ISSN 2278-3091

Volume 8, No.6, November – December 2019

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse70862019.pdf

https://doi.org/10.30534/ijatcse/2019/70862019

Breast Cancer Predictive Analytics Using Supervised Machine Learning Techniques

Jide E. T. Akinsola[1], Moruf A. Adeagbo[2], Ayomikun A. Awoseyi[3]

[1]Department of Computational Sciences, First Technical University, Ibadan, Nigeria, akinsolajet@gmail.com
[2]Department of Computational Sciences, First Technical University, Ibadan, Nigeria, adedegy@gmail.com
[3]Department of Computational Sciences, First Technical University, Ibadan, Nigeria, awoseyiayomikun@gmail.com

ABSTRACT

Breast cancer unarguably has been the very prominent disease amongst women as well as the next most dangerous after lung cancer. Early diagnosis and prevention is of paramount importance. Several methods such as micro-array analysis and network analysis have been proffered but they are somewhat expensive and time consuming. There is a need to develop an automated system based on Machine learning techniques to detect breast cancer early. Benign and Malignant tumors were classified using Logistic Regression (LRO), Bayes Network (BNK), Multilayer Perceptron (MLP), Sequential Minimal Optimization (SMO), J48, Naive Bayes (NBS) and Instance Based Learner (IBK) algorithms, which were implemented in Waikato Environment for Knowledge Analysis (WEKA). The breast cancer database for this study was collected from the University of Wisconsin Hospitals, published on California College, Irvive (UCI) website. The five most critical performance metrics when selecting an algorithm in model building in the health related domain are Area Under the ROC Curve (AUC), Receiver Operating Characteristics Curve (ROC), Mean Absolute Error (MAE), Accuracy and Kappa Statistic. In relation to the results of Accuracy, Precision and Kappa Statistic which were evaluated and compared, BNK has best predictive accuracy of 97.14%, followed by SMO with 96.71%, then LOR with 96.57%. On the other hand, LOR has the highest AUC of 99.3%, followed by BNK with 99.2%, then SMO with 96.5%. Beyond accuracy, AUC should be keenly considered in algorithm selection and model building. Therefore, Logistic Regression should be chosen as the best classifier instead of Bayes network for breast cancer optimal prediction.

Key words: Breast Cancer, Classification Algorithm, Machine Learning, Predictive Analytics

1. INTRODUCTION

Cancer can be regarded as a complicated global health-related issues, and sources of high mortality among people[1]. It involves abnormal cell growth which is characterized as a set of connected diseases with the potential to divide continuously and unfold into nearby tissues[2]. Of all the types of cancer, breast cancer which is most prominent among women is the most dangerous cancer aside lung cancer, and therefore, the major reason for recording a high mortality rate among women[3]. Breast cancer is characterized as different kinds of tumor that have varieties of biologically different subtypes in term of behavior, clinicopathological and characteristics of molecules. It must be given serious attention [4].

The reason behind breast cancer is dependent on a number or combination of genetic and environmental factors. These days, many risk factors for breast cancer known can be grouped into modifiable, non-modifiable and environmental risk factors. Examples of modifiable risk factors are menstrual and factors responsible for reproduction, exposure to radiation, medical aid for internal secretion replacement, intoxicating liquor, and a diet with high fat; non-modifiable risk factors are gender, age, and hereditary factors (5-7%) and environmental risk factors are smoking, exposure to organochlorine, and magnetic field attraction[5], [6]. In the years 2018, the report has it that over two million of fresh incidence of cancer were recorded out of which death was approximated to be 626,679[7]. In this report, fresh incidence of breast cancer was estimated to be 11.6% which was about 24.2% among women. This menace is associated with a lack of awareness and inadequate health services [8].

The frequent rise in the demand globally for the untimely identifying the presence of breast cancer at many screening centers and clinics in last few years has created a new avenue for the need to conduct research. Breast cancer is often diagnosed by taking patients through: undertaking thorough medical history, Physical assessment of both the breasts and conjointly check for swelling or hardening of any bodily fluid nodes in the armpit. As indicated by the World Health Organization (WHO), the discovery of cancer early enormously expands the odds of choosing the correct choice on a fruitful treatment method [9]. This will along these lines encourage the early finding and categorization of a cancer with a view to manage patients appropriately[10]. In such manner, a significant level of breast cancer could be completely healed provided it is early identified[11]. This will in the long run prompts an expansion in counteractive action, identification and treatment methodologies in breast cancer patients[6], [12], [13].

Accordingly, with the quick improvement on the sequence of innovation with high throughput, and the utilization of different techniques in machine learning that had unfolded as of late, advancement in disease forecast has been steadily made based on gene expression, giving knowledge into powerful and precise treatment decision-making. Thus, creating machine learning techniques, which can effectively identify health and malignant growth patients is of big concern. However, based on the application of a different characterization techniques for cancer prediction up to this moment, no one approach outperforms all the others. With the quick improvement of computer-aided techniques as of late, utilization of machine learning techniques is assuming an inexorably significant role in the cancer detection, and different forecasting algorithms are being considered continuously by the researchers[1].

Consequently, many researchers have focused majorly on using Accuracy as the main performance metric in the prediction of breast cancer without consideration to other metrics and this may not give true reflection of the prediction model. Therefore, this study examines the significance of Area Under the Curve (AUC), Kappa Statistics (KPS) and Mean Absolute Error (MAE) in relation to Accuracy and Precision.

1.1 Types of Breast Cancer

Breast cancers are categorized based on the area it starts, that is, the ducts, the lobules, and the tissue in the middle[14]. However, they can be generally classified as Benign and Malignant [15]–[17]. These categories are:

A. Benign

Benign Cancer which is also referred to as Non-invasive. This is a kind of breast cancer that do not transform or attack normal tissues inside or beyond the breast

B. Malignant

Malignant which is sub-divided into*:*

1. Invasive

It a type of breast cancer that occurs when malignancy cell emanate from the milk ducts or lobules released into the nearby breast tissues. This malignant growth cell can move via bloodstream or the lymphatic system from the breast to different parts of the body.

