Comparison of Machine Learning Algorithms in Breast Cancer Prediction Using the Coimbra Dataset

Article *in* International Journal of Simulation: Systems, Science & Technology · July 2019

DOI: 10.5013/IJSSST.a.20.S2.23

6 authors, including:

Yolanda Austria
Adamson University
16 PUBLICATIONS **108** CITATIONS
SEE PROFILE

Marie Luvett Goh
FEU Institute of Technology
11 PUBLICATIONS **96** CITATIONS
SEE PROFILE

Jay-Ar Lalata
FEU Institute of Technology
11 PUBLICATIONS **138** CITATIONS
SEE PROFILE

Joselito Eduard E. Goh
De La Salle-College of Saint Benilde
5 PUBLICATIONS **38** CITATIONS
SEE PROFILE

Comparison of Machine Learning Algorithms in Breast Cancer Prediction using the Coimbra Dataset

Yolanda D. Austria [1], Jay-ar P. Lalata [2], Lorenzo B. Sta. Maria, Jr. [3], Joselito Eduard E. Goh [4]
Marie Luvett I. Goh [2], Heintjie N.Vicente [2]

[1] Adamson University, San Marcelino St. Ermita, Manila, Philippines.
[2] FEU Institute of Technology, P. Paredes St. Sampaloc, Manila, Philippines.
[3] Asian Institute of Management, Paseo de Roxas, Legazpi Village, Makati, Philippines.
[4] De La Salle – College of St. Benilde, Taft Ave., Malate, Manila, Philippines.

yolanda.austria@adamson.edu.ph; jayar_030181@yahoo.com; lorenzo_stamaria@yahoo.com; joedgoh@gmail.com; luvett.goh@gmail.com; hnvicente@feutech.edu.ph

***Abstract* - In the medical field, machine learning (ML) techniques are playing a significant and growing role because of their high potential in helping health practitioners make decisions and diagnosis. This inspective research aims to review ML models that may predict breast cancer in women and to compare their performances. A number of clinical features were measured among the 116 participants in the dataset of this study including insulin, glucose, resistin, adiponectin, homeostasis model assessment (HOMA), leptin, monocyte chemoattractant protein-1 (MCP-1), along with their age and body mass index (BMI). The researchers implemented 11 classification algorithms and their variations 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), in the detection of breast cancer on the publicly available Coimbra Breast Cancer Dataset (CBCD). Each classifier applies a unique hyper-parameter setting to perform prediction and their performances are compared in identifying breast cancer. As a conclusion of this study, Gradient Boosting (GB) machine learning algorithm is the best classifier in predicting breast cancer using the Coimbra Breast Cancer Dataset (CBCD) with an accuracy of 74.14%. k-Nearest Neighbor (kNN) classifier produces the fastest training time at 0.000598 seconds while Nonlinear Support Vector Machine (SVM) classifier gives with the fastest testing time at 0 seconds. Another conclusion of this paper is that the body mass index (BMI) is the top predictor, with 50% of the classifiers observing it as their top predictor and Glucose comes in second. This recommends that they may be a good pair of variables, which may predict breast cancer in women.**

Keywords - breast cancer, machine learning algorithm, classifier, Logistic Regression (LR), k-Nearest Neighbor (kNN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Gradient Boosting Method (GBM), Naive Bayes (NB)

I. INTRODUCTION

According to the World Health Organization (WHO) one of the primary causes of death worldwide in 2018 is cancer. In the estimated 9.6 million deaths in cancer, breast cancer is the second most prevalent cancer, subsequent to lung cancer, with 2.09 million cases. It is also the fifth most common reason of cancer death, with an approximated 627 000 deaths, that is estimated 15% of all cancer deaths among women. [1] And in all new cancer diagnoses for women, breast cancer alone accounts for 30% all these new cases [2].

At present, X-ray mammography is the lone procedure that has the capability of detecting early-stage breast cancer, or before the cancer is self-evident. It is also the basis of the most systematized breast screening programs to detect breast cancer in an asymptomatic population. To successfully detect breast cancer in its beginning phase, however, mammography must sufficiently differentiate small masses and micro-calcifications, which in principle can only produce subtle contrast differences in mammography images [3].

