Machine Learning Methods for Breast Cancer Diagnosis

Khwanchanok Chumkhun*

August 2, 2022

Abstract

Worldwide, breast cancer is the most common form of cancer. There were 2.3 million new diagnoses of breast cancer in 2020 and 685,000 deaths [22]. The treatment of breast cancer, however, can be highly effective, with survival probabilities of as much as 90%, especially when the disease is detected early [25]. In order to reduce the mortality rate for breast cancer, it is essential to make a diagnosis in an early stage which will allow for an earlier and more precise treatment procedure. There are several clinical practices to identify breast cancer, including mammography, magnetic resonance imaging (MRI), ultrasound, computerized tomography, positron emission tomography and biopsy [27]. The biopsy is a technique that a small sample of body tissue is examined under the microscope [6]. Despite the fact that biopsy provides high prognosis accuracy, they require a high cost and a long lead time[27], and not every woman will get these screening methods. To strengthen the breast cancer prognosis, the biomarkers approach is studied. This technique is to measure biological parameters such as blood pressure, cholesterol level, and glucose level which are utilized to monitor and predict health states [20]. Since biomarkers are examined as part of the routine consultation and blood analysis, these unravel the high expense and time-consuming limitations [27]. But besides the rapid prognosis, prediction efficiency is also required. This study aims to study the performance of machine learning (ML) techniques applied to the dataset obtained from the fine-needle aspiration (FNA) biopsy method in comparing with those accuracies from dataset obtained from the biomarkers dataset. As a result, implementing machine learning in the dataset fron biopsy method (Wisconsin) gives a better result than in the biomarkers dataset(Coimbra). However, the Coimbra can be used as a biomarker for breast cancer and its result can be a pre-process of breast cancer analysis by biopsy.

*Department of Mathematical Sciences, University of Essex, CO4 3SQ Colchester, United Kingdom (e-mail: kc21747@essex.ac.uk).

1 Introduction

Ordinarily, the new healthy cells replace the old dead ones during the cell growth process. Nonetheless, it is possible that the mutated cells grow unconstrained and uncontrollably, resulting in a tumour. A tumour can be unharmful to health (benign) or can be cancerous, which has the potential to spread away from the initial tumour to other parts of the body (malignant), and this becomes dangerous for health. The term "breast cancer" refers to a malignant tumour that originally developed from cells in the breast. 85% of breast cancer arises in the lining cells (epithelium) of the milk ducts while 15% of it initially arises in lobules as shown in the Figure 1 from [25]. 85-90% of breast cancers are caused by a genetic abnormality, however, 5-10% of cancers are inherited from the parents[4].

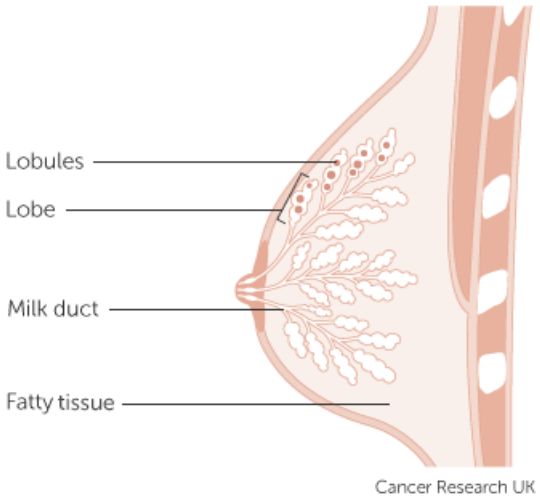

Figure 1: Breast cancer

Source: reference[25]

There are many factors that increase the risk of breast cancer, including age, obesity, extreme consumption of alcohol, genetics, smoking and so on. Over 40 years old women who do not have a known breast cancer risk factor make up approximately half of the cases [22][25].

These days breast cancer become the most prevalent cancer in the world, which has been diagnosed in 7.8 million women in the past 5 years, according to the World Health Organization. In 2020, the global breast cancer mortality rate is at 685 000 deaths, with 2.3 million women diagnosed with the disease[22]. Focusing on the UK, records from 2016 to 2018 show that approximately 55,900 new breast cancer cases are diagnosed each year, or more than 150 cases daily,

reckoning for 15% of all new cancer cases [25].

As shown in Figure 2, breast cancer treatment can be highly effective, especially when the disease is identified early. Treatment of breast cancer often consists of a combination of surgical removal, radiation therapy and medication (hormonal therapy, chemotherapy and/or targeted biological therapy) to treat microscopic cancer that has spread from the breast tumour through the blood. Such treatment, can prevent cancer growth and spread, thereby saving lives[22].

