Journal of Theoretical and Applied Information Technology
31st July 2022. Vol.100. No 14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN EFFICIENT OPTIMIZED FRAMEWORK FOR ANALYZING THE PERFORMANCE OF BREAST CANCER USING MACHINE LEARNING ALGORITHMS

MAGDY ABD-ELGHANY ZEID [1], KHALED EL-BAHNASY [2], S.E.ABU-YOUSSEF[3]

[1,2] Computer Science department Obour High Institute for Management and Informatics, Cairo, Egypt.

[3] Mathematics and Computer Science Department Faculty of Science Al-Azhar University, Cairo, Egypt.

Email: [1]magdy_zeid83@oi.edu.eg, [2]khaled.bahnasy@oi.edu.eg, [3]abuyousf@hotmail.com

ABSTRACT

Breast cancer is a significant issue for women worldwide and a leading cause of death. This disease can be detected by differentiating malignant and benign tumors. As a result, physicians require a dependable diagnostic process for differentiating malignant from benign tumors. So automated detection of tumors is required. This research aims to introduce an optimized framework for identifying breast cancer types and predicting breast cancer recurrence using Seven Machine Learning algorithms: Logistic Regression (LR), eXtreme Gradient Boost (XGboost), Multi-Layer Perception (MLP) of Neural Network, Naive Bayes (NB), Random Forest (RF), K-Nearest neighbor (KNN) and Decision Tree (DT). We use Grid Search to optimize the machine learning algorithms. The performance of the framework was compared to determine which classifier performs the best on the Wisconsin datasets as follows Wisconsin Breast cancer (WBC) dataset, Wisconsin Diagnosis Breast cancer (WDBC) dataset, and Wisconsin Prognosis Breast Cancer (WPBC) data set. Our work presents a significant increase in cancer prediction accuracy, with the highest value being 98.3 % in the WBC dataset, 99.2% in the WDBC dataset, and 78.6% of accuracy in the WPBC dataset for cancer recurrence prediction. These results show significant progress in the area of breast cancer classification and recurrence prediction as compared to the existing state of art results of baseline machine learning models.

Keywords: *Breast Cancer, Machine Learning, Classification algorithms*

1. INTRODUCTION

Cancer kills one in every six people worldwide. Cancer is the world's second leading cause of death, expected to claim approximately 10 million lives in 2020ly 70% of cancer deaths occur in low- and middle-income countries [1]. According to the World Health Organization (WHO), breast cancer is the most common type of cancer in women in Egypt, accounting for 32.4 percent of cancers in this population, with nearly 22,038 cases estimated in 2020 [2] and expected to reach approximately 46,000 in 2050 [3].

Cancer research has made tremendous strides in the last few decades. Breast cancer is becoming more prevalent, and numerous researchers have examined various treatment options. Experts are developing new methods for detecting and predicting cancer early. Image processing and machine learning techniques determine the cancer stage in a particular patient before the onset of symptoms. One of the most difficult tasks facing researchers is accurate cancer prediction with new detection techniques. This cancer is caused by the uncontrolled and rapid growth of benign and malignant breast tissues.

On the other hand, benign breast tissue abnormalities do not always result in death; on the other hand, malignant breast tissue is a type of tumour tissue. Its early detection can significantly increase the patient's mortality rate. The accuracy of classification techniques is calculated as the proportion of correctly classified test sets [4].

Numerous studies have attempted to determine the survivability of carcinoma in humans using machine learning approaches. As a result, an automatic method for detecting tumours has also

demonstrated that these techniques are more effective at diagnosing carcinoma [5].

Machine learning is critical for breast cancer classification. Artificial intelligence is a subfield of machine learning. Numerous developers retrain existing models and optimize performance using machine learning. Three distinct machine learning techniques are used to train the model. With the assistance of a supervisor, supervised machine learning operates on known data. Without supervision, unsupervised machine learning is carried out. Machine learning with reinforcement is losing popularity. These algorithms mine the most useful data from prior knowledge to make the best decisions possible. Machine learning (ML) techniques automate data analysis and extract key relationships and datasets. Additionally, it generates a computational model that best fits the data. According to cancer research, machine learning techniques can aid in the early detection and diagnosis of cancer [5].

Early breast cancer detection improves treatment outcomes and potentially saves lots of lives. In addition, cancer recurrence probability prediction is vital for patient follow-up and treatment. The detection and prognosis of breast cancer are challenging and require expertise and time. Therefore, to save time and reduce the chance of human error, automation procedures are required. As mentioned later in related work section, many studies were performed for this purpose applying different ML algorethms [8-21], as shown in table 1. This research aims to introduce an efficient optimized framework for identifying breast cancer types and predicting breast cancer recurrence using various Machine Learning approaches.

The paper employs a variety of machine learning techniques, including Extreme Gradient Boost (XGboost), Multi-Layer Perception (MLP) of Neural Networks, Random Forest, Naive Bayes (NB), and Instance-Based for K-Nearest Neighbor (KNN). This technique applies to the Wisconsin diagnosis Breast Cancer dataset (WDBC) [6].

The paper contributions of the proposed approches can be summarized as follows:

- ✓ We proposed a novel breast cancer Detection system using optimized machine learning algorithms
- ✓ Improving the accuracy of existing breast cancer detection using optimized machine learning algorithms
- ✓ We optimized the results of the machine learning Algorithms with hyper-parameters optimization methods.
- ✓ We perform a comparison-based study with other cutting-edge machine learning techniques for ensuring the results of the proposed framework
- ✓ The proposed framework Achieved 99.2% accuracy in breast cancer Detection on the WDBC dataset.

The remaining sections of this work are structured as follows: The related work is described in Section 2. In section 3, the proposed methodology is outlined. The results and discussion of the experiment are reported in section 4. Section 5 concludes with a presentation of the conclusions.

2. RELATED WORK

Many studies have been published in the literature describing breast cancer detection. Several machine learning algorithms have been created to extract knowledge from databases, including supervised learning techniques. These algorithms are most frequently used for the categorization the breast cancer detection. This section summarizes various recent studies on this problem.

In 2017 Nilashi et al. [8] applied Classification and Regression Trees (CART) to the WBC dataset in order to produce fuzzy rules for the categorization of breast cancer disease in a knowledge-based system employing fuzzy rule-based reasoning. They achieve 93.20% accuracy.

To construct prediction models, Chaurasia et al. (2018) [9] used the WBC dataset and three algorithms (RBF Network, Naive Bayes, and J48). The results revealed that Nave Bayes is the most accurate predictor with 97.36 % accuracy on the holdout sample, followed by RBF Network with 96.77 % accuracy and J48 with 93.41 % accuracy.

Wang et al. [10] reported in 2020 a strategy based on a multilayer fuzzy expert system for the identification of breast cancer utilizing an extreme learning machine (ELM) classification model combined with a radial basis function (RBF) kernel termed ELM-RBF achieving 95.39% accuracy.

