International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 06, June-2022

Towards Ensemble Classification Algorithms for Breast Cancer Diagnosis in Women

[1]Akampurira Paul, [2]Semaluulu Paul(Ph.D), [3]Elly Gamukama (Ph.D).
[1]Kampala International University (Teaching Assistant, Computing, Science & Technology, Kampala, Uganda)
[2] Kabale University (Faculty of Computing, Library and Information Science, Kabale, Uganda)
[3] Kampala International University (School of Mathematics and Computing, Kampala, Uganda)

***Abstract* - Despite a spike in growth rate of modern techniques towards breast cancer diagnosis where a perfect diagnostic system would discriminate between benign and malignant findings perfectly, flawless discrimination has not been realized, so radiologists' decisions are founded on their best judgment of breast cancer risk amidst substantial uncertainty. And in low developed countries where adoption of computer based diagnostics for decision support is low, given the variety of options in the artificial intelligence and machine learning perspective, we endeavored to perform simulations on the breast cancer dataset and 5 classification algorithms that are supported for best performance given small datasets and low computational complexity needs, towards achieving an optimal ensemble model that would nearly perfectly discriminate between cancerous and non-cancerous breast tumors.**

Keywords: Machine learning, breast cancer, artificial intelligence, benign, malignant, supervised learning, ensembles, model optimization, cross validation.

I. INTRODUCTION

According to World Health Organization (WHO), (2018), Cancer is the second leading cause of death, responsible for approximately 9.6 million deaths in 2018 in the whole world. Among the many types of cancer, breast cancer is one of the leading causes of deaths and permanent body effects and leads to about 2.09 million deaths per year. And Approximately 70% of deaths from cancer occur in low- and middle-income countries.

In 2018, at least 22,000 Ugandans died of cancer and at least 350 new cases detected per 100,000 people compared to 2008 where statistics were 250 per 100,000 people. Dr. Jackson Orem, the director Uganda cancer institute (UCI) vividly reported that the great mortality rates are majorly because of late diagnosis. "when cancer is on the increase, it means so are the deaths because at any one time it is estimated that 80% of cancer patients die because of late diagnosis," he said, adding that 30% of all cancer cases are curable if detected early (Uganda Cancer Institute, 2021)

A key challenge against its detection is how to classify tumors into malignant (cancerous) or benign (non-cancerous). A tumor in the breast can be discriminated into malignant or benign. It is said to be malignant if the cells are likely to grow into surrounding tissues or spread to distant areas of the body.(Hamsagayathri & Sampath, 2017) A benign tumor is one that is unlikely to spread into the surrounding tissue or to propagate itself to other parts of the body like the cancerous tumors can. Therefore, screening is very important and vital and should be carefully done as it can have potential harm or benefits. When the diagnosis is done, it may be decided by the radiologists that the patient is okay or requires further treatment, that is, chemotherapy, mastectomy, and or even surgery need to be done (Henry, 2020).

Despite current challenges in medical diagnosis around the world today and especially in low-income countries, AI provide incredible potential for altering the course towards provision of healthcare services in resource-poor settings(Chaurasia, Pal, and Tiwari 2018). Many health system questions in such settings could be answered with the use of AI and other complementary emerging technologies, such as E-Systems and machine learning systems.(Blümel et al. 2020)

While different research studies have endeavored to assess different classification algorithms, including SVM, Naïve Bayes, Random Forests, Decision trees and neural networks, the percentage accuracy attained still lack a lot with a measurably significant error. For example, accuracy of data mining algorithms SVM, IBK, BF Tree as compared by (Thakur et al., 2017), showed a performance of SMO to have achieved higher accuracy rates compared with other classifiers. (Hamsagayathri and Sampath 2017) analyzed the performance of the four different decision tree algorithms for Breast cancer classification. The simulation results showed Priority based decision tree classifier classifies the data with 93.63% accuracy and confirmed that a Priority based decision tree algorithm is better than other classification algorithms for Wisconsin original, diagnostic and prognostic breast cancer dataset (P. Hamsagayathi, 2017).

With respect to these works, individual methods of classification still have low strengths as compared to a combination of algorithms called ensembles (AdnanO.M.Abuassba, 2017). To tackle the weakness in the most current works aiming at breast cancer diagnosis, our approach in this study aimed to improved prediction accuracy first, through thorough data preparation and second, through advanced modeling procedures of cross validation and ensemble approaches.

Ensemble learning is a branch of machine learning that seeks to use multiple learning algorithms so that better predictive performance can be acquired. Ensemble learning is a promising field for improving the performance of base classifiers. (Pavlicko, 2021) There are several classification models including Naïve Bayes, Logistic Regression, Multilayer Perceptron, Random Forest, Stochastic Gradient

Descent. The performance of different state-of-the-art machine learning classification algorithms were evaluated for the Wisconsin Breast Cancer Dataset (WBCD) and the best four were be used for ensemble classification.

1.1 Motivation

Early breast cancer diagnosis and the ability to discriminate malignant breast lesions from benign ones and accurately predicting the risk of breast cancer for individual patients are critical in successful clinical decision-making. In 2020, 9.2m cases of which 24.5% was breast cancer worldwide, 29.5% of cases in Africa, and in Uganda, 32617 new cases were recorded that lead to 21829 (66.9%) deaths in 2018. Various techniques are being used to detect cancer at an early stage. The major challenge in cancer diagnosis is the number of patients who are incorrectly diagnosed and thus increasing mortality and other late called for procedures and false assurance or where non-sick patients are wrongly recommended for treatment and undergo unnecessary treatment and face risky side effects, and wrong interventions that lead to irreversible damages including unnecessary surgeries. Moreover, investigations show that there are surgical interventions and treatment done while there is no need in the range of 65% and 80% of patients.(World Health Organization., 2019.)

