Breast Cancer Detection Using Multi Machine Learning Models

Banan Jamil Awrahman
Information Technology
Halabja Technical Institute
Sulaimani Polytechnic University
Sulaimani, Iraq
banan.awrahman@spu.edu.iq
Orcid.org/0009-0005-4231-3587

Chya Fatah Aziz
Food Science and Quality Control
Technical College of Applied Science
Sulaimani Polytechnic University
Sulaimani, Iraq
chia.aziz@spu.edu.iq

DOI: http://dx.doi.org/10.31642/JoKMC/2018/120110
Received Aug. 10, 2024. Accepted for publication Oct. 22, 2024

Abstract— *Breast cancer prognosis remains an important area in health care. Breast cancer affects individuals of both sexes, but a higher prevalence is observed in the female population. To reduce this rate, machine learning techniques are being used, which have clear potential to improve the accuracy of cancer sensitivity and treatment. This paper presents a comparative study of the effectiveness of various machine learning algorithms with principal component analysis feature selection in the study to enhance model accuracy, and the same machine learning algorithm without principal PCA is used for breast cancer prediction. The application of PCA as a dimension reduction technique includes comparative analysis of common machine learning algorithms such as random forest (RF), support vector machine (SVM), logistic regression, and K-Nearest Neighbor models. Evaluation criteria including accuracy and sensitivity are used to evaluate the prediction performance of each method. In addition, it explores the effects of PCA on feature representation and model interpretation to provide insight into the trade-off between prediction accuracy and dimensionality reduction. The results of the PCA implementation generated by different ML algorithms which ranked SVM as the best classifier, obtained classification accuracies of 99.1%, 97.0%, and 99.3% on the three sets 75%-25%, 70%-30%, and 80 respectively sequence, although the implementation outcomes without PCA changed to... The classification accuracy was 97.9%, 97.6%, and 97.3% on the three sets 75%-25%, 70%-30%, and 80%- 20 respectively for the LR method, at Between methods of studying classification method, through a systematic comparison.*

***Keywords*: PCA, ML, RF, LR, KNN and SVM.**

I. INTRODUCTION

Breast cancer a health issue arises from irregular cell growth, in the blood vessels or ducts of the breast [1]. This growth exhibits a nature. One being aggressive and capable of spreading to parts of the body [2]. Breast cancer is more common after the age of 40 [3]. Affecting both women and less frequently than men, it is considered the second leading cause of death among women due to abnormal cell growth [4],[5]Despite high mortality rates since 2007, progress has emerged in detection and treatment. This study presents a strategy employing logistic regression (LR), whose accuracy rate is 92.98% [6] . Advances in machine learning, especially in deep learning (DL), have shown promise in improving breast cancer diagnosis accuracy, especially in medical imaging. This study employs a three-step approach utilize deep learning models and combine classifiers to improve accuracy of the classification, showing yielding encouraging outcomes, across measures. Furthermore, it introduces a learning-based framework, for detecting and diagnosing breast abnormalities showcasing medical expertise in an asynchronous manner. Using MobileNetV2 and single-layer perceptron's, it achieves high performance, surpass previous work and overcomes limitations[7].

In Addition, it compares random forest (RF), support vector machine (SVM), LR, and K-nearest neighbors (KNN) algorithms to predict the outcome of breast cancer, aimed at enhancing prognostic accuracy across cancer staging and emphasizing the importance of early detection in saving lives and improving healthcare systems through machine learning (ML) [2]. Deep learning, especially in conventional neural networks (CNNs), provides high potential for analyzing complex data sets, especially in image-based tasks as breast cancer detection. It emphasizes the important role of early detection in the fight against breast cancer[8], which is a leading cause of female mortality globally. It focuses on AI and ML techniques, especially MRI and CNN, for effective diagnosis and enhanced treatment. The article proposes the CNNI-BCC for breast cancer identify in mammograms and also tests six classification models on the Wisconsin breast cancer dataset [9][10][11].