Table 1 summarizes the related work for breast cancer classification on Wisconsin datasets (WBC and WDBC) and WPBC for cancer recurrence prediction.

Table 1: Comparison Between Previous Work For Breast Cancer Detection Using Different Datasets

Dataset	Ref	Year	Classifiers	ACC
WBC	Nilashi [8]	2017	CART	93.20%
	Chaurasia et al. [9]	2018	NB	97.36%
	Wang et al. [10]	2020	RIPPER	95.39%
	Mojrian et al. [11]	2020	ELM-RBF	95.69%
	Bayrak et al. [12]	2019	SMO	96.90%
WDBC	Ramos et al. [13]	2019	LDA	98.82%
	Najmu et al. [14]	2020	DT	97.29%
	Sharma et al. [15]	2018	KNN	94.00%
	Rufai et al. [16]	2020	SVM	94.30%
	Salama et al. [17]	2013	SMO	97.7%.
WPBC	Ojha & Goel[18]	2017	SVM	68.00%
	Pritom et al. [19]	2017	SVM	75.70%
	Salama et al. [17]	2013	fusion of MLP, J48, SMO and IBK	77.00%
	Kiage [20]	2015	NB, KNN, RT	73.00%
	Chi et al. [21]	2007	ANN	64.90%

3. METHODOLOGY

Figure 1 demonstrates the suggested framework for Breast cancer prediction, which consists of six major steps:

- ✓ Data collection
- ✓ Data pre-processing
- ✓ Data partitioning
- ✓ Parameters optimization for ML algorithms
- ✓ Classification based on the proposed ML
- ✓ Metrics for prediction and evaluation.

3.1 Data Collection

Before gathering our data, we established many criteria. The Wisconsin Breast Cancer dataset (WBC), Wisconsin Diagnostic Breast Cancer dataset (WDBC), and Wisconsin Prognostic Breast Cancer dataset (WPBC) from the UCI Machine Learning Repository [7] were obtained from the University of Wisconsin Hospitals and are used by many researchers conducting breast cancer research. These datasets are tiny needles of data mass comprised of features taken from scanned images. Each feature corresponds to the visible cell nuclei characteristic in the image. The following table 2 provides an overview of various datasets. Each dataset contains examples or classification patterns with numerical features or attributes.

3.1.1 Wisconsin breast cancer dataset (WBC)

The Wisconsin Breast Cancer (Original) (WBC) dataset was used in this investigation. It contains 699 benign and malignant breast cancer cases. Additionally, the dataset has 11 attributes with integer values.

Each instance contains nine cytology features that quantify the exterior appearance and internal chromosomal alterations on nine scales. The nine characteristics are scored on an interval scale ranging from 1 to 10, with ten being the most eccentric. Each is kept as an ordinal data type (ordered set). As indicated in Table 3, the class attribute is of the flag type, with two states: 2 for benign and 4 for malignant. There are 458 benign cases (65.52 %) and 241 malignant cases (34.48 %).

Table 2: Wisconsin Dataset Description

Dataset name	WBC	WDBC	WPBC
Instances	699	569	198
Attributes	10	32	34
Attribute Type	Integer	Real	Real
Classes	Benign (B) and Malignant (M)	Benign (B) and Malignant (M)	Non-Recurrence (N) and Recurrence (R)
Classes distribution	B= 458 and M= 241	B= 357 and M= 212	N= 151 and R= 47
Missing Values	19	No missing value	4

Journal of Theoretical and Applied Information Technology
31st July 2022. Vol.100. No 14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

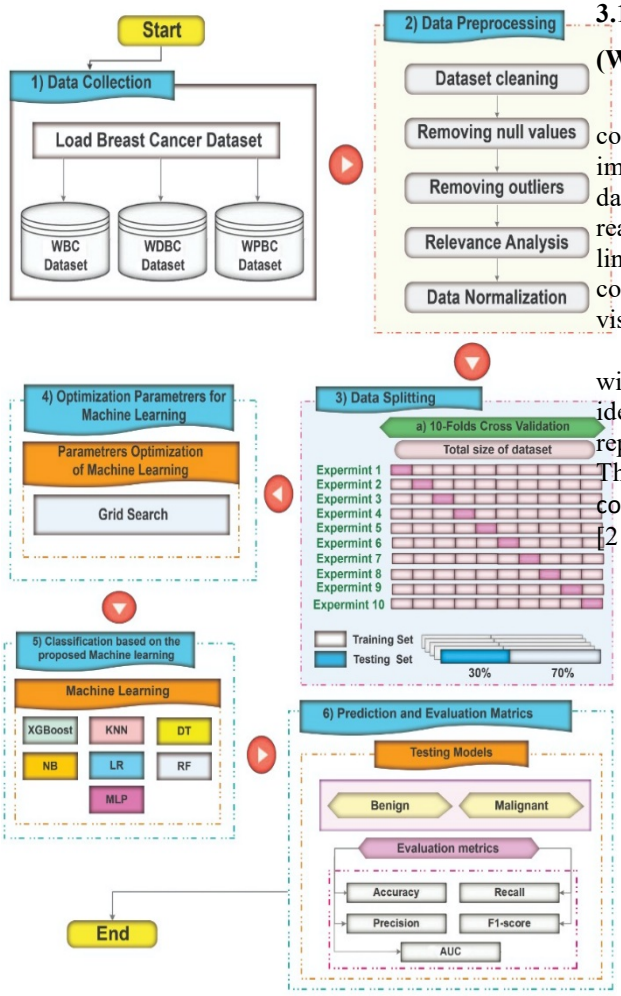

Figure. 1: The proposed system of breast cancer Model

Table 3: Wisconsin Breast Cancer Dataset (WBC) Features

NO.	Feature	Domain
1	instance code number	id number
2	Thickness of Clump	From one to ten
3	Size Uniformity	From one to ten
4	Shape Uniformity	From one to ten
5	Marginal Adhesion	From one to ten
6	Single Epithelial Cell Size	From one to ten
7	Bare Nuclei	From one to ten
8	Bland Chromatin	From one to ten
9	Normal Nucleoli	From one to ten
10	Mitoses	From one to ten
11	Class	2 for benign, and 4 for malignant

3.1.2 Wisconsin diagnostic breast cancer dataset (WDBC)

WDBC dataset is a fine needle of data mass consisting of extracted features from digitized images. There are 569 sample records in this dataset, each with 32 attributes (ID, Diagnosis, 30 real-valued features). The data set can be separated linearly using all 30 input features. All the features correspond to the properties of the cell nuclei visible in the image.

Table 4 summarizes the attribute information, with the first attribute representing the unique identifier for each patient and the second representing the class label of malignant or benign. The attribute range of 3-32 corresponds to computed characteristics for each cell nucleus [21].