Even though mammography is currently the widely used standard screening process for breast cancer, the incidents of incorrect classifications of mammograms, is still one of the areas for improvement in breast cancer forecasting. Thus, there is still a challenge to discover effective predictors, which may come from cheap and easily accessible methods. Bodily parameters, such as those obtainable from blood samples, may provide alternative ways to better diagnose breast cancer among women [4].

Alternative ways of detecting breast cancer, specifically, ones that are non-invasive are evident in several recent studies. Exhaled breath and urine analysis, for instance, were used in a study of a non-invasive early discovery of breast cancer using an Artificial Neural Network (ANN) model. [5] The combination of age, body mass index (BMI), and metabolic parameters, in another paper, was concluded as a potential inexpensive and effective predictor for breast cancer [6].

This research paper used the same dataset that was employed in [6], wherein instead of using mammography images in breast cancer prediction, it used age, body mass index (BMI), and clinical features that may be extracted from a routine blood analysis only. However in the said study, the researchers only applied ML algorithms to the four parameters which are Glucose, Resistin, Age and BMI. This paper used all of the nine attributes to verify which among may be considered as the top predictor. Also, the researchers used several classifiers and their variations using different hyper-parameters, namely, k-Nearest Neighbor (kNN), Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), and Naive Bayes (NB), in the detection of breast cancer on the publicly available Coimbra Breast Cancer Dataset (CBCD) using codes created in Python.

The rest of this research paper is structured as follows. In Section 2, the risk factors for breast cancer and the theory of different machine learning (ML) algorithms are discussed, and the related literature are cited. In Section 3, the description of the CBCD and the experimental work are presented. In Section 4, the performance evaluation of the different models is discussed. Finally, in Section 5, the researchers provided the conclusion on the paper.

II. BACKGROUND OF THE STUDY

A. Risk Factors

Currently, there is an increasing number of studies that suggest that a weakened absorption of glucose can be a threat for the progression of several types of cancer. Among diabetic patients, numerous medical studies indicated rising occurrences of different kinds of cancers, such as those of the kidney, liver, colorectal, pancreas, and breast. In a continuous study that has spun for ten years, it has shown that a non-working glucose absorption or diabetes has a strong positive correlation to death rate related to breast cancer [7, 10].

High Body Mass Index (BMI) is another risk factor for cancer. In the result of a study in Brazil, high BMI accounts for 15 000 (3.8%) of all new cancer cases that are diagnosed in the country. And it was observed that high BMI was greater in women, specifically in breast cancer cases. [8]

Another study that was made links several adipocytokines with breast cancer cells in obesity. The study shows that development of cancer is most likely when there is an increase in the levels of leptin, resistin, and a decrease of adiponectin secretion [9, 10].

In a different study, women after menopause have large possibility in having a breast cancer through the Metabolic Syndrome, specifically insulin resistance and abdominal fat. The researchers have proposed that to identify the patients with subclinical insulin resistance, Homeostasis Model Assessment – Insulin Resistance (HOMA-IR) can be used. This is important for prevention and testing patient which are at high-risk [11].

Being obese is an ailment that is also considered as a dangerous risk for breast cancer. Through the molecular mechanisms such as compensatory hyperinsulinemia to insulin resistance and high levels of insulin and insulin-like growth factor (IGF-1), changes in the leptin and the hypoxia, and the release of larger pro-inflammatory cytokines, obesity is now considered as a possible influence for breast cancer in premenopausal women [22].

B. Machine Learning Algorithms

Classification is one of the most essential tasks in ML and data mining. Through the years, it has shown that ML algorithms have significantly improved their accuracy and correctness in the method of classification. Despite the fact that the amount of data is continuously increasing and becoming more complex, ML algorithms still continue to mature and improve. In the year 2004, the statistical model Linear Regression has been tested and compared with 2 ML models for classification, DT and ANN, to guess the survival rate of patients with breast cancer, utilizing a huge database which has 200 000 cases above.

In 2004, two ML classification methods, Decision tree (DT) and Artificial Neural Network (ANN), were compared with a statistical method, linear regression, to predict the breast cancer survival using a large dataset which has more than 200 000 cases. The study showed that for real-world usage, ML algorithms could be highly possible classification methods. The outcomes revealed that DT was the best classifier with a precision of 93.6%, ANN attained a 91.2% accuracy, and both were superior than linear regression reaching only 89.2% accuracy. In addition, an evaluation of current studies tells that almost all the ML algorithms employed in the breast cancer analysis and prediction are supervised [13].