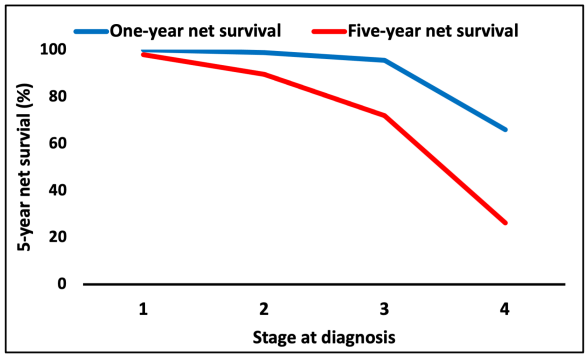

Figure 2: Breast cancer net survival by stage

Data source: reference [25]

There are various approaches to breast cancer determination, including mammography, magnetic resonance imaging (MRI), ultrasound, computerized tomography, positron emission tomography and biopsy [27] and so on. The biopsy approach is a technique that a small sample of body tissue is examined under the microscope. The prognosis of breast cancer using this data provides high accuracy, but this process is applied in the later step of breast cancer diagnosis and its cost is expensive. Other methods are studied to make the prognosis become earlier, biomarker is one of them. This technique is to measure biological parameters such as blood pressure, cholesterol level, and glucose level which are utilised to monitor and predict health states [20]. The biomarkers are done at the routine consultation and blood analysis, these unravel the high expense and time-consuming limitations. The main purpose of this study is to compare the performance of applying machine learning to 2 datasets obtained from these 2 different clinical approaches.

Continuing with the paper, the remaining parts are organized as follows. The Methodology part includes the Data portion where the description of 2 datasets used in this study, Statistical Analysis where Exploratory Data Analysis of these 2 datasets have proceeded, and Machine Learning Model and Evaluation Methods where the machine learning algorithms and evaluation metrics that can be applied to this kind of problem are explained. The next part is the Analysis of results,

results from papers studied in past will be discussed and compared in this section. And then the finding will be concluded in the Conclusion part.

2 Methodology

2.1 Data

2 datasets will be analyzed in this study, one is the Wisconsin Breast Cancer Diagnostic (WBCD) dataset. This dataset is from the University of California Irvine (UCI) Machine Learning Repository. The data was obtained from the University of Wisconsin Hospitals, Madison by Dr William H. Wolberg. This 569 instances x 11 columns dataset is the result of the computing of a digitized image of the fine needle aspirate (FNA) of a breast mass and this is the samples received by the biopsy method. One label column which indicates whether the test tissue is malignant or not is called Class. And other 10 columns which are Sample code number, Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, and Mitoses describe the input variables[14].

Another dataset which is mentioned in this study is the Coimbra Breast Cancer Dataset which is also retrieved from UCI Machine Learning Repository. This dataset is representative of the biomarkers clinical approach. Portuguese women from the University Hospital Centre of Coimbra (CHUC) participate in this data collection between 2009 and 2013. In each patient, the diagnosis was confirmed histologically after positive mammography. Data were collected before surgery and any treatment on every breast cancer patient, excluding those who had received prior cancer treatment. While the study included healthy female volunteers as controls [13]. The dataset contains 10 columns with 116 instances, classification column is a response variable to display whether the participant's breast cancer test result is positive or not. While remaining 9 columns which are Age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, and MCP1 represent the predictors.

2.2 Statistical Analysis

From the study of Madhu Kumari et al., the univariate analysis of the WBCD dataset is exected. The dataset consists of 65.0% of cancerous samples and 35.0% of non-cancerous samples. The statistical analysis of all the nine features (Mean and Standard deviation), excluding the response variable, is presented in Table 1 [18].

Table 1: Description of the WBCD

Source: reference [18]

Attribute numbers	Attribute description	Values of attribute	Mean	Standard deviation
1	Clump thickness	1–10	4.44	2.83
2	Uniformity of cell size	1–10	3.15	3.07
3	Uniformity of cell shape	1–10	3.22	2.99
4	Marginal adhesion	1–10	2.83	2.86
5	Single epithelial cell size	1–10	2.23	2.22
6	Bare nuclei	1–10	3.54	3.64
7	Bland chromatin	1–10	3.45	2.45
8	Normal nucleoli	1–10	2.87	3.05
9	Mitoses	1–10	1.6	1.73

Moreover, in the same paper, bivariate analysis is implemented in order to understand the correlation between each predictor and label.

