International Science and Technology Journal
المجلة الدولية للعلوم و التقنية

عدد خاص بالمؤتمر الليبي الدولي للعلوم
التطبيقية و الهندسية دورته الثانية
LICASE -2
2024 / 10 / 30-29

تم استلام الورقة بتاريخ:27/ 9/2024م وتم نشرها على الموقع بتاريخ: 30/ 10/2024م

Mitigating Missed Diagnoses: A Patch-Based Deep Learning Framework for Breast Cancer Classification

http://www.doi.org/10.62341/mmam2101

Mahmoud M.Alakrimi[1] , Abdelsalam M. Ahmed[2]

1-Computer Department, College of Education, Abu Issa, University of Zawia, Libya
E-mail: M.Elakrami@zu.edu.ly GSM:+218918791453
2-School of applied science and engineering, Janzoor, Tripoli, Libya.
Email: Abdussalam.alfthi@gmail.com GSM:+218926448380

Abstract:
Breast cancer remains a significant public health concern, with early and accurate diagnosis being crucial for effective treatment. Machine learning (ML) algorithms offer promising support in breast cancer classification. This study evaluates the performance of various ML techniques (GNB & LR 97%, SVM 96%, DT 95%), using well known a Kaggle dataset with 699 instances and 10 attributes. The findings highlight the potential of these algorithms to enhance diagnostic accuracy, aiding in early detection and treatment.

Keywords: Breast cancer, Machine learning, Classification, Decision trees, Logistic regression, Naive Bayes, Support vector machines.

التخفيف من التشخيصات الخاطئة: إطار عمل التعلم العميق القائم على التصحيح لتصنيف سرطان الثدي

محمود م. العكرمي[1]، عبد السلام م. أحمد[2]

(1) قسم الحاسب الآلي، كلية التربية أبو عيسى، جامعة الزاوية، ليبيا
البريد الإلكتروني: M.Elakrami@zu.edu.ly GSM:+218918791453
(2) مدرسة العلوم التطبيقية والهندسة، جنزور، طرابلس، ليبيا.
البريد الإلكتروني: Abdussalam.alfthi@gmail.com GSM:+218926448380

الملخص— يظل سرطان الثدي مصدر قلق كبير للصحة العامة، حيث يعد التشخيص المبكر والدقيق أمرًا بالغ الأهمية للعلاج الفعال. تقدم خوارزميات التعلم الآلي (ML) دعمًا واعدًا في تصنيف سرطان الثدي. حيث ان نتائج هذه الدراسة بالاستعانة بنماذج تقنيات التعلم الآلي المختلفة على التوالي هي (GNB & LR 97 ٪, SVM 96٪, DT 95٪) باستخدام مجموعة بيانات Kaggle المعروفة جيدًا والتي تحتوي على 699 حالة و10 سمات.

تسلط النتائج الضوء على إمكانات هذه الخوارزميات لتعزيز دقة التشخيص والتقليل من النتائج الخاطئة، والمساعدة في الكشف المبكر والعلاج.

الكلمات الرئيسية— سرطان الثدي، التعلم الآلي، التصنيف،GNB ,DT,SVM, LR

Introduction

Breast cancer is the most prevalent cancer among women globally [1]. Early detection is paramount for successful treatment and improved patient survival rates. However, traditional diagnostic methods like mammography can have limitations, including human error leading to missed cancers or unnecessary biopsies [2]. This paper explores the potential of machine learning (ML) algorithms to enhance breast cancer classification accuracy and efficiency.

Various machine learning algorithms commonly employed include Gaussian Naïve Bayes (GNB), which employs statistical techniques to estimate the probabilities of each class based on the distribution of features, yielding Mitigating Missed Diagnoses outcomes for specific datasets [3]. Logistic Regression (LR), despite its name, is used for classification problems. In the context of breast cancer, it models the probability that a tumor is malignant based on specific features [5]. Decision Tree (DT) is a classification algorithm that makes decisions based on a set of predefined rules in the context of breast cancer classification [4]. It maps features such as tumor size, shape, and density to make informed decisions. And subsequently, based on the problems previously based on and with the support of the result of research that has been done previously related to the implementation of feature selection for optimizing machine learning algorithms in breast cancer classification, we improved on this research using different famous supervised learning classifiers. Each classifier is compared against each other based on performance metrics, especially ROC (receiver operating characteristic) and confusion matrix. With the result of the classification by supervised algorithm, patients with existing parameters can be classified between benign and malignant cancer. So that this pattern can be used for benchmark diagnosis so that can be detected early and is expected to be able to reduce mortality and cancer rates in breast cancer.