A large number of researches have already been conducted in the application of data mining and ML on different medical datasets to classify breast cancer. Numerous studies show remarkable accurateness in classification. For the longest time, a standard dataset has been widely used in the literature: Wisconsin breast cancer diagnosis (WBCD). This database has 11 attributes, namely, Sample code number, Clump thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoil, Mitoses, and Class [13].

The difference of the CBCD from WBCD is that the former's attributes are clinical features that may be extracted from a routine blood analysis only.

Based from the ML algorithms that included the use of SVMs, DTs, ANNs, and k-NNs in WBCD using the entire 699 samples, it shows that the top six algorithms which resulted in the highest accuracies are Genetically Optimized Neural Network (GONN) with 100%, Deep Belief

Network–Neural Network (DBN-NN) with 99.68%, F-Score-SVM with 99.51%, PSO-SVM with 99.31%, and Artificial Meta-plascticity Multilayer Perceptron (AMMLP) at 99.26% [13].

A GONN for classifying breast cancer cases was used in a research that resulted to classification accuracy of 98.24%, 99.63% and 100% for 50-50, 60-40, 70-30 training-testing partition respectively and 100% for a 10-fold cross validation. [14] A Computer-Aided Diagnosis (CAD) scheme for detection of breast cancer has been developed using DBN unsupervised path followed by back propagation supervised path. It resulted in an overall accuracy of 99.68% with 100% sensitivity and 99.47% specificity, using a 54.9% of training data and 45.1% of testing data. [15] In the study which obtained a classification accuracy of 99.51%, the breast cancer diagnosis was grounded on the SVM-based method combined with feature selection using a 10-fold cross validation. [16] A swarm intelligence technique-based SVM is proposed for breast cancer diagnosis that obtained a classification accuracy of 99.31%, Compared with existing methods when this study was conducted, it had a higher accuracy percentage via 10-fold cross validation analysis. [17] The performance of AMMLP algorithm was tested using classification accuracy and it was obtained at 99.26% using a 60-40 training-testing sampling strategy [18].

In this study, the researchers used a number of ML algorithms, namely, k-Nearest Neighbor (kNN), Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), and Naive Bayes (NB).

III. METHODOLOGY

A. Coimbra Breast Cancer Dataset (CBCD)

The dataset used in this study is the CBCD from the UC Irvine (UCI) Machine Learning Repository. Between 2009 and 2013, female patients who have been identified with breast cancer were recruited by the Gynecology Department of the Coimbra Hospital and University Center (CHUC) in Portugal. The finding was derived from a positive mammography and was histologically validated. All samples were collected before any operation or management has been done on the patient. For controls, healthy women volunteers were chosen and included in the research. [6]

Part of the limitation of this study is the small sample size, however, there are a number of existing researches that employed ML techniques which also used limited number of instances. A paper describing an operational decision making system in healthcare based on ML classifiers to forecast the decisions in comparison to the actual judgements of doctors during the healthcare procedures used a dataset that only has 80 samples.[25] Another study employed a dataset which only has 76 instances, wherein the authors developed a method to study computers capability in diagnosing gastrointestinal lesions from regular colonoscopic videos compared to two levels of clinical knowledge [26].

Data gathered from the participants comprised of their age, weight, height, BMI, and menopausal status. In addition, several measurements – glucose, insulin, Homeostasis Model Assessment (HOMA), Leptin, Adiponectin, Resistin, and Monocyte Chemoattractant Protein-1 (MCP-1). were extracted at the Laboratory of Physiology of the Faculty of Medicine of University of Coimbra from peripheral venous blood vials collected in the hospital for all participants [6].

Utilizing the same publicly available CBCD in an earlier study, prediction models are developed and assessed by using clinical features that are extracted from a routine blood analysis. Several models have been proposed for breast cancer detection, including LR, RF and SVM. Based from the results of the research, SVM classifier using the group of age, glucose, Resistin, together with BMI as forecasters allowed predicting the breast cancer existence in women with specificity stretching between 85 and 90% and sensitivity extending from 82 to 88%. The 95% confidence interval for the area under the curve (AUC) was [0.87,0.91] [6].

