Accuracy Analysis of Supervised and Unsupervised Techniques on Breast Cancer Datasets

Elham Jahanpeikar, Mahdi Ebrahimi
Computer Science Department
CSU Northridge
Northridge, United States
elham.jahanpeikar.889@my.csun.edu, mahdi.ebrahimi@csun.edu

Abstract— **Modern medicine benefits extensively from technology as a valuable assistant to diagnose diseases at early stages and cure them with less complication and more success. In the health sector, cancer diseases undeniably, possess the most variety and complexity of all diseases. Therefore, this paper focuses on breast cancer and applies supervised, unsupervised, and deep learning as one candidate from each major machine learning field to evaluate different models and analyze their outcomes. We aim to answer the following questions: Which dataset produces the best result and improves its metrics and performance using machine learning optimization techniques? Will there be any definite decision on which machine learning models or algorithms have superiority over the others? The original and diagnostic datasets are trained by Logistic Regression, K-means clustering, and multi-layer perceptron or artificial neural network model (ANN). Then, unsupervised techniques, heatmaps, and Principal Component Analysis (PCA) are used to reduce dimensionality and concise the dataset for any probable improvements. The original dataset produced better results for the machine learning models, and ANN obtained the best accuracy score. The comprehensive and systematic calculation of the metrics and indexes of the breast cancer datasets and the thorough optimization by unsupervised technics is the novelty of this research. The comparison between these two datasets has not been approached before. The clustering by K-means creates novel visualization of the datasets, which could give the experts in the field ideas of the cancerous mass's characteristics.**

Keywords—Supervised machine learning, Unsupervised machine learning, Accuracy analysis, Observation, Breast Cancer

I. INTRODUCTION

Machine learning and Neural Networks are two central subfields of artificial intelligence (AI). One way to use AI is machine learning. In the 1950s ML was defined by AI pioneer Arthur Samuel as an area of study that gives computers the ability to learn without explicitly being programmed [16]. There are three subdivisions of machine learning: supervised, unsupervised, and reinforcement. In supervised methods, models will be trained with labels, and the model's accuracy grows over time. Unsupervised learning could find patterns or trends without human intervention after implementing the models. The reinforcement, specifically Neural networks, mimic intelligent human behavior. Artificial intelligence systems, like ANN, perform their tasks like the way human brains solve problems [16]. The market for artificial intelligence in health care and the life sciences is projected to grow by 40 percent a year, to $6.6 billion in 2021, according to estimates from Frost & Sullivan. Accuracy in the health sector, is as important as the market benefit. Dr. Andy Beck, a pathologist at Harvard Medical School and Beth Israel Deaconess Medical Center, and Aditya Khosla, a computer scientist trained at MIT and Caltech, are challenging cancer diagnosis through images. They formed a startup named PathAI in Cambridge, Massachusetts, after their win in a competition in detecting breast cancer. Beck mentioned that earlier in a challenge, an expert pathologist did the same task as the computational teams, achieved an error rate of about 3.5 percent. The error rate they achieved was closer to 7.5 percent, which was the winner in the competition. Beck believed that putting the computer and pathology together was the most interesting part of the experience[17]. “The combination of human plus AI in this example reduced the expert’s error rate by 85 percent,” Beck said. “That was a really exciting finding”[17].

The importance of saving lives and detecting breast cancer early on is so significant that we decided to do a comprehensive analysis of machine learning detection of breast cancer.

The following research challenges are the reasons that we chose to approach our study: 1) the gap of implementation on the original dataset, 2) the absence of K-means clustering models and evaluating its performance indexes, and 3) comparing the two existing breast cancer datasets to find a probable superiority of one dataset over the other one.

Machine learning methods have never been applied to both the original and diagnostic datasets in one place. Therefore, this paper compares the metrics of two available breast cancer datasets, original and diagnostic. We create one model for each major category of machine learning: supervised, unsupervised, and deep learning. In addition, Logistic Regression, K-means clustering, and deep learning model’s metrics and indexes are extensively analyzed.