By noting the great importance of FN and FP values in evaluating the performance of the model and making medical decisions for breast cancer diagnosis, where the FN value (False Negative): refers to the case in which the model fails to classify a positive case (i.e. the presence of cancer) as positive, and classifies it as negative. In other words, a patient with cancer is classified as healthy, which may lead to delayed treatment and worsening of the condition.

FP (False Positive): Its value refers to the case in which the model classifies a negative case (i.e. the absence of cancer) as positive. In other words, a healthy patient is classified as having cancer, which may lead to unnecessary additional medical tests and psychological harm to the patient.

Therefore, the higher the FN value, the lower the sensitivity of the model (Sensitivity), which means that the model fails to effectively detect cancer cases. It means that there is an increased risk of delayed diagnosis of breast cancer cases, which may negatively affect the chances of recovery. The higher the FP value, the lower the model specificity, which means that the model tends to incorrectly classify negative cases as positive.

That is why our research is distinguished by focusing on having a zero FN value, which means that no infected case is diagnosed as healthy, so that the infected person can treat himself as quickly as possible and doctors can make appropriate medical decisions.

In conclusion, we used data processing methods in the research before passing it to machine learning tools. We found that the GNB technique is the best, most accurate and reliable in not missing any positive value. and this is our main contribution for research. This model is anticipated to aid pathologists in conducting examinations with greater consistency and efficiency in order to detect breast cancer diagnoses.

Literature review

(K. Amril M.Siregar, S.Faisal , 2023) investigated feature selection using PCA for dimension reduction before applying machine learning models. An SVM with RBF kernel achieved the highest accuracy (unspecified), followed by Logistic Regression (97.3%). Notably, the SVM exhibited perfect precision and recall, and the ROC curve favored SVM over LR. These findings suggest PCA and SVM's potential for accurate classification [6].

(R.Hridoy, 2024) Rashidul Studies confirm hyperparameter optimization's impact on machine learning model performance. This research explores its effectiveness in breast cancer diagnosis. Grid search yielded a 100% recall for k-nearest neighbors and 99.42% accuracy for other models (kNN, logistic regression, MLP). Only XGBoost showed no improvement. These findings support the proposed technique's potential for breast cancer diagnosis [7].

(M.Awan , 2024) explores the use of seven machine learning models for breast cancer prediction. KNN achieved the highest accuracy (99%) on the Wisconsin dataset, while Logistic Regression performed best (83%) on the breast cancer dataset. This suggests model performance may vary depending on the data source [8].

(J.Purnomo, 2024) investigates machine learning for breast cancer stage classification. Imbalanced data is addressed using SMOTE oversampling. Neural networks outperform K-Nearest Neighbors (82.3% vs. 80.8% AUC) for stage detection [9].

(A.Bansal , 2024) explores how machine learning, particularly CNNs, excel in breast cancer classification. It highlights limitations in mammography and proposes a method

وتم نشرها على الموقع بتاريخ: 30/ 2024/10م
تم استلام الورقة بتاريخ:27/ 2024/9م

combining whole-image and patch-based analysis, reducing dependence on manual ROIs. Emphasizing Python libraries for model development, the review suggests machine learning's potential to improve diagnostic accuracy, CAD systems, and ultimately, patient care [10].

Aim of the Project

The purpose of this research is to determine which features are yielding missed diagnoses when predicting benign or malignant cancer, as well as to discover general tendencies that could help with model and hyperparameter selection and to develop machine learning models to classify between malignant and benign breast tumors. The study was conducted using the following steps shown in Figure 1. We divided our research study into five sections: “Data Collection, Data Pre-processing, Exploratory Data Analysis, Model Selection, and Model Evaluation." These sections will be briefly explained and discussed one by one.