The importance of accurate prediction of cancer, especially in breast cancer, using SVM, random forest, logistic regression, and KNN. Evaluating them from working on the Wisconsin breast cancer diagnostic dataset, SVM emerges as the more effective with an accuracy of (97.2%). The study uses Python with Scikit-learn programming but acknowledges limitations, suggesting further research to improve accuracy[12],[13]. The ability of ML to analyze complex data sets for pattern classification is essential. The study uses six algorithms, including decision trees and random forests, which have achieved high accuracy rates[14].Most cancer research emphasizes screening and timely treatment because of the various subtypes. ML and DL techniques, including ANNs, SVMs, and DTs [15], play important roles in cancer patient classification. Attention to experimental design, validation, and quality of is crucial for the advancement of ML and DL in cancer research and practice[16] .

The techniques of feature selection may introduce fluctuations in the generation of predictive feature lists. This work especially presents studies that used machine learning techniques for model cancer diagnosis and prognosis[17].

It is clear that the use of ML methods can improve the accuracy of cancer sensitivity, reproducibility, and prognosis of multiples. In the years there has been a notable 15% 20% enhancement, in the precision of cancer prognosis, with the application of ML techniques. Woman cancer mortality is 15 %. While cancer rates are higher among women in more developed regions cancer rate are increasing in every region worldwide [1]

The research aims to develop a novel computational approach for the interpretation of breast pathology, different ML techniques are used to develop a cancer prediction, including KNN, SVM, LR, Rf, using PCA classifier to reduce features and to enhance the dataset as the preprocessing step and Normalization, also using the UCI dataset. Works on improving experimental design and biological validation to advance accurate ML-based cancer prediction. For versatility, the study anticipates further exploration of ML algorithms.

II. MACHINE EARNING ALGORITHM

Machine learning methods guarantee accurate disease prediction, as seen in breast cancer detection, where data sets discriminate between malignant and benign tumors. This subset of artificial intelligence allows systems to learn independently from data, significantly improving healthcare utilization, especially in breast cancer detection. essential algorithms like Support Vector Machine (SVM), KNN, Logistic Regression (LR), and Naive Bayes (NB) play key roles in disease prediction, deployed across various domains[18], [19] .

In this paper four models such as(SVM,KNN,LR,RF) for improving the accurate in the breast detection.

A. Support Vector Machine

It is a supervised learning technique widely used in cancer detection. It selects important samples called support vectors to separate cluster using a linear function. The aim is to find the best hyperplane that maximizes the distinction between classes. In the breast cancer segment, SVM helps classify abnormalities as malignant or nonmalignant, ensuring accurate predictions[20].

Preprocessing techniques like normalization to enhance features and standardize datasets. specific steps for analysis come through SVM implementation in medical data.

B. Logistic Regression

A commonly used method in data mining to construct a classifier using a classification algorithm is the Logistic Regression (LR) algorithm[21] LR, classified among linear classifiers, is an essential technique in classification tasks. Unlike linear regression and polynomial, it offers fast and straightforward modeling while delivering accurate and easy-to-understand outcomes. Known for its expertise in binary clustering, it can also handle multi-class scenarios. Logistic regression is characterized in scenarios where the outcome is binary (0 or 1), facilitated by the. There sigmoid function[17].

C. K-nearest neighbor algorithm (KNN)

The K-nearest neighbor algorithm (KNN) is a nonparametric statistical learning method known for its simplicity and minimal impact factors. The effect depends on how the distance is determined and the value of K is chosen,[22], typically computed using Euclidean distance in our study. The selection of K significantly impacts classification, that is the number of neighbors used to get the best outcome with the kNN method, and with different K values potentially leading to varied categorizations for the same test sample.

We determined the optimal K value using cross-validation our K value in this study is 3, which impacts the best accuracy among other values.

D. Random Forest(RF)

The RF extension has seen usage in the decision tree (DT) algorithm by merging a larger number of decision trees [23]. In this study, we randomly select samples from the data set, each sample has a decision tree generated by the RF algorithm. Then there is a result for each DT, then the RF algorithm uses the voting process to combine the results of several trees, finally choosing the best prediction, which makes it accurate prediction for breast cancer because it is the result of multiple predictions.

III. METHODS AND MATERIALS

Machine learning will be used to detect breast cancer. In general, four classification models were meticulously chosen for evaluation, in this study. The diversity, in selecting classifiers based on techniques and scenarios aims to enhance our research outcomes and ensure results. Each classifier underwent training and testing using three train test sets derived from the expanded dataset outlined below.