2. Metastatic Breast Cancer

It is breast malignancy that has spread to another part of the body, mostly the liver, mind, bones, or lungs.

3. Intrinsic or Sub-Atomic Subtypes of Bosom Malignant Growth

It describes the smaller groups that a type of cancer can be divided into, in light of specific characteristics of the malignant growth cells.

1.2 Techniques for Breast Cancer Prediction

Various cancer prediction techniques have been used in the past by different researchers. Some of the techniques are:

1. Deep Learning-Based Multi-model Ensemble Strategy

[1] proposed a technique that utilization deep learning-based multi-model ensemble strategy which was tried on three public RNA-seq data collections of three sorts of cancers with the outcome demonstrated to be precise and viable for cancer prediction.

2. Network Learning

Kim et al. [13], proposed an improved strategy for predicting cancer prognosis by using network learning. This technique was carried out by indicating the candidate prognostic gene module by graph learning to use the Generative Adversarial Networks (GANs) model, and scores genes using a PageRank algorithm. This technique was applied to multiple-omics data for five cancer types with the outcome demonstrating preferable prediction accuracy over the current strategies.

3. Microarray Analysis

[18] utilized microarray analysis to assess gene prognosis profile that has been previously established. The power of prediction of the prognosis profile was assessed utilizing uni-variable and multi-variable statistical analyses. Utilizing this technique, the outcome demonstrated that the profile of gene-expression is far powerful in predicting disease in the youthful patient that has breast cancer than formalized frameworks based on clinical and histology criteria.

4. Statistical Methods

[19] utilized statistical methods in medical research with a copula-based system to examine the inclination brought about by dependent censoring on gene selection, and subsequently utilized copula-based reliance model to build up an alternative procedure for gene selection. This was then used to investigate non-small cell lung cancer data to exhibit its uses.

2. LITERATURE REVIEW

It is a known fact that early detection of breast cancer increases the survivability of patients. This informs the interest of many researchers in the area of detecting as well as predicting breast cancer. Various Machine Learning (ML) paradigms has been adopted to predict breast cancer, like supervised and unsupervised algorithms [20]–[31].

[20], explored personalized breast cancer prediction using a set of machine learning algorithms comparing them with two statistical models currently in use, namely Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm BODICEA) and Breast Cancer Risk Assessment Tool (BCRAT). The results showed that machine learning techniques improves classification accuracy for women with breast cancer as well as those without breast cancer using the same independent variables as the statistical models. This lends credence to the reliability of Machine

Learning Algorithms (MLAs). Breast Cancer Diagnosis with Neuro Fuzzy Inference Systems (BCD-NFIS) contributes to decreased use of data-sets features by using the Fuzzy networks. This ends, in effect, with an improved accuracy of 98.24% compared to previous methodologies [43].

Tobacco smoking, alcohol consumption, among others cause cancer, especially mouth cancer. Nevertheless, the physicians remain confused and unsure about the origins of these diseases, and their effects are ambiguous. Therefore, the Max-Min composition approach is remarkable to handle this issue [44]

[3], compared three MLAs, these are Support Vector Machine (SVM), Logistic Regression (LOR) and K-Nearest Neighbor (KNN), with Dimensionality Reduction Technique, with Dimensionality Reduction Technique. The results showed SVM as the best classifier based on the accuracy metric with a value of 92.7%, KNN was second with an accuracy of 92.23% and Logistic Regression, an accuracy of 92.10%.

[21] noted the fuzzy nature of features in breast cancer datasets, this informed the use of a fuzzy inference system for Breast cancer prediction. This system was compared with other machine learning techniques such as Decision Table, RBF-Network, Naive Bayes, Random Tree etc. The fuzzy system performed best with an accuracy of 84.64%, Decision Table was second best with accuracy of 79.02%, while the worst performance was from Random Forest with an accuracy of 72.38. According to literature, some popular ML algorithms were not considered in the study for example SVM, KNN, etc. There is no singular algorithm that can out-perform other algorithms considering the varieties of metrics to be used in decision making in model building. [22] explored the possibility of decision trees in breast cancer prediction. It was compared with other supervised learning algorithms such as, ID3, Naïve Bayes, Random Forest, C4.5 and CART using Wisconsin data set. The comparison showed that the Random Forest outperformed other supervised learning algorithms. Precision and Recall were used as their evaluation metrics. Consideration for more ML metrics is essential considering the nature of the problem domain.

[23] compared the performance of the following data mining algorithms namely, C4.5, RIPPER, and PART algorithms on two data set, breast cancer and heart disease. These datasets were analyzed using number of rules generated. [24] utilized over 7,000 breast cancer datasets of histopathology images acquired from 82 patients. Using different classification algorithm, the study recorded between 80% to 85% accuracy. [25], the author compared six machine learning algorithms using the Wisconsin's diagnostic breast cancer dataset. The algorithms are, Softmax Regression, Multilayer Perceptron (MLP), Linear Regression, Nearest Neighbor (NN) search, and Support Vector Machine (SVM), Gated Recurrent Unit SVM (GRU-SVM). MLP gave the highest accuracy at 99.04%.

[26], analysis of histopathological images of breast cancer tissues is another approach to breast cancer diagnosis. They automated the classification of benign and malignant tissue images using machine learning models. [27], [28], many papers have done some reviews on MLAs, comparing different machine learning algorithms to determine the best algorithm for breast cancer prediction and classification. [29] the authors discussed and compared the performances Bayes classifiers. Boosted Augmented Naive (BAN) Bayes, Tree Augmented Naive (TAN) Bayes and Bayes Belief Network (BBN) were considered for the study. TAN with gradient boosting gave the best performance in terms of accuracy, sensitivity and specificity. Beyond these metrics, MAE, TTB are also very essential in the performance evaluation of the algorithms.