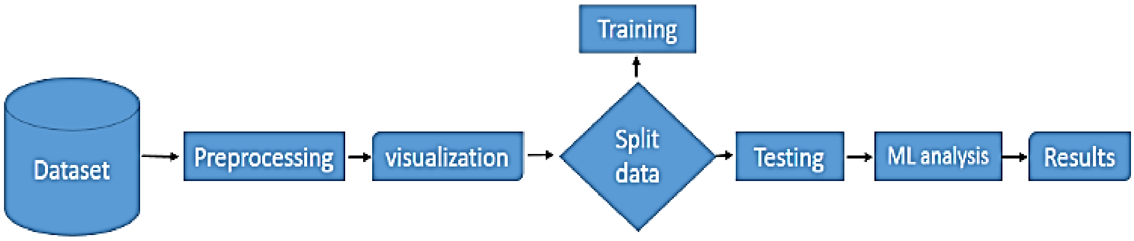


Figure 1. Proposed model by using breast cancer datasets

Motivation

Machine learning offers immense potential in this area. This study explores the use of various algorithms for Mitigating Missed Diagnoses breast cancer classification to potentially enhance performance. By comparing these models using metrics like ROC curves and confusion matrices, we hope to identify the most effective approach. Ultimately, this research aspires to develop a reliable tool that can assist pathologists in achieving consistent, accurate, and efficient breast cancer diagnoses. This, in turn, could lead to earlier detection, reduced mortality rates, and improved patient outcomes.

Dataset and Attribute Description

The study is based on a dataset that is publicly available from the UCI Machine Learning Repository (Asuncion and Newman, 2007) [http://mlearn.ics.uci.edu/MLRepository.html]. The dataset consists of 699

وتم نشرها على الموقع بتاريخ: 30/ 2024/10م
تم استلام الورقة بتاريخ:27/ 2024/9م

human cell sample records, each described by 10 features, which contain the values of a set of cell characteristics. However, the Kaggle dataset offers a unique opportunity to explore an alternative data source often used in breast cancer research. A detailed description of the Kaggle dataset is provided in Table 1.

Table No. (1) Dataset features

Field name	Description	Field name	Description
ID	Clump thickness	SingEpiSize	Single epithelial cell size
Clump	Clump thickness	BareNuc	Bare nuclei
UnifSize	Uniformity of cell size	BlandChrom	Bland chromatin
UnifShape	Uniformity of cell shape	NormNucl	Normal nucleoli
MargAdh	Marginal adhesion	Mit	Mitoses
		Class	Benign or malignant

For the purposes of this study, we're using a dataset that has a relatively small number of predictors in each record. To download the data, we will download it from IBM Object Storage. (https://cf-courses-data.s3.us.cloud-object-storage.appdomaincloud/IBMDeveloperSkillsNetwork-ML0101EN-SkillsNetwork/labs/Module%203/data/cell_samples.csv)

The ID field contains the patient identifiers. The characteristics of the cell samples from each patient are contained in fields Clump to Mit. The values are graded from 1 to 10, with 1 being the closest to benign.

The Class field contains the diagnosis, as confirmed by separate medical procedures, as to whether the samples are malignant (M = 4) or benign (B = 2), representing malignant and benign tumor cells, respectively. There are no missing values in the dataset. Among the samples, 444 are benign and 239 are malignant.

Exploratory data analysis

Exploratory data analysis helps to investigate the critical decision for further processing to build data modeling. We use Python libraries Plotly seaborn, and Matplotlib to plot a scatter plot and heatmap of both datasets. Multi-variable scatter plots help display interactions between more than two variables in a single plot, whereas heatmaps, which synthesize data and present it graphically, give a practical visual overview of information. In the heatmap (Figure 2) of the dataset, we can see that the variables uniformity of cell size, uniformity of cell shape, and bare nuclei are highly correlated with the target variable "class."