II. RELATED WORK

Machine learning models are present widely in research. In [5], Logistic Regression provided the best scores in almost all metrics: precision, recall, accuracy, and f1 score, with ML models: Naïve Bayes, K-Nearest Neighbors, and Support Vector Machines. Their research concluded that the Support Vector Machine outperformed all other classifiers and achieved the highest accuracy (97.2%) when this accuracy was measured for Support Vector Machine (SVM), Random Forest, Logistic

979-8-3503-2028-2/22/$31.00 ©2022 IEEE
DOI 10.1109/CSCI58124.2022.00168

Regression, Decision tree (C4.5) and K-Nearest Neighbors (KNN) [6]. In [7], ensemble learning outperforms Logistic Regression with an accuracy of 97.90. The models in ensemble learning are Logistic Regression, K-Nearest Neighbor, Linear Discriminant Analysis, Support Vector Classifier, and Random Forest Classifier. Another study claims that the Decision Tree is the winner between the two training models of the Decision Tree Classifier and Logistic Regression [8]. It could be concluded there is no single machine learning model that all papers agree on its performance.

K-means clustering was almost absent from the research and studies. An unverified reason is that because the K-means clustering analysis does not provide accuracy metrics, the researchers are avoiding this model altogether. Looking at deep learning models, in [9], the authors claim the DNN classifier had a great performance in accuracy (92%) when Multilayer Perceptron, Decision Tree, Random Forest, Support Vector Machine, and Deep Neural Network are used for evaluation. Similar to the previous paper, the artificial neural network gives better prediction: 97.85%, by ML models: Support Vector Machine, K-Nearest Neighbor, Naïve Bayes, Decision Tree, and K-means [10].

The K-means prediction raises the question of how a clustering analysis, whose only prediction could be predicting the cluster a new data entry belongs to, was compared with the other models. This prediction could not be compared with the other algorithms' predictions, which their outcomes are metrics like precision and f1 score and utterly different from K-means clustering prediction.

The related works to the PCA and dimension reduction methods are as follows: when applying Principal Components, the accuracies surpass 99% across the machine learning models [11]. A paper result shows that the best model is the random forest classifier which achieved the best accuracy when the number of features was reduced in the Wisconsin Diagnostic dataset[12]. The most relevant to the present work is that the researcher has tested different features to obtain better accuracy through feature elimination or noise reduction and focuses on the Logistic Regression algorithm [13].

The paper claims the accuracy of the Logistic Regression model could vary from 90.48% to 96.5% depending on which features will be used for evaluation and making the model. Among all the papers related to the breast cancer machine learning study, some papers were unclear about whether the dataset was original or diagnostic. Some papers mentioned the dataset, and they were original datasets, and the rest of the papers used the diagnostic dataset. None of these papers had machine learning in common with what this paper is about unless using cancer datasets [14][15].

III. Methodology

A. Datasets

The Wisconsin original dataset contains 699 rows and 11 columns. The Wisconsin diagnosis dataset contains 569 rows and 33 columns [2].

B. Models

1) Logistic Regression

Logistic Regression is one of the supervised classification algorithms in machine learning. It predicts the categorical dependent variables using a given set of independent variables. In other words, the statistical method predicts the outcome of a dependent variable based on previous observation. This predictive analysis algorithm is based on the concept of probability. The following techniques will be applied to both original and diagnostic datasets, and the accuracy, f1 score, and the number of incorrect prediction(s) will be obtained. Logistic Regression model will be utilized to the datasets. An optimization method, named hyperparameter tunning, will be used to optimize the mentioned metrics. A dimension reduction method, heatmaps, will be used to obtain any improvement in the metrics. Variance inflation factor (VIF) and Principal Component Analysis (PCA) will be used to see any possible change in the accuracy and f1 scores. Tables 1 and 2 illustrate the confusion matrix of both datasets after optimizations.

2) K-means clustering

One of the unsupervised machine learning algorithms is K-means clustering. This technique assigns the objects or data points into clusters that have similarities. The 'K' in K-means clustering is the number of such clusters [3]. The system needs to get K as the number of the cluster. For example, K = 3 refers to three clusters.

