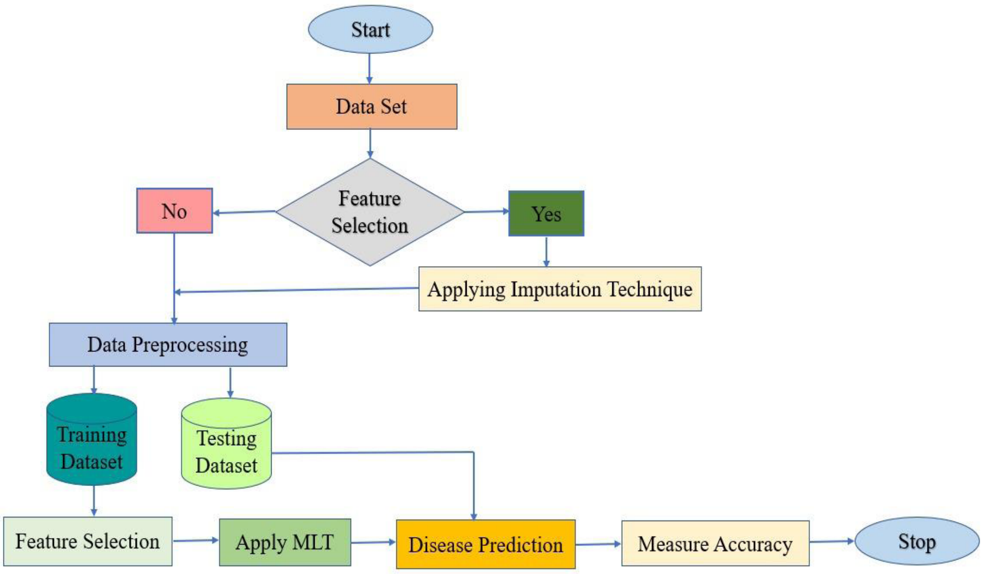


Figure 1. The Proposed Flow chart of breast cancer in women

Step 1 Data Set: the dataset which is a standard dataset was taken from the UCI repository. There are 520 attributes and 11 instances in the WBCD. Table 1 below displays all nine characteristics in their entire.

Table 1. Data Set Attributes

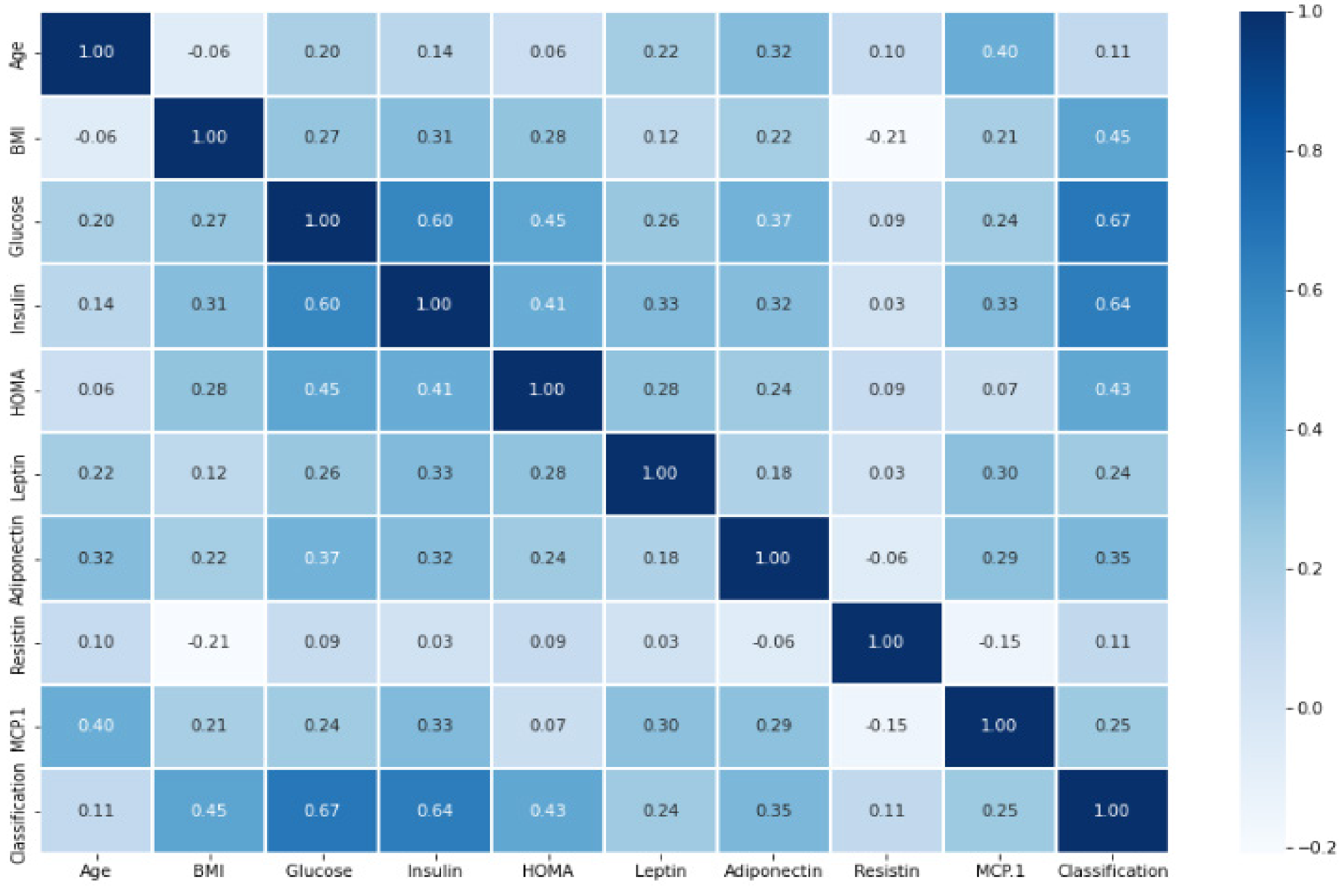
Sl.No	Attributes	Count	Mean	Std	Min	25 %	50 %	75 %	Max
1	Age	520,000	48,028	12,151	16,000	39,000	47,500	57,000	90,000
2	BMI	520,000	00,369	0,4830	0,0000	0,0000	0,0000	1,0000	1,0000
3	Glucose	520,000	00,496	0,5004	0,0000	0,0000	0,0000	1,0000	1,0000
4	Insulin	520,000	00,448	0,4977	0,0000	0,0000	0,0000	1,0000	1,0000
5	HOMA	520,000	00,417	0,4935	0,0000	0,0000	0,0000	1,0000	1,0000
6	Leptin	520,000	00,586	0,4929	0,0000	0,0000	1,0000	1,0000	1,0000
7	Adiponectin	520,000	00,455	0,4985	0,0000	0,0000	0,0000	1,0000	1,0000
8	Resistin	520,000	00,223	0,4167	0,0000	0,0000	0,0000	0,0000	1,0000
9	MCP.1	520,000	00,448	0,4977	0,0000	0,0000	0,0000	1,0000	1,0000
10	Classification	320,000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