[30] conducted a study where different data mining techniques for breast cancer prediction were explored. The Wisconsin dataset from UCL containing ten attributes and 699 instances. Sixteen instances with missing values were removed leaving 683 instances left. The authors used three supervised learning algorithms IBK, BF Tree and Sequential Minimal Optimization (SMO). Comparison showed Sequential Minimal Optimization (SMO) gave the highest prediction accuracy (96.2%). SMO had 0.92 of Kappa statistic (KS) and less of mean absolute error (MAE). [31] conducted a research which objective was to find the performance of different classification algorithms by analyzing the mammogram images. Three algorithms were used, J48, CART and ADTree, measured based on a couple of metrics which includes specificity, kappa statistics and MAE. [42] opined that Multi-Criteria Decision Method (MCDM) methods can be used to find the optimal classification and regression models in relation to supervised machine learning algorithms.

Table 1 shows the comparative analysis of various machine learning algorithms with the performance metrics to ascertain consideration being given to each performance metric in predictive analytics. Eleven performance metrics were considered. The comparative analysis revealed that most studies do not focus on Area Under the ROC Curve, Receiver Operating Characteristics Curve (ROC), Kappa Statistic and Mean Absolute Error. According to [41] to have supervised predictive machine learning, ML algorithms require precise accuracy and minimum errors in addition to putting several factors into consideration. Also, it may be difficult or impossible to find a single classifier doing as well as a good group of classifiers if the only performance metric being utilized is best possible classification accuracy.

Table 1: Comparative Analysis of Various Machine Learning Algorithms and Performance Metrics

Author (s)	ML Algorithm	ACR	PRC	ER	TTB	Recall	TNR	*F*-M	AUC	ROC	KPS	MAE
[20]	BOADICAE	✕	✕	✕	✕	✕	✕	✕	✓	✕	✕	✕
	BCRAT	✕	✕	✕	✕	✕	✕	✕	✓	✕	✕	✕
	AdaBoost	✕	✕	✕	✕	✕	✕	✕	✓	✕	✕	✕
	Random Forest	✕	✕	✕	✕	✕	✕	✕	✓	✕	✕	✕
[21]	Naïve Bayes	✓	✓	✕	✕	✓	✕	✓	✕	✕	✕	✕
	RBF Network	✓	✓	✕	✕	✓	✕	✓	✕	✕	✕	✕
	Logistic	✓	✓	✕	✕	✓	✕	✓	✕	✕	✕	✕
	LWL	✓	✓	✕	✕	✓	✕	✓	✕	✕	✕	✕
	Logit Boost	✓	✓	✕	✕	✓	✕	✓	✕	✕	✕	✕
	Decision Table	✓	✓	✕	✕	✓	✕	✓	✕	✕	✕	✕
	OneR	✓	✓	✕	✕	✓	✕	✓	✕	✕	✕	✕
	RandomTree	✓	✓	✕	✕	✓	✕	✓	✕	✕	✕	✕
	FIS	✓	✓	✕	✕	✓	✕	✓	✕	✕	✕	✕
[3]	SVM	✓	✕	✕	✕	✕	✕	✕	✕	✕	✕	✕
	KNN	✓	✕	✕	✕	✕	✕	✕	✕	✕	✕	✕
	Log R	✓	✕	✕	✕	✕	✕	✕	✕	✕	✕	✕
[22]	ID3	✓	✓	✓	✕	✓	✕	✕	✕	✕	✕	✕
	C4.5	✓	✓	✓	✕	✓	✕	✕	✕	✕	✕	✕
	CART	✓	✓	✓	✕	✓	✕	✕	✕	✕	✕	✕
	Random tree	✓	✓	✓	✕	✓	✕	✕	✕	✕	✕	✕
	Naive Bayes	✓	✓	✓	✕	✓	✕	✕	✕	✕	✕	✕
[25]	L2-NN	✓	✕	✕	✕	✕	✕	✕	✕	✕	✕	✕
	Linear	✓	✕	✕	✕	✕	✕	✕	✕	✕	✕	✕
	L1-NN	✓	✕	✕	✕	✕	✕	✕	✕	✕	✕	✕
	GRU-SVM	✓	✕	✕	✕	✕	✕	✕	✕	✕	✕	✕
	Soft Regression	✓	✕	✕	✕	✕	✕	✕	✕	✕	✕	✕
	SVM	✓	✕	✕	✕	✕	✕	✕	✕	✕	✕	✕
	MLP	✓	✕	✕	✕	✕	✕	✕	✕	✕	✕	✕
	Regression	✓	✕	✕	✕	✕	✕	✕	✕	✕	✕	✕
[26]	Linear	✓	✓	✕	✕	✓	✕	✓	✕	✕	✕	✕
	Regression	✓	✓	✕	✕	✓	✕	✓	✕	✕	✕	✕
	SVM	✓	✓	✕	✕	✓	✕	✓	✕	✕	✕	✕
	K-NN	✓	✓	✕	✕	✓	✕	✓	✕	✕	✕	✕
[29]	Bayes Network	✓	✕	✕	✕	✓	✓	✕	✕	✕	✕	✕
	BAN	✓	✕	✕	✕	✓	✓	✕	✕	✕	✕	✕
	TAN	✓	✕	✕	✕	✓	✓	✕	✕	✕	✕	✕
[30]	BFTree	✓	✓	✕	✕	✓	✕	✕	✕	✕	✓	✓
	IBK	✓	✓	✕	✕	✓	✕	✕	✕	✕	✓	✓
	SMO	✓	✓	✕	✕	✓	✕	✕	✕	✕	✓	✓
[31]	CART	✓	✓	✕	✓	✓	✕	✓	✕	✓	✓	✓
	AD Tree	✓	✓	✕	✓	✓	✕	✓	✕	✓	✓	✓
	J48	✓	✓	✕	✓	✓	✕	✓	✕	✓	✓	✓