Table 4: WDBC Dataset Features Information

No	Features
1	Id
2	Diagnosis
3-32	3-32). Each nucleus is described by ten computed characteristics.
Texture	
Radius	
Area	
Perimeter	
Compactness	
Smoothness	
Concavity	
Concave points	
Symmetry	
Fractal dimension	

The radius is the average of the distances between the entrance and each point on the perimeter. The texture parameter is defined as the standard deviation of grayscale values. Smoothness is the degree to which the radius length varies locally. The compactness factor is calculated as follows: (perimeter power 2/area-1.0).Concavity refers to the degree to which a contour is concave, and the fractal dimension is (an approximation to the coast)-1. The mean, standard error, and worst of 30 features are computed. For example, field 3 represents the mean radius, field 13 represents the standard deviation of the radius, and field 23 represents the worst radius. According to the WDBC dataset's features, these attributes have three values (mean, standard error, and worst) and three columns.

- Eq. (1) calculates the Mean as the following:

$$mean = \frac{\sum_{i=1}^{n} x_i}{n} \quad (1)$$

- Eq. (2) calculates the Standard Error as:

$$S_e = \delta \frac{e}{n} \quad (2)$$

- Worst mean or largest mean.

3.1.3 Wisconsin prognostic breast cancer dataset (WPBC)

Breast cancer prognosis is determined by the kind and stage of the disease. The WPBC dataset consists of 198 observations and 47 recurrences, 151 of which are not. The WPBC dataset contains benign and malignant cases, as with the other datasets. This dataset has the following characteristics:

- ✓ The patient ID is the first feature.
- ✓ The class (output) is the second feature; R denotes "recurrence," whereas N denotes "non-Recurrence."
- ✓ The third characteristic is time, which refers to the recurrence period for "R" and "being healthy" for N.
- ✓ The attributes from 3 to 33 identify ten computed real values for cell nuclei: radius, area, perimeter (dimensions and shape of a nucleus), concavity, concave points, symmetry, fractal dimension (approximation of a coastline), compactness, texture (standard deviation of grayscale values), and smoothness (local variation in radius lengths).
- ✓ The 34th characteristic is the size of tumour , which is expressed in centimeters. The tumour size is classified into four categories: T-1 is between 0 and 2 centimeters. T-2 ranges in size from 2 to 5 centimeters. T-3 exceeds 5 centimeters in length. T-4 is a term that refers to any tumour has pierced the skin (ulcerated) or is linked to the chest wall. The lymph node status refers to the number of auxiliary lymph nodes diagnosed with cancer during surgery.
- ✓ The 35th characteristic is the status of lymph node, which refers to the number of malignant auxiliary lymph nodes detected throughout surgical procedure. Breast cancer is most likely to spread through the lymph nodes in the armpit (axillary lymph nodes). The value of the property' Lymph node' status was missing in four records.
- ✓ Missing in four records.

3.2 Data Pre-processing

Data pre-processing is critical for every classification system because it converts picture data into a form that machine learning models can understand. We process the data set via Data Pre-processing to ensure that high-quality data is supplied without errors. Because classification performance is dependent on data quality, data should be unambiguous, correct, and full. Data pre-processing eliminates discrepancies and fills in missing values in the data set. Pre-processing step is used to boost the quality of a dataset to obtain clean data suitable for modeling [22]. The data set was compiled from multiple sources and contained repetitive and useless information. We use data cleaning techniques to eliminate discrepancies of this nature from the data collection. Before performing classification tasks using machine learning techniques, numerous pre-processing techniques were employed to the Breast cancer dataset. The dataset was cleaned, null values were removed, and layers were removed during pre-processing. These pre-processing steps, including the cleaning phase, are used to prepare the dataset with machine learning models. It removes redundant features from the data and then achieves improved performance results. The pre-processing procedure is divided into several sub-phases, as detailed below.

3.2.1 data cleaning:

Removing or reducing noise and dealing with missing values. Delete null values: We analyzed the dataset and used it in this work. While the WDBC dataset is error-free, the WBC and WPBC datasets contain some missing and irrelevant data; therefore, we clean the data by replacing missing values with relevant ones. Sixteen WBC instances and four WPBC instances include a single missing attribute value, represented by"?". The attribute means filling in missing values for all instanes belonging to the equivalent class.

3.2.2 removing outlier:

Extremely harmful are outliers. They have a considerable impact on the output of a machine learning model. Typically, researchers assess outliers to determine if each record results from a data collection error or a unique phenomenon considered during data processing. An outlier is a data point that appears to be out of step with the rest of the data. Eliminating outliers may result in a dataset that is smaller than the original but retains the original data's integrity.

3.2.3 relevance analysis:

Statistical correlation analysis excludes redundant features from further investigation. The WBC, WPBC, and WDBC all share one superfluous characteristic called 'Sample code number,' which has no bearing on the classification operation; thus, the feature is ignored.

3.2.4 data normalization:

Reduces training time by initiating the process with features of a similar scale. The purpose of normalization is to reduce the range of feature values to a manageable size.

3.3 Data Spliting

The fundamental idea behind 10-Fold CV is to divide datasets into ten sections/folds, nine of which are used for training and testing. The dataset is divided into training and testing datasets via hold out (70 % for training and 30 % for testing) and 10-fold cross-validation (CV). The process of data partitioning is repeated k times (k = 10).

3.4 Hyperparameters Optimization Methods for Standard ML Techniques

Grid search is a technique for hyper-parameter tuning that can determine the optimal value for an ML algorithm. It assesses the machine learning model for each combination of algorithm parameters defined in a grid and then returns the model hyper-parameters optimal answer. The hyper-parameters optimization approaches (i.e., Grid Search with stratified 10-fold cross-validation) are utilized in this step to determine each parameter's ideal rang in machine learning models.

3.5 Classification Based ON ML Models

During this stage, we implement seven common ML algorithms, including Logistic Regression (LR), Extreme Gradient Boost (XGboost), Multi-Layer Perception (MLP) of Neural Network, instance Based for K-Nearest Neighbor (KNN), Naive Bayes (NB), Decision Tree (DT) and Random Forest (RF).

3.5.1 K-nearest neighbor (k-NN)

K-Nearest Neighbor (k-NN) classifier is utilized in this research and is one of the extremely well-known machine learning strategies for classification [23]. The classification method Neighbor K-Nearest is used for non-parametric trick learning. This classification scheme organizes the items that your immediate neighbors refer to as "k." It is concerned with the neighbors of the object, not with the essential data allocation.

3.5.2 random forest (RF)

Random Forest (RF) classification combines the output of numerous decision-making trees to create an ensemble of trees. This is supported by the argument that a single decision tree can generate either a specific or a very simple model [24].