The purpose of this research is to employ the same dataset in comparing the performances of different ML algorithms in detecting breast cancer. This dataset is composed of ten quantitative predictors and a dependent variable “classification”, which indicates the manifestation of breast cancer. If these classifier algorithms based on the attributes are validated to be correct, they may potentially be used as an instrument or a screening tool for the detection of breast cancer [6].

The authors of the paper employed statistical analysis on each attribute of the dataset. To further visualize the spread of data, standard deviation was obtained. The minimum and maximum values were also acquired for each feature, as well as the mean or the average.

B. Data Classification

Data classification is a method where an assumption is selected from a set of choices that best fits a set of interpretations. It includes two steps: training and testing [19].

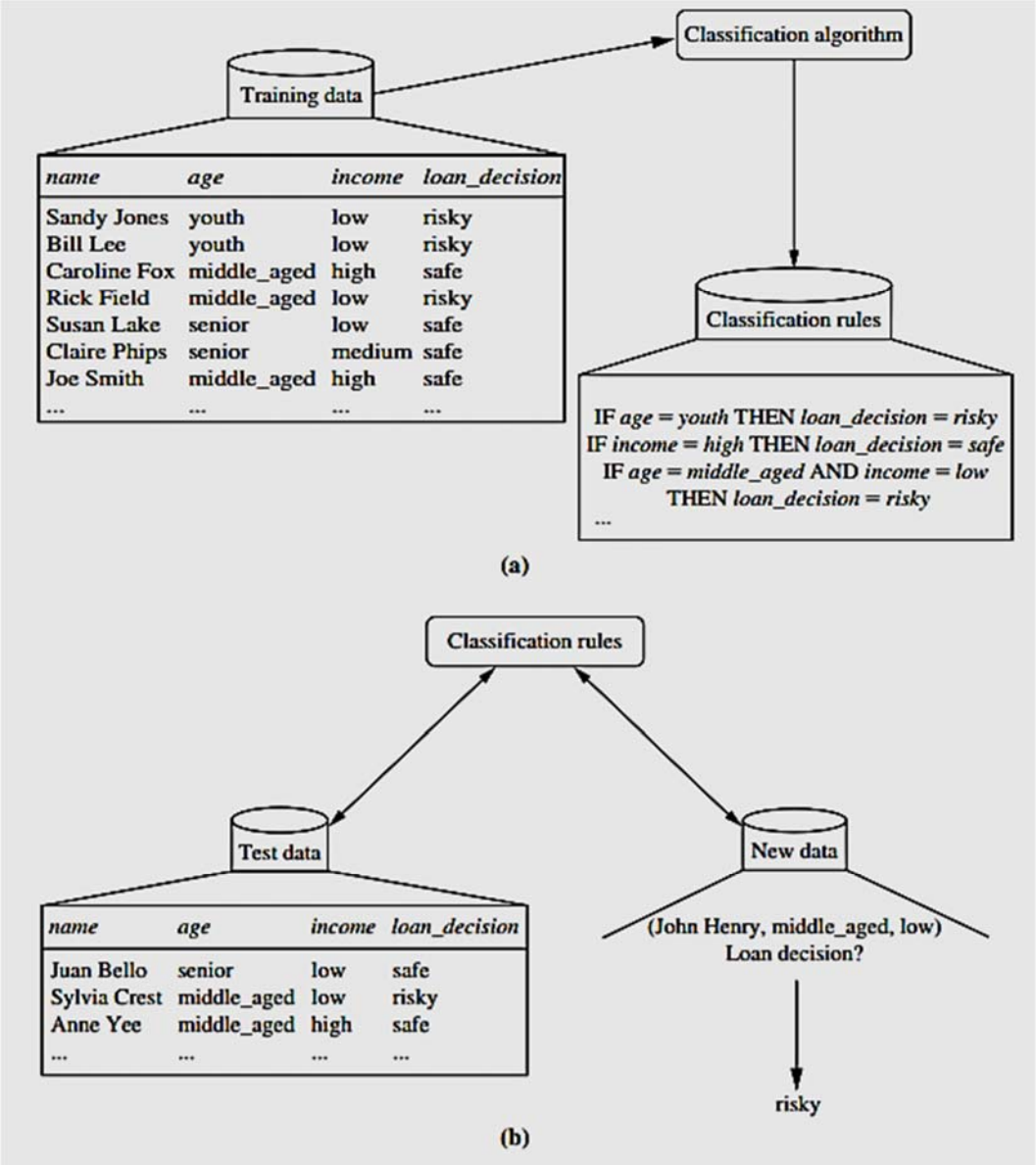

Figure 1. Data classification process: (a) Learning (b) Classification [20]