ML =Machine Learning, ACR = Accuracy, PRC = Precision, ER = Error Rate, TTB = Time To Build, KPS = Kappa Statistic, TPT Rate = True Positive Rate, TNR = True Negative Rate, FPT Rate = False Positive Rate, *F*-M = F-Measure, AUC = Area Under Receiver Operating Characteristic Curve, Receiver Operating Characteristic Curve = ROC, MAE= Mean Absolute Error

✕ = Metrics Not Implemented ✓= Metric Implemented

3. MACHINE LEARNING ALGORITHMS

Machine Learning (ML) field focuses on helping computing systems to know from the data how to carry out the required task automatically. ML has been largely utilized in the fields of medicine, research, innovation, business and finance. It is a crucial tool throughout information management and large data mining technologies, including decision making, modeling and/or forecasts[32]. ML is designed to allow a machine to learn from the past or the present and use that information to forecast or predict for the future uncertain occurrences.

The study examined seven classification algorithms for analysis of performance metrics, which belong to the following four classes: Function (Sequential Minimal Optimization (SMO), Logistic Regression and Multilayer Perceptron (MLP)), Bayes (Bayes Network (BNK) and Naive Bayes (NBS)), Tree (J48) and Lazy (IBK). It should be noted that SMO is a variant of the SVM. MLP is also an Artificial Neural Network (ANN) variant with the utilization of WEKA data mining tool on Breast Cancer database on all the algorithms.

3.1 Bayes Classifiers

The classes BNK and NBS in the Bayes classifications are the Bayes Classifiers (BC). BC are probabilistic classifiers focused on Thomas Bayes ' basic probability principle alluded to in (1) that is regarded as the Bayes Theorem.

$$P\ (B/A) = \frac{P(B/A)\ \ x\,P(A)}{P(B)} \qquad (1)$$

Where *A* and *B* are events and *P(B) ≠ 0. P(A/B)* is a conditional probability, that is, the likelihood of event *A* occurring given that *B* is true. *P(B/A)* is a conditional probability, that is, the likelihood of event *B* occurring given that *A* is true The relationship between *A* and *B*, which is contingent likelihood and chance, is shown in (1). As an uncomplicated algorithm, a classifier named Naïve Bayes, means that an algorithm does not consider the features to be equally probable. Comparatively advanced algorithms such as the Bayesian networks that evaluate the likelihood of uncertainty require more complex information from the analyzed data.

3.2 Function Classifiers

The function classifiers include Sequential Minimum Optimization (SMO), Multi-Layer Perceptron (MLP) and Logistic Regression (LOR) classifiers.

1. LOR

LOR is a classification function with a single multinomial model of a building class and a single logistic regression. In fact, the logistics show where the category cap resides. Based on how far from the limit, the group chances are also calculated in a particular approach [33]. If the dataset is larger, it will pass to (0 and 1) ends. Such likelihood assumptions not only describe logistic regression but also accurately define it. This makes better, more accurate projections and can match differently; but these good forecasts sometimes go wrong.

2. SMO

SMO is a version of SVM. For classical Multilayer Perceptron networks, neural networks are strongly connected with SVM algorithms. SVMs rely on the idea of a breach between the two data categories on both sides of the hyperplanes[34]. It has been shown to reduce the maximum limit of the expected generalization error[35], by maximizing the margin consequently, therefore the maximum likely distance among the divisive hyper-planes as well as installations on both sides.

3. MLP

MLP is the ANN variation. MLP is an element that categorizes the weights of the network, not by creating a non-convex, uncompromising minimization problem, such as for conventional Neural Network training, but by tackling a linear limitation quadratic programme. ANN is a category problem solving learning algorithm. An ANN model includes several collimate dynamic in addition to interlinked neuron network systems. Neuron is used for the production of results using inputs by a given computational processor[36]. The existence of local solutions is one of the problems in addressing optimization in ANN. In a single-objective search area there is only one best solution, often known as the Global Optimum [45].

3.3 Tree Classifiers

The J48 is the 3 (ID3) extension of the Iterative Dichotomiser. J48 also includes features for the correction of missing values, decision-taking trees, the collection of continuous values, derivation of rules, etc. This is a decision tree algorithm that is utilized to evaluate in multiple instances the demeanour of the attributes / vectors. The programs for the recent instances were identified also on the basis of the instances of teaching [37]. This algorithm produces the projection essential for the goal factor prediction. Using the algorithm for tree classification the important data distribution can be readily understood [38].

3.4 Lazy Classifiers

IBK is categorized as a Lazy Classifier for learners based on instances. It's a k-Nearest Neighbor (k-NN) algorithm. This approach is a fast and simple way to define a certain dataset with defined apriori K-means (suppose k-clusters). K-means algorithms are used[39]. If information on the labels are not available. This utilizes a certain way of converting rough thumb laws into a very specific prediction law. For weak training algorithms classifications may be at least slightly better than arbitrarily (thumb rules) continuously and with an accuracy of approximately 55%.

4. METHODOLOGY

There are several free open source software that can be used in data mining as well as machine learning problems of which WEKA is readily available with users' friendly capabilities. Nine performance metrics which comprise of seven benefit criteria such as Accuracy, Kappa Statistic, Precision, True Positive (TPT) Rate, *F*-measure, False Positive (FPT) Rate and Area Under the Curve (AUC) and two cost criteria performance measures such as Mean Absolute Error (MAE), and Time To Build Model (TTB) were evaluated using the following seven machine learning algorithms such as LOR, SMO, BNK, IBK, MLP, NBS and J48 implemented in WEKA in this study.