3.5.3 logistic regression (LR)

Logistic Regression (LR) classifier is a supervised learning process utilized to classify data items. Typically, the target variable in logistic regression is binary, containing only samples classified as 1 or 0, which in our situation indicates a positive or negative breast cancer patient. LR is used to determine whether an instance belongs to a class. If the estimated probability exceeds the threshold, the model predicts that the instance is a class member; otherwise, it predicts that it is not a class member [23].

3.5.4 decision tree (DT)

A Decision Tree (DT) algorithm is a form of supervised learning primarily applied to solve classification problems. It works for both discrete and continuous parameters in the outputs and inputs. The algorithm infers simple decision-making principles from its data characteristics and then indicates target data values [25]. In other words, based on the most significant major differences between the input values, the population or sample is divided into two or more homogeneous sets (or sub-populations). DT divides a node into two or more sub-nodes using several algorithms. The existence of sub-nodes improves the homogeneity of the resulting sub-nodes.

3.5.5 extreme gradient boost (XGboost) classifier

XGboost is used to improve the performance of short works via constructing a sequence of weak decision-makers, with each tree correction attempting to decrease the error of the prior one. Chen and Guestrin [26] proposed it as a mountable machine learning classification. Latest research has demonstrated that some classifiers perform better than others at classification tasks. XGboost is a classifier designed for these types of applications [27]. This technique is intended to enhance the calculation rate and effectiveness of the machine during the test. The most critical parameters for the XGboost classifier are as follows:

- ✓ base_estimator: A relatively inexperienced learner is used to train the model. It uses the support vector machine (SVM) as the default

weak learner for training purposes. Additionally, you have the option of specifying multiple machine learning algorithms.

- ✓ n_estimators: Number of weak learners to train iteratively.
- ✓ learning_rate: It contributes to the weight of weak learners. It uses one as a default value.

3.5.6 naive bayes (NB) classifier

Naive Bayes (NB) classifier is a Bayes-based probabilistic classifier. The Naive Bayes classifier generates probability estimates rather than predictions. For each class value, they assess the probability that a particular instance belongs to a specific class. The Naive Bayes classifier has the advantage of requiring a small quantity of training data to calculate classification parameters. It is presumptuous to believe that an attribute value's effect on a given class is independent of the values of the other attributes. This is referred to as the assumption of class conditional independence. The Naive Bayes Technique is based on the Bayesian approach, which is extremely straightforward and useful for rapid classification. This technique takes mutually independent features into account and is used in various domains to achieve significant results in machine learning [28].

3.5.7 the multilayer perceptron (MLP)

The Multilayer Perceptron (MLP), A feed-forward backpropagation network, is the most frequently used technique for pattern recognition in artificial neural networks (ANNs) [29] MLP is a supervised learning technique comprised of three components; an input layer, an output layer, and one or more hidden layers that harvest meaningful information during the training process and allocate adaptable weighting coefficients to input layer components [30].

3.6 Performance Metrics

In Table 5, five standard performance metrics; Accuracy (ACC), Recall (REC), Precision (PREC), F1-score (F1), and Area under the receiver operating characteristics curve (AUC) are calculated as follows:

The accuracy of an classifier is calculated as the proportion of accurately classified instances (TP+TN) to the total number of instances (TP+TN+FP+FN). Calculated using the Eq. (3)

$$Accuracy\,(AC) = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Precision is the proportion of accurately classified samples with the disease (TP) to all predicted patients (TP+FP). Eq. (4) was used to calculate.

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

The recall evaluation metric is identified as the proportion of accurately classified samples (TP) to the total number of diseased patients. The recall perception is based on the number of patients categorized as having the disease. Sensitivity is another term for recall.

The recall or true malignant (positive) rate (TM) or (TP) is the proportion of malignant cases correctly identified, or in our case, sick people diagnosed correctly as sick. Calculated following Eq. (5):

$$recall = \frac{TP}{TP+FN} \quad (5)$$

F1 is also referred to as the F Measure. The F1 score indicates the balance of precision and recall. In practice, increasing the precision of our model decreases the recall and vice versa. The F1-score is a single value that encapsulates both trends with Eq. (6):

$$F1\_score = \frac{2TP}{2TP+FP+FN} \quad (6)$$

The area under the receiver operating characteristics curve (AUC) has been utilized commonly to assess numerous machine learning methods.

$$AUC = = \frac{S0-n0(n0+1)/2}{n0n1} \quad (7)$$

Table 5: Correlation Matrix

		actual values	
		Malignant (Positive)	Benign (Negative)
predicted values	Malignant (Positive)	TP	FN
	Benign (Negative)	FP	TN

4. EXPERIMENTAL RESULTS

To analyze the efficiency of Machine Learning (ML) approaches in detecting Breast Cancer, we evaluated three datasets: the WBC dataset, the WDBC dataset, and the WPBC dataset, using two learning strategies: hold-out validation (70 % for training and 30 % for testing) and 10-folds cross-validation. It is critical to emphasize the machine learning approaches (KNN, XGboost, RF, MLP, NB, DT, and LR) employed to achieve more accurate findings for each dataset. Accuracy, AUC, Precision, F1-score, and Recall are applied to evaluate our model.

4.1 Case Study I (WBC Dataset)

4.1.1 hold out validation

Journal of Theoretical and Applied Information Technology
31st July 2022. Vol.100. No 14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

We investigated the split impact of breast cancer detection using 70 Training and 30 Testing based on Seven machine learning models. MLP outperformed competitors in terms of AUC, accuracy, precision, F1-score, and recall, with a performance of 99.2%, 98.2%, 96.4%, 96.4%, and 96.4%, respectively. This tendency results from the kernel's strength and the unique ability of MLP to address binary challenges. Additionally, the KNN classifier is the lowest performer on the WBC dataset. Table 6 and Figure 2 show the MLP classifier's results in basic machine learning.

Table 6: The performance results of Machine Learning for WBC dataset using HOLD OUT VALIDATION

Algorithm	70 Training and 30 Testing				
	ACC	AUC	PREC	REC	F1-score
KNN	53.8	54.9	52.5	52.1	52.1
XGboost	97.5	99	97.2	97.1	97.1
RF	96.8	98.8	96.2	96.2	96.2
MLP	**98.2**	**99.2**	**96.4**	**96**	**96.4**
NB	97	98.7	96.1	96.1	96.4
DT	93.2	92.1	93.5	93.6	93.4
LR	54.5	55.2	54	53.7	53.7