Computer aided tools and Machine learning technologies have been adopted in some cases and have seen improvements in cancer diagnostics in breast cancer up to 97 % accuracy (Abuassba et al., 2017). However, precise and expert analysis on which ML model to employ on which data for the different algorithms or a combination of algorithms that perform differently on individual data sets for a given problem is required. Most proposed algorithms like SVMs greatly depend on the kennel and have high computational complexity and hence too expensive for medical centers in developing countries, and still some algorithms like t-SNE can work well only on a current dataset and cannot apply well to new data and hence not very useful for deployment in real world scenarios. The researcher undertook the task and came up with a more appropriate model that, if applied, could greatly benefit developing countries and significantly reduce misdiagnosis of the disease.

1.2 General objective

To develop an ensemble model for detecting breast cancer to reduce the error rate on diagnosis, and accurately predict a future risk of the disease.

1.2.2 Specific objectives

To establish requirements for designing machine learning model for diagnosing breast cancer in women with abnormal breast masses.

To design and develop classification models for diagnosing breast cancer in women with abnormal breast masses.

To Evaluate the performance of the individual models and establish a better model for diagnosing breast cancer in women with abnormal breast masses.

To develop an ensemble model from the evaluated classification algorithms for performance optimization towards breast cancer diagnosis in women.

1.3 Research questions

RQ1: what are the requirements for designing classification models for diagnosis of breast cancer in women?

RQ2: how do we develop the classification algorithms for the diagnosis of breast cancer in women?

RQ3: How do we evaluate the developed algorithms to establish a better model?

RQ4: How can we combine the different developed models to achieve a better classification?

1.4 Conceptual modeling

The researcher established that a model is a translation into a mathematical form of a system placed under study and in this case a breast cancer diagnosis system, and once there is a mathematical, or logical form that would describe system responses under different levels of precision, hence we would be able to make predictions about its development and responses to certain inputs. The formal challenge of establishing a mathematical model for an unknown system (also referred to as target system) by observing its input and output data pairs, is generally referred to as system identification which involved structure identification and parameter identification.

Under structure identification, we considered a parameterized function y = f (u, t) where y is the output, u is the input and t, is a parameter vector. In this case, where our system is predetermined, the input variables were the independent variables and the output the dependent parameter which is classified into cancerous or not cancerous, that is, malignant or benign. Thereafter, optimization techniques were applied to determine parameter vector t such that the resulting model y* = f (u, t*) could be applied for a more optimized and more accurate model. In the parameter identification, a process of identifying the parameters that best fit the available dataset was done (difference y-y* is minimal).

Furthermore, since the problem at hand was to accurately discriminate between cancerous and non-cancerous masses with good accuracy, this therefore became a classification problem. Based on a binary classification, for data of the form D= {(x1, y1),(x2, y2), ..., (xn, yn)} where x ER and y = ±1. considering x as the independent variable and y as the dependent variable and for feature engineering assuming that we could represent the features of the sample mass or biopsied cell as x ER and the target variable as y.

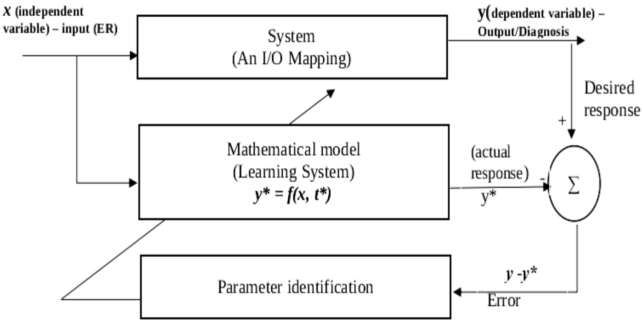


Figure 1: Conceptual model

1.5 Justification

When breast cancer diagnosis is done, it may be decided by the radiologists that the patient is okay or not recommending further treatment, that is, chemotherapy, mastectomy, and or even surgery need to be done(Macaulay et al., 2021). According to National Cancer institute, 2018, screening tests can have false-positive results where the test indicates that cancer may be present even when it actually is not. False-positive test results can cause anxiety and are usually followed by additional tests and procedures that are always expensive and also have potential harms including unnecessary surgeries and mothers have lost single or both breasts as a result of poor diagnosis. false negative can give false assurance and lead to late diagnosis and makes complications for a case which would have been simple. Late interventions are normally intended to nurse the patient with less potential of recovery and the procedures are very expensive. According to WHO, 2018, Early diagnosis is necessary and accurate results are mandatorily required and action is needed urgently to reduce such cases.

Several data mining classification approaches such as Neural Networks, Support Vector Machine, Random forests Decision Tree, Naïve Bayes were implemented by researchers to diagnose breast cancer disease. But there is a challenge to ascertain which of these data mining techniques perform effectively. It has been also identified that most time single data mining method may not provide desired result. In order to find a solution to this problem, the study conducted a performance evaluation on the most commonly data mining algorithms that would require less computing power to cater for low-income communities: A combination of different classifiers could help to achieve better results. In addition, the importance accurate diagnosis is in finding ways to improve patient outcomes, it can reduce the medical cost and enhances early disease discovery(Abuassba et al., 2017)

It was therefore, imperative that models that can easily learn from small dataset such as meta learners or ensembles be studied and designed to solve this issue to benefit medical research especially in developing countries where data collection is still young. With meta ensemble learning one can minimize generalization error to some extent irrespective of the data distribution, number of classes, choice of algorithm, number of models, complexity of the datasets, etc. So, in summary, the predictive models will be able to generalize better(Pavlicko, 2021; Perlich & Świrszcz, 2011).