The breast cancer databases for this research was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg which was published at California College, Irvive (UCI) website and made available online at https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)[40]. The datasets are selected for its reliability which has been also anonymised (de-identified), meaning anonymity is guaranteed. The number of attributes is 10, with a single class representing the dependent variable making it 11 fields, which is the estimation outcome of the machine-learning algorithm. All values of the attributes in the domain are numerical as presented in Table 2.

Class value applicable in the distribution of Benign and Malignant used for the breast cancer prediction is shown in Table 2 where class value 2 is interpreted as "Benign" that signifies Non-invasive. The type of breast cancer that don't develop into or attack normal tissues inside or beyond the breast while class value 4 is interpreted as "Malignant" that signifies any of Invasive, Metastatic or Intrinsic type of breast cancer. 10-fold cross validation was applied for data analysis as testing option for both test data and train data with 699 instances.

Table 2: Attributes of Breast Cancer Dataset

Number	Attribute	Domain
1	Sample code number	id number
2	Clump Thickness	1 – 10
3	Uniformity of Cell Size	1 – 10
4	Uniformity of Cell Shape	1 – 10
5	Marginal Adhesion	1 – 10
6	Single Epithelial Cell Size	1 – 10
7	Bare Nuclei	1 – 10
8	Bland Chromatin	1 – 10
9	Normal Nucleoli	1 – 10
10	Mitoses	1 – 10
11	Class	(2 for benign, 4 for malignant)

Table 3: Breast Cancer Class Distribution

Class Number	Breast Cancer Type	Number of Instances	Converted Class
2	Benign	458	B
4	Malignant	241	M

4.0 RESULTS AND DISCUSSION

The results of the seven supervised machine learning algorithms performance' evaluation using WEKA as the Machine Learning tool are depicted in Table 4.

Table 4: Supervised Machine Learning Algorithms Performance Evaluation

MLA	Classifier Category	ACR %	PRC %	KPS %	TPT Rate %	FPT Rate %	*F*-M %	AUC %	ROC %	MAE %	TTB sec
LOR	Function	96.5665	96.6	92.4	96.6	4.2	96.6	99.3	0.2300	4.88	0.27
SMO	Function	96.7096	96.7	92.74	96.7	3.7	96.7	96.5	0.2614	3.29	0.21
BNK	Bayes	97.1388	97.2	93.74	97.1	2.3	97.2	99.2	0.4222	2.89	0.22
IBK	Lazy	95.1359	95.1	89.19	95.1	6.3	95.1	94.5	0.1520	5.01	0.00
MLP	Function	95.8512	95.9	90.86	95.9	4.5	95.9	98.9	0.2131	4.72	3.81
NBS	Tree	95.9943	96.2	91.27	96.0	3.3	96.0	98.6	0.2909	4.03	0.15
J48	Tree	94.5637	94.6	87.99	94.6	6.4	94.6	95.5	0.1478	6.91	0.27

MLA = Machine Learning Algorithm, ACR = Accuracy, PRC = Precision, KPS = Kappa Statistic, TPT Rate = True Positive Rate, FPT Rate = False Positive Rate, *F*-M = F-Measure, AUC = Area Under Receiver Operating Characteristic Curve, Roc = Receiver Operating Characteristic Curve, MAE= Mean Absolute Error, TTB = Time To Build.

LOR = Logistic Regression, BNK = Bayes Network, NBS = Naive Bayes, SMO = Sequential Minimal Optimization and MLP = Multilayer Perceptron and IBK = Instance Based Learner

The study reveals that out of the seven supervised machine learning algorithms namely LOR, SMO, BNK, IBK, MLP, NBS and J48 considered, Bayes Network (BNK) has the highest accuracy of 97.14%, followed by Sequential Minimal Optimization (SMO) with 96.71% and then Logistic Regression (LRO) with 96.67% as shown in Figure 1. There is a close relationship between Accuracy and Precision.

Beyond Accuracy (ACR) there are other metrics such as AUC, ROC, MAE and Kappa Statistic which play significance role in the choice of algorithm for model building. It has been deduced that ACR may not give true reflection of the prediction model as shown in Table 3. This is because the algorithm with the highest ACR such as BNK does not have corresponding highest AUC. This study emphasis the importance of other measures for MLAs performance evaluation in relation to several Machine Learning techniques beyond Accuracy to arrive at reliable predictive analytics. This is highly essential due to the fact that over-fitting can be misleading. Apart from this, the heterogeneity of dataset attributes is another factor for considering other metrics.

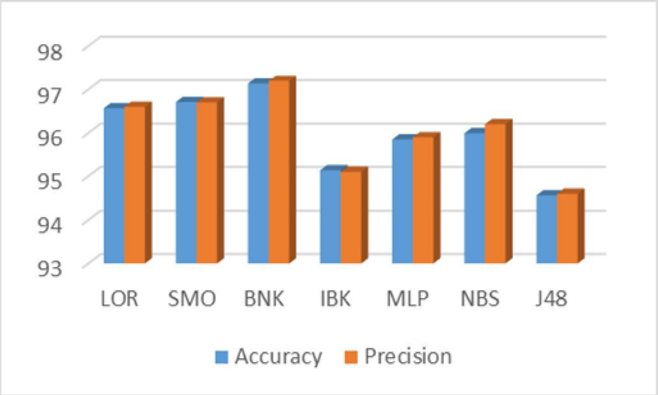

Figure 1: Comparison of Performance Metrics Based on Accuracy and Precision

The result of relationship between Accuracy and Kappa Statistic is not clearly evident as the BNK algorithm with the highest Accuracy also has the highest Kappa Statistic as shown in Figure 2.