Step 2 Data Pre-processing: pre-processing is the method of transforming the raw data into the format needed for the analysis of the result. Here, the preparation for imputation and data cleaning have been done. There will be a number of problems without pre-processing, including irregularities, errors, Noise, missing data, over-fitting of the model, etc. To evaluate the influence of these processes the diagnosis of breast cancer was assessed both with and without pre-processing using the results of the classification algorithms.[18] we used two feature selection techniques for pre-processing, and we choose the one that performed the best. Data cleaning involves taking unneeded information out of the collected raw data or formatting it. Since the data we gathered cannot be stored in a CSV file, we cleaned the data by deleting the symbols that the CSV format does not permit and converted the dataset to a CSV file.

Utilizing Correlation Analysis to Reduce Dimensionality

A method for removing characteristics that are insignificant for outcome prediction is dimensional reduction. In this research dimensionality reduction was carried out by looking at the relationships between the characteristics of the input data and eliminating those with a large variance. Heat map was used to examine relationships between the dataset characteristics, as shown in figure 2. The Age, HOMA which all provide information regarding size of breast cancer cells, were shown to be highly correlated. Only the "MCP-1" characteristic was chosen as a result to better depict the data about breast cancer cell size.

However, there is a relationship between BMI, Glucose and MCP1. The best result for the SVM technique was obtained in the subsequent phase with a 97,12 % accuracy due to selecting algorithms.



	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Classification
Age	1.00	-0.06	0.20	0.14	0.06	0.22	0.32	0.10	0.40	0.11
BMI	-0.06	1.00	0.27	0.31	0.28	0.12	0.22	-0.21	0.21	0.45
Glucose	0.20	0.27	1.00	0.60	0.45	0.26	0.37	0.09	0.24	0.67
Insulin	0.14	0.31	0.60	1.00	0.41	0.33	0.32	0.03	0.33	0.64
HOMA	0.06	0.28	0.45	0.41	1.00	0.28	0.24	0.09	0.07	0.43
Leptin	0.22	0.12	0.26	0.33	0.28	1.00	0.18	0.03	0.30	0.24
Adiponectin	0.32	0.22	0.37	0.32	0.24	0.18	1.00	-0.06	0.29	0.35
Resistin	0.10	-0.21	0.09	0.03	0.09	0.03	-0.06	1.00	-0.15	0.11
MCP.1	0.40	0.21	0.24	0.33	0.07	0.30	0.29	-0.15	1.00	0.25
Classification	0.11	0.45	0.67	0.64	0.43	0.24	0.35	0.11	0.25	1.00

Figure 2. is a heat map plot that displays the relationships between the input characteristics in WBCD dataset

Principal Component Analysis to Reduce Dimensionality

The variance of the chosen characteristics was further investigated. In order to achieve diminishing the dimensions Based on their variance, we used the well-known principal component analysis (PCA) technique. For the purpose of feature selection, we used PCA (linear dimensionality reduction employing singular value decomposition of the data to reduce it to a lower dimensional space). Using the sklearn library's sklearn decomposition algorithm.[19]

Figure 3 displays the variance of the dataset's various attributes. This graph demonstrates that just 10 characteristics can adequately capture the majority of variance such as The 10 parameters listed are age, BMI, glucose, insulin, HOMA, leptin, adiponectin, resistin, MCP.1 and categorization are all important factors.