2. METHODOLOGY

Our study conforms to the Data Science Methodology (DSM), which helped us to keep track of which phase of the analysis we were performing. A better industry standard process encouraged for computer scientists and data scientists is the cross-industry process for data mining (CRISP-DM). Broadly, CRISP-DM recognizes six phases which include; Problem understanding, Data understanding, Data preparation, Modeling, Evaluation, and Deployment. CRISP-DM is an idealized sequence of events. In practice many of the tasks could be performed in a different order and it would often be necessary to backtrack to previous tasks and repeat certain actions(Kantardzic, 2020). The researcher therefore based on the above methodology and phases therein to align this research project.

2.1 Overview of Methodology

A descriptive study with quantitative data in this study was done through simulation and experimentation involving iterative processes aimed to discover appropriate models and derive values from breast cancer data set. Our methodology therefore was performed through the following processes of problem identification, data understanding and preparation, data modeling, model evaluation, model validation and optimization, ensembles and finalization and documentation as per the CRISP-DM

2.2 Problem Identification

Our data-based modeling methodologies were performed in particular to the problem of breast cancer diagnosis as stated in our chapter one of this study. Our study performed operations to implement our conceptual study through modeling and analysis of the breast cancer dataset features that are used to discriminate between malignant and benign tumors. The aim was to determine a model that is capable of reducing the error related to false negatives and false positives in the diagnosis results and thus a good accuracy level of a model whose results can easily be interpreted.

2.3 Understanding the data

Data Collection

In this phase, an observation approach was followed since the researcher could not influence the data generation process unlike design experiments where the data generation process is under the control of the researcher or expert. In our setting, the data used was secondary data collected from existing online databases which provide the required standard and authenticated datasets to be used for our experimentation and simulation modeling.

Data exploration, definition, and preparation

After collecting the data, we imported it into R studio for exploration and visualization. We explored the data structures, the feature and examples and realized the peculiarities within our data. We did this to better understand our data and match appropriate machine learning models towards our learning problem. The organization of our dataset was studied and where necessary reorganized or restructured it to our preference to make it easy to work with.

The major exploration and visualization studies in our work included measuring central tendency of the data, measuring the spread of the data, visualizing numeric variables, understanding numeric data through uniform and normal distributions, and exploring and visualizing and examining, relations between the features also referred to as variables. In our observational setting, the collected data underwent several tasks of preparation that include; outlier detection, dealing with missing data, and data normalization.

Outlier detection (neutralization and or removal):

We can define outliers as some unusual data or data values which are not consistent with most observations. In most cases, outliers can come up due measurement errors and coding and recording errors and, sometimes, are natural, abnormal values. Such non-representative samples can with great significance affect the model produced later and we therefore studied our data to identify any outliers research neutralized them or removed them as deemed necessary.

Dealing with missing data

The simplest solution for a missing data problem would be the reduction of the data set and the elimination of all samples with missing values. That can be done especially with large data sets where missing values occur only in a small percentage of samples as compared to the whole data set. If the researcher does not choose to drop the samples with missing values, then we have to find values for them.

2.2 Data Normalization

There are several methodologies that we can use for data normalization including decimal scaling, Min-Max normalization, Z-score normalization but the researcher used the former for this research since most algorithms are accommodated in the normalization process.

Data normalization was a significant step performed by the researcher and was majorly done to remove bias where absolute quantities are less meaningful than relative ones due to differences in scale and the normalization step ensures that all variable would hold same weight during modeling.

We applied the min-max normalization which would transform a feature such that all of its values fall in range between 0 and 1. The formula for normalizing a feature is as follows;

$$Xnew = \frac{X - min(X)}{\max(X) - \min(X)}$$

Where, for each value of feature X, the formula subtracts the minimum X value and divides by the range of X. The resulting normalized feature values can be interpreted as indicating how far, from 0 percent to 100 percent, the original value fell along the range between the original and maximum.

Checking for multicollinearity among the variables in our dataset

A multicollinearity check was done to look for correlation in the variables. This was done because most ML algorithms assume that the predictor variables are independent from each other for an analysis to be robust, and hence the researcher performed an analysis that led to checking and removing multicollinearity. We used Pearson correlation to check for relationships among our dataset features. Mathematically, the Pearson correlation coefficient (ρ) between two random variables x and y is denoted as follows:

$$\rho_{x,y} = \frac{Cov(x, y)}{\sigma_x \sigma_y}$$

where Cov (x y), is the covariance of x; y, σx is the standard deviation of x; and σy is the standard deviation of y.

2.3 Data reduction (Feature selection and extraction)

In this phase, the researcher further explored the features in the dataset to establish their importance towards the outcome or target variable and their relationships among the variables. Unimportant features were removed, Collinearity checks were done and highly correlated features were dealt with appropriately. Further, the dimensional space was reduced, following standard methodology of Principal component analysis (PCA) which is a general-purpose technique to reduce the dimensionality of the data and enhance our feature selection extracting criteria.

There are various methodologies for dimensionality reduction including, relief algorithm, entropy measure of ranking features, principal component analysis, Chi Merge, value reduction, case reduction etc, but the researcher employed Principal component analysis (PCA) in this study for its simplicity and yet comprehensive techniques. PCA is a method of transforming the initial data set represented by vector samples into a new set of vector samples with derived dimensions.