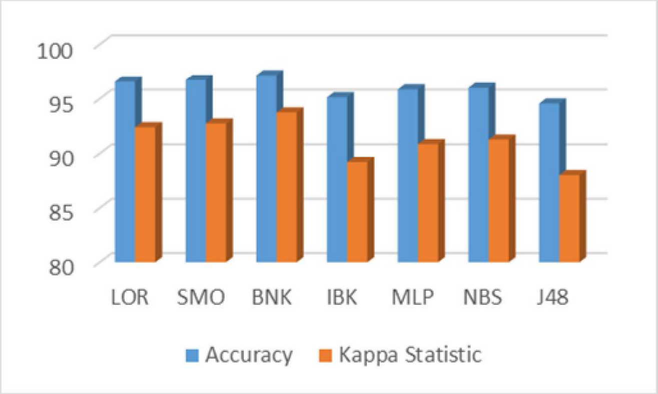

Figure 2: Performance Comparison Metrics Based on Accuracy and Kappa Statistic

The degree or measure of separability is represented by AUC. This shows how many models will discriminate between categories. When the AUC is larger, the best 0s and 1s are expected as 1s. The higher the AUC, the more the model differentiates people with a breast cancer from those without a breast cancer.

Figure 3 shows the connection between Accuracy and AUC. The highest Accuracy which is BNK does not correspond to highest AUC instead, LOR has the highest AUC.

Also, choosing Accuracy (ACR) as the benchmark for prediction could be erroneous as it has inability to find relevant cases within a dataset. The fact that an algorithm has the highest precision does not make it an optimal classifier especially when dealing with very sensitive classification problems.

Such is the case in this study where BNK has the highest precision of 97.2%, followed by SMO with 96.7% as well as LOR with 96.6%.

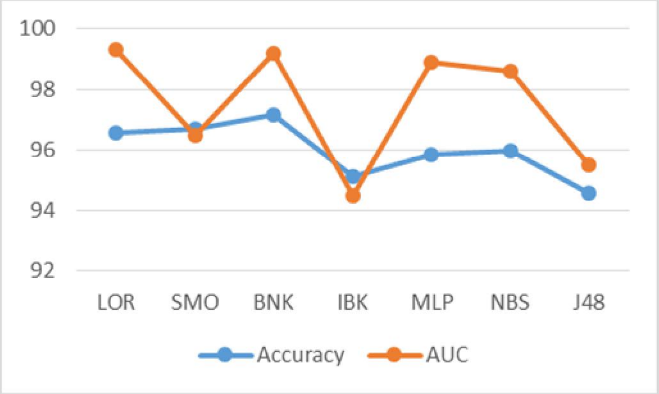

Figure 3: Comparison of Performance Metrics Based on Accuracy and AUC

Figure 4 shows that algorithm with lowest Time To build which is IBK does not correspond to the one with highest Accuracy. Likewise, SMO with the lowest MAE does not guarantee highest performance in terms of Accuracy. That is why, other metric such as AUC and Kappa Statistic must be utilized in the comparative analysis for optimal selection of the best algorithm to build breast cancer predictive analytics model. Consideration in relation to Time to build the model (TTB) is of no significance. IBK has the least TTB value of 0.00 second and MAE of 5.01% which is one of the highest MAE, and higher MAE value signifies unreliability of the model, yet with lowest ACR of 95.14%. The selection of machine learning algorithm for predictive analytics must ensure thorough consideration of the metrics beyond accuracy.

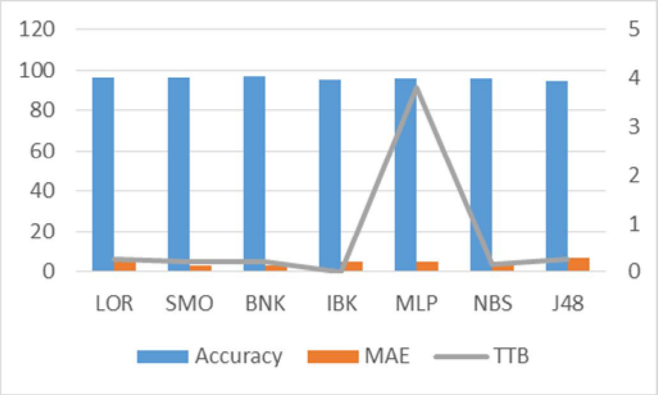

Figure 4: Comparison of Performance Metrics Based on Accuracy MAE and TTB

Hence, TTB does not connote an efficient algorithm. Likewise, higher TTB does not indicate inefficient algorithm. LOR has the highest AUC of 99.3%, followed by BNK with 99.2%, then SMO with 96.5%. Therefore, Logistic Regression with the highest AUC of 99.3% should be consider as the best classifier instead of BNK.

5. CONCLUSION

Every performance metric must be considered holistically before choosing an optimal algorithm for predictive analytics. When dealing with classification problems, consideration should be beyond accuracy only, and special attention must also be paid to Area Under the Curve, Mean Absolute Error, as well Kappa Statistics especially when addressing a multi-classifier system. The higher the value of AUC, the more reliable the model. The lower the MAE, the more dependable the model. Kappa Statistic as a performance measure is utilized for comparing an observed accuracy with an expected accuracy (Random Chance) takes into cognizance comparison and similarity. The No Free Lunch theorem is highly essential because good number of correctly classified instances in predicting valid disease outcomes using supervised machine learning techniques is not just a function of accuracy. Further research should be carried out using different cross validation hold-out ratios on different datasets with higher number of instances. Also, Multi-Criteria Decision Making techniques could be implemented with machine learning performance evaluation procedures in selecting an optimal unbiased model for predicative analytics decision making.