https://doi.org/10.56294/dm2024.346

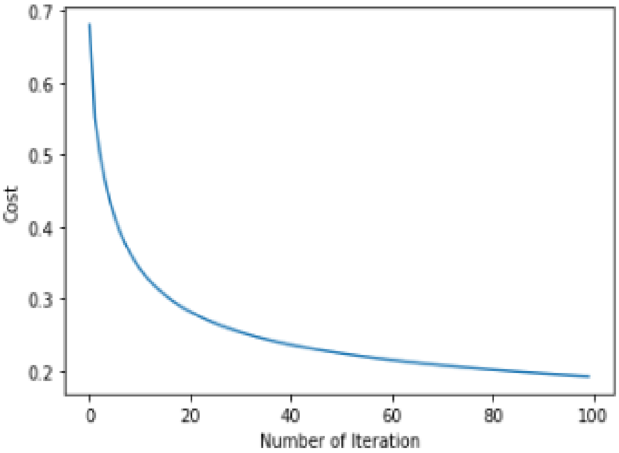


Figure 3. WBCD datasets feature cost and Number of iteration

Step 3 Feature Selection: finding technique is the goal of feature selection a crucial stage in the pre-processing of data. By eliminating irrelevant and unneeded information, it might potentially increase accuracy. Filter-based, wrapper-based, and embedded feature selection methods are the three main categories. Certain metrics are supplied as filters in filter-based techniques, and features are chosen in accordance with those filters. Examples of this feature selection methodology include the chi-square test and information gain. The wrapper-based solutions, which use a search approach, deal with the issue of feature selection. Algorithms like the best-first search are used in the search process. Recursive feature removal is one of these techniques that uses the wrapper approach. The idea behind embedded techniques is to identify the characteristics that improve the accurateness while it is being developed. The normalisation algorithm is an example of an algorithm that makes use of embedded methods.

Step 4 Machine Learning Techniques: This paper predicts the various types of breast cancer based on LR, KNN and SVM. The ML algorithms that have been researched in this research are briefly presented in this section.

a) Breast Cancer Prediction Based on Logistic Regression: It is one of the most used techniques for analysing medical data.(20) On the basis of multiple linear regression, logistic regression is a nonlinear function. A dataset is analysed using the statistical technique of logistic regression, where one or more independent variables are used to derive the outcome. It is a linear regression-like supervised learning technique.

b) Breast Cancer Prediction Based on K –Nearest Neighbour: According to variable values, samples are projected into higher-dimensional space via the K-nearest neighbour method.(21) Comparable samples demonstrate local aggregation in high-dimensional. Due to its reliance on the closest training data points. Since it does not consider the dataset's dimensionality when making a diagnosis, KNN is a non-parametric technique. The "GridSearch CV" was created to determine the total number of neighbours needed for KNN training to improve performance.(22)

c) Breast Cancer Prediction Based on Support Vector Machine: It is one of the most commonly used MLT. To create a division between two groups of observations, it makes use of an input technique. In order to increase the distance between the nearest members of classes the SVM algorithm uses the hyper planes between the two classes when determining a decision boundary. The SVM algorithm's primary goal is to clearly classify the data points. SVM operates like logistic regression when combined with a linear kernel. Since linear SVM performs poorly nonlinear SVM is favoured over linear SVM. Typically, a kernel function called the Radial Basis Function is utilised to assist translate the feature vector into a high-dimensional feature space.(23) The ability of the SVM algorithm to represent nonlinear decision boundaries is hence one of its greatest strengths. This method is also quite strong and resistant to overfitting. However, if the wrong kernel is used, it is very challenging to adapt and scale this technique to a new, larger dataset.

Step 5 Measure Accuracy: by taking into consideration the actual and expected classification, the suggested system's performance is evaluated. The confusion matrix produced for the selected classifier is used to calculate the system's efficiency. The confusion matrix for a two-class classifier is shown in table 2. It is possible to evaluate classification accuracy, sensitivity, specificity, positive predictive value and negative predictive value using the confusion matrix component parts. Classification performance: The following equation is used to determine classification accuracy:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} * 100 \qquad (1)$$

Where:
TP: properly diagnosed with breast cancer.
TN: properly identified as also being free of breast cancer.
FP: although they are listed as having breast cancer, they truly don't. (Illogical operation)
FN: although they are listed as not having breast cancer they really have Inaccuracy of type.

Table 2. Confusion Matrix representation

Actual	Predicted	
Positive	Positive True Positive	Negative False Negative
Negative	False Positive	False Positive

RESULT AND DISCUSSION

For each of the algorithms separately the sensitivity, specificity, positive predictive value (PPV), and negative predictive value were all calculated (NPV). Figure 4 displays a bar chart comparing the various machine learning methods accuracies. Table 3 displays the classification report for the proposed framework for accuracy, Precision, Recall and ROC area.