Figure 1 shows the two phases in data classification. First is the learning step or the training phase, where a classifier model is generated by analyzing the training set. The training set consists of a number of examples, each of which has a number of attribute values and one label. If the class label of each training tuple or sample is given, it is known as supervised learning because it is told which class each training tuple belongs. The result of this analysis is a model that attempts to make generalizations about how the attributes relate to the label [20, 27].

In the second phase, the model is applied to the test data, where the labels are unknown. The accuracy of a model on a given test set is the percentage of test set tuples that it was able to correctly categorize [20, 27].

Given the circumstance that the CBCD is composed of attributes that are already categorized into class labels and this study is aiming to construct a model that will predict the existence of breast cancer among women, the researchers used a supervised classification task.

For the distribution of the training and the data set, the CBCD dataset has been randomly allocated to a 75% training set and a 25% testing set. The set of data for people that were used to construct the training set are not used for the testing set.

This distribution is different from the previous study on the CBCD dataset wherein the researchers chose 70% for the training set and 30% for the testing set. The training set came from 70% of controls – 36 out of the 52 healthy

women, and 70% of patients – 45 out of the 64 patients diagnosed with cancer [6].

The CBCD constitutes of nine numerical predictors and a binary dependent variable, indicative of the presence of breast cancer. The predictors are anthropometric features and variables that were collected through a routine blood analysis. Clinical predictors were distinguished and collected from 64 patients with breast cancer and 52 fit and well women as controls, for a total of 116 instances [12].

TABLE 1. ATTRIBUTES OF THE COIMBRA BREAST CANCER DATASET

Quantitative Attributes	Count	Status	Data Type	Unit
Age	116	non-null	int64	years
BMI	116	non-null	float64	kg/m^2
Glucose	116	non-null	int64	mg/dL
Insulin	116	non-null	float64	µU/mL
HOMA	116	non-null	float64	
Leptin	116	non-null	float64	ng/mL
Adiponectin	116	non-null	float64	µg/mL
Resistin	116	non-null	float64	ng/mL
MCP.1	116	non-null	float64	pg / dL
Classification	116	non-null	int64	

Table I shows the count, data type, if empty or with a value, and the unit of each of the ten attributes of CBCD. It shows that there is a total of 116 rows for all attributes, which means that there are no null or missing values in the dataset used. Since there are no missing values present in the dataset, this states that we do not need to perform additional processing in the dataset and can be explored. Age, glucose, and classification, are all of integer data type and the rest of the attributes are of data type float.

The given dataset does not require extensive processing prior to analysis. The only necessary step needed is to map the target variable which is the classification.

For the settings of the imports in Python, the authors used NumPy library for fast numeric array computations, pandas for data analysis, matplotlib.pyplot for plotting, and seaborn for data visualization. Also, the researchers imported the needed libraries from Scikit-learn, one of Python's free machine learning libraries for the various classification algorithms including kNN, LR, SVC, NB, GB, DT, and RF.

The researchers then applied a heat-map analysis using a Spearman correlation matrix of the attributes. It is a visualization method in two-dimension where numerical values are displayed in colors and arranged in rows and columns. [23] The heat-map was coded in Python using the function heat-map from the Seaborn library.