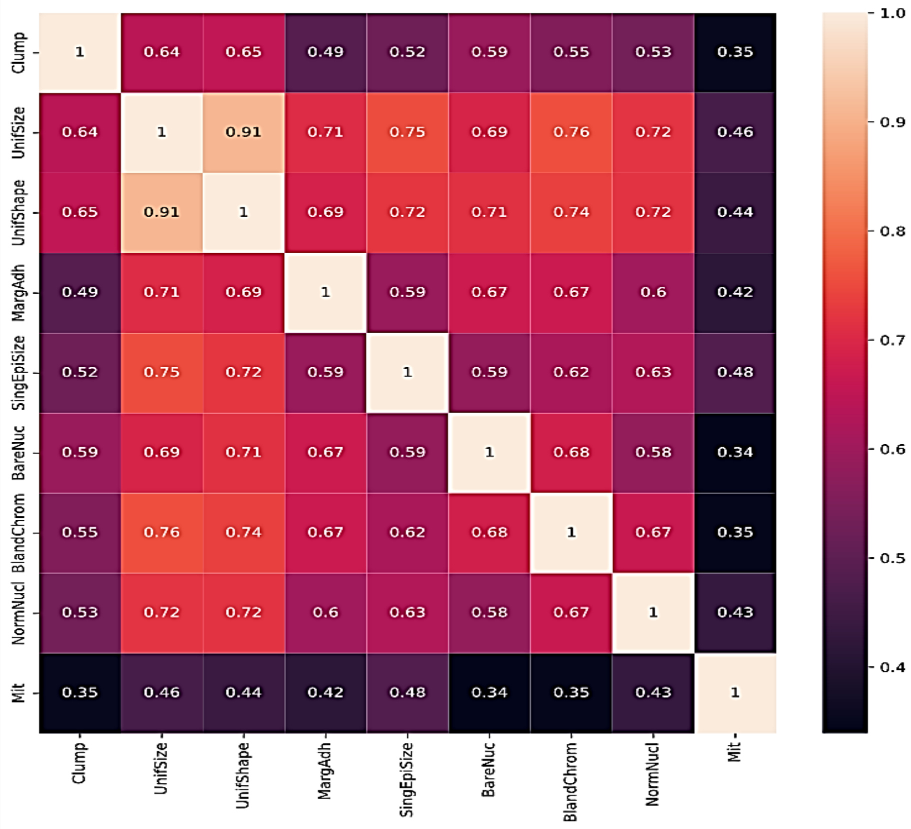


Figure 2. Heatmap of dataset

Dataset Pre-processing.

Data Cleaning and Transformation:

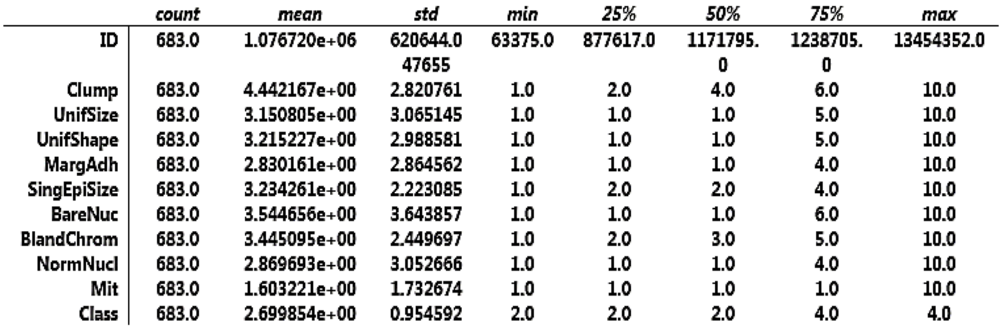
To ensure high-quality data, we employed two key techniques:

Data Cleaning: The dataset underwent a meticulous cleaning process to remove any inconsistencies or errors. This involved eliminating duplicate entries, identifying and removing unnecessary columns (like IDs), and ensuring consistent formatting. Fortunately, the chosen dataset contained no missing values or duplicates. [11]. Table 2 summarizes the results of the feature data explanatory analysis for the dataset.

وتم نشرها على الموقع بتاريخ: 30/ 2024/10م

تم استلام الورقة بتاريخ:27/ 2024/9م

Table 2. Statistical summary of the dataset.



	count	mean	std	min	25%	50%	75%	max
ID	683.0	1.076720e+06	620644.0 47655	63375.0	877617.0	1171795. 0	1238705. 0	13454352.0
Clump	683.0	4.442167e+00	2.820761	1.0	2.0	4.0	6.0	10.0
UnifSize	683.0	3.150805e+00	3.065145	1.0	1.0	1.0	5.0	10.0
UnifShape	683.0	3.215227e+00	2.988581	1.0	1.0	1.0	5.0	10.0
MargAdh	683.0	2.830161e+00	2.864562	1.0	1.0	1.0	4.0	10.0
SingEpiSize	683.0	3.234261e+00	2.223085	1.0	2.0	2.0	4.0	10.0
BareNuc	683.0	3.544656e+00	3.643857	1.0	1.0	1.0	6.0	10.0
BlandChrom	683.0	3.445095e+00	2.449697	1.0	2.0	3.0	5.0	10.0
NormNucl	683.0	2.869693e+00	3.052666	1.0	1.0	1.0	4.0	10.0
Mit	683.0	1.603221e+00	1.732674	1.0	1.0	1.0	1.0	10.0
Class	683.0	2.699854e+00	0.954592	2.0	2.0	2.0	4.0	4.0

Data Transformation:

To improve model performance, we transformed categorical variables into numerical values. For instance, the BareNuc label's data type was object and transferred to numerical format. This allows machine learning algorithms to better understand and utilize the data. It's important to note that all other features besides BareNuc were already in numerical format.