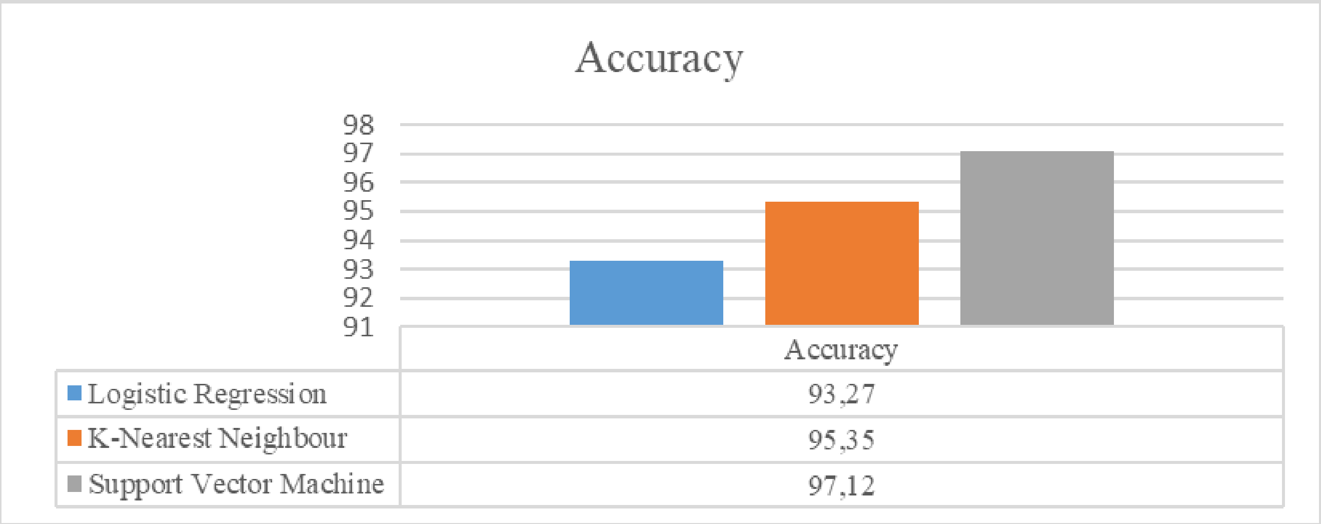


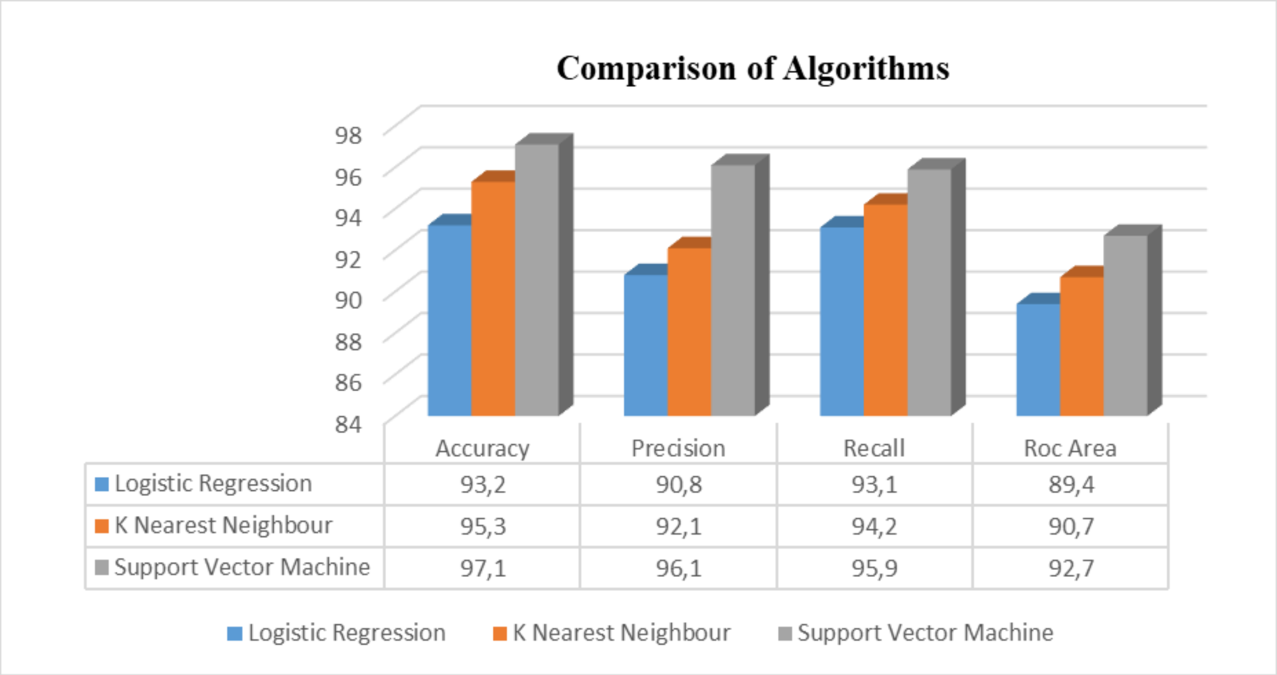
Figure 4. Accuracy of Algorithms

Table 3. Comparison of Algorithms, Accuracy, Precision, Recall and Roc Area

S. No.	Algorithms	Accuracy	Precision	Recall	Roc Area
1	Logistic Regression	93,2	90,8	93,1	89,4
2	K Nearest Neighbour	95,3	92,1	94,2	90,7
3	Support Vector Machine	97,1	96,1	95,9	92,7

The comparison between various methods is shown in figure 5 comparative chart. The graph showed that SVM was more accurate than any other methods currently being used in this paper.

The histogram of the full dataset is presented in figure 6 for improved results visualisation. The model was trained using ten and 100 estimators and the accuracy of the trained model was 97,12 % for the SVM.