A. Dataset :

This study utilized a dataset to train and evaluate the classifiers being tested. The enhanced dataset was created using the WBCD, a recognized dataset found on the UCI machine learning repository website[24] .

The WBDC includes 32 features and 569 instances. Principal Component Analysis, or Principal Component Analysis (PCA) is employed to improve this dataset resulting in a reduction, in the number of features. A significant emphasis has been on criteria while other aspects such, as unimportant factors have been disregarded. This approach is an aspect highlighted in this paper leading to enhanced accuracy, in classification while also minimizing the burden involved in the classification process. Achieving both accuracy and decreased overhead simultaneously is indeed a clever strategy.

We utilized three versions of our updated data collection, for training. Evaluating all classifiers. These versions consist of; version 1 with a split of 70% for training and 30% for testing; version 2 with a split of 75% for training and 25% for testing; version 3 with a split of 80% for training and 20%, for testing. Its widely understood that using training sizes results, in conducting experiments that yield reliable outcomes.

B. Data Preproceccing

To achieve the best outcomes in the dataset data pre-processing is operated. We improve the data quality through preprocessing. Data preparation and data visualization are elements of the recommended approach. Before running the machine learning model it's important to address issues, like missing data, noise and data imbalance, in the dataset as they can impact the accuracy of the results. Make sure to clean up these aspects before proceeding with the analysis.

See figure 1 to show M and B represent malignant and benign in the diagnosis columns in this data set. These string data are converted to the numeric data 0, and 1, where 1 indicate to benign and 0 refers to malignant.

Following that the process of normalization is carried out to guarantee scaling for all features while data standardization ensure that each feature has an impact, on the distance metric.

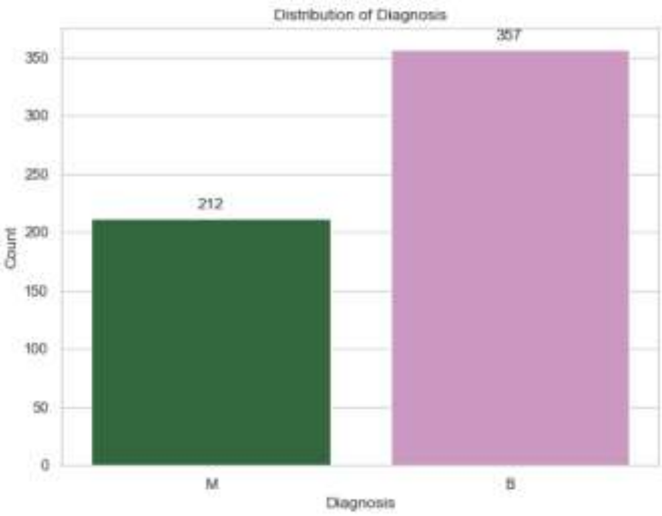


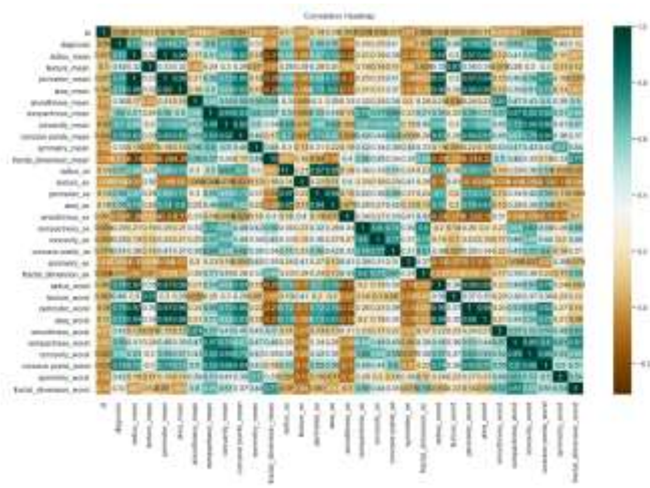
Fig. 1. Distribution of Diagnosis

The dataset utilized in this study possesses high dimensionality and consists of 32 features. These large feature numbers create overfitting also hinder reaching the optimum result. As a result, using PCA, to increase the performance of the result, to reduce the features in a dataset. The major advantages of using PCA are to reduce overfitting, remove relevant features, and improve the efficiency of machine learning algorithms. Data is normalized using the PCA method [25].