To train the datasets with the K-means model, a random number of K will be selected, then, with the help of the elbow method, the optimum number of clusters will be determined (Fig. 1) and the inertia value, which shows the quality of the clusters will be calculated. The performance indexes for K-means clustering will be measured. These indexes will be compared to the dimensionally reduced dataset performance indexes, and the result will show the effectiveness of the optimization techniques. The performance indexes are Rand, Adjusted Rand, Mutual Information, Calinski – Harabasz, and Davies - Bouldin Index.

After dimension reduction, the clusters of the datasets (Fig. 2 and 3) illustrate the difference between the original and diagnostic datasets.

TABLE 1: The Original Dataset Confusion Matrix

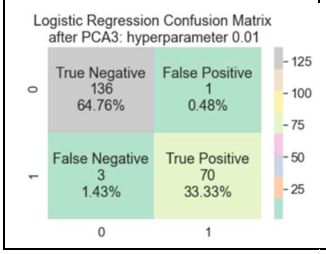
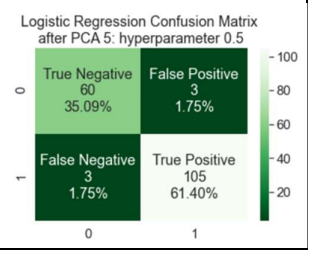


TABLE 2: The Diagnostic Dataset Confusion Matrix



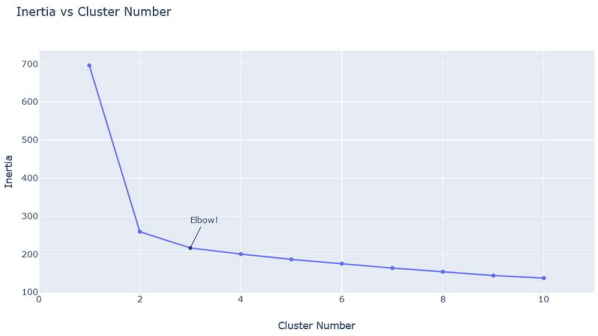


Fig. 1: Elbow Plot

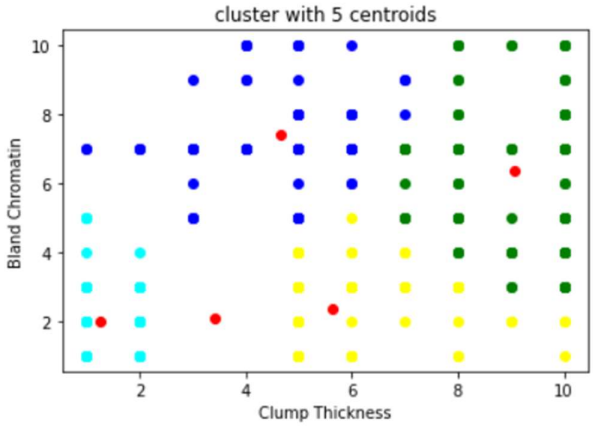


Fig. 2: Original reduced dataset scatter plot with five clusters

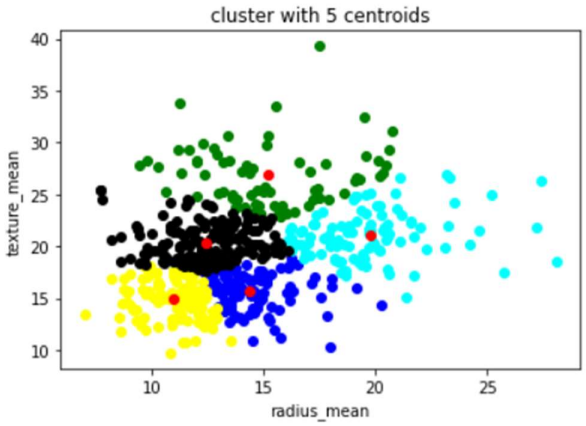


Fig. 3: Diagnostic reduced dataset scatter plot with five clusters

3) Deep Learning - Artificial Neural Network

Automation in industries gets the advantage of deep learning in various services and applications. A few instances of deep learning models that make human intervention and analysis less needed are voice-enabled TV remotes, credit card fraud detection, and evolving technologies such as self-driving vehicles. [4].

Artificial Neural Network (ANN) is one method of implementing deep learning that simulates the human brain and could consist of three or more layers. The neural networks in ANN models mimic the behavior of the human brain by letting the model learn from substantial amounts of data. A neural network with only one layer is capable of predictions, but extra hidden layers will optimize and enhance the accuracy [4].