	Accuracy	Precision	Recall	Roc Area
🟦 Logistic Regression	93,2	90,8	93,1	89,4
🟧 K Nearest Neighbour	95,3	92,1	94,2	90,7
⬜ Support Vector Machine	97,1	96,1	95,9	92,7

Figure 5. Comparisons of Algorithms

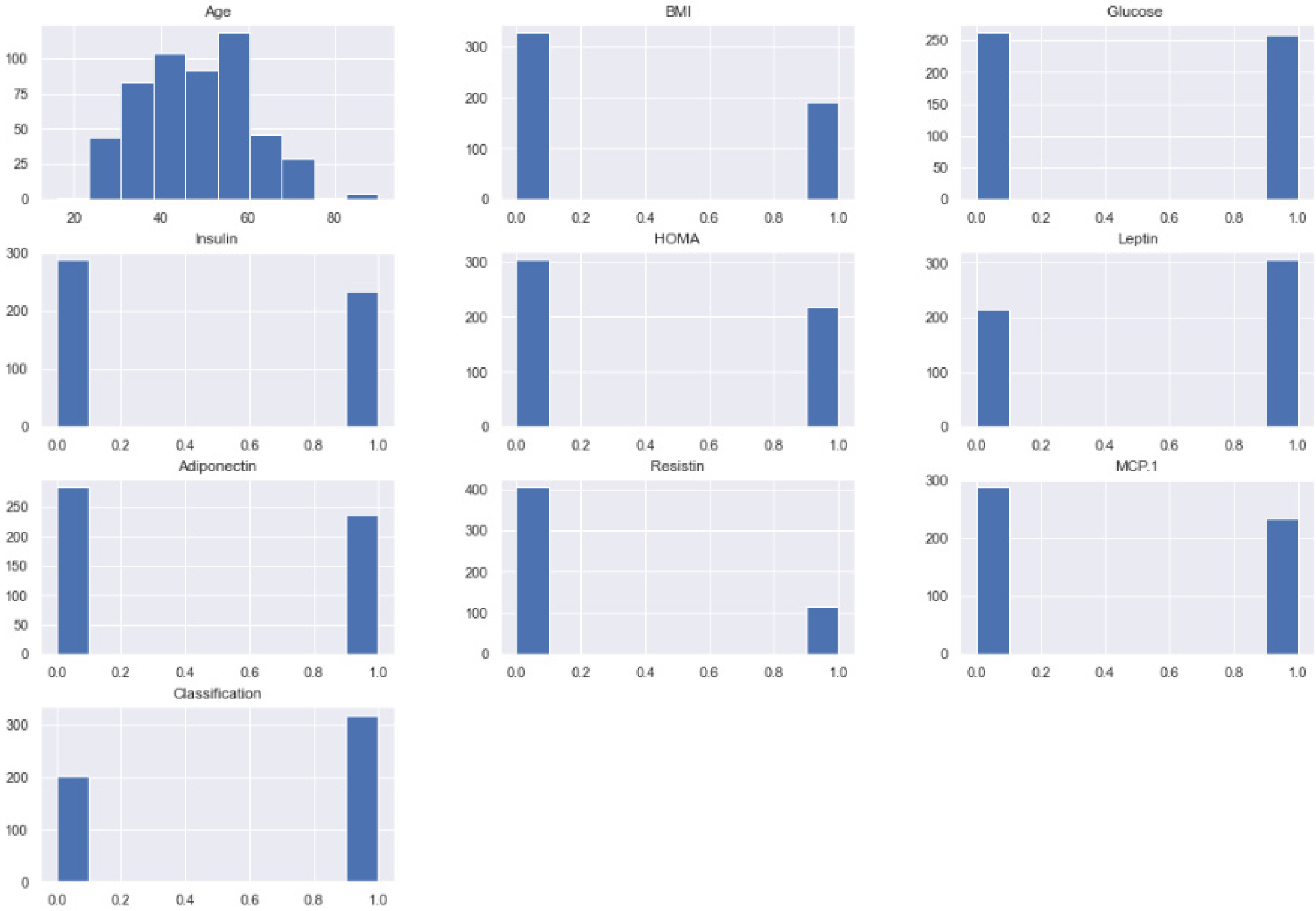


Figure 6. Histogram of entire dataset

The confusion Matrix is plotted and depicted in figure 7, 8 and 9.