C. Data Visualization :

When data and information are shown in graphs data

Fig. 2. Visualizing relationships or patterns in a dataset



visualization helps the human mind process the data easily. We utilized a scatter plot to display the correlation, between two variables in a dataset indicating whether they are related or not. The pair plot illustrates the connections between pairs of variables, within a dataset. The following Figure 2 represents the pair plot graph. The correlation is utilized to find possible relationships between attributes and indicates the intensity of these connections, with green denoting positivity and brown indicating negativity. The features and overall data can be visualized using the process of data visualization as depicted in the correlation matrix. Some variables' relationships like Smothness_Mean is strongly related to Smothness_worst, Concavity_mean, Concave Point Mean, Concavity Wrost, and Concave Points Worst has also strong relations These all-feature relationships are more than 90%.

The data in features are distributed, not scaled and they may be transformed, Figure.3 checks distributed data in features that may be visualized in the below pair plot.

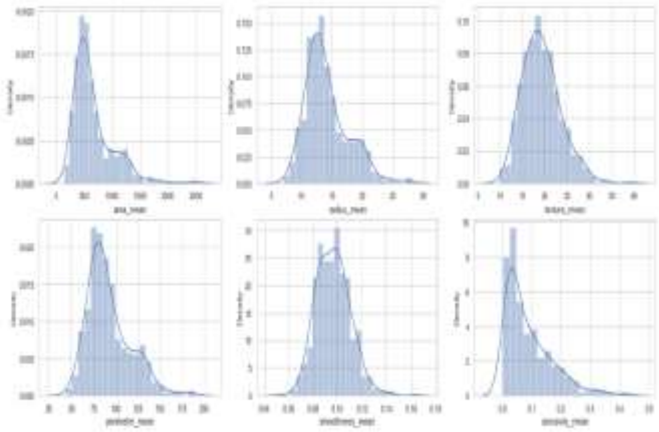


Fig. 3. Checking Distribution of Data in Features

D. Dimensionality Reduction and PCA

Principal Component Analysis, commonly known as PCA is a method used for data preprocessing for use with machine learning algorithms and applications. PCA is used for feature selection and dimensionality reduction in a smart way and produce new factors these factors are called Principal Component Analysis (PCA), and it is often used to reduce the complexity of data sets effectively by creating factors known as principal components.

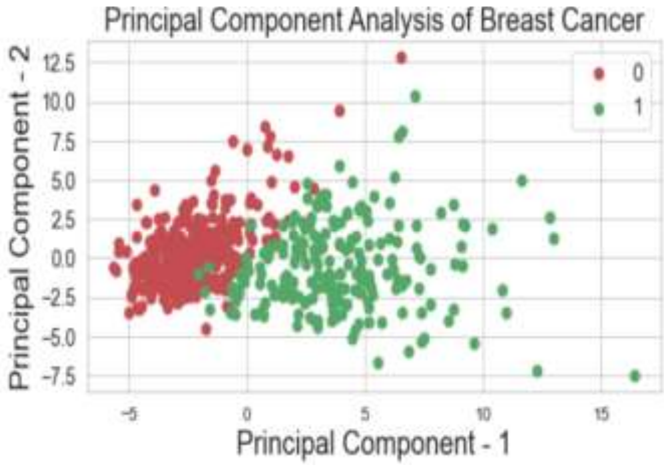


Fig. 2. Visualizing PCA

It focuses on small components where you will keep most of the information from the dataset PCA is quite versatile. Can be used to analyze datasets that might have issues, like missing values, multicollinearity, categorical data, and imprecise measurements. The aim is to uncover insights, from the data. So, to sum up, the concept, behind PCA is straightforward; decrease the variables in a dataset, and a benefit of using PCA is a use case when PCA can remove noise or redundant Information from data by focusing on the principal components that capture the underline pattern, while retaining an information as feasible as shown in figure 4.