Splitting the Dataset:

To train and evaluate the models effectively, we divided the dataset of 699 data points into training and testing sets. A 70:30 split was chosen, allocating 70% (489 points) of the data for training the models and 30% (210 points) for testing their performance. Since the dataset has minimal outliers, this 30% split is sufficient for objective evaluation. An 80:20 split led to an overfitting problem.

Machine learning Techniques

1. Naïve Bayes

The algorithm discussed in this study is a fundamental outcome in the fields of probability and statistics. It can be defined as a conceptual framework used for decision-making. In the context of Naive Bayes (NB), the variables are conditionally independent. NB can be employed to analyze data that have direct influence on each other, in order to establish a model.

Naïve Bayes is one of the most efficient yet straightforward classifiers. It is based on the Bayes theorem, which describes how event probability is calculated using prior knowledge of circumstances that could be pertinent to the occurrence. In this particular research, the default formulation of the NB equation is presented as follows:

وتم نشرها على الموقع بتاريخ: 30/ 2024/10م
تم استلام الورقة بتاريخ:27/ 2024/9م

$$P(y \mid x_1, x_2, \ldots, x_n) = \frac{P(y)P(x_1|y)P(x_2|y)\ldots P(x_n|y)}{P(x_1), P(x_2), \ldots, P(x_n)}$$

where $P(y \mid x1 , x2 , \ldots , xn)$ is the posterior probability of class y given features of x, $P(y)$ is the prior probability of class y,$P(xi\,|y)$ is the likelihood of feature xi given class y, and $x1 , x2 , \ldots , xn$ are the features. The Naïve bayes classifier then predict the class with the highest posterior probability [12].

2. **Logistic regression**

Logistic regression is a machine learning algorithm most frequently employed under supervised learning [13]. It's used to predict categorical dependent variables (0 or 1, yes or no, true or false) from a set of independent variables [14]. The prediction is made by converting the unobserved data to the built-in logit function. Predict 0 and 1 for the logistic regression modeling utilizing the standard logistic function and linear probability function.

$$p(x) = \frac{e^{ax+b}}{1+e^{ax+b}} = \frac{1}{1+e^{-ax+b}}$$

Logistic regression gives linear classifier results, predicting y=1 when $p \geq 0.5$ and $y = 0$ when <0.5. Logistic function in general

$$f(x) = \frac{1}{1+e^{-ax+b}}$$

3. **Decision tree classifier**

Decision tree is one of the most used classification methods. The classifier is tree-structured, where the internal nodes correspond to the dataset's properties, and each leaf node signifies the classification result. Decision trees classify depending on the values of the features. The information gain approach determines which aspect of the dataset provides the most information, designates that as the root nodes, and so on until they can classify each dataset entity. Using:

Information Gain=Entropy(S)−[(Weighted Avg)∗Entropy (Each feature)]

Entropy is a metric used to quantify the impurity in a particular characteristic. It describes data randomness. Calculating entropy is as:

$$Entropy(S) = -P(0)\,log2\;P(0) - P(1)\;log2\;P(1)$$

Where,

− S is the total number of samples.
− P(2) is the probability of Benign.
− P(4) is a probability of malignant

4. **Support Vector Machine (SVM): To** improve security and service quality, this method uses supervised machine learning for pattern recognition. Support Vector Machines (SVMs) excel at classification tasks by creating a dividing line (hyperplane) that best separates different data groups. This line maximizes the distance between the data and the line, leading to strong classification accuracy. The SVM formula is:

$$f(x) = \sum_{i=1}^{N} a_i y_i K(x_i, x_j) + b$$

In the context, N represents the number of training samples, ai denotes the weights calculated during the training process, yi corresponds to the class label of the i training sample, $K(xi\ ,xj\ )$ is stands for kernel, and b is represent terms bias in.

Experimental Results.