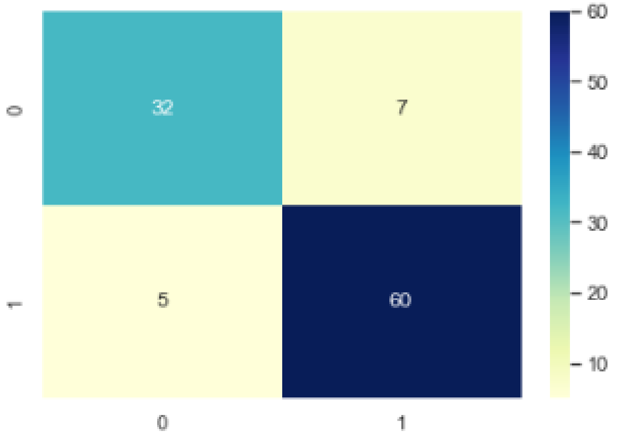


Figure 7. Confusion Matrix for LR

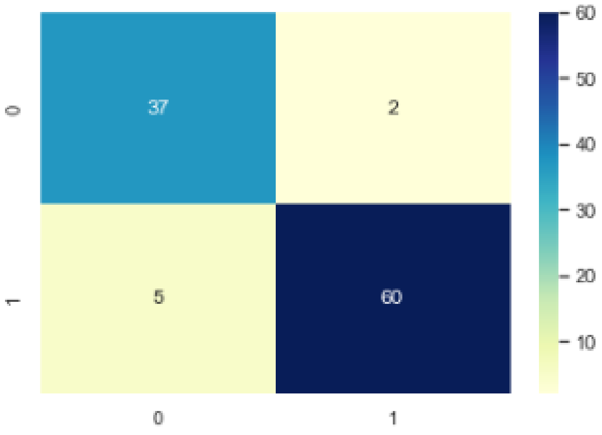


Figure 8. Confusion Matrix for KNN

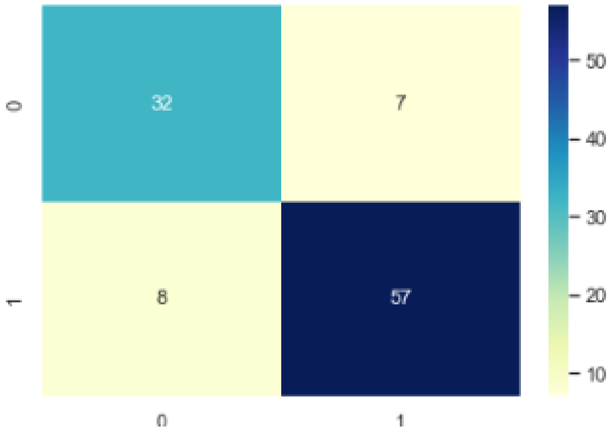


Figure 9. Confusion Matrix for SVM

Table 4. comprehensive review of various computer techniques (for predicting breast cancer)

Techniques	Description	Benefits	Limitations
Prediction of Breast Cancer	Classification Model SVM, DT, RF was considered for the Evaluation of three types of data that consists of DM	SVM Is Based on Parallel computation have strength to analyze the multiple data at the same time It provides the highest accuracy rate on different tools.	Data collection is difficult task .To Achieve good results, Accuracy, precision, Sensitivity of Large Samples are needed.
Comparison of SVM and ANN For Breast Cancer Prediction	Evaluation of SVM and ANN was done through performance Such as AUC, Roc, Precision.	The prediction of breast cancer was found through SVM because the classes are separated through hyperline and provides more accuracy in ANN.	Expected Probability of Accuracy and non-accuracy are calculated through K Fold Cross Validation.
Optimization of Algorithms through Genetic Programming Techniques	Data was founded in the images, feature selection and extraction method were applied to get information	Comparative Analysis of different Machine Learning Algorithms is performed after selecting some features operators. Extra tree classifier obtained the highest accuracy other than algorithms	It took too much time during the Evaluation process and model training model was designed to solve the problem but algorithm process was slow.
Comparative Analysis of Data Mining Classifier for cancer prediction and detection	Classification of Algorithms i.e. RF, CART Was analyzed through K Fold cross Validation.	RF provides the highest accuracy during Evaluation, this algorithms requires less effort. RF does not require the standardization and normalization of data can handle non-linear data.	Separate model was designed to check that Weather there is a tumor or not. The model took much processing time. k Fold cross validation techniques are applied for n number of Iterations. Each iteration took much time

According to the Confusion Matrix only 7 occurrences come under class 1 of the anticipated class whereas 32 instances belong to class 0 of the predicted class. This shows clearly how well our algorithm performed on the

https://doi.org/10.56294/dm2024.346

dataset used to forecast breast cancer in LR. Similarly, in KNN 37 Occurs under 0 and 60 instances under 1 and finally in SVM confusion matrix 0 occurrences comes under 32 and 57 under 1. The authors made a comparison with numerous research that use the same methodology in the same area. The comparison study shows that our suggested framework comes in second place after the other frameworks that are already in use. Table 4 provides an overview of benefits and drawbacks of a number of important research studies that have undergone evaluation. Predicting breast cancer can be done using main methods: Deep Learning, Machine Learning, and Ensemble Learning Techniques.

The comparative study using comparable frameworks as shown in table 5 provides a comparison of similar approaches.

Table 5. Comparison using similar methodologies

Methodology	Dataset Used	Accuracy
Ref[23]	UCI Machine Learning	72,3
Ref[24]	Wisconsin Breast Data set	89,3
Ref[25]	Breast Cancer Coimbra Dataset	90,2
Ref[26]	UCI Machine Learning	78,9
Ref[27]	UCI Machine Learning	75,8
Ref[28]	Wisconsin Breast Data set	90,2
Ref[29]	Breast Cancer Coimbra Dataset	91,9
Ref[30]	Wisconsin Breast Data set	76,8
Ref[31]	Breast Cancer Coimbra Dataset	81,4
Ref[32]	DDSM	86,7
Ref[33]	MRI from radiologists of the University of Bari Aldo Moro	89,77
Ref[34]	Wisconsin Breast Data set	94,01
Proposed Framework	Wisconsin Breast Data set	97,1

Based on a number of variables in the dataset, tried to use machine learning algorithms to determine whether the patient has cancer. If the presence of the illness can be predicted, cervical cancer can be found earlier.[28] The random forest performed best after taking into account a number of well-known machine learning classifiers. Furthermore, the recommended cervical cancer forecasting model fared better than previously released models for cervical cancer prediction.[29,30] Some aspects in the dataset that the authors in[31] aimed to use MLT to determine if the patient had breast cancer or not. If breast cancer presence can be anticipated, it may be discovered earlier. The authors[32,33] discovered that the SVM performed the best after evaluating several well-known machine learning classifiers. The proposed breast cancer forecasting model also performed better than previously reported breast cancer prediction models.[34,35] Furthermore, a software tool is being developed that might collect information on breast cancer potential risks and offer results from a breast cancer forecasting model for rapid and efficient treatment at the earliest stages of breast cancer. Our proposed model worked on a Machine Learning Techniques supported by SVM, KNN and LR in which SVM algorithm are offered the best performance amongst the other algorithms.