The mechanism of deep learning, in a concise explanation, is that the deep learning will achieve its prediction and precision through the processes of gradient descent and backpropagation and adjusts and fits itself for accuracy. Data preprocessing has been done more thoroughly for the deep learning model in this research. A close look at the datasets shows that size of the datasets is small. Datasets have missing and unbalanced data, and various ranges of data values exist. Replacing the missing data with zeros (the number of missing data is not significant compared to the whole dataset), balancing the number of malignant and benign labels by normal distribution random noise generation to the existing data samples, and using the same method for making the dataset larger, and also scaling the larger dataset by Numpy sqrt function (to address the data skewness) will produce a balanced, augmented, scaled dataset, ready to be fed to the artificial neural network model.

Although we have Logistic Regression model to compare the deep learning model to, the scikit-learn "Random Forest Classifier" is also used to set a benchmark for comparing its result with the deep learning model. Random Forest Classifier score is 0.978 for original dataset.

The open-source Keras deep learning library is applied to implement the new deep learning model. The execution begins with defining a function that creates the new model instances. This function is reusable. The model will be trained. To validate the robustness of the deep learning model, a 10-fold cross-validation will be performed [3]. The model is tested. The accuracy score for various optimizers for the original dataset are as follows:

SGD Optimizer	Adam Optimizer	RMSProp Optimizer
accuracy: 99.29%	accuracy: 99.29%	accuracy: 99.29%
accuracy: 98.58%	accuracy: 98.58%	accuracy: 98.58%
accuracy: 97.87%	accuracy: 97.87%	accuracy: 97.87%
accuracy: 97.16%	accuracy: 99.29%	accuracy: 97.87%
accuracy: 95.04%	accuracy: 94.33%	accuracy: 92.91%
accuracy: 97.16%	accuracy: 97.87%	accuracy: 96.45%
accuracy: 98.58%	accuracy: 98.58%	accuracy: 97.16%
accuracy: 97.87%	accuracy: 98.58%	accuracy: 99.29%
accuracy: 98.58%	accuracy: 97.87%	accuracy: 98.58%
accuracy: 97.16%	accuracy: 97.87%	accuracy: 97.16%
97.73% (+/- 1.13%)	98.01% (+/- 1.34%)	97.52% (+/- 1.77%)

The number of epochs (when an entire dataset is passed forward and backward through the neural network only once) and model loss (loss = 0 is desirable) is shown in Fig. 4.
Fig. 5, with the help of the ROC curve, illustrates a graph of sensitivity over specificity. For both datasets, hyperparameter refinement will be considered and adjusted to obtain the optimum value for the depth of the network, dropout, batch size, and epochs.

IV. Experimental Results and Discussion

A. Logistic Regression

The Accuracy and f1 score for Logistic Regression on the original and diagnostic datasets are summarized in tables 3 and 4.