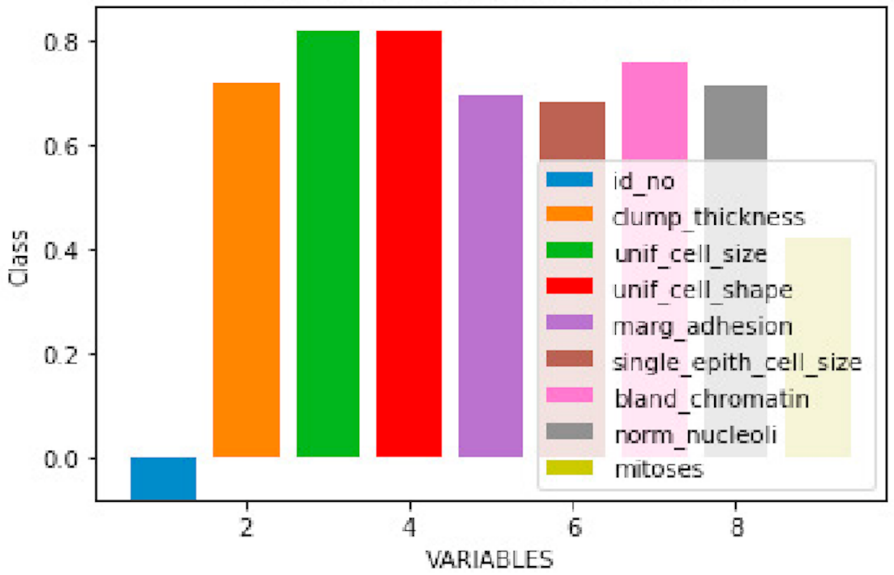

Figure 3: Correlation between WBCD variables

Source: reference [18]

For the Coimbra dataset, the univariate analysis is implemented in the study of Austria et al. The statistical analysis of each attribute, including the mean, standard deviation, minimum, and maximum values are studied.

Table 2: Statistical analysis of attributes

Source: reference [2]

Quantitative Attributes	Mean	Standard Deviation	Min	Max
Age	57.302	16.113	24	89
BMI	27.582	5.02	18.37	38.579
Glucose	97.793	22.525	60	201
Insulin	10.012	10.068	2.432	58.46
HOMA	2.695	3.642	0.467	25.05
Leptin	26.615	19.183	4.311	90.28
Adiponectin	10.181	6.843	1.656	38.04
Resistin	14.726	12.391	3.21	82.1
MCP-1	534.647	345.913	45.843	1698.44

Table 2 displays the analysis result of each attribute, which demonstrates that HOMA, BMI, and Adiponectin are the attributes with the lowest SD, respectively. The mean or average BMI is equal to 27.582 kg/m^2 which is slightly higher than the healthy range.

In the same study by Austria et al., a heat map is utilized to interpret the correlation between each predictor variable. Figure 3 shows how the colours indicate the association between each parameter. The high correlation between the 2 features is displayed by the light colour, while darker colours illustrate that the 2 features are weakly related [3].

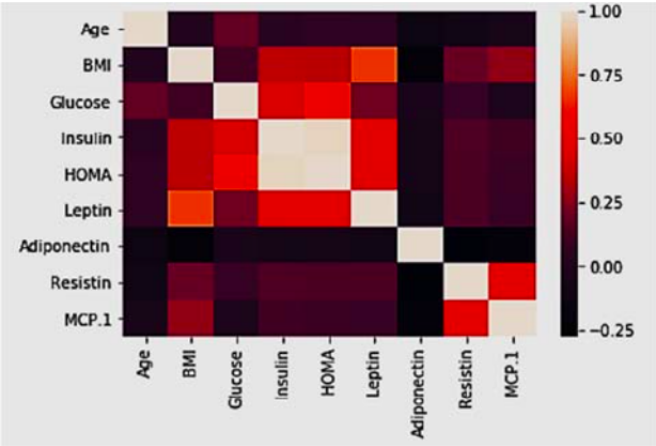

Figure 4: Heat-map Analysis of Features

Source: reference [2]

More bivariate analysis has proceeded in the study of Patrício et al., a comparison of the quantitative characteristics of patients and healthy controls are analyzed as shown in Table 2.

Table 3: Descriptive statistics of the clinical features

Source: reference [23]

	Patients		Control		p-value
	median	interquartile range	median	interquartile range	
Age ($years$)	53	23	65	33.2	0.479
BMI (kg/m^2)	27	4.6	28.3	5.4	0.202
Glucose (mg/dL)	105.6	26.6	88.2	10.2	¡0.001
Insulin ($\mu U/mL$)	12.5	12.3	6.9	4.9	0.027
HOMA	3.6	4.6	1.6	1.2	0.003
Leptin (ng/mL)	26.6	19.2	26.6	19.3	0.949
Adiponectin ($\mu g/mL$)	10.1	6.2	10.3	7.6	0.767
Resistin (ng/mL)	17.3	12.6	11.6	11.4	0.002
MCP-1 (pg/dL)	563	384	499.7	292.2	0.504

The medians of age in the two groups of participants did not differ statistically (p = 0.479), and it is the same for BMI (p = 0.272). According to the p-value, there are statistically significant differences in Glucose, Insulin, HOMA and Resistin, which are higher in the patient group. While there is no difference between the two groups in leptin, adiponectin, or MCP-1 levels [3].