CONCLUSIONS

Classification of studies used in this research was determined by risk of breast cancer. In this research, we recommended using several MLT including Logistic Regression, SVM and KNN to develop a breast cancer diagnostic model used in Anaconda Python language. In order to assess the efficacy of risks utilising predictive models based on accuracy, recall, Precision and Roc Area the authors analysed data and carried out comprehensive evaluation. The Python programming language and its tools and libraries were used to create the suggested machine learning model. Basic data analysis, followed by data standardisation and visualisation, were carried out on the WBDC dataset. In the conclusion, the model was trained to predict breast cancer accurately and its performance and accuracy were evaluated. The dataset was specifically selected to assess factors such as Classification, Age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, etc. which are major risk of breast cancer. In addition to increasing cancer detection accuracy compared to manual diagnosis, the use of machine learning algorithms will also assist diagnose cancer more quickly and with less money. Future work will concentrate on investigating additional dataset values and producing more enhanced. By lowering total cost, time, and mortality rate, this research may aid in developing more accurate and reliable illness prediction and diagnostic systems, which will assist to create a better healthcare system.

ACKNOWLEDGMENT

This project is sponsored by Prince Sattam Bin Abdulaziz University (PSAU) as part of funding for its SDG Roadmap Research Funding Programme project number PSAU-2023/SDG/110.

BIBLIOGRAPHIC REFERENCES

1. El Sharif N, Khatib I. Healthy Lifestyle and Breast Cancer Risk in Palestinian Women: A Case-Control Study. Nutr Cancer. 2023;75(3):901–11.

2. Rezakhani L, Darbandi M, Khorrami Z, Rahmati S, Shadmani FK. Mortality and disability-adjusted life years for smoking-attributed cancers from 1990 to 2019 in the north Africa and middle east countries: a systematic analysis for the global burden of disease study 2019. BMC Cancer. 2023;23(1):80.

3. Cui Z, Kawasaki H, Tsunematsu M, Cui Y, Rahman MM, Yamasaki S, et al. Breast cancer screening and perceptions of harm among young adults in Japan: results of a cross-sectional online survey. Curr Oncol. 2023;30(2):2073–87.

4. Gupta A, Sagar G, Siddiqui Z, Rao KVS, Nayak S, Saquib N, et al. A non-invasive method for concurrent detection of early-stage women-specific cancers. Sci Rep. 2022;12(1):2301.

5. Md. Alimul Haque, Shameemul Haque, Samah Alhazmi DNP. Artificial Intelligence and Covid-19: A Practical Approach [Internet]. Bentham Science Publisher; 2022. 92-109 (18) p. Available from: https://www.eurekaselect.com/chapter/18168

6. Zeba S, Haque MA, Alhazmi S, Haque S. Advanced Topics in Machine Learning. Mach Learn Methods Eng Appl Dev. 2022;197.

7. Laumer S, Maier C, Gubler FT. Chatbot acceptance in healthcare: Explaining user adoption of conversational agents for disease diagnosis. 2019;

8. Whig V, Othman B, Gehlot A, Haque MA, Qamar S, Singh J. An Empirical Analysis of Artificial Intelligence (AI) as a Growth Engine for the Healthcare Sector. In: 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). IEEE; 2022. p. 2454–7.

9. Mokni R, Gargouri N, Damak A, Sellami D, Feki W, Mnif Z. An automatic Computer-Aided Diagnosis system based on the Multimodal fusion of Breast Cancer (MF-CAD). Biomed Signal Process Control. 2021;69:102914.

10. breast-cancer-coimbra-classification-with-eda-ml.

11. Chen H, He Y. Machine learning approaches in traditional Chinese medicine: a systematic review. Am J Chin Med. 2022;50(01):91–131.

12. Ibrahim S, Nazir S, Velastin SA. Feature selection using correlation analysis and principal component analysis for accurate breast cancer diagnosis. J imaging. 2021;7(11):225.

13. Ansari GA, Bhat SS. Exploring a link between fasting perspective and different patterns of diabetes using a machine learning approach. Educ Res. 2022;12(2):500–17.

14. Dar RA, Rasool M, Assad A. Breast cancer detection using deep learning: Datasets, methods, and challenges ahead. Comput Biol Med. 2022;149:106073.

15. Krajnc D, Papp L, Nakuz TS, Magometschnigg HF, Grahovac M, Spielvogel CP, et al. Breast tumor characterization using [18F] FDG-PET/CT imaging combined with data preprocessing and radiomics. Cancers (Basel). 2021;13(6):1249.

16. Shinde S, Kalbhor M, Wajire P. DeepCyto: A hybrid framework for cervical cancer classification by using deep feature fusion of cytology images. Math Biosci Eng. 2022;19(7):6415–34.