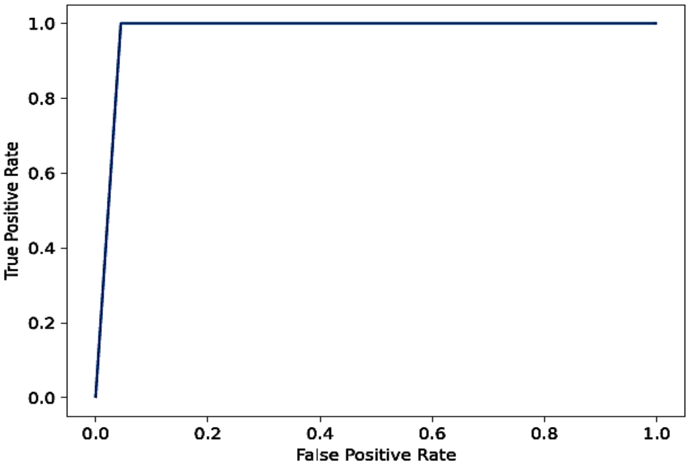
We apply four machine learning algorithms to the dataset to see how the models perform. We analyze models' performance based on accuracy, precision, and more (Table 3). We analyze and find that the Gaussian Naive Bayes model and Logistic Regression model outperform with 97% accuracy in our predictive analysis. The SVM model is the second-best performer with 96% accuracy and the optimized decision tree is the lowest performer with 95% accuracy.

Table 3. Performance each algorithm based on confusion matrix

ML algorithm	Overall accuracy - Train	Overall accuracy - Test	F1-Score Test	AUC score
Optimized DT	0.97	0.95	0.93	0.96
LR	0.98	0.97	0.95	0.96
Gussian NB	0.96	0.97	0.96	0.98
SVM	0.99	0.96	0.94	0.94

The area under the ROC curve or AUC-ROC measures a classification model's effectiveness and potential classification thresholds. The categorization threshold changes from 0 to 1, illustrating the trade-off between genuine and false positive rates. A perfect model has an AUC-ROC value of 1, whereas a mediocre model has an AUC-ROC value of 0.5 [15]. The ROC curves in Figures 3 and 8 demonstrate the relationship between the true positive rate and the false positive rate. In both the figures, since curves of Gaussian Naive Bayes, Logistic Regression model, Naïve Bayes, and logistic regression are closely following the left and the top border of ROC space, it can be said that these classifiers are comparatively more accurate than decision tree and SVM for the data set under study for this research

وتم نشرها على الموقع بتاريخ: 30/ 2024/10م
تم استلام الورقة بتاريخ:27/ 2024/9م



Figures 3.Gaussian Naive Bayes AUC score: 0.977

Conclusion and Future Work

This study investigated the effectiveness of various supervised machine learning algorithms as valuable tools in important medical applications such as breast cancer classification. The analysis revealed that both Gaussian Naive Bayes (GNB) and Logistic Regression.

(LR) achieved high accuracy, reaching 97%. However, GNV demonstrated a clear advantage based on the Receiver Operating Characteristic (ROC) curve. This suggests that GNB is superior in differentiating between positive and negative cases.

This superiority can be further explained by analyzing the confusion matrix. GNB achieved a false negative (FN) value of 0, indicating no instances were missed classified as positive (cancerous). also Minimizing FP is critical in breast cancer diagnosis, as it ensures the model doesn't mistakenly identify healthy patients as having cancer.

Therefore, Gaussian Naive Bayes is the most effective algorithm among those tested.

Future research We recommend studying a database with a large number of cases and processing the features using (PCA) tools , fine-tuning the model's hyperparameters, and comparing with different machine learning techniques may lead to further progress in this field.

تم استلام الورقة بتاريخ:27/ 9/2024م وتم نشرها على الموقع بتاريخ: 30/ 10/2024م

References

[1] WHO, “WHO launches new roadmap on breast cancer,” *World Health Organization*, 2023. https://www.who.int/news/item/03-02-2023-who-launches-new-roadmap-on-breast-cancer (accessed Jan. 20, 2023).

[2] A. A. Abdul Halim et al., “Existing and emerging breast cancer detection technologies and its challenges: a review,” Applied Sciences, vol. 11, no. 22, Nov. 2021, doi: 10.3390/app112210753.

[3] H. Aljuaid, N. Alturki, N. Alsubaie, L. Cavallaro, and A. Liotta, “Computer Methods and Programs in Biomedicine Computer-aided diagnosis for breast cancer classification using deep neural networks and transfer learning,” Comput. Methods Programs Biomed., vol. 223, p. 106951, 2022, https://doi.org/10.1016/j.cmpb.2022.106951.