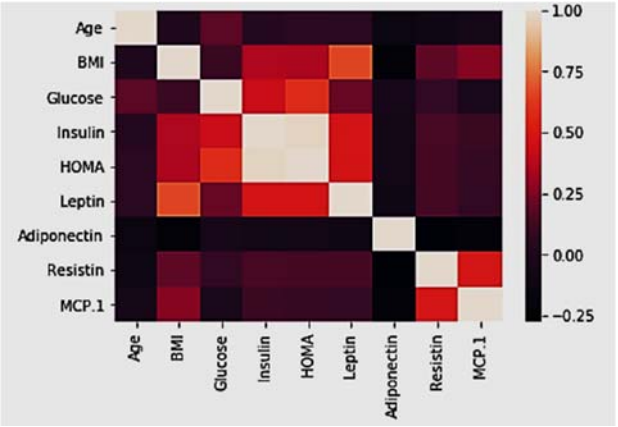

Figure 2. Heat-map Analysis of Features

Figure 2 shows the heat-map analysis of the nine attributes to show their correlation. The colors show how one parameter is associated with another parameter through the colors displayed. Lighter colors show that two attributes are highly correlated, white being the most correlated with a value of 1.00. Darker colors, on the other hand, show that two parameters are poorly correlated.

TABLE II. HYPER-PARAMETER USED FOR EACH MODEL

Classifier Name	Hyper-parameter	Type
k-Nearest Neighbor	n_neighbors	similarity-based learning
Logistic Regression	C	error-based learning
Support Vector Machine	C	error-based learning
Decision Tree	max_depth	information-based learning
Random Forest	n_estimators, max_features, max_depth	information-based learning
Gradient Boosting	n_estimators, max_features, learning_rate	information-based learning
Naive-Bayes	alpha	probability-based learning

Table II shows the hyper-parameters that are used in each ML model in this study, as well as their type of learning. The value of this hyper-parameter is a configuration external to the model, which is set before the learning process starts. Its value is often set by the person doing the study and is tuned for a certain predictive modeling ptoblem. Given these hyper-parameters, the training algorithm learns the parameters from the data. [24]

IV. RESULTS AND DISCUSSION

The researchers run the code in Python with the number of trials set at 100 for each ML model. It was processed in a computer with the following specifications – 4 Intel® Core™ i3-6006U CPU running at 2.00GHz Processor, 11.6 GiB of memory storage and a 64-bit operating system.

TABLE III. STATISTICAL ANALYSIS OF ATTRIBUTES

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 III shows the mean, standard deviation, minimum, and maximum value of each attribute. Mean is simply the average of the values for each attribute, Min is the lowest value, Max is the highest value, and Standard Deviation (SD) is a measure of how spread the numbers are. In this case, the attributes with the lowest SD are HOMA, BMI, and Adiponectin, respectively.

Each model has resulted to its test accuracy, best hyper-parameter, top predictor variable, training time, and test time.

TABLE IV. CLASSIFICATION ACCURACY

Machine Learning Method	Accuracy
kNN	58.14%
Logistic (L2)	72.48%
Logistic (L1)	72.10%
Linear SVM (L2)	69.59%
Linear SVM (L1)	72.52%
Nonlinear SVM	60.38%
Decision Tree	69.28%
Random Forest	70.31%
Gradient Boosting	74.14%
Naive Bayes Gaussian	62.38%

Table IV above shows the accuracy percentage result of each ML algorithm. The accuracy of the model is to calculate the ratio of the total correct predictions out of all predictions made, called the classification accuracy. The accuracy of each machine learning algorithm was derived from a confusion matrix. Accuracy is calculated as the number of all correct predictions divided by the total number of the dataset. Gradient Boosting produced the highest accuracy at 74.14%. Second highest accuracy is the Linear Support Vector Machine, using L1 norm, at 72.52%. The third highest accuracy result is from the ML Logistic Regression, using L2 norm, at 72.48%.

Table V below shows the best hyper-parameter that was used by each of the classifier. For kNN, the best hyper-parameter was the value 1 for its k, C=10 for LR using L2 norm, C=5 for LR using L1 norm, C=0.001 for Linear SVM using L2 norm, C=3 for Linear SVM using L1 norm, C=0.0001 for Nonlinear SVM, maximum depth of 37 for DT, 100 number of estimators and maximum depth of 12 for RF, 200 number of estimators and maximum depth of 12 and learning rate of 0.1 for GB, and NB was not able to result in its best hyper-parameter.