2.3 Machine Learning Model and Evaluation Methods

To predict whether an individual has breast cancer or not is a kind of binary classification problem and there are several machine learning algorithms which can solve this kind of problem.

The following definitions are adapted from the lecture notes of Modules:

- MA334: Data analysis and statistics with R
- MA335: Modelling experimental and observational data
- MA336: Artificial intelligence and machine learning with applications

They are also complemented with external sources such as [24] [10] [26] [12] [5] [9] [16] [11] [7].

Binomial logistic regression is one of them, it is a simple model to perform on a binary classification problem in which the dependent variable can be only 2

possible values either 1 or 0. The represented equation of the logistic regression algorithm very much resembles the one of linear regression but instead of giving the probability of an event occurring, it has the log odds for the event [10][26]. The traditional logistic regression equations for multiple explanatory variables is

$$log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 x_1 + \beta_2 + x_2 + ... + \beta_p x_p \tag{1}$$

where x_1, x_2,..., x_p are the predictor variables, $\pi = Pr(Y = 1|X) = Pr("Success")$ and β_0, β_1,..., β_p are the unknown parameters that need to be estimated.

K-Nearest Neighbours (KNN) is an algorithm which works both on classification and linear problems. The concept of KNN is to assign the observation to the group or class by a plurality vote of its neighbours. *k* is the number of nearest neighbours used to vote. According to the data, *k* should be selected depending on its effect on classification; generally, larger values of *k* reduce the effect of noise on classification, but they lead to a less distinct boundary between classes. A proper *k* can be defined by various heuristic techniques. The distance between the observation and its neighbours can be calculated using Euclidean distance [24].
Euclidean distance in 2D (using Pythagoras's Theorem) :

$$d = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2} \tag{2}$$

where (x_1, y_1) is the coordinate of the neighbour and (x_0, y_0) is the coordinate of the observation.
Euclidean distance in 3D:

$$d = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + (z_1 - z_0)^2} \tag{3}$$

where (x_1, y_1, z_1) is the coordinate of the neighbour and (x_0, y_0, z_0) is the coordinate of the observation.

Decision Tree is another member of the supervised machine learning algorithm family. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too. In this method, the data is split into two branches iteratively to make predictions. Gini Impurity and Entropy are utilized to optimize the split of each node. The Gini Impurity concept is to maximize the Gini Gain, while the Entropy concerns on how to maximize the Information Gain [24].

Random forest is a widely used supervised machine learning algorithm in both classification and regression problems. It is constructed from a decision tree algorithm. For classification tasks, the output of the random forest is the class or group chosen by most trees. With a small tweak, it improves on the simple bagging method of the decision tree method. Each tree is tweaked to use a subset of features based on a random selection [12] [24].

Gradient boosting is a machine learning algorithm used in regression and classification tasks. Decision trees are typically used as weak prediction models in gradient-boosted. There is a good chance that it will perform better than random forest. As with other boosting methods, gradient-boosted tree models are constructed in a stage-wise manner, but they allow optimization of any differentiable loss function [5][9].

Support Vector Machine (SVM) is another supervised machine learning algorithm which can be utilized for both classification and regression problems but it is generally used in classification problems. It divides the data into categories by an n-dimensional hyperplane or called decision boundary. The observation is grouped according to which side of the boundary it belongs to. The best hyperplane is the one that maximizes the margin between 2 levels (the middle of the 2 groups)[24].

Naive Bayes is a simple technique is classification technique based on Bayes Theorem (a mathematical formula to calculate conditional probabilities), which allows us to determine probability based on a set of other probabilities. A Naive Bayes model is based on the assumption that each attribute contributes equally to the result. The classification of the response variable is done by assigning a class label to each instance based on some finite set of feature values. Several algorithms are used to train these classifiers, but not a single algorithm [16] [11].

Artificial neural network (ANN) is a deep learning model which can also be used for both regression and classification problems. ANN consists of the first layer, hidden layers, and last layer. Artificial neurons have inputs and outputs that can be sent to multiple other neurons. The weights of the connections from the inputs to the neuron are used to find the output of the neuron. A bias term is also added to this sum. This weighted sum is called the activation. To produce the output, this weighted sum is crossed through an activation function [7][24].