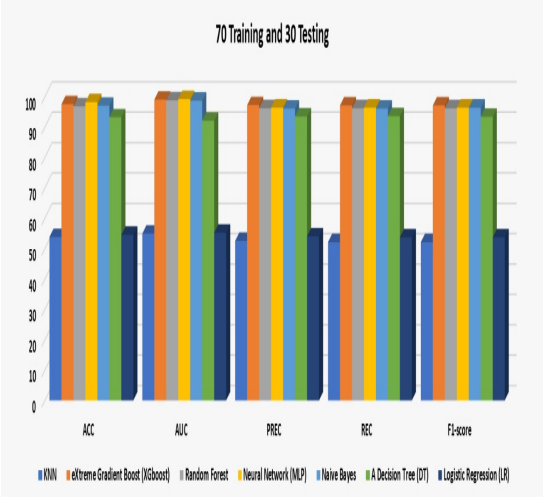

Figure 2: The performance of hold-out results of Machine Learning for WBC dataset

4.1.2 10-folds cross-validation

5. According to the findings in Table 7 and Figure 3 the strongest machine learning classifier is Neural Network (MLP) (ACC (98.3%), AUC (99.3%), PREC (96.6%), REC (96.6%), and F1 (96.7%)). This performance can be viewed in terms of the kernel MLP efficiency, which enables the problem of binary classifications (Benign and Malignant opinions) to be treated flawlessly. The worst machine learning classifier is KNN, which achieved an accuracy value (53.8%), AUC, precision, recall, and F1-score of 55.2%, 52.1%, 51.7%, and 51.9%, respectively. KNN is based on Euclidean distance.

Table 7: The performance results of Machine Learning for WBC dataset using 10-Fold Cross-Validation

Algorithm	10-FOLDS CROSS-VALIDATION				
	ACC	AUC	PREC	REC	F1-score
KNN	53.8	55.2	52.1	51.7	51.9
XGboost	97.9	99	97.1	97	97
RF	97	98.9	95.9	95.9	95.9
MLP	**98.3**	**99.3**	**96.6**	**97**	**96.7**
NB	97.5	98.8	96.4	96.3	96.3
DT	93.9	94.4	94.7	94.7	94.7
LR	54.2	55.8	52.7	52.9	52.9

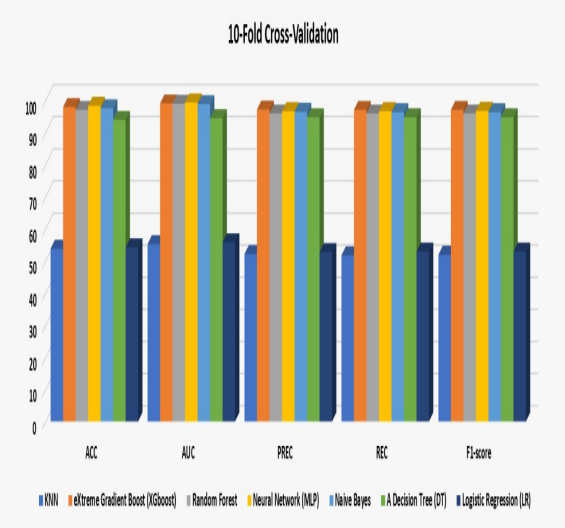

Figure 3 The performance of 10 fold Cross-validation results of Machine Learning for WBC dataset

4.2 Case Study II (WDBC Dataset)

4.2.1 hold out validation

This section discusses the performance of machine learning using hold-out validation on the WDBC dataset. The values of five metrics are

Journal of Theoretical and Applied Information Technology
31st July 2022. Vol.100. No 14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

shown in Table 8 and Figure 4 including accuracy (ACC), the area under the curve (AUC), precision (PREC), recall (REC), and F1-score (F1). By examining the findings of machine learning algorithms, it is evident that XGboost is the best classifier in terms of (ACC 99 %, AUC 99.4 %, REC 97.2 %, PREC 97.1 %, and F1 97.1 %) when 70 Training and 30 Testing is used. The KNN classifier achieves the lowest performance (ACC of 58.3 %, AUC 57%, PREC of 52.6%, REC of 53.1 %, and F1 of 51.9 %). This performance is obvious and understandable, given KNN's nature as a lazy learner. Thus, the categorization assignment is accomplished solely by the computation of Euclidean distance, which affects performance in the case of a high-dimensional representation space.

Table 8: The Machine Learning performance results of WDBC dataset using HOLD OUT VALIDATION

Algorithm	70 Training and 30 Testing				
	ACC	AUC	PREC	REC	F1-score
KNN	58.3	57	52.6	53.1	51.9
XGboost	**99**	**99.4**	**97.1**	**97.2**	**97.1**
RF	95.2	98.7	94.7	95.3	95.3
MLP	89.8	99.3	97.1	97.2	97.1
NB	93	98.3	93.6	93.5	93.5
DT	92.4	92.1	92.1	92.5	91.8
LR	75	79.2	76	75.1	54.9

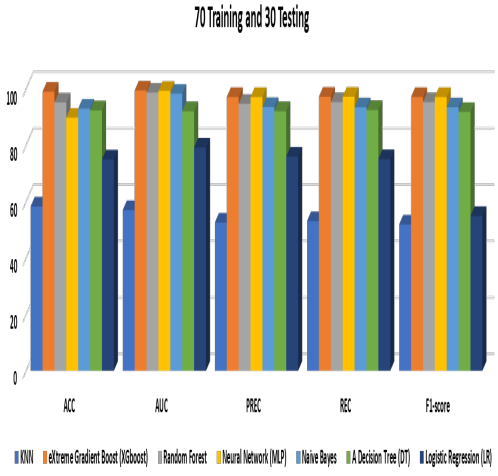

Figure 4: The performance of hold-out results of Machine Learning for WDBC dataset

4.2.2 10-folds cross-validation

This section examines the effect of partitioning the WDBC dataset using 10-fold cross-validation utilizing machine learning methods, as demonstrated in Table 9 and Figure 5. By analyzing the results of 10-fold cross-validation, XGboost gives a significant increase with a performance of 99.2% accuracy, AUC, recall, precision, and F1-score accomplish a great performance of 99.5%, 97.4%, 99.4%, and 97.4%, respectively. The XGboost model obtained an improved result.

Because of the simple Euclidean distance utilized to distinguish between groups, the KNN continues the most terrible classifier for the WDBC dataset when 10-fold cross-validation is used.