2.4 Designing classification models for diagnosing breast cancer.

In this phase, the researcher designed the classification models and train them on the learning data here by referred to as the training data prepared from the last step of feature engineering and dimensionality reduction which provide an optimal set for training the learners. The different models underwent different methodologies as per their requirements each model was trained on the prepared data. The performance of the models in terms of speed, resource usage in terms of machine power required, the accuracy (considering error rate), specificity and sensitivity were considered. Confusion matrix methodology and ROC Curves were majorly used to determine the accuracy of the models.

Decision Tree classifier modeling and evaluation

The decision tree has versatile features that help to actualize both categorical and continuous dependent variables, it is a type of supervised learning algorithm mostly used for classification problems. The decision tree splits the population into two or more homogeneous sets based on the most significant attributes making the groups as distinct as possible.

The CART method in R produces decision trees that are strictly binary, containing exactly two branches for each decision node. CART recursively partitions the records in the training data set into subsets of records with similar values for the target attribute. The CART algorithm grows the tree by conducting for each decision node, an exhaustive search of all available variables and all possible splitting values, selecting the optimal split according to the Gini Index.

Let Φ(s|t) be a measure of the “goodness” of a candidate split s at node t, where

$$\phi(\mathrm{s|t}) = 2P_L P_R \sum_{j=1}^{\#classes} |P(j|t_L) - P(j|t_R)|$$

And where;

tL = left child node of node t

tR = right child node of node t

$$\mathrm{P_L} = \frac{\text{number of records at } \mathrm{t_L}}{\text{number of records in training set}}$$

$$\mathrm{P_R} = \frac{\text{number of records at } \mathrm{t_R}}{\text{number of records in training set}}$$

$$\mathrm{P(j|t_L)} = \frac{\text{number of class j records at } \mathrm{t_L}}{\text{number of records at t}}$$

$$\mathrm{P(j|t_R)} = \frac{\text{number of class j records at } \mathrm{t_R}}{\text{number of records at t}}$$

Random Forest classifier building and evaluation

With random forests, we built a series of decision trees and combine the trees disparate classifications of each record into one final classification. Random forests are an example of an ensemble method which seek to improve performance of the model.

Partial least squares-discriminant analysis

Partial least squares-discriminant analysis (PLS-DA) is a versatile algorithm that can be used for predictive and descriptive modelling as well as for discriminative variable selection. Partial Least Squares are examples of such methods of dimensionality reduction and they provide crucial datasets while dealing with medical data, since it is necessary to compress patient information and retain only the most useful in order to discriminate subjects into benign and malignant classes as in our case.

Logistic Regression

Logistic regression models the probability of a particular response value. Applying this idea to our stated problem, we try predict the probability that a patient has a cancerous tumor or not. We use a logistic function below to simulate our model

$$\mathrm{P(X)} = \Pr(Y = 1|X) = \frac{e^{B_0+B_1X}}{1 + e^{B_0+B_1X}}$$

The output for the logistic function is always between 0 and 1 for all possible values of X.

2.5 Developing an ensemble model

We intended to combine the different algorithms as a means of optimization where two or more algorithms could be more robust and more accurate than individual algorithms. We used three methods of ensembles including bootstrap aggregation, stacking and boosting.

Decision tree ensemble through boot strap Aggregating (bagging)

Bootstrap aggregating, also called bagging (from bootstrap aggregating), is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. One advantage with bagging is that it reduces variance. However, it does not reduce bias. Although it is usually applied to decision tree methods, it can be used with any type of method. Bagging is a special case of the model averaging approach. Bagging, involves having each model in the ensemble vote with equal weight. In order to promote model variance, bagging trains each model in the ensemble using a randomly drawn subset of the training set.

Given a training set D = {(x1, y1), . . . (xn, yn)}, sample T sets of n elements from D (with replacement) D1, D2, . . . DT → T quasi replica training sets; train a machine on each Di, i = 1, ..., T and obtain a sequence of T outputs f1(x), . . . fT (x). Hence for our classification $\dot{\mathrm{f}}(\mathrm{x}) = sign(\sum_{i=1}^{T} f_i(x))$ the standard function “Standard” bagging: each of the T subsamples has size n and created with replacement.

Stochastic Gradient Boosting with Random Forest

Random forests are in themselves and ensemble of decision tree and hence further boosting of random forest can generate more improvement on our model and we employed Gradient boosting that sprouts from Gradient descent. Gradient descent can often have slow convergence because each iteration requires calculation of the gradient for every single training example. Our model therefore endeavored to update the parameters each time by iterating through each training example, so that we could get excellent estimates.

Both boosting and bagging randomly generate a number of data subsets from our bc_training dataset through sampling with replacement. Boosting, unlike bagging, further integrates a weighting strategy in the sampling process that assigns higher weights to the incorrectly classified examples. This is done to increase the diversity among the different classification trees(classifiers) in our forest model. Here, the classification error is measured after each classifier is trained, and the samples that are classified incorrectly by the first classifier receive a larger weight in the subsequent training subsets. Since our stochastic gradient boosting model can be derived with regard to gradient descent, Gradient descent

$$\nabla J(\Theta) = \frac{1}{N}(y^T - \theta X^T)X$$

becomes stochastic gradient descent $\nabla J(\Theta)_i = \frac{1}{N}(y_i - \theta^T X_i)X_i$, Where i is each row of the of the breast cancer data set. This is the stochastic gradient descent algorithm proceeds as follows for the case of linear regression:

```
Step 1: Randomly shuffle the data
Step 2:  repeat
{
for
i:=1,⋯,N{
θ:=θ−η∇ J(θ)i
}
}
```

We applied adaboost library with the gbm packages that implement the above in r.