REFERENCES

[1] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, “A deep learning-based multi-model ensemble method for cancer prediction,” *Comput. Methods Programs Biomed.*, vol. 153, pp. 1–9, 2018. https://doi.org/10.1016/j.cmpb.2017.09.005

[2] National Cancer Institute, “Understanding Cancer.” [Online]. Available: https://www.cancer.gov/about-cancer/understanding/what-is-cancer. [Accessed: 07-Dec-2019].

[3] C. H. Shravya, K. Pravalika, and S. Subhani, “Prediction of breast cancer using supervised machine learning techniques,” *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 6, pp. 1106–1110, 2019.

[4] L. A. Carey, C. M. Perou, and C. A. Livasy, “Race, breast cancer subtypes, and survival in the carolina breast cancer study,” *JAMA*, vol. 295, no. 2, pp. 492–502, 2006. https://doi.org/10.1001/jama.295.21.2492

[5] M. Clemons and P. Goss, “Estrogen and the risk of breast cancer,” *N Engl J. Med.*, vol. 344, pp. 276–285, 2001. https://doi.org/10.1056/NEJM200101253440407

[6] M. Heidari, A. Z. Khuzani, and A. B. Hollingsworth, “Prediction of breast cancer risk using a machine learning approach embedded with a locality preserving projection algorithm,” *Phys Med Biol*, vol. 63, 2018. https://doi.org/10.1088/1361-6560/aaa1ca

[7] World Cancer Research Fund, “Breast cancer statistics,” *World Cancer Research Fund and American Institute for Cancer Research*. [Online]. Available: https://www.wcrf.org/int/cancer-facts-figures/data-specific-cancers/breast-cancer-statistics. [Accessed: 29-Nov-2019].

[8] A. Shariff, J. Kangas, L. P. Coelho, S. Quinn, and R. F. Murphy, “Automated image analysis for highcontent screening and analysis,” *J. Biomol. Screen.*, vol. 15, no. 7, pp. 726–734, 2010. https://doi.org/10.1177/1087057110370894

[9] A. M. Abdel-Zaher and A. M. Eldeib, “Breast cancer classification using deep belief networks,” *Expert Syst. Appl.*, vol. 46, pp. 139–144, 2016. https://doi.org/10.1016/j.eswa.2015.10.015

[10] W. William, A. Ware, A. H. Basaza-Ejiri, and J. Obungoloch, “A review of image analysis and machine learning techniques for automated cervical cancer screening from pap-smear images,” *Comput. Methods Programs Biomed.*, vol. 164, pp. 15–22, 2018.

[11] G. I. Salama, M. B. Abdelhalim, and M. A. Zeid, “Breast Cancer diagnosis on three different data sets using multi-classifiers,” *Int. J. Comput. In-formation Technol.*, vol. 1, pp. 36–43, 2012.

[12] A. Jemal, F. Bray, and J. Ferlay, “Global cancer statistics,” *CA Cancer J Clin*, vol. 61, pp. 69–90, 2011. https://doi.org/10.3322/caac.20107

[13] M. Kim, I. Oh, and J. Ahn, “An Improved Method for Prediction of Cancer Prognosis by Network Learning,” *Genes (Basel).*, vol. 9, no. 10, p. 478, 2018. https://doi.org/10.3390/genes9100478

[14] Breast Cancer Org, “BreastCancer.org.” [Online]. Available: https://www.breastcancer.org. [Accessed:

09-Dec-2019].

[15] Mayo Clinic, “Breast Cancer: Symptoms and causes.” [Online]. Available: https://www.mayoclinic.org/diseases.../ breast-cancer/symptoms-causes/syc-20352470. [Accessed: 09-Jul-2019].

[16] Roche, “Breast cancer a guide for journalists on breast cancer and its treatment.” [Online]. Available: https://www.roche.com/dam/jcr:5260dc48-ffc1-4991-9f3f-0bdae3e42128/en/med-breast-cancer.pdf. [Accessed: 07-Nov-2019].

[17] Genentech, “Types and Features of Breast Cancer,” 2015. [Online]. Available: https://www.gene.com/media/press-releases/14594/2015-05-31. [Accessed: 09-Dec-2019].

[18] M. J. Van de Vijver *et al.*, “A Gene-Expression Signature as a Predictor of Survival in Breast Cancer,” *N. Engl. J. Med.*, vol. 347, no. 25, pp. 1999–2009, 2002. https://doi.org/10.1056/NEJMoa021967

[19] T. Emura and Y.-H. Chen, “Gene selection for survival data under dependent censoring: A copula-based approach,” *Stat. Methods Med. Res.*, vol. 25, no. 6, pp. 2840–2857, 2016. https://doi.org/10.1177/0962280214533378

[20] C. Ming, V. Viassolo, N. Probst-Hensch, P. O. Chappuis, I. D. Dinov, and M. C. Katapodi, “Machine learning techniques for personalized breast cancer risk prediction: Comparison with the BCRAT and BOADICEA models,” *Breast Cancer Res.*, vol. 21, no. 1, pp. 1–11, 2019.

[21] S. Dutta, S. Ghatak, A. Sarkar, R. Pal, R. Pal, and R. Roy, “Cancer Prediction Based on Fuzzy Inference System,” *Adv. Intell. Syst. Comput.*, vol. 851, pp. 127–136, 2019. https://doi.org/10.1007/978-981-13-2414-7_13

[22] S. S. Shajahaan, S. Shanthi, and V. Manochitra, “Application of Data Mining Techniques to Model Breast Cancer Data,” *Int. J. Emerg. Technol. Adv. Eng.*, vol. 3, no. 11, pp. 1–8, 2013.

[23] Vijayarani Divya, “An Efficient Algorithm for Generating Classification Rules,” *Int. J. Comput. Sci ence Technol.*, vol. 2, no. 4, pp. 512–515, 2011.