Table 9: The Machine Learning performance results of 10-Fold Cross-Validation for WDBC dataset

Algorithm	**10-FOLDS CROSS-VALIDATION**				
	ACC	**AUC**	**PREC**	**REC**	**F1-score**
KNN	58.9	59.2	53.2	53.4	53.1
XGboost	**99.2**	**99.5**	**97.4**	**97.4**	**97.4**
RF	95.5	98.3	95.3	94.7	95.7
MLP	99	99.2	97.2	97.2	97.2
NB	93.5	94.5	94.03	94.02	94.03
DT	92.7	93.1	92.6	92.7	92.6
LR	75.1	80.7	76.8	76	75.7

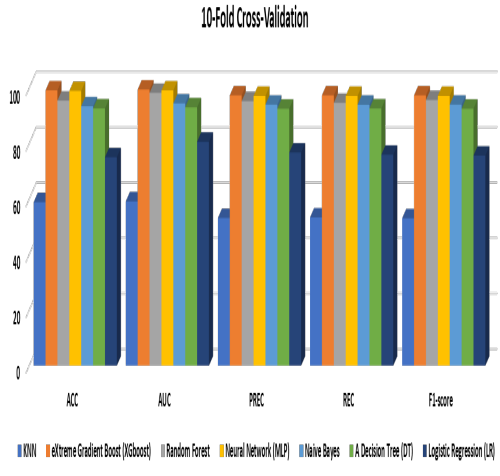

Figure 5: The performance of 10-fold Cross-validation results of Machine Learning for the WDBC dataset

4.3 Case Study III (WPBC Dataset)

4.3.1 hold out validation

With studying the acquired findings of basic machine learning, shown in Table 10 and Figure 6, the main observation is that the accuracy ranges between 57.8% and 78.2%, whereas the F1-score ranges between 58.2% and 77.8%. As with the third dataset (WPBC), MLP ranks first in terms of accuracy (78.2%), AUC (78.7%), precision (77.1 %), recall (76.8%), and F1-score (77.8%) when compared to fundamental machine learning techniques (KNN, XGboost, RF, MLP, NB, DT, and LR). KNN, on the other hand, is placed bottom in terms of accuracy (57.8%), AUC (58.5%), precision (58.8%), recall (58.2%), and F1-score (58.2%).

Table 10: The performance results of Machine Learning for WPBC dataset using HOLD OUT VALIDATION

Algorithm	70 Training and 30 Testing				
	ACC	AUC	PREC	REC	F1-score
KNN	57.8	58.5	58.8	58.2	58.2
XGboost	73.7	74	68	67.8	67.2
RF	60.4	57.1	68.9	69.4	69.8
MLP	**78.2**	**78.7**	**77.1**	**76.8**	**77.8**
NB	60.7	61.3	62.5	65.4	65.9
DT	59.8	62.3	68	68	68
LR	59	60	71.8	70.5	70

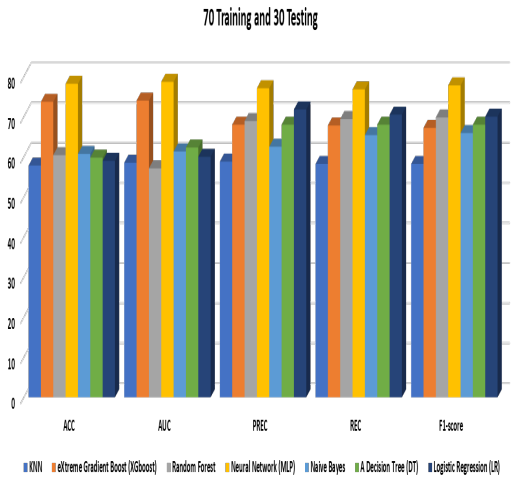

Figure 6: The performance of hold-out results of Machine Learning for WPBC dataset

4.3.2 10-folds cross-validation

We employed 10-folds cross-validation to address the issue of imbalanced data distribution. Inaccuracy, AUC, recall, precision, and F1-score all show considerable improvement. In machine learning, accuracy values range between 58% and 78.6%, whereas the F1-score has a lower and upper limit of 62.8% and 78%, respectively. The MLP classifier is placed first, achieving an overall performance accuracy of 78.6%, AUC 78.9%, precision 77.7%, recall 77.2%, and F1-score 78%). The least accurate machine learning classifier for identifying breast cancer subtypes is KNN, which achieves 58 %, AUC 60%, 61.8% precision, 62.9% recall, and 62.8% F1-score. In conclusion, imbalanced data significantly impacts the performance of machine learning systems, as shown in Figure 7 and Table 11.

Table 11: The performance results of Machine Learning for WPBC dataset using 10-Fold Cross-Validation

Algorithm	CROSS-VALIDATION				
	ACC	AUC	PREC	REC	F1-score
KNN	58	60	61.8	62.9	62.8
XGboost	73.8	74.8	70	74.5	74.2
RF	60.7	63	76.6	78.9	75
MLP	**78.6**	**78.9**	**77.7**	**77.2**	**78**
NB	60.9	61.6	59.1	59.2	59.4
DT	60.1	62.2	66.3	68.6	67.3
LR	59.2	60	75	77.3	70

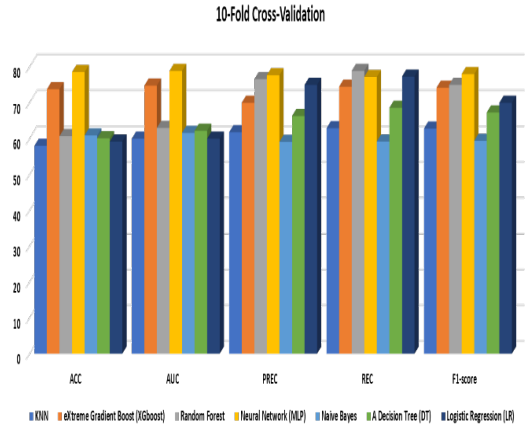

Figure 7: The performance of 10-fold Cross-validation results of Machine Learning for WPBC dataset

Journal of Theoretical and Applied Information Technology
31st July 2022. Vol.100. No 14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5 DISCUSSION

5.1 Discussion for Case study I

From the findings achieved in our experiments for the first dataset, Figure 8 illustrates the overall practical outcomes for the cross-validation performances and the testing outcomes, respectively. They demonstrate the performance of the most effective models for each feature extraction technique. To recap the performance of the compared models, we explore the average cross-validation and the testing results of each model using different baseline machine learning Extreme Gradient Boost (XGboost), Multi-Layer Perception (MLP) of Neural Networks, Naive Bayes (NB), Random Forest, and Instance-Based for K-Nearest Neighbor (KNN).

The MLP model has achieved the highest average of cross-validation, contrasted to other standard machine learning methods. For cross-validation outcomes, the MLP technique has accomplished accuracy of 98.3%, AUC of 99.3%, recall of 96.6%, precision of 96.6%, and F1-score of 96.7%.