Ensembling through stacking

Stacking (short for stacked generalization, also known as meta ensembles, meta-learning, stacking meta-learning, or stacked ensembles) is often based on heterogeneous learning algorithms. Stacking obtains the final ensemble decision by stacking different classifier layers, hence the name. As in bagging, stacked classifiers in the base learning pool have a parallel structure. The difference between stacking and bagging is the algorithm used by the classifiers in the base learning pool. In bagging, each classifier uses the same classification learning algorithm (such as a decision tree), while stacking uses different algorithms to train different classifiers (such as decision trees, logistic regression, and random forests).

In other words, each classifier in stacking uses a heterogeneous learning algorithm in contrast to bagging and boosting. In terms of the combination method, stacking also combines the predictions of different classifiers by training classifiers (Hazel, B., et al, 2017).

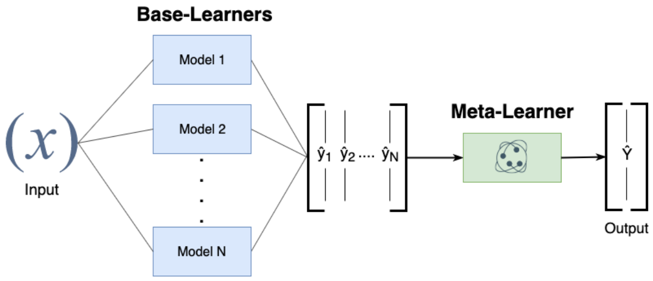


Fig 2: Model stacking

Therefore, we used our weak learners (base-learners) including 'rf', 'treebag', 'gbm', 'rpart', 'glm', 'pls' and logistic regression as a meta learner.

Preparing the models for stacking: We combined the different models of random forests, decision trees, logistic regression, partial least squares, and also included the already designed to ensemble models of bagged trees and boosted random forests. The models were trained on the normalized set and we validated them through 10-fold cross validation

Stacking using the generalized linear model (glm): Here, we have two layers of machine learning models; bottom layer models (rf, treebag, rpart, glm, gbm, pls) which receive the input features from our cross validated training dataset, and a top layer model, glm, which implements logistic regression as a meta learner which takes the output of the bottom layer models as its input and predicts the final output.

Stacking using random forest: We also applied random forests (rf) as a meta learner to use predictions from the base learners of rf, treebag, rpart, glm, gbm, pls, which also received input from a cross validated set.

2.6 Evaluating the performance of the designed models and establish a better model

Model evaluation phase involved a comparative study of the algorithms' performance achieved in the modeling and evaluation phases. This was majorly done to establish the strengths and weaknesses of the algorithms and determine which algorithm can be most appropriate for breast cancer diagnosis. A more appropriate model should be affordable in terms of cost to an institution, should be easily explainable, that is, the results from the algorithm should well understood and simple enough to help the decision-making process, and should exhibit robustness in terms of capability to deploy in real life cases without affecting its performance.

In this phase, the researcher performed experiments to make sure that the chosen algorithm performs better on unknown instances and parameter tuning was done to finally optimize the model and make it ready for deployment. The accuracy, sensitivity and specificity and the Area Under the Curve (AUC) of the model performance on test data set were considered to compare the performance of individual learners.

Model evaluation metrics (Confusion matrix)

Accuracy: We developed classification evaluation measures for the case where we have a binary target variable. In order to apply the measures, we would need to denote (arbitrarily, if desired) one of the two target outcomes as positive and one as negative (Timothy Masters, 2020). And in this study a sample being malignant would mean positive and benign would mean negative.

$$Accuracy = \frac{TN + TP}{TN + FN + FP + TP} = \frac{TN + TP}{GT}$$

We also determined the error rate as follows;

$$Error\ Rate = 1 - Accuracy = \frac{FN + FP}{TN + FN + FP + TP} = \frac{FN + FP}{GT}$$

Accuracy represents an overall measure of the proportion of correct classifications being made by the model, while error rate measures the proportion of incorrect classifications, across all cells in the contingency table. However, these measures do not distinguish between the various types of errors or the various types of correct decisions. And since we are interested in the true positive rate versus the true negative rate to determine how well a model would we discriminate between sick and non-sick patients, we therefore had to do so using sensitivity and specificity, as follows.

Sensitivity and Specificity: Sensitivity determines the ability of the model to classify a record positively, while specificity determines the ability to classify a record negatively. Sensitivity measures what proportion of all positive records are captured by a model, while specificity measures what proportion of all the negative records that are captured by your model. Of course, a perfect classification model would have sensitivity= 1.0 = 100%. However, a model which simply classified all cases as positive would also have sensitivity = 1.0=100%.

$$Sensitivity = \frac{Number\ of\ true\ positives}{Total\ actually\ positive} = \frac{TP}{TAP} = \frac{TP}{TP+FN}$$

$$Specificity = \frac{Number\ of\ true\ negatives}{Total\ actually\ negative} = \frac{TN}{TAN} = \frac{TN}{FP+TN}$$

Clearly, it is not sufficient to identify the positive responses alone. A classification model also needs to be specific, meaning that it should identify a high proportion of the cases which are negative. Of course, a perfect classification model would have specificity = 1.0. But, so would a model which classifies all cases as negative. A good classification model should have acceptable levels of both sensitivity and specificity, but what constitutes acceptable varies greatly from domain to domain (Kantardzic, 2020). And hence in our case, we would not want to have more false negatives as compared to false positives. Both situations have dire consequences, however a patient who is sick should not be wrongly classified as not sick. We would there prefer a model that greater sensitivity levels reaching 100 %.