After complete modelling, the model performance evaluation is an important fundamental step. There are several ways to measure the performance of the binary classification models.

The accuracy is used to explain how much in per cent the model success to predict the result by comparing the number of correct predictions with all predicted cases. But accuracy is not always the best solution, for the dataset with an imbalance class of labels requires other methods to evaluate the models, such as Recall, Specificity, and Precision. These can also be visualized in the metric called Confusion metric.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{4}$$

Recall can also be called True Positive Rate (TPR) or Sensitivity, it is the ratio between the number of correctly classified as positive and all number of correct predictions, it is used to measure the ability of the model to define the positive samples.

$$Recall = \frac{TP}{(TP + FN)} \tag{5}$$

Specificity or True Negative Rate (TNR) is another way used to evaluate model performance, it is often compared to Recall. In contrast with Recall, Specificity is used to calculate how good the model can correctly predict the negative class.

$$Accuracy = \frac{TN}{(TN + FP)} \tag{6}$$

Precision or Positive Predictive Rate (PPR) is another way to indicate a machine learning model's performance. It uses to find the proportion between the number of correctly classified as positive and all number of positive class predictions.

$$Accuracy = \frac{TP}{(TP + FP)} \tag{7}$$

where The implication of the terms is given below:

- TP = True Positive
- FP = False Positive
- TN = True Negative
- FN = False Negative

F-score combines Recall and Precision metrics into a single metric by computing the average of them. It works well on an imbalanced class of response variables. $F_1 - score$ gives equivalent weight to precision and recall. $F_\beta - score$ is a more generic version which applies additional weights, making one of precision or recall more important than the other. The possible value of the $F - score$ varies from 0 to 1, 0 implies that either the precision or recall is 0 while the perfect precision or recall is indicated by the value of the $F - score$ equal to 1 [8].[17] The formula for the $F_1 - score$ is:

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{8}$$

While the general $F_\beta - score$ is represented by:

$$F_\beta = (1 + \beta^2) \times \frac{Precision \times Recall}{(\beta^2 \times Precision) + Recall} \tag{9}$$

where β represents the number of times the recall is considered more important than precision

Logistic Loss or Cross-Entropy loss (Log loss) is an important evaluation metric for binary classifiers. Classification models use this loss function to determine the subjective probability of a model based on the training data. Binary or multiple labels are required to define this loss [1].

$$Log\ loss_(y, p) = -(y \log(p) + (1 - y) \log(1 - p)) \tag{10}$$

The figure of Agarwal shown below explains that the increase in prediction certainty reduces the log loss[1].

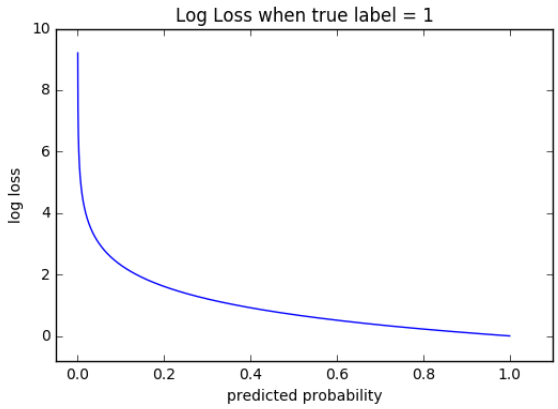

Figure 5: Logistic Loss or Cross-Entropy loss

Source: reference [1]

Area Under The Curve or AUC is defined as an area under the ROC curve. ROC curve shows the probability and AUC describe the measure of distinctness. It indicates how well the model can distinguish between classes. The higher predictions are correct, the larger AUC values become [19]. As explained by Sarang Narkhede, the ROC curve is plotted with True Positive Rate or Recall or Sensitivity on the y-axis against the Fault Positive Rate or 1- Specificity on the x-axis[19].

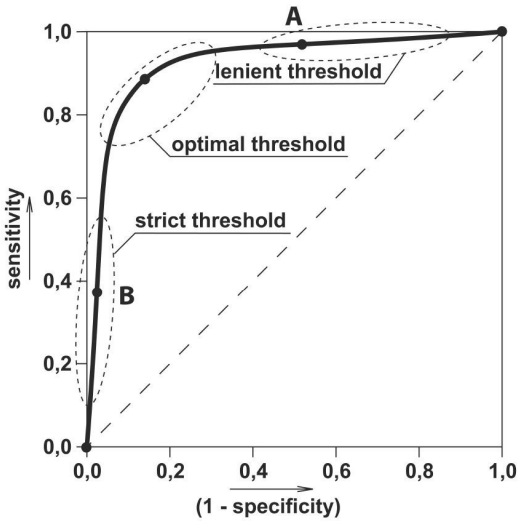

Figure 6: AUC - ROC Curve

Source: reference [19]

3 Analysis of results

A Comparative Study Using Machine Learning Techniques is conducted on the Wisconsin Breast Cancer dataset. 5 supervised machine learning techniques named support vector machine (SVM), K-nearest neighbours, random forests, artificial neural networks (ANNs) and logistic regression are compared in the paper of Islam et al., in 2020. The performances of each model are demonstrated in Table 4.