[24] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, “A Dataset for Breast Cancer Histopathological Image Classification,” *IEEE Trans. Biomed. Eng.*, vol. 63, no. 7, pp. 1455–1462, 2016. https://doi.org/10.1109/TBME.2015.2496264

[25] A. F. M. Agarap, “An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset,” in *International Conference on Machine Learning and Soft Computing*, 2018.

[26] M. Nuruddin Qaisar Bhuiyan, M. Shamsujjoha, S. H. Ripon, F. H. Proma, and F. Khan, *Transfer Learning and Supervised Classifier Based Prediction Model for Breast Cancer*. Elsevier Inc., 2019.

[27] 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. https://doi.org/10.3390/designs2020013

[28] S. Eltalhi and H. Kutrani, “Breast Cancer Diagnosis and Prediction Using Machine Learning and Data Mining Techniques : A Review Breast Cancer Diagnosis and Prediction Using Machine Learning and Data Mining Techniques : A Review,” *J. Dent. Med. Sci.*, vol. 18, no. April, pp. 85–94, 2019.

[29] A. Bazila Banu and T. Ponniah, “Comparison of bayes classifiers for breast cancer classification,” *Asian Pacific J. Cancer Prev.*, vol. 19, no. 10, pp. 2917–2920, 2018.

[30] V. Chaurasia and S. Pal, “A Novel Approach for Breast Cancer Detection using Data Mining Techniques,” *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 2, no. 1, pp. 2456–2465, 2014.

[31] B. Padmapriya and T. Velmurugan, “Classification Algorithm Based Analysis of Breast Cancer Data,” vol. 5, no. 1, 2016.

[32] A. Ali, J. Qadir, and R. . Rasool, “Big data for development: applications and techniques. Big data for development: applications and techniques,” *Big Data Anal.*, vol. 1, no. 2, 2016. https://doi.org/10.1186/s41044-016-0002-4

[33] Statistics Department CMU, “Logistic Regression,” *Carnegie Mellon University*. [Online]. Available: https://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch12.pdf. [Accessed: 10-Nov-2019].

[34] S. Ali and K. A. Smith-Miles, “A meta-learning approach to automatic kernel selection for support vector machines,” *Neurocomputing*, vol. 70, pp. 173–186, 2006.

[35] A. Unagar, “Support Vector Machines. Unwinded.” [Online]. Available: https://medium.com/data-science-group-iitr/support-vector-machines-svm-unraveled-e0e7e3ccd49b. [Accessed: 15-Nov-2019].

[36] M. Ko, A. Twari, and J. Mehmen, “A review of soft computing applications in supply chain management,” *Appl. Soft Comput.*, vol. 10, pp. 661–664, 2010. https://doi.org/10.1016/j.asoc.2009.09.004

[37] T. S. Korting, “C4. 5 algorithm and Multivariate Decision Trees. Image Processing Division,” *National Institute for Space Research*.

[38] A. Nadali, E. N. Kakhky, and H. E. Nosratabadi, “Evaluating the success level of data mining projects based on CRISP-DM methodology by a Fuzzy expert system,” in *3rd International Conference on Electronics Computer Technology*, 2011, pp. 161–165.

[39] S. Alex and S. V. N. Vishwanathan, *Introduction to Machine Learning*. Cambridge University Press , Cambridge, United Kingdom, 2008.

[40] O. L. Mangasarian and W. H. Wolberg, “Cancer diagnosis via linear programming,” *SIAM News*, vol. 23, no. 5, pp. 1–18, Sep-1990.

[41] F. Y. Osisanwo, J. E. T. Akinsola, O. Awodele, J. O. Hinmikaiye, O. Olakanmi and J. Akinjobi. Supervised Machine Learning Algorithms: Classification and Comparison. *International Journal of Computer Trends and Technology (IJCTT) – Volume 48 Number 3*, 2017, https://doi: 10.14445/22312803/IJCTT-V48P126

[42] J. E. T. Akinsola, S. O. Kuyoro, O. Awodele & F. A. Kasali, Performance Evaluation of Supervised Machine Learning Algorithms Using Multi-Criteria Decision Making Techniques. *International Conference on Information Technology in Education and Development (ITED) Proceedings,* 17 – 34, 2019.

[43] S. H. Nallamala, P. Mishra and S. V, Koneru. Qualitative Metrics on Breast Cancer Diagnosis with Neuro Fuzzy Inference Systems. *International Journal of Advanced Trends in Computer Science and Engineering*. Volume 8, No.2, pp. 259 – 264, March – April 2019. Available at: http://www.warse.org/IJATCSE/static/pdf/file/ijatcse26822019.pdf, https://doi.org/10.30534/ijatcse/2019/26822019

[44] R. Jothikumar, S. G. Shanmugam, M. Nagarajan, S. Premkumar and A. Asokan. Analyzes of Mouth Cancer Using Max-Min Composition in Soft Computing. *International Journal of Advanced Trends in Computer Science and Engineering*. Volume 8, No.3, pp. 825 – 830, May - June 2019. Available at: http://www.warse.org/IJATCSE/static/pdf/file/ijatcse76832019.pdf. https://doi.org/10.30534/ijatcse/2019/76832019

[45] Yeow Haur Teng, Louis, Kuok, King Kuok, Imteaz, Monzur, Lai, Wai Yan and Kuok Xiong Ling, Derrick. Development of Whale Optimization Neural Network for Daily Water Level Forecasting. *International Journal of Advanced Trends in Computer Science and Engineering*. Volume 8, No.3, pp. 354 – 362. May - June 2019. The World Academy of Research in Science and Engineering Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse04832019.pdf. ISSN 2278-3091
DOI: ttps://doi.org/10.30534/ijatcse/2019/04832019.