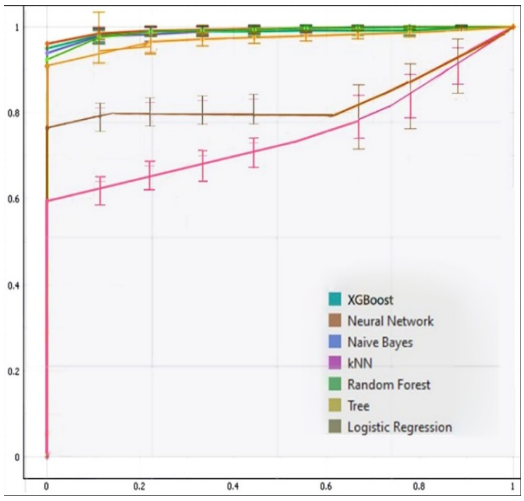

Figure 8: Cross-validation result for WBC dataset

5.2 Discussion for Case Study II

From the findings achieved in our experiments for the first dataset, Figure 9 depicts the overall practical outcomes for the cross-validation performances and the testing results, respectively. The XGboost method has achieved the highest average of cross-validation contrasted to other standard machine learning techniques. For cross-validation outcomes, the XGboost metod has accomplished AUC of 99.5%, accuracy of 99.2%, precision 99.4%, recall of 97.4% and F1-score 97.4%).

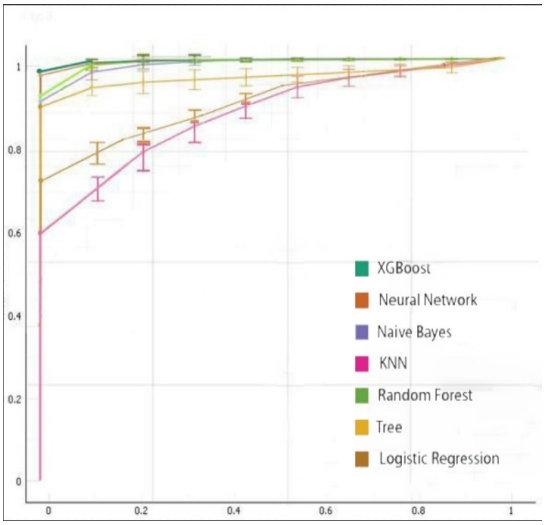

Figure 9: Cross-validation result for WDBC dataset

5.3 Discussion for Case study III

From the results obtained in our experiments for the first dataset, Figure 10 shows the overall practical results for the cross-validation performances and the testing results, respectively.

The MLP method has achieved the highest cross-validation average compared to other regular machine learning tecnniques. For cross-validation outcomes, the MLP method has accomplished AUC 78.9%, accuracy of 78.6%, recall of 77.2 %, precision 77.7%, and F1-score 78%.

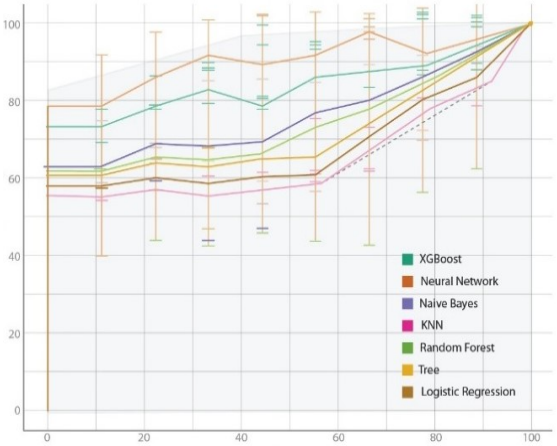

Figure 10: Cross-validation result for WPBC dataset

6 CONCLUSION

This paper proposed an effecient framework for breast cancer diagnosis and recurrence prediction. Our proposed framework for breast cancer detection mainly contains six steps: Data collection, data pre-processing, Data partitioning, hyper-parameters tuning for Machine Learning models, classification based on the proposed ML, and Prediction and evaluation metrics. We applied a comparative study of seven machine learning techniques to detect breast cancer, including XGboost, Multi-Layer Perception, Naive Bayes, KNN, and Random Forest. The framework's performance was compared to determine which classifier performs the best on the Wisconsin datasets. The primary contribution of this study was the introduction of an optimized framework for both both breast cancer detection and recurrence prediction, which provided superior results compared to the state of art algorithms. The combination of different features provided our method with increased accuracy, the area under the curve (AUC), precision, recall, and f1 measure compared to several state-of-the-art methods.

In the WBC dataset for brest cancer classification, The Multilayer Perceptron (MLP) is placed in the first spot accordinding to accuracy (98.3\%), AUC (99.3%), precision (96.6\%), recall 96.6\%), and F1-score (96.7\%) based on ten folds cross-validation. On the other hand, KNN is placed in the last spot accordinding to accuracy (53.8\%), AUC (55.2%), precision (52.1%), recall (51.7%), and F1-score (51.9%). In conclusion, the performance of machine learning algorethms deeply affected by unbalanced data.

As the results on the WDBC dataset for breast cancer detection, the XGboost model obtained the highest performance accuracy (99.2\%), AUC (99.5%), precision (97.4\%), recall (97.4\%), and F1-score (97.4\%) based ten folds cross-validation. The next model is MLP and the last one is that KNN obtained the lower performance,

Regarding the third dataset (WPBC) for cancer recurrence prediction, the top classifiers according to accuracy are MLP as they achieved 78.7% % using 10-folds cross-validation and 78.2% using hold out, respectively. For the this dataset, the performance is deeply affected by unbalanced data distribution. The worst classifier was KNN due to the simple concept of classification using Euclidian distance and it require large dataset and less features for better results.

The limitation we faced was a small size and imbalanced datasets, especially with the WPBC dataset; it has 198 instances only, among which 151 Non-Recurrence class instances and 47 instances only for the Recurrence class. Therefore, the result of cancer recurrence prediction is low compared with cancer classification on the other two datasets. Applying ML algorithms to a larger dataset can improve the accuracy.

For future work: We plan to use swarm algorithms for selection of the best hyperparameters. Also transforming the chosen models into a feasible and practical tool for supporting and assisting physicians with breast cancer diagnosis. Future research may also compare other machine learning techniques, additional illness possibilities, and alternative types of datasets can be examined.

REFERENCES:

[1] "Cancer." https://www.who.int/en/news-room/fact-sheets/detail/cancer (accessed Jan. 17, 2022).

[2] J. Ferlay *et al.*, "Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer," *Lyon, France: International Agency for Research on Cancer*, 2020.

[3] A. S. Ibrahim, H. M. Khaled, N. N. Mikhail, H. Baraka, and H. Kamel, "Cancer Incidence in Egypt: Results of the National Population-Based Cancer Registry Program," *Journal of Cancer Epidemiology*, vol. 2014, pp. 1–18, 2014, doi: 10.1155/2014/437971.

[4] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates," *Cancer Lett*, vol. 77, no. 2–3, pp. 163–171, Mar. 1994, doi: 10.1016/0304-3835(94)90099-X.

[5] A. Lg, A. T. Eshlaghy, A. Poorebrahimi, M. Ebrahimi, and R. Ar, "Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence," *undefined*, vol. 04, no. 02, 2013, doi: 10.4172/2157-7420.1000124.

[6] "Breast Cancer Wisconsin (Diagnostic) Data Set | Kaggle." https://www.kaggle.com/uciml/breast-cancer-wisconsin-data (accessed Jan. 17, 2022).