Table 4: Performances of breast cancer prediction system

Source: reference [15]

	SVM	K-NN	RF	ANN	LR
Accuracy (%)	97.14	97.14	95.71	98.57	95.71
Sensitivity (%)	100	97.82	95.65	100	95.74
Specificity (%)	92.3	95.83	95.83	96	95.65
Precision (%)	95.65	97.82	97.77	97.82	97.82
NPV (%)	100	95.83	92	100	91.66
FPR (%)	7.69	4.16	4.16	4	4.34
FNR (%)	0	2.17	4.34	0	4.25
F1 score	0.9777	0.9782	0.967	0.989	0.9677
MCC	0.9396	0.9365	0.9062	0.969	0.9043

It was found that ANNs produced the highest accuracy of 98.57%, while RFs and LRs produced the lowest accuracy of 95.7%.

Table 5: The comparison result of machine learning models

Source: reference [21]

		Accuracy	Time
SVM Kernel	linear	0.979	4 sec
	quadratic	0.981	3sec
	cubic	0.97	4 sec
k-NN Types	Fine	0.954	2 sec
	Medium	0.967	2 sec
	Coarse	0.932	2 sec
DT Types	Complex	0.937	5 sec
	Medium	0.937	3 sec
	Simple	0.923	2 sec

Another study on the Wisconsin dataset is proceeded by Obaid et al. in 2018, 3 machine learning algorithms including Support Vector Machine, K-nearest neighbours, and Decision tree, are implemented and compared in order to detect which model gives the better performance to diagnose the breast cancer. The accuracy and simulation time from these 3 models with 3 adjusted parameters are shown in Table 5. The best performance in this study was achieved from the SVM model with an accuracy of 98.1% [21].

Logistic regression, random forests, and support vector machine algorithms are implemented in the study of Patrício et al. The research is done using the Coimbra dataset. Different combinations of features are used as predictors. The AUC, sensitivity and specificity with confidence intervals of 95% were calculated in the test data set which illustrates in Table 6. Using SVM with the best combination of 4 predictors, which are glucose, resistin, age, and body mass index, provides the best sensitivity with 95% CI [82.2%, 87.5%] and specificity with 95% CI [84.5%, 89.7%][23].

Table 6: Multivariate analysis of how well the parameters allow distinguishing between patients with Breast Cancer and controls

Source: reference [23]

Variables	Figures of interest	Classifier		
		LR	RF	SVM
V1-V2	AUC	[0.76, 0.81]	[0.70, 0.75]	[0.76, 0.81]
	Sensitivity	[0.75, 0.81]	[0.75, 0.82]	[0.81, 0.86]
	Specificity	[0.73, 0.80]	[0.63, 0.70]	[0.70, 0.76]
V1-V3	AUC	[0.76, 0.80]	[0.81, 0.85]	[0.82, 0.86]
	Sensitivity	[0.74, 0.81]	[0.85, 0.90]	[0.87, 0.92]
	Specificity	[0.74, 0.80]	[0.72, 0.78]	[0.78, 0.83]
V1-V4	AUC	[0.79, 0.83]	[0.84, 0.88]	[0.87, 0.91]
	Sensitivity	[0.72, 0.78]	[0.80, 0.86]	[0.82, 0.88]
	Specificity	[0.80, 0.87]	[0.81, 0.87]	[0.84, 0.90]
V1-V5	AUC	[0.79, 0.83]	[0.82, 0.87]	[0.86, 0.90]
	Sensitivity	[0.73, 0.79]	[0.79, 0.85]	[0.84, 0.90]
	Specificity	[0.81, 0.87]	[0.77, 0.83]	[0.81, 0.87]
V1-V6	AUC	[0.78, 0.83]	[0.82, 0.86]	[0.83, 0.88]
	Sensitivity	[0.74, 0.80]	[0.79, 0.85]	[0.81, 0.86]
	Specificity	[0.79, 0.85]	[0.76, 0.82]	[0.80, 0.86]
V1-V9	AUC	[0.76, 0.81]	[0.78, 0.83]	[0.81, 0.85]
	Sensitivity	[0.70, 0.76]	[0.78, 0.85]	[0.75, 0.81]
	Specificity	[0.80, 0.86]	[0.70, 0.77]	[0.78, 0.84]

Another study on Coimbra Dataset is [2], 7 classification algorithms including Logistic Regression (LR), k-Nearest Neighbor (kNN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Gradient Boosting Method (GBM), and Naive Bayes (NB) are implemented in the research. The parameters used in each model are tuned as shown in Table 7.