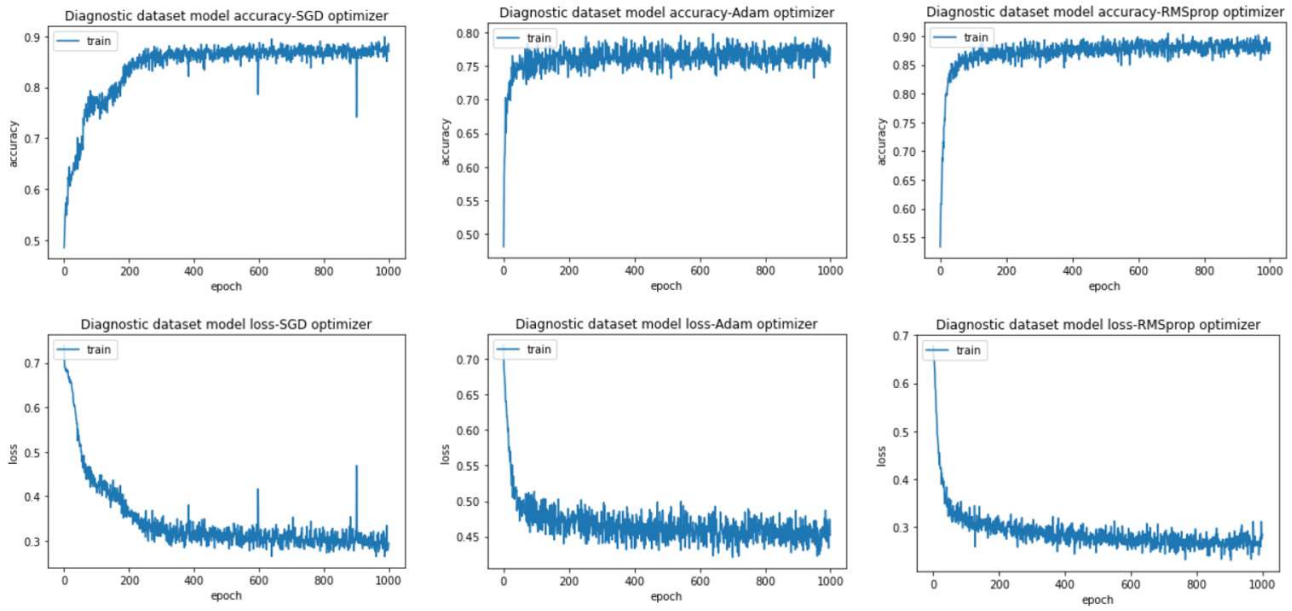


Fig. 4: The original dataset accuracy and loss relation with epochs

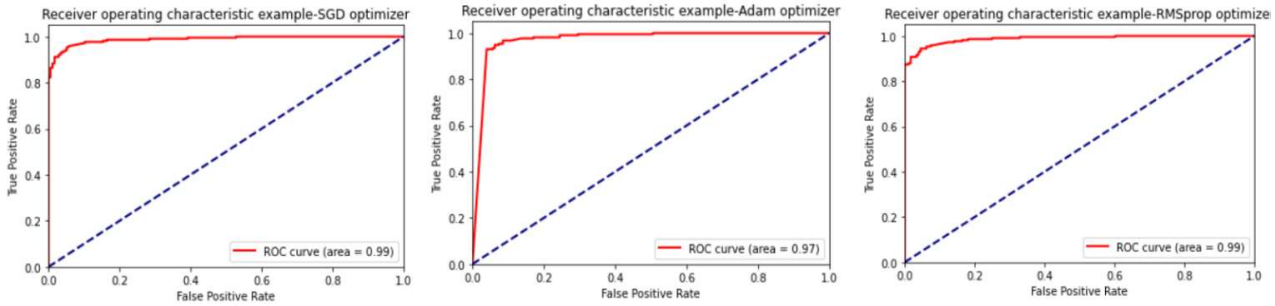


Fig. 5: The original dataset ROC Curve

TABLE 3: THE ORIGINAL DATASET F1 SCORE AND ACCURACY

	Logistic Regression F1 Score					
Wisconsin Breast Cancer Original Dataset	Base model	After Hyperparameter tuning	after Heatmap	After heatmap and hyperparameter tuning	After PCA	After PCA and hyperparameter tuning
	95.8%	95.8%	77.1%	77.2%	95.8 %	96.2%
	Logistic Regression Accuracy					
	97.1%	97.1%	86.1%	86.1%	97.1 %	98.0%

TABLE 4: THE DIAGNOSTIC DATASET F1 SCORE AND ACCURACY

	Logistic Regression F1 Score					
Wisconsin Breast Cancer Diagnostic Dataset	Base model	After Hyperparameter tuning	after Heatmap	After heatmap and hyperparameter tuning	After PCA	After PCA and hyperparameter tuning
	93.7%	97.6%	94.4%	97.6%	96.2%	98.3%
	Logistic Regression Accuracy					
	95.3%	96.4%	95.9%	95.9%	95.3%	96.4%

B. K-Means Clustering

The performance indexes of the original and the diagnostic datasets are summarized in tables 5 and 6 by Considering the smallest inertia value and its relative number of clusters.

C. Deep Learning Model

Applying 10-fold cross-validation, best, worst, and average scores of the ANN model by various optimizers are shown in tables 7 and 8 for the original and the diagnostic datasets.