Model evaluation phase involved a comparative study of the algorithms' performance achieved in the modeling and evaluation phases. This was majorly be done to establish the strengths and weaknesses of the algorithms and determine which algorithm can be most appropriate for breast cancer diagnosis. A more appropriate model should be affordable in terms of cost to an institution, should be easily explainable, that is, the results from the algorithm should well understood and simple enough to help the decision-making process, and should exhibit robustness in terms of capability to deploy in real life cases without affecting its performance. In this phase, the researcher performed validation experiments to make sure that the chosen algorithm performs better on unknown instances and parameter tuning was done to finally optimize the model and make it ready for deployment. The accuracy, sensitivity and specificity and the Receiver Operating characteristic (ROC) curve with the Area Under the Curve (AUC) taken into consideration for the model performance evaluation on test data set were considered to compare the performance of individual learners.

There are a variety of methodologies for model validation but the researcher used k-fold cross validation for this phase. Cross-validation is a technique for ensuring that the results uncovered in an analysis are generalizable to an independent, unseen, data set (Timothy Masters, 2020). This was done through applying the model to cross-validated dataset.

3.0 RESULTS

3.1 Data Understanding and Data Preparation

The researcher downloaded the Wisconsin Breast Cancer Database (WBCD) dataset which has been widely used in research experiments. The WBCD dataset for breast cancer diagnosis is comprised of feature values calculated from digitized image of a Fine Needle Aspirate (FNA) of a breast mass. These features describe the characteristics of the cell nuclei present in the image. This database is also available through the UW CS ftp server: ftp ftp.cs.wisc.edu cd math-prog/cpo-dataset/machine-learn/WDBC. This standard dataset is publicly available and recommended for data science and machine learning experiementation.

Data Description:

The data was originally created by Dr. William H. Wolberg, General Surgery Department at the University of Wisconsin, Clinical Sciences Center in Madison, WI 53792 (wolberg '@' eagle.surgery.wisc.edu)

Table 1 : data set detail

Data Set Characteristics:	Multivariate	**Number of Instances:**	569	**Area :**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	32	**Date Donated**	1995-11-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	1534523

Understanding our data: Attribute Information:

The different attributes include: ID number, Diagnosis (M = malignant, B = benign), and Ten real-valued features are computed for each cell nucleus: radius (mean of distances from center to points on the perimeter); texture (standard deviation of gray-scale values),

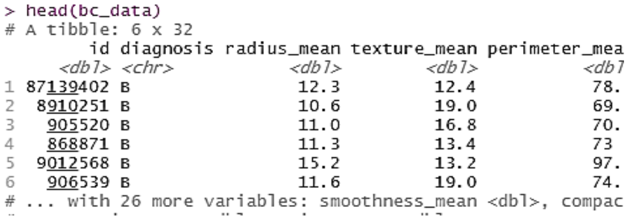
perimeter, area, smoothness (local variation in radius lengths); compactness (perimeter^2 / area - 1.0); concavity (severity of concave portions of the contour); concave points (number of concave portions of the contour); and symmetry, fractal dimension ("coastline approximation" - 1).

Also, the mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius. All feature values are recorded with four significant digits.

The downloaded data is in a comma separated values format and can be opened form most data editing tools and our environment in RStudio. We imported our dataset into the RStudio and saved it as bc_data.

Viewing our data

We looked at the structure of the data by using head(), which also show a detailed view of the data in terms of data structures of the features, as follows;



```
> head(bc_data)
# A tibble: 6 x 32
         id diagnosis radius_mean texture_mean perimeter_mea
      <dbl> <chr>           <dbl>        <dbl>         <dbl
1 87139402 B                12.3         12.4           78.
2  8910251 B                10.6         19.0           69.
3   905520 B                11.0         16.8           70.
4   868871 B                11.3         13.4           73 
5  9012568 B                15.2         13.2           97.
6   906539 B                11.6         19.0           74.
# ... with 26 more variables: smoothness_mean <dbl>, compac
```

Table 2: summary for the data and structure

From the head() results as shown above, we discovered that most features were stored as double float (dbl) and character(chr). The dataset consists of a total of 32 columns and 569 entries or examples. It includes an id column and labels or target values as B and M for Benign and Malignant respectively.

A raw count of the data after initial preprocessing showed 30 features or predictors and 569 observations. We also discover that all the predictors have continuous values for observations and there are no missing values. We noted that the observations were all recorded as continuous numerals in decimals.

Checking for multi-collinearity among the variables in our dataset

Person's correlation values range from -1 to 1 and any feature with a value of 0.9 and above from our plot above shows a very strong positive correlation and features with -0.9 or below show a strong negative correlation and need to be removed for better modeling. Area_se, texture_mean, texture_worst are some of the highly positively correlated feature. In the step below, the researcher demonstrated how to check for the highly correlated values using the caret package.

We managed to get the detailed view of the relationships between features and the correlation values showing how some features are highly correlated with each other which may hinder the robustness of our modeling results and hence this helped the researcher to identify and remove or harmonize such features, as area mean and radius mean. The researcher chose to harmonize these by applying principal component analysis as seen in the next sections. But first, we looked at a further detail of the correlations using scatter diagrams as follows.

The Correlation plots visually show how the different features are correlated. We note that correlation does not mean causation and hence this merely demonstrate an observed association. We noticed a strong positive trend between radius mean, area mean and perimeter mean. We also noted a positive correlation between compactness mean and concavity mean, radius mean and concavity mean. The scatter diagrams also show the distribution graphs of the features which further show the skewness of the data.