[4] H. U. A. Chen, K. Mei, Y. Zhou, N. A. N. Wang, and G. Cai, “Auxiliary Diagnosis of Breast Cancer Based on Machine Learning and Hybrid Strategy,” IEEE Access, vol. 11, pp. 96374-96386, 2023, https://doi.org/10.1109/ACCESS.2023.3312305.

[5] A. S. Elkorany, M. Marey, K. M. Almustafa, and Z. F. Elsharkawy, “Breast Cancer Diagnosis Using Support Vector Machines Optimized by Whale Optimization and Dragonfly Algorithms,” IEEE Access, vol. 10, pp. 69688–69699, 2022, https://doi.org/10.1109/ACCESS.2022.3186021.

[6] K. Amril M.Siregar, S.Faisal” Optimized Machine Learning Performance with Feature Selection for Breast Cancer Disease Classification”, Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI) Vol. 9, No. 4, December 2023, pp. 1131-1143 ISSN: 2338-3070, DOI: 10.26555/jiteki.v9i4.27527.

[7] R.Hridoy, A.Arni, S.Ghosh, N.Chakraborty, Imran Mahmud,” Performance enhancement of machine learning algorithm for breast cancer diagnosis using hyperparameter optimization”, International Journal of Electrical and Computer Engineering (IJECE) Vol. 14, No. 2, April 2024, pp. 2181~2190 ISSN: 2088-8708, DOI: 10.11591/ijece.v14i2.pp2181-2190

[8] M.Awan, M.Arif, M.Ul Abideen, Kamaleldin Abodayeh, ” Comparative analysis of machine learning models for breast cancer prediction and diagnosis: A dual-dataset approach”, Indonesian Journal of Electrical Engineering and Computer Science Vol. 34, No. 3, June 2024, pp. 2032~2044 ISSN: 2502-4752, DOI: 10.11591/ijeecs.v34.i3.pp2032-2044

[9] J. Purnomo, D.Augustina Pratiwi,” Breast Cancer Classification Procedure Using Machine Learning Techniques” , BIO Web of Conferences 117, 01029 (2024) ,https://doi.org/10.1051/bioconf/202411701029 ICoLiST 2023.

[10]A.Bansal, D.Arora, K.Soni, R.Chugh, S.Vardhan,” Breast Cancer Classification Using Machine Learning”, International Journal of Scientific Research in Computer Science, Engineering and Information Technology, doi : https://doi.org/10.32628/CSEIT2410274,2024.

تم استلام الورقة بتاريخ:27/ 9/2024م   وتم نشرها على الموقع بتاريخ: 30/ 10/2024م

[11] L. B. V. de Amorim, G. D. C. Cavalcanti, and R. M. O. Cruz, "The choice of scaling technique matters for classification performance," Applied Soft Computing, vol. 133, Jan. 2023, doi: 10.1016/j.asoc.2022.109924.

[12] N. S. Alfaiz and S. M. Fati, "Enhanced Credit Card Fraud Detection Model Using Machine Learning," Electronics (Switzerland), vol. 11, no. 4, p. 662, Feb. 2022, doi: 10.3390/electronics11040662.

[13] O. Karasoy and S. Ballı, "Spam SMS Detection for Turkish Language with Deep Text Analysis and Deep Learning Methods," Arabian Journal for Science and Engineering, vol. 47, no. 8, pp. 9361–9377, Aug. 2022, doi: 10.1007/s13369-021-06187-1.

[14] F. S. de Menezes, G. R. Liska, M. A. Cirillo, and M. J. F. Vivanco, "Data classification with binary response through the Boosting algorithm and logistic regression," Expert Systems with Applications, vol. 69, pp. 62–73, Mar. 2017, doi: 10.1016/j.eswa.2016.08.014.

[15] R. E. Whisnant, "A Novel Data Analytics-derived Metric (Nearest Cluster Distance) Is Easily Implemented in Routine Practice and Correctly Identifies Breast Cancer Cases for Quality Review," J. Pathol. Inform., vol. 13, p. 100005, 2022, https://doi.org/10.1016/j.jpi.2022.100005.

[16] M. Javaid, A. Haleem, R. Pratap Singh, R. Suman, and S. Rab, "Significance of machine learning in healthcare: Features, pillars and applications," International Journal of Intelligent Networks, vol. 3, pp. 58–73, 2022, doi: 10.1016/j.ijin.2022.05.002.