TABLE 5: THE ORIGINAL DATASET PERFORMANCE INDEXES AND INERTIA

Wisconsin Breast Cancer Original Dataset						
	K-Means Performance Indexes Before Dimension Reduction by Heatmaps					
	Rand	Adjusted Rand	Mutual Information	Calinski-Harabasz	Davis-Bouldin	Best Inertia
	0.941	0.793	0.509	1264.788	0.684	(K = 5) 2598.982
	K-Means Performance Indexes After Dimension Reduction by Heatmaps					
	Rand	Adjusted Rand	Mutual Information	Calinski-Harabasz	Davis-Bouldin	Best inertia
	0.988	0.946	0.610	770.071	0.889	(K = 2) 277.419

TABLE 6: THE DIAGNOSTIC DATASET PERFORMANCE INDEXES AND INERTIA

Wisconsin Breast Cancer Diagnostic Dataset						
	K-Means Performance Indexes Before Dimension Reduction by Heatmaps					
	Rand	Adjusted Rand	Mutual Information	Calinski-Harabasz	Davis-Bouldin	Best Inertia
	0.869	0.634	0.410	253.382	1.279	(K = 5) 9262.165
	K-Means Performance Indexes After Dimension Reduction by Heatmaps					
	Rand	Adjusted Rand	Mutual Information	Calinski-Harabasz	Davis-Bouldin	Best inertia
	0.872	0.644	0.416	212.279	1.390	(K = 2) 6969.793

TABLE 7: THE ORIGINAL DATASETIEDDEEP LEARNING MODEL ACCURACY

Wisconsin Breast Cancer Original Dataset	10 fold cross validations			Optimizers					
				SGD		Adam		RMSprop	
	Best Accuracy	Worst Accuracy	Average Accuracy	Accuracy	loss	Accuracy	loss	Accuracy	loss
	99.29%	92.91%	98.01%	97.2%	1.61%	97.8%	7.5%	97.6%	9.21%

TABLE 8: THE DIAGNOSTIC DATASET DEEP LEARNING MODELS ACCURACY

Wisconsin Breast Cancer Diagnostic Dataset	10 fold cross validations			Optimizers					
				SGD		Adam		RMSprop	
	Best Accuracy	Worst Accuracy	Average Accuracy	Accuracy	loss	Accuracy	Loss	Accuracy	Loss
	98.31%	91.45%	95.56%	92.8%	16.6%	93.3%	22.4%	94.3%	16.5%

V. CONCLUSION

In this paper, we applied supervised, unsupervised, and deep learning to evaluate different models and analyze their outcomes by focusing on breast cancer. The diagnostic dataset had a better value for the f1 score than the original dataset in the logistic regression model. Considering the f1 score, the diagnostic dataset performed better, while the accuracy score deemed the original dataset superior.

The smallest inertia value belonged to the original dataset with two clusters after dimension reduction. The original dataset produced better results than the diagnostic dataset for performance indexes of clustering models before and after dimension reduction. Therefore, the original dataset had better outcomes for the clustering model.

For the deep learning model, the average accuracy of the original dataset was higher than the diagnostic dataset. Adam optimizer produced the best accuracy for the original dataset, and RMS prop made the best accuracy for the diagnostic dataset. Overall, the original dataset performed better for the deep learning model with the Adam optimizer.

Considering the best score, the deep learning model obtained a better score. However, using the average score for comparison resulted in a tie for the original dataset between the two models, while Logistic Regression won for the diagnostic dataset.

The scope of this research was not limited to announcing a winner of the machine learning models, as the related papers have done. If the above was the goal, no particular result could be obtained, as it could be seen that the models on the related works have obtained contradictory results. In some papers, Logistic Regression is the winner and in others is not. Also, machine learning models in supervised and unsupervised areas produce different outcomes and results that are not comparable.

Therefore, the present work thoroughly analyzes the three models: Logistic Regression, K-means clustering, and Deep Learning on two breast cancer datasets. This work's outcome is visualizing the datasets and clusters, calculating the metrics and performance indexes, and optimization.