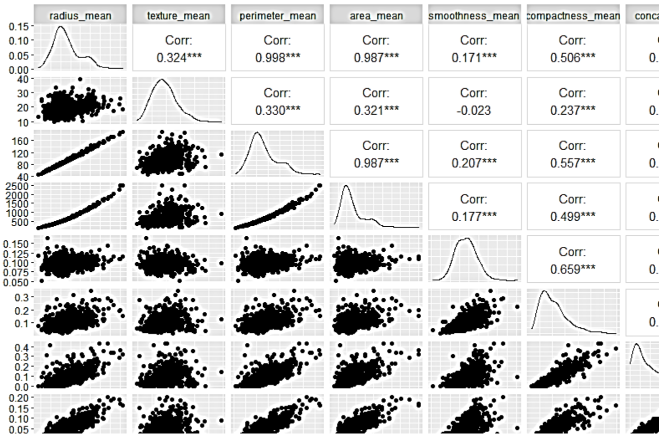


Fig 3: correlation plot

The findcorrelation() function from the caret package was used to remove highly correlated predictors. The function uses a heuristic algorithm to determine which variable should be removed. The researcher applied the function to remove features that are highly correlated with a Pearson's correlation coefficient of 0.9 or more as follows and saved the new data into a new dataset, bc_data_corr1. The resulting dataset from the above transformation is 10 variables shorter and is only comprised of 22 predictors in the dataset bc_data_Corr1. We noted however that some algorithm may work well despite being applied to highly correlated features or not.

3.2 Ensemble Modeling with Bagging, Boosting and Stacking

Decision tree ensemble through boot strap Aggregating (bagging):

```
RESULTS
Bagged CART

456 samples
 22 predictor
  2 classes: 'B', 'M'
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 410, 411, 410, 411, 411,
410, ...
Resampling results:
  ROC        Sens       Spec
  0.9891445  0.9685961  0.9176471
Confusion Matrix and Statistics
          Reference
Prediction  B  M
         B 69  2
         M  2 40
```

The resulting confusion matrix for the bagged tree model showed a great improvement from the original Decision tree classifier of error rate of 0.0531 and reduce to 0.0354. Only two malignant patients were misclassified and only 2 benign patients were misclassified and hence the ensemble algorithm did better than the individual decision tree.

```
                   Accuracy : 0.9646
                     95% CI : (0.9118, 0.9903)
        No Information Rate : 0.6283
        P-Value [Acc > NIR] : <2e-16

                      Kappa : 0.9242

 Mcnemar's Test P-Value : 1

                Sensitivity : 0.9524
                Specificity : 0.9718
             Pos Pred Value : 0.9524
             Neg Pred Value : 0.9718
                 Prevalence : 0.3717
             Detection Rate : 0.3540
       Detection Prevalence : 0.3717
          Balanced Accuracy : 0.9621

           'Positive' Class : M
```

The bagged classification tree managed to achieve a prediction accuracy of 96% and hence only 4 of 100 patients would be misclassified. The sensitivity and specificity of the model also improved to 95% and 97% respectively. The kappa value of 0.92 showed a very highly reliable model.

Random Forest ensemble with Stochastic Gradient Boosting

```
cm_gbm_bc
Confusion Matrix and Statistics
          Reference
Prediction  B  M
         B 71  0
         M  2 40
```

The ensemble random forests were designed to automatically manage overfitting through parameter tuning and ROC was used to select the optimal model suing the largest value. A total of 150 tree were reached. The Stochastic Gradient Boosted model achieved good results and beautifully classified the non-sick patients without any false positive and misclassified two sick patients as non-sick.

```
                   Accuracy : 0.9823
                     95% CI : (0.9375, 0.9978)
        No Information Rate : 0.646
        P-Value [Acc > NIR] : <2e-16
                      Kappa : 0.9617
 Mcnemar's Test P-Value : 0.4795
                Sensitivity : 1.0000
                Specificity : 0.9726
             Pos Pred Value : 0.9524
             Neg Pred Value : 1.0000
                 Prevalence : 0.3540
             Detection Rate : 0.3540
       Detection Prevalence : 0.3717
          Balanced Accuracy : 0.9863
           'Positive' Class : M
```

The ensemble model achieved an accuracy of 98% with an error rate of 0.0177. This an almost accurate model and its sensitivity is at 100% meaning all non-sick patients are well discriminated from sick patients and no sick patients are told they not. The specificity of 97% is also a fairly good measure where only about three patients are told they are sick when actually they are.

Ensembling through stacking

We combined the different models of random forests, decision trees, logistic regression, partial least squares, and also included the already designed to ensemble models of bagged trees and boosted random forests. The models were trained on the normalized set and we validated them through 10-fold cross validation.