Table 7: Best hyper-parameter

Source: reference [2]

Machine Learning Method	Best Hyper-parameter
kNN	N Neighbor = 1
Logistic (L2)	C = 10
Logistic (L1)	C = 5
Linear SVM (L2)	C = 0.001
Linear SVM (L1)	C = 3
Nonlinear SVM	C = 0.0001
Decision Tree	max depth = 37
Random Forest	n estimators = 100, max depth = 12
Gradient Boosting	n estimators = 200, max depth = 12, learning rate = 0.1
Naive Bayes Gaussian	Not Now

The accuracy obtained from each model is shown in the table below. The highest accuracy is obtained from Gradient Boosting with 74.14% of accuracy. While SVM with L1 norm and Logistic Regression, using L2 norm provide the second and third accuracy scores at 72.52% and 72.48% respectively.

Table 8: Classification accuracy

Source: reference [23]

Machine Learning Method	Accuracy
kNN	0.5814
Logistic (L2)	0.7248
Logistic (L1)	0.721
Linear SVM (L2)	0.6959
Linear SVM (L1)	0.7252
Nonlinear SVM	0.6038
Decision Tree	0.6928
Random Forest	0.7031
Gradient Boosting	0.7414
Naive Bayes Gaussian	0.6238

4 Conclusion

Several studies confirm that breast cancer prediction by the biopsy method gives a better result than the one received from the biomarkers clinical approach. The study by Islam et al. showed that ANNs formed the highest accuracy of 98.57% with the Sensitivity, Specificity and Precision at 100%, 96% and 97.82%, respectively [15]. While the best performance from the Coimbra dataset by the SVM model gives AUC with 95% of CI at [0.87, 0.91] in the study of Patricio et al [23]. These demonstrate that although there are weak points (expensive and long lead time) in the biopsy approach, it provides the satisfaction result as better performance. For the biomarkers, the clinical approach which is reachable by any individual, even though the implementation of machine learning algorithms on it might not give the best performance, its result is useful to predict the potential breast cancer patient so they can proceed next consultation or treatment processes. In the conclusion of this study, the Coimbra dataset can be utilized as the biomarker of breast cancer prognosis. Implementing machine learning to this dataset can be applied as a pre-process to screen for breast cancer before using the biopsy approach.

References

[1] Rahul Agarwal. *The 5 Classification Evaluation metrics every Data Scientist must know*. Medium, Sept. 2019. URL: https://towardsdatascience.com/the-5-classification-evaluation-metrics-you-must-know-aa97784ff226 (visited on 07/29/2022).

[2] Yolanda Austria et al. "Comparison of Machine Learning Algorithms in Breast Cancer Prediction Using the Coimbra Dataset". In: *International journal of simulation: systems, science technology* (July 2019). DOI: 10.5013/IJSSST.a.20.S2.23.

[3] Yolanda Austria et al. "Comparison of Machine Learning Algorithms in Breast Cancer Prediction Using the Coimbra Dataset". In: *International journal of simulation: systems, science technology* (July 2019). DOI: 10.5013/IJSSST.a.20.S2.23.

[4] Breastcancer.org. *Breast Cancer Facts and Statistics*. Breastcancer.org, 2022. URL: https://www.breastcancer.org/facts-statistics?gclid=Cj0KCQjw2MWVBhCQARIsAIjbwoOQbFgpw3930CpAFwaZp9C-Qa4vKO-oa_gz6950aOf9dEXYGtOeZhMaAj_OEALw_wcB (visited on 06/21/2022).

[5] Jason Brownlee. *A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning*. Machine Learning Mastery, Sept. 2016. URL: https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/ (visited on 08/02/2022).

[6] NHS Choices. *Overview - Biopsy*. 2022. URL: https://www.nhs.uk/conditions/biopsy/#:~:text=A%20biopsy%20is%20a%20medical,skin%2C%20organs%20and%20other%20structures. (visited on 08/02/2022).

[7] Wikipedia Contributors. *Artificial neural network*. Wikipedia, July 2022. URL: https://en.wikipedia.org/wiki/Artificial_neural_network (visited on 07/29/2022).