[7] W. H. Wolberg, W. N. Street, D. M. Heisey, and O. L. Mangasarian, "Computerized breast cancer diagnosis and prognosis from fine-needle aspirates," *Arch Surg*, vol. 130, no. 5, pp. 511–516, 1995, doi: 10.1001/ARCHSURG.1995.01430050061010.

[8] M. Nilashi, O. Ibrahim, H. Ahmadi, and L. Shahmoradi, "A knowledge-based system for breast cancer classification using fuzzy logic method," *Telematics and Informatics*, vol. 34, no. 4, pp. 133–144, Jul. 2017, doi: 10.1016/J.TELE.2017.01.007.

[9] V. Chaurasia, S. Pal, and B. B. Tiwari, "Prediction of benign and malignant breast cancer using data mining techniques:," *https://doi.org/10.1177/1748301818756225*, vol. 12, no. 2, pp. 119–126, Feb. 2018, doi: 10.1177/1748301818756225.

[10] S. Wang, Y. Wang, D. Wang, Y. Yin, Y. Wang, and Y. Jin, "An improved random forest-based rule extraction method for breast cancer diagnosis," *Applied Soft Computing*, vol. 86, p. 105941, Jan. 2020, doi: 10.1016/J.ASOC.2019.105941.

[11] S. Mojrian *et al.*, "Hybrid Machine Learning Model of Extreme Learning Machine Radial basis function for Breast Cancer Detection and Diagnosis; A Multilayer Fuzzy Expert System," *Proceedings - 2020 RIVF International Conference on Computing and Communication Technologies, RIVF 2020*, Oct. 2020, doi: 10.1109/RIVF48685.2020.9140744.

[12] E. A. Bayrak, P. Kirci, and T. Ensari, "Comparison of machine learning methods for breast cancer diagnosis," *2019 Scientific Meeting on Electrical-Electronics and Biomedical Engineering and Computer Science, EBBT 2019*, Apr. 2019, doi: 10.1109/EBBT.2019.8741990.

[13] M. Ramos *et al.*, "Machine Learning Classification Techniques for Breast Cancer Diagnosis," *IOP Conference Series: Materials Science and Engineering*, vol. 495, no. 1, p. 012033, Apr. 2019, doi: 10.1088/1757-899X/495/1/012033.

[14] N. Najmu, J. Sanjay, and M. Shahid, "Early Detection of Cardiovascular Disease using Machine learning Techniques an Experimental Study 636," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 9, no. 3, 2020, doi: 10.35940/ijrte.C46570.99320.

[15] S. Sharma, A. Aggarwal, and T. Choudhury, "Breast Cancer Detection Using Machine Learning Algorithms," *Proceedings of the International Conference on Computational Techniques, Electronics and Mechanical Systems, CTEMS 2018*, pp. 114–118, Dec. 2018, doi: 10.1109/CTEMS.2018.8769187.

[16] M. A. Rufai, A. S. Muhammad, S. Garba, and L. Audu, "MACHINE LEARNING MODEL FOR BREAST CANCER DETECTION," *FUDMA JOURNAL OF SCIENCES*, vol. 4, no. 1, pp. 55–61, Apr. 2020, Accessed: Jan. 17, 2022. [Online]. Available: https://fjs.fudutsinma.edu.ng/index.php/fjs/article/view/16

[17] G. I. Salama, M. B. Abdelhalim, and M. Abd-elghany Zeid, "Experimental Comparison of Classifiers for Breast Cancer Diagnosis," Jan. 2013.

[18] U. Ojha and S. Goel, "A study on prediction of breast cancer recurrence using data mining techniques," *Proceedings of the 7th International Conference Confluence 2017 on Cloud Computing, Data Science and Engineering*, pp. 527–530, Jun. 2017, doi: 10.1109/CONFLUENCE.2017.7943207.

[19] A. I. Pritom, M. A. R. Munshi, S. A. Sabab, and S. Shihab, "Predicting breast cancer recurrence using effective classification and feature selection technique," *19th International Conference on Computer and Information Technology, ICCIT 2016*, pp. 310–314, Feb. 2017, doi: 10.1109/ICCITECHN.2016.7860215.

[20] B. N. Kiage, "A Data Mining Approach for Forecasting Cancer Threats," Kenyatta , 2015.

[21] C. L. Chi, W. N. Street, and W. H. Wolberg, "Application of Artificial Neural Network-Based Survival Analysis on Two Breast Cancer Datasets," *AMIA Annual Symposium Proceedings*, vol. 2007, p. 130, 2007, Accessed: Mar. 06, 2022. [Online]. Available: /pmc/articles/PMC2813661/

[22] M. D. , W. N. S. Ph. D. , D. M. H. Ph. D. , O. L. M. Ph. D. William H. Wolberg, "COMPUTERIZED BREAST CANCER DIAGNOSIS AND PROGNOSIS FROM FINE NEEDLE ASPIRATES," *Human Oncology, and Computer Sciences University of Wisconsin, Madison, WI* , Nov. 1994.

[23] Lan Guo, Yan Ma, B. Cukic, and H. Singh, "Robust Prediction of Fault-Proneness by Random Forests," in *15th International Symposium on Software Reliability Engineering*, pp. 417–428. doi: 10.1109/ISSRE.2004.35.

[24] L. Xiong and Y. Yao, "Study on an adaptive thermal comfort model with K-nearest-neighbors (KNN) algorithm," *Building and Environment*, vol. 202, Sep. 2021, doi: 10.1016/J.BUILDENV.2021.108026.

[25] C. Zhu, C. U. Idemudia, and W. Feng, "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques," *Informatics in Medicine Unlocked*, vol. 17, p. 100179, Jan. 2019, doi: 10.1016/J.IMU.2019.100179.

[26] D. Xue, A. Frisch, and D. He, "Differential Diagnosis of Heart Disease in Emergency Departments Using Decision Tree and Medical Knowledge," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11721 LNCS, pp. 225–236, 2019, doi: 10.1007/978-3-030-33752-0_16.

[27] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, doi: 10.1145/2939672.

[28] M. Gumus and M. S. Kiran, "Crude oil price forecasting using XGBoost," *undefined*, pp. 1100–1103, Oct. 2017, doi: 10.1109/UBMK.2017.8093500.

[29] I. H. Witten, E. Frank, and J. Geller, "Data mining," *ACM SIGMOD Record*, vol. 31, no. 1, pp. 76–77, Mar. 2002, doi: 10.1145/507338.507355.

[30] M. Riedmiller, "Advanced supervised learning in multi-layer perceptrons - From backpropagation to adaptive learning algorithms," Computer Standards and Interfaces, vol. 16, no. 3, pp. 265–278, 1994, doi: 10.1016/0920-5489(94)90017-5.