IV. RESULT

After examining four used classifiers, for analyzing medical images this research paper presents various findings highlighting Logistic Regression (LR) and Support Vector Machine (SVM) as the top two classifiers for breast cancer diagnosis. LR emerges as the classifier achieving the highest classification accuracy compared to SVM. The results displayed in Table 1 and Table 2 indicate the accuracy of each classifier across three training testing sets with using PCA and without using PCA. SVM consistently outperforms all classifiers across testing scenarios achieving classification accuracies of 99.3%, 97.0%, and 99.1%, for the 75% 25% 70% 30% and 80% 20% sets respectively by using PCA, also LR archives performance accuracies of 97.9%, 97.6% and 97.3%, for the 75% 25% 70% 30% and 80% 20% sets respectively with using PCA or Personal Component Analyses. It is noteworthy that the highest accuracy is achieved with a training test size of 75%-%25 of a dataset with machine learning models with PCA for SVM that achieved 99.1%.

Table 1. Accuracy of ML Methods with PCA

ML Method with PCA	**%80-%20 Of Training testing size**	**%70-%30 Of Training testing size**	**%75-%25 Of Training testing size**
KNN	%96.4	%96.4	%97.2
RF	%98.2	%95.3	%97.9
LR	%96.4	%95.3	%98.6
SVM	%99.1	%97.0	%99.3

Table 2. Accuracy of ML algorithm without PCA

ML Method without PCA	**%80-%20 Of Training testing size**	**%70-%30 Of Training testing size**	**%75-%25 Of Training testing size**
KNN	%94.7	%96.4	%95.8
RF	%95.8	%95.9	%94.4
LR	%97.3	%97.6	%97.9
SVM	%95.6	%95.9	%95.8

The classification errors shows in Table 3 and Table 4 respectively for using PCA with machine learning models and without using PCA, false positives (FP), and false negatives (FN) for all classifiers under three training-testing sets. The findings suggest that LR performed better, than classifiers by producing the classification errors,”1” for FN and “0” for FP. After examining our results and conducting an assessment the paper ultimately determines that employing PCA the LR machine learning methods rank at the top; SVM emerges as the contender surpassing the stack classifier—a collective classifier built on the stacking technique. In Figure 5 a bar graph displays the accuracy distribution, across machine learning algorithms using Principal Component Analysis (PCA) for three Training testing sets.

Table 3. The classification errors of each machine learning models using PCA

Training testing size	%80-%20		%70-%30		%75-%25	
Methods	***FP***	***FN***	***FP***	***FN***	***FP***	***FN***
KNN	0	4	2	4	1	3
RF	0	2	4	3	3	0
LR	1	3	5	3	0	2
SVM	0	1	3	2	0	1

Table 4. Classification errors without using PCA with machine learning methods

Training testing size	%80-%20		%70-%30		%75-%25	
Methods without PCA	***FP***	***FN***	***FP***	***FN***	***FP***	***FN***
KNN	3	3	3	3	2	4
RF	3	2	4	3	4	4
LR	1	2	2	2	1	2
SVM	4	1	5	2	4	2

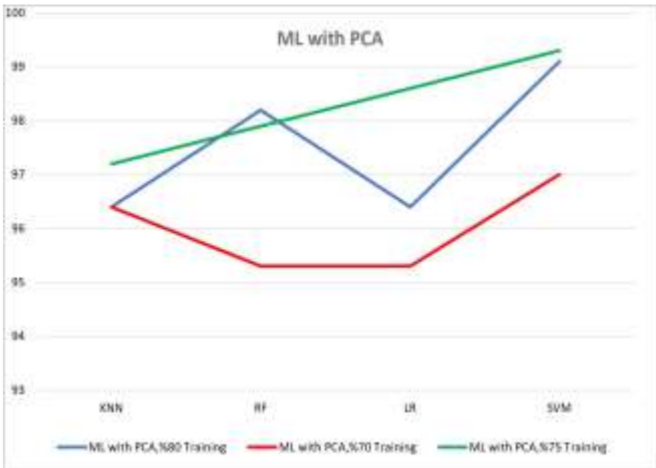


Fig. 3. The Distribution of Accuracies Among Different Machine Learning Algorithm By Using Principal Component Analysis (PCA)

In Figure 6 a bar graph displays the accuracy distribution, across machine learning algorithms KNN, RF, Lr, and SVM without using Principal Component Analysis (PCA) for three Training testing sets. In this research by using for methods of Machine Learning as the same with using PCA, in here without using PCA Logistic Regression (LR) rank at the top after that SVM, KNN and RF prospectively.