The comparison between two datasets of one area of the health field, breast cancer, is the superiority of this paper to the existing research. This could give researchers ideas of what to look for when selecting a dataset to get a meaningful result. This thesis sheds light on what makes a dataset more practical than others.

Future work could focus more on optimizing the heatmaps and working with a different number of PCAs. With a combination of PCA and choosing the right features, future work could result in different and likely better accuracy scores and performance indexes.

VI. REFERENCES

[1] Y. Zhang, "Deep Learning in Wisconsin Breast Cancer Diagnosis", https://towardsdatascience.com/deep-learning-in-winonsin-breast-cancer-diagnosis-6bab13838abd (accessed Nov. 13, 2022).

[2] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: Uni[18versity of California, School of Information and Computer Science.

[3] S. Lloyd, "Least squares quantization in PCM," in IEEE Transactions on Information Theory, vol. 28, no. 2, pp. 129-137, March 1982, doi: 10.1109/TIT.1982.1056489.

[4] IBM Cloud Learn Hub, https://www.ibm.com/cloud/learn/cloud (accessed Nov. 23, 2022).

[5] V.P.C. Magboo, and M.S.A. Magboo, "Machine Learning Classifiers on Breast Cancer Recurrences." Procedia Computer Science, 192, pp.2742-2752, 2021.

[6] M. A. Naji, S. El Filalib, K. Aarikac, E. H. Benlahmard, R. A. Abdelouhahi, O. Debauche, "Machine Learning Algorithms for Breast Cancer Prediction And Diagnosis", International Workshop on Edge IA-IoT for Smart Agriculture (SA2IOT) August 9-12, 2021, Leuven, Belgium.

[7] R. MurtiRawat, S. Panchal, V. K. Singh, Y. Panchal "Breast Cancer Detection Using K-Nearest Neighbors, Logistic Regression and Ensemble Learning", in the Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC) 2020.

[8] P. P. Sengar, M. J. Gaikwad, A. S. Nagdive, "Comparative Study of Machine Learning Algorithms for Breast Cancer Prediction", in the Proceedings of the Third International Conference on Smart Systems and Inventive Technology (ICSST), 2020.

[9] F. Teixeira; J. L. Z. Montenegro; Cristiano André da Costa; Rodrigo da Rosa Rigi, "An Analysis of Machine Learning Classifiers in Breast Cancer Diagnosis", XLV Latin American Computing Conference (CLEI), 2019.

[10] T. Thomas, N. Pradhan, V. S.h Dhaka, "Comparative Analysis to Predict Breast Cancer using Machine Learning Algorithms: A Survey", in the Proceedings of the Fifth International Conference on Inventive Computation Technologies (ICICT), 2020.

[11] A. Ahuja, L. Al-Zogbi, A. Krieger, "Application of noise-reduction techniques to machine learning algorithms for breast cancer tumor identification", National library of medicine, national center for biotechnology information.

[12] N. F. Omran, S. F. Abd-el Ghany, H. Saleh, A. Nabil and A. M. Khalil, "Breast cancer identification from patients' tweet streaming using machine learning solution on spark", Complexity, vol. 2021, pp. 12, Jan. 2021.

[13] L. Liu, "Research on Logistic Regression Algorithm of Breast Cancer Diagnose Data by Machine Learning," 2018 International Conference on Robots & Intelligent System (ICRIS), 2018, pp. 157-160, doi: 10.1109/ICRIS.2018.00049.

[14] W. Yue, Z. Wang, H. Chen, A. Payne, and X. Liu, "Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis", Department of Computer Science, Brunel University London, Uxbridge, Middlesex UB8 3PH, UK, School of Mathematics, Southeast University, Nanjing 210096, China.

[15] E. A. Bayrak, P. Kırcı, T. Ensari, "Comparison of Machine Learning Methods for Breast Cancer Diagnosis", Scientific Meeting on Electrical–Electronics & Biomedical Engineering and Computer Science (EBBT), 2019.

[16] MIT management Sloan school, machine learning explained, https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained (accessed Nov. 10, 2022).

[17] From coding to cancer: How AI is changing medicine, modern medicine CNBC. https://www.cnbc.com/2017/05/11/from-coding-to-cancer-how-ai-is-changing-medicine.html (accessed Nov. 23, 2022).