TABLE V. BEST HYPER-PARAMETER

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

TABLE VI. TOP PREDICTOR VARIABLE OF EACH MACHINE

Machine Learning Method	Top Predictor
kNN	Not Now
Logistic (L2)	BMI
Logistic (L1)	BMI
Linear SVM (L2)	BMI
Linear SVM (L1)	BMI
Nonlinear SVM	Not Now
Decision Tree	BMI
Random Forest	Glucose
Gradient Boosting	Glucose
Naive Bayes Gaussian	Not Now

Table VI above shows the top predictor for each of the ML model used in the research. The top predictor is BMI using the classifiers LR, Linear SVM, and DT. The second top predictor is Glucose using the ML models RF and GB. The classifiers kNN, Nonlinear SVM, and NB did not result in a top predictor.

TABLE VII TRAINING TIME AND TEST TIME

Machine Learning Method	Training Time (sec.)	Testing Time (sec.)
kNN	0.000598	0.000802
Logistic (L2)	0.001618	0.000269
Logistic (L1)	0.010455	0.000328
Linear SVM (L2)	0.007651	0.000298
Linear SVM (L1)	0.005616	0.000341
Nonlinear SVM	1	0
Decision Tree	0.000912	0.000302
Random Forest	0.14128	0.00809
Gradient Boosting	0.120435	0.000822
Naive Bayes Gaussian	0.000996	0.000405

Table VII above illustrates the training time and testing time of each classifier used in the study. The top three models which are able to learn the fastest, listed chronologically, are kNN with 0.000598 sec., DT with

0.000912 sec., and NB with 0.000996 sec. The top three models with the fastest testing time are Nonlinear SVM 0 sec., LR(L2) with 0.000269 sec., and Linear SVM with 0.000298 sec.

V. CONCLUSIONS AND RECOMMENDATIONS

We concludes that:

- Gradient Boosting (GB) machine learning algorithm is the best classifier in predicting breast cancer using the Coimbra Breast Cancer Dataset (CBCD) with an accuracy of 74.14%.
- Body Mass Index (BMI) is the top predictor observed in this study, with 50% of the classifiers observing it as their top predictor. The second top predictor observed is Glucose at 20%. This supports the previous study [6], wherein these two top predictors were also used by the researchers to predict breast cancer on the same dataset.
- The classifier with the fastest training time at 0.000598 sec. is the k-Nearest Neighbor (kNN).
- Nonlinear Support Vector Machine (SVM) is the classifier with the fastest testing time at 0 seconds.

Recommendations

Further researches may use different percentage distribution of training, validation, and test groups. The top three ML models which produced the highest accuracy, namely, Gradient Boosting, Linear SVM (L1), and Logistic Regression (L2), may be the algorithms of choice for future improvement of the study. An ensemble of multiple ML algorithms may also be used in the future to increase accuracy. Scaling may also be applied to the features and future studies may compare if there is an improvement in their performance.

ACKNOWLEDGMENT

The authors wish to thank Mr. Miguel Patricio of University of Coimbra, Portugal, for providing the dataset used in this study.

AUTHORS' CONTRIBUTIONS

All authors contributed to the conception and design. LS and JL contributed to the creation and execution of the classifiers. LG and JG contributed to the interpretation of data. YA and HV contributed to the review of the manuscript and the writing of the article. All authors have read and approved the final manuscript.

REFERENCES

[1] "Breast cancer", World Health Organization, 2018. [Online]. Available: http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/. [Accessed: 24- Sep- 2018].

[2] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2018," CA. Cancer J. Clin., vol. 68, no. 1, pp. 7–30, 2018.

[3] P. Mora et al., "Improvement of early detection of breast cancer through collaborative multi-country efforts: Medical physics component," Phys. Medica, vol. 48, no. December 2017, pp. 127–134, 2018.

[4] S. Y. Loke and A. S. G. Lee, "The future of blood-based biomarkers for the early detection of breast cancer," Eur. J. Cancer, vol. 92, pp. 54–68, 2018.

[5] O. Herman-Saffar, Z. Boger, S. Libson, D. Lieberman, R. Gonen, and Y. Zeiri, "Early non-invasive detection of breast cancer using exhaled breath and urine analysis," Comput. Biol. Med., vol. 96, no. February, pp. 227–232, 2018.

[6] M. Patrício et al., "Using Resistin, glucose, age, and BMI to predict the presence of breast cancer," BMC Cancer, vol. 18, no. 1, pp. 1–8, 2018.