17. Lee SW. Regression analysis for continuous independent variables in medical research: statistical standard and guideline of Life Cycle Committee. Life cycle. 2022;2.

18. Vindas Y, Guépié BK, Almar M, Roux E, Delachartre P. Semi-automatic data annotation based on feature-space projection and local quality metrics: An application to cerebral emboli characterization. Med Image Anal. 2022;79:102437.

19. Alam A, Muqeem M, Ahmad S. Comprehensive review on Clustering Techniques and its application on High Dimensional Data. Int J Comput Sci Netw Secur. 2021;21(6):237–44.

20. Bhat SS, Selvam V, Ansari GA, Ansari MD, Rahman MH. Prevalence and early prediction of diabetes using machine learning in North Kashmir: a case study of district bandipora. Comput Intell Neurosci. 2022;2022(1):2789760.

21. Jha S, Ahmad S, Arya A, Alouffi B, Alharbi A, Alharbi M, et al. Ensemble Learning-Based Hybrid Segmentation of Mammographic Images for Breast Cancer Risk Prediction Using Fuzzy C-Means and CNN Model. J Healthc Eng. 2023;2023(1):1491955.

22. Haq AU, Li JP, Saboor A, Khan J, Wali S, Ahmad S, et al. Detection of breast cancer through clinical data using supervised and unsupervised feature selection techniques. IEEE Access. 2021;9:22090–105.

23. Agbley BLY, Li JP, Haq AU, Bankas EK, Mawuli CB, Ahmad S, et al. Federated fusion of magnified histopathological images for breast tumor classification in the internet of medical things. IEEE J Biomed Heal Informatics. 2023;

24. Ubaidillah SHSA, Sallehuddin R, Ali NA. Cancer detection using aritifical neural network and support vector machine: a comparative study. J Teknol. 2013;65(1).

25. Haq AU, Li JP, Khan I, Agbley BLY, Ahmad S, Uddin MI, et al. DEBCM: deep learning-based enhanced breast invasive ductal carcinoma classification model in IoMT healthcare systems. IEEE J Biomed Heal Informatics. 2022;28(3):1207–17.

26. Gupta SR. Prediction time of breast cancer tumor recurrence using Machine Learning. Cancer Treat Res Commun. 2022;32:100602.

27. Iqbal MS, Ahmad W, Alizadehsani R, Hussain S, Rehman R. Breast cancer dataset, classification and detection using deep learning. In: Healthcare. MDPI; 2022. p. 2395.

28. Gómez JPZ. Breast Cancer Diagnosis Using Machine Learning Techniques. Universidad Autonoma del Caribe; 2019.

29. Eltalhi S, Kutrani H. Breast cancer diagnosis and prediction using machine learning and data mining techniques: A review. IOSR J Dent Med Sci. 2019;18(4):85–94.

30. Davey MG, Hynes SO, Kerin MJ, Miller N, Lowery AJ. Ki-67 as a prognostic biomarker in invasive breast cancer. Cancers (Basel). 2021;13(17):4455.

31. Bevilacqua V, Brunetti A, Triggiani M, Magaletti D, Telegrafo M, Moschetta M. An optimized feed-forward artificial neural network topology to support radiologists in breast lesions classification. In: Proceedings of the 2016 on Genetic and Evolutionary Computation Conference Companion. 2016. p. 1385–92.

32. Das AK, Biswas SK, Mandal A, Bhattacharya A, Sanyal S. Machine learning based intelligent system for breast cancer prediction (MLISBCP). Expert Syst Appl. 2024;242:122673.

33. Veeranjaneyulu K, Lakshmi M, Janakiraman S. Swarm Intelligent Metaheuristic Optimization Algorithms-Based Artificial Neural Network Models for Breast Cancer Diagnosis: Emerging Trends, Challenges and Future Research Directions. Arch Comput Methods Eng. 2024;1-18.

34. Sharma A, Goyal D, Mohana R. An ensemble learning-based framework for breast cancer prediction. Decis Anal J. 2024;10:100372.

35. Haque MA, Ahmad S, Sonal D, Haque S, Kumar K, Rahman M. Analytical Studies on the Effectiveness of IoMT for Healthcare Systems. Iraqi J Sci. 2023;4719–28.

FINANCING

This project is sponsored by Prince Sattam Bin Abdulaziz University (PSAU) as part of funding for its SDG Roadmap Research Funding Programme project number PSAU-2023/SDG/110.

CONFLICT OF 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.

AUTHOR CONTRIBUTIONS

Conceptualization: Gufran Ahmad Ansari, Salliah Shafi Bhat, Mohd Dilshad Ansari, Sultan Ahmad, Hikmat A. M. Abdeljaber.

Investigation: Gufran Ahmad Ansari, Salliah Shafi Bhat, Mohd Dilshad Ansari, Sultan Ahmad, Hikmat A. M. Abdeljaber.

Methodology: Gufran Ahmad Ansari, Salliah Shafi Bhat, Mohd Dilshad Ansari, Sultan Ahmad, Hikmat A. M. Abdeljaber.

Writing - original draft: Gufran Ahmad Ansari, Salliah Shafi Bhat, Mohd Dilshad Ansari, Sultan Ahmad, Hikmat A. M. Abdeljaber.

Writing - review and editing: Gufran Ahmad Ansari, Salliah Shafi Bhat, Mohd Dilshad Ansari, Sultan Ahmad, Hikmat A. M. Abdeljaber.