[8] Wikipedia Contributors. *F-score*. Wikipedia, July 2022. URL: https://en.wikipedia.org/wiki/F-score (visited on 08/02/2022).

[9] Wikipedia Contributors. *Gradient boosting*. Wikipedia, July 2022. URL: https://en.wikipedia.org/wiki/Gradient_boosting (visited on 08/02/2022).

[10] Wikipedia Contributors. *Logistic regression*. Wikipedia, Aug. 2022. URL: https://en.wikipedia.org/wiki/Logistic_regression#Problem (visited on 08/02/2022).

[11] Wikipedia Contributors. *Naive Bayes classifier*. Wikipedia, July 2022. URL: https://en.wikipedia.org/wiki/Naive_Bayes_classifier (visited on 07/28/2022).

[12] Wikipedia Contributors. *Random forest*. Wikipedia, June 2022. URL: https://en.wikipedia.org/wiki/Random_forest#:~:text=Random%20forests%20or%20random%20decision,class%20selected%20by%20most%20trees. (visited on 08/02/2022).

[13] Joana Crisóstomo et al. "Hyperresistinemia and metabolic dysregulation: a risky crosstalk in obese breast cancer". In: *Endocrine* 53 (Feb. 2016), pp. 433–442. DOI: 10.1007/s12020-016-0893-x. URL: https://link.springer.com/article/10.1007/s12020-016-0893-x (visited on 07/25/2022).

[14] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: http://archive.ics.uci.edu/ml.

[15] Md. Milon Islam et al. "Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques". In: *SN Computer Science* 1 (Sept. 2020). DOI: 10.1007/s42979-020-00305-w. URL: https://link.springer.com/article/10.1007/s42979-020-00305-w (visited on 06/03/2022).

[16] KDnuggets. *Naive Bayes Algorithm: Everything You Need to Know - KDnuggets*. KDnuggets, 2022. URL: https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html (visited on 07/28/2022).

[17] Joos Korstanje. *The F1 score — Towards Data Science*. Medium, Aug. 2021. URL: https://towardsdatascience.com/the-f1-score-bec2bbc38aa6 (visited on 08/02/2022).

[18] Madhu Kumari and Vijendra Singh. "Breast Cancer Prediction system". In: *Procedia Computer Science* 132 (2018), pp. 371–376. DOI: 10.1016/j.procs.2018.05.197. URL: https://reader.elsevier.com/reader/sd/pii/S1877050918309323?token=1CC0F3317A6315E6085694781C originRegion=eu-west-1&originCreation=20220726145934 (visited on 07/26/2022).

[19] Sarang Narkhede. *Understanding AUC - ROC Curve - Towards Data Science*. Medium, June 2018. URL: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 (visited on 07/29/2022).

[20] News-Medical. *What is a Biomarker?* News-Medical.net, June 2010. URL: https://www.news-medical.net/health/What-is-a-Biomarker.aspx (visited on 08/02/2022).

[21] Omar Ibrahim Obaid et al. "Evaluating the Performance of Machine Learning Techniques in the Classification of Wisconsin Breast Cancer". In: *International Journal of Engineering Technology* 7 (Dec. 2018), pp. 160–166. DOI: 10.14419/ijet.v7i4.36.23737. URL: https://www.sciencepubco.com/index.php/ijet/article/view/23737.

[22] World Health Organization. *Breast cancer*. Who.int, Mar. 2021. URL: https://www.who.int/news-room/fact-sheets/detail/breast-cancer (visited on 06/02/2022).

[23] Miguel Patrício et al. "Using Resistin, glucose, age and BMI to predict the presence of breast cancer". In: *BMC Cancer* 18 (Jan. 2018). DOI: 10.1186/s12885-017-3877-1. URL: https://bmccancer.biomedcentral.com/articles/10.1186/s12885-017-3877-1 (visited on 07/26/2022).

[24] Stuart J Russell and Peter Norvig. *Artificial intelligence : a modern approach*. Prentice-Hall, 2010.

[25] Cancer Research UK. *Breast cancer statistics*. Cancer Research UK, May 2015. URL: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer#heading-Zero (visited on 06/22/2022).

[26] Rakshith Vasudev. *How are Logistic Regression Ordinary Least Squares Regression (Linear Regression) Related? Why the "Regression" in Logistic?* Medium, June 2018. URL: https://towardsdatascience.com/how-are-logistic-regression-ordinary-least-squares-regression-related-1deab32d79f5 (visited on 08/02/2022).

[27] Lulu Wang. "Early Diagnosis of Breast Cancer". In: *Sensors* 17.7 (2017). ISSN: 1424-8220. DOI: 10.3390/s17071572. URL: https://www.mdpi.com/1424-8220/17/7/1572.