Fig. 6. The Distribution of Accuracies Among Different Machine Learning Algorithms without Using Principal Component Analysis (PCA)

Ensuring that a model is validated is crucial for it to be accepted. In our experiment we have repeated the sampling validation technique 12 times for each classification model to ensure results. The accuracy of each model’s classification has been determined using the accuracy formula (1)

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

In classification models a true positive (TP) occurs when positive instances are predicted accurately while a true negative (TN) happens when negative cases are predicted correctly. On the hand a false positive (FP) represents a prediction of a positive case, by the model and a false negative (FN) is an inaccurate prediction of a negative case. The confusion error is determined based on the occurrences of positives (FP) and false negatives (FN).

Table 5. Comparison of performance models of ML methods with PCA and without PCA, with different train test size

Train Test Size	ML without PCA	F1 Score	Precision	Recall
%80-%20	KNN	%93	%93	%93
	RF	%94	%95	%93
	LR	%97	%95	%98
	SVM	%94	%97	%91
%70-%30	KNN	%95	%95	%95
	RF	%94	%95	%94
	LR	%97	%97	%97
	SVM	%94	%97	%92
%75-%25	KNN	%95	%93	%96
	RF	%93	%93	%93
	LR	%97	%96	%98
	SVM	%94	%96	%93
Train Test Size	**ML with PCA**	**F1 Score**	**Precision**	**Recall**
%80-%20	KNN	%97	%95	%100
	RF	%99	%97	%100
	LR	%97	%96	%99
	SVM	%99	%99	%100
%70-%30	KNN	%97	%97	%98
	RF	%96	%96	%96
	LR	%96	%97	%96
	SVM	%98	%98	%97
%75-%25	KNN	%98	%97	%99
	RF	%98	%100	%97
	LR	%99	%98	%100
	SVM	%99	%99	%100

V. CONCLUSION

Breast cancer survival rates may be achieved with early diagnosis, the utilization of machine learning has guaranteed the optimization of their effectiveness. in defining patterns and discovering from a huge number of medical images. The dataset pertaining to breast cancer is subjected to data mining utilizing various training/testing splits including 75% training/25% testing, 70% training/30% testing, and 80% training/20% testing. By means of comparative analysis, this research incorporates four unique types of classifiers, each streamlining the process of classification. The primary significance of this study is that the logistic regression model is identifying the optimal classifier. It achieved an accuracy classification of 99%, 99%, and 97% over the three sets 75%-25%, 70%-30%, and 80%-20 respectively by using PCA, and LR achieved the best performance result without using PCA for three sets of train test split and for four machine learning models which is %98, %98, and %99 for three sets 75%-25%, 70%-30%, and 80%-20 respectively. Some enhancements have been made, encompassing various phases. These enhancements involve the exclusion of features with no weighted impact and the selection of features with the highest influence on the prediction process to enhance prediction accuracy. Utilizing PCA to reduce the number of features results in decreased computational burdens. Consequently, this improvement phase highlights three key enhancements: a) it decreases the model computation cost and enriches classification accuracy; b) to boost the trustworthiness applying three sets of training-test scenarios; and c) in this study used some variant evaluation methods.

REFERENCE

[1] A. Alloqmani, Y. B. Abushark, and A. I. Khan, "Anomaly Detection of Breast Cancer Using Deep Learning," *Arab J Sci Eng*, pp. 1–26, 2023.

[2] S. Prasath Alias Surendhar, "A Novel Approach for Breast Cancer Detection Using Neural Networks".

[3] S. M. McKinney *et al.*, "Addendum: International evaluation of an AI system for breast cancer screening," *Nature*, vol. 586, no. 7829, pp. E19–E19, 2020.

[4] V. Apoorva, H. K. Yogish, and M. L. Chayadevi, "Breast Cancer Prediction Using Machine Learning Techniques," in *3rd International Conference on Integrated Intelligent Computing Communication & Security (ICIIC 2021)*, Atlantis Press, 2021, pp. 348–355.

[5] A. Altameem, C. Mahanty, R. C. Poonia, A. K. J. Saudagar, and R. Kumar, "Breast cancer detection in mammography images using deep convolutional neural networks and fuzzy ensemble modeling techniques," *Diagnostics*, vol. 12, no. 8, p. 1812, 2022.

[6] C. Srividya, M. Madhubala, U. Neelaveni, A. Reddy, and D. D. Kommareddy, "BREAST CANCER DETECTION USING LOGISTIC REGRESSION".