[7] S. M. Samuel, E. Varghese, S. Varghese, and D. Büsselberg, "Challenges and perspectives in the treatment of diabetes associated breast cancer," Cancer Treat. Rev., vol. 70, no. August, pp. 98–111, 2018.

[8] L. F. M. de Rezende et al., "The increasing burden of cancer attributable to high body mass index in Brazil," Cancer Epidemiol., vol. 54, no. February, pp. 63–70, 2018.

[9] J. Li and X. Han, "Adipocytokines and breast cancer," Curr. Probl. Cancer, vol. 42, no. 2, pp. 208–214, 2018.

[10] Crisóstomo, J., Matafome, P., Santos-Silva, D. et al. Endocrine (2016) 53: 433. https://doi.org/10.1007/s12020-016-0893-x

[11] I. Capasso et al., "Homeostasis model assessment to detect insulin resistance and identify patients at high risk of breast cancer development: National Cancer Institute of Naples experience," J. Exp. Clin. Cancer Res., vol. 32, no. 1, p. 1, 2013.

[12] "UCI Machine Learning Repository: Breast Cancer Coimbra Data Set", Archive.ics.uci.edu, 2018. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra. [Accessed: 20- Sep- 2018].

[13] W. Yue, Z. Wang, H. Chen, A. Payne, and X. Liu, "Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis," Designs, vol. 2, no. 2, p. 13, 2018.

[14] Bhardwaj, A.; Tiwari, A. Breast cancer diagnosis using genetically optimized neural network model. Expert Syst. Appl. 2015, 42, 4611–4620.

[15] A. M. Abdel-Zaher and A. M. Eldeib, "Breast cancer classification using deep belief networks," Expert Syst. Appl., vol. 46, pp. 139–144, 2016.

[16] Akay, M.F. Support vector machines combined with feature selection for breast cancer diagnosis. Expert Syst. Appl. 2009, 36, 3240–3247.

[17] Chen, H.-L.; Yang, B.; Wang, G.; Wang, S.-J.; Liu, J.; Liu, D.-Y. Support vector machine based diagnostic system for breast cancer using swarm intelligence. J. Med. Syst. 2012, 36, 2505–2519.

[18] Marcano-Cedeño, A.; Quintanilla-Domłnguez, J.; Andina, D. WBCD breast cancer database classification applying artificial metaplasticity neural network. Expert Syst. Appl. 2011, 38, 9573–9579.

[19] A. Naik and L. Samant, "Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange, and Knime," Procedia Comput. Sci., vol. 85, no. Cms, pp. 662–668, 2016.

[20] J. Han et al., "Classification: Basic Concepts," in Data Mining Concepts and Techniques, 3rd ed. 225 Wyman Street, Waltham, MA 02451, USA: M. Kauffman.

[21] "Classification and clustering", IBM Developer, 2018. [Online]. Available: https://www.ibm.com/developerworks/library/os-weka2/os-weka2-pdf.pdf. [Accessed: 20- Sep- 2018].

[22] D. Laudisio et al., "Obesity and Breast Cancer in premenopausal women: Current evidence and future perspectives," European Journal of Obstetrics & Gynecology and Reproductive Biology.2018.

[23] J. Gu et al., "Selection of key ambient particulate variables for epidemiological studies - Applying cluster and heat-map analyses as tools for data reduction," Sci. Total Environ., vol. 435–436, pp. 541–550, 2012.

[24] J, Brownlee. (2017, July 26). What is the Difference between a Parameter and a Hyper-parameter?. [Online]. Available: https://machinelearningmastery.com/difference-between-a-parameter-and-a-hyper-parameter. [Accessed: September. 25, 2018].
[25] M, Amin and A. Ali, "Performance Evaluation of Supervised Machine Learning Classifiers for Predicting Healthcare Operational Decisions," Machine Learning for Operational Decision Making, Wavy Artificial Intelligence Research Foundation, 2018
[26] P. Mesejo et al., "Computer-Aided Classification of Gastrointestinal Lesions in Regular Colonoscopy," IEEE Transactions on Medical Imaging, vol. 35, no. 9, pp. 2051-2063, Sept. 2016.
[27] Murphy, Chris et al. "An Approach to Software Testing of Machine Learning Applications." SEKE, 2007.