[7] "13".

[8] A. Sharma and P. K. Mishra, "Performance analysis of machine learning based optimized feature selection approaches for breast cancer diagnosis," *International Journal of Information Technology*, pp. 1–12, 2022.

[9] A. Khalid *et al.*, "Breast Cancer Detection and Prevention Using Machine Learning," *Diagnostics*, vol. 13, no. 19, p. 3113, 2023.

[10] N. Rane, J. Sunny, R. Kanade, and S. Devi, "Breast cancer classification and prediction using machine learning," *International Journal of Engineering Research and Technology*, vol. 9, no. 2, pp. 576–580, 2020.

[11] K. S. Priyanka, "A review paper on breast cancer detection using deep learning," in *IOP conference*

series: materials science and engineering, IOP Publishing, 2021, p. 012071.

[12] O. Iparraguirre-Villanueva, A. Epifanía-Huerta, C. Torres-Ceclén, J. Ruiz-Alvarado, and M. Cabanillas-Carbonel, "Breast cancer prediction using machine learning models," 2023.

[13] M. A. Naji, S. El Filali, K. Aarika, E. L. H. Benlahmar, R. A. Abdelouhahid, and O. Debauche, "Machine learning algorithms for breast cancer prediction and diagnosis," *Procedia Comput Sci*, vol. 191, pp. 487–492, 2021.

[14] S. A. Mohiddin, T. P. Sri, K. Sharmila, and P. Bhavya, "BREAST CANCER PREDICTION USING MACHINE LEARNING".

[15] F. J. Shaikh and D. S. Rao, "Prediction of cancer disease using machine learning approach," *Mater Today Proc*, vol. 50, pp. 40–47, 2022.

[16] F. J. Shaikh and D. S. Rao, "Prediction of Cancer Disease using Machine learning Approach," *Mater Today Proc*, vol. 50, pp. 40–47, 2022, doi: https://doi.org/10.1016/j.matpr.2021.03.625.

[17] Mohit Agrawal, "Cancer Prediction Using Machine Learning Algorithms," *International Journal of Science and Research (IJSR)*, vol. 9, no. 8, 2020.

[18] M. U. Sarwar and A. R. Javed, "Collaborative health care plan through crowdsource data using ambient application," in *2019 22nd international multitopic conference (INMIC)*, IEEE, 2019, pp. 1–6.

[19] M. Usman Sarwar, A. Rehman Javed, F. Kulsoom, S. Khan, U. Tariq, and A. Kashif Bashir, "Parciv: Recognizing physical activities having complex interclass variations using semantic data of smartphone," *Softw Pract Exp*, vol. 51, no. 3, pp. 532–549, 2021.

[20] R. U. Khan, X. Zhang, M. Alazab, and R. Kumar, "An improved convolutional neural network model for intrusion detection in networks," in *2019 Cybersecurity and cyberforensics conference (CCC)*, IEEE, 2019, pp. 74–77.

[21] C. F. Aziz and B. J. Awrahman, "Prediction Model based on Iris Dataset Via Some Machine Learning Algorithms," *Journal of Kufa for Mathematics and Computer*, vol. 10, no. 2, pp. 64–69, 2023.

[22] A. Delgado-Bonal and A. Marshak, "Approximate entropy and sample entropy: A comprehensive tutorial," *Entropy*, vol. 21, no. 6, p. 541, 2019.

[23] K. Lemons, "A comparison between Naïve bayes and random forest to predict breast cancer," *International Journal of Undergraduate Research and Creative Activities*, vol. 12, no. 1, p. 10, 2023.

[24] M. Elsadig, A. Altigani, and H. Elshoush, "Breast cancer detection using machine learning approaches: a comparative study," *International Journal of Electrical and Computer Engineering*, vol. 13, pp. 736–745, Mar. 2023, doi: 10.11591/ijece.v13i1.pp736-745.

[25] K. M. M. Uddin, N. Biswas, S. T. Rikta, and S. K. Dey, "Machine learning-based diagnosis of breast cancer utilizing feature optimization technique," *Computer Methods and Programs in Biomedicine Update*, vol. 3, p. 100098, 2023, doi: https://doi.org/10.1016/j.cmpbup.2023.100098.