We used summary to get a deeper insight of how well our models would learn from the breast cancer dataset and the results were as follows;

Table 2: Model training summary results

```
summary(results)
Call:
summary.resamples(object = results)
Models: rf, treebag, gbm, rpart, glm, pls
Number of resamples: 30
Accuracy
             Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
rf      0.9122807 0.9466635 0.9562808 0.9566474 0.9649123 1.0000000    0
treebag 0.8947368 0.9464286 0.9482759 0.9531073 0.9649123 1.0000000    0
gbm     0.9310345 0.9649123 0.9821429 0.9747886 0.9826830 1.0000000    0
rpart   0.8245614 0.8933271 0.9298246 0.9209903 0.9473684 0.9824561    0
glm     0.8965517 0.9475953 0.9652148 0.9637582 0.9824561 1.0000000    0
pls     0.7931034 0.8653017 0.8937970 0.8893329 0.9122807 0.9649123    0

Kappa
             Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
rf      0.8169557 0.8846154 0.9076595 0.9069319 0.9246032 1.0000000    0
treebag 0.7738095 0.8846154 0.8929889 0.8998378 0.9254329 1.0000000    0
gbm     0.8560794 0.9230769 0.9615385 0.9456736 0.9629610 1.0000000    0
rpart   0.6374046 0.7720238 0.8490509 0.8312629 0.8877099 0.9619238    0
glm     0.7841191 0.8883048 0.9264189 0.9225324 0.9626719 1.0000000    0
pls     0.5180055 0.6926914 0.7594937 0.7484432 0.8016701 0.9230769    0
```

Table:

The table above showed that the models learned well with a mean accuracy of 96% for random forests, 95% for bagging, 97% stochastic gradient boosting, 92% for decision trees, 96% for logistic regression, and 89% for partial least squares models. Following the requirements for model combinations under stacking, we performed a correlation analysis of the models' prediction to determine that they all act independent of each other as follows;

```
# correlation between results
> modelCor(results)
               rf   treebag       gbm       rpart         glm         pls
rf      1.0000000 0.6702288 0.6911679  0.49846352  0.19282288 0.51363167
treebag 0.6702288 1.0000000 0.6239820  0.64556274  0.14227945 0.32973031
gbm     0.6911679 0.6239820 1.0000000  0.42665378  0.13550714 0.25723194
rpart   0.4984635 0.6455627 0.4266538  1.00000000 -0.07352882 0.35750322
glm     0.1928229 0.1422795 0.1355071 -0.07352882  1.00000000 0.02198567
pls     0.5136317 0.3297303 0.2572319  0.35750322  0.02198567 1.00000000
```

Table 3: collinearity check for model stacking

After checking for collinearity, the results were satisfying since the results showed independence of the predictions. The model's predictions that were highly correlated at 0.69 were gbm and random forests, and treebag and random forest. This could be attributed to the fact that they are all based on decision trees classification algorithm. The results however were good enough for us to continue and build ensemble with the stacking technique.

Stacking using the generalized linear model (glm)

Here, we have two layers of machine learning models; bottom layer models (rf, treebag, rpart, glm, gbm, pls) which receive the input features from our cross validated training dataset, and a top layer model, glm, which implements logistic regression as a meta learner which takes the output of the bottom layer models as its input and predicts the final output.

```
Ensemble results:
Generalized Linear Model

1707 samples
   6 predictor
   2 classes: 'B', 'M' 

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 
times) 
Summary of sample sizes: 1536, 1537, 1536, 1536, 
1537, 1537, ... 
Resampling results:

  Accuracy   Kappa    
  0.9767663  0.9500492
```

The stacked ensemble with logistic regression meta learner achieved an accuracy of 98% with an error rate of only 0.0232. The reliability of model based on its predictors is very high shown by the kappa value of 0.95.

Stacking using random forest

We also applied random forests (rf) as a meta learner to use predictions from the base learners of rf, treebag, rpart, glm, gbm, pls, which also received input from a cross validated set. The rf meta learner then gave us the final prediction and it accuracy was measured as shown below.

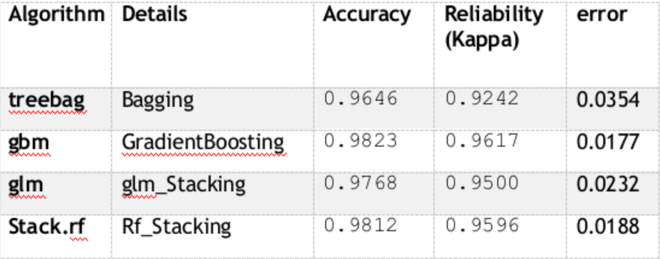
```
Ensemble results:
Random Forest 
1707 samples
   6 predictor
   2 classes: 'B', 'M' 
No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 
times) 
Summary of sample sizes: 1536, 1536, 1537, 1537, 
1536, 1536, ... 
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa    
  2     0.9812441  0.9596511
  4     0.9808554  0.9588414
  6     0.9812441  0.9597022

Accuracy was used to select the optimal model using 
the largest value.
The final value used for the model was mtry = 2.
```

The ensemble model designed with rf as the meta learner also managed to achieve an accuracy of 98% and the model proved to be more reliable with a kappa value of 0.96.

Comparative Analysis of the models

The table below shows the difference model performance measures of our designed ensemble models based on accuracy, reliability and error rates.

Table 4: Comparative analysis of ensemble models



Algorithm	Details	Accuracy	Reliability (Kappa)	error
treebag	Bagging	0.9646	0.9242	0.0354
gbm	GradientBoosting	0.9823	0.9617	0.0177
glm	glm_Stacking	0.9768	0.9500	0.0232
Stack.rf	Rf_Stacking	0.9812	0.9596	0.0188

A comparative analysis shows that stochastic gradient boosting for random forests achieved the highest accuracy with least error of 0.0177, and model stacking with rf meta learner also achieved an accuracy of 98% with an error of 0.0188. bagging with decision tree achieved the least accuracy and reliability of 96% and 0.9242 respectively.

Stacking with logistic regression as a meta learner achieved an accuracy of 98% with an error rate of 0.0232. even though it performed better that the bagged classification trees, both boosting with random forest and stacking with logistic regression performed better that all, with stochastic gradient boosting being the winner.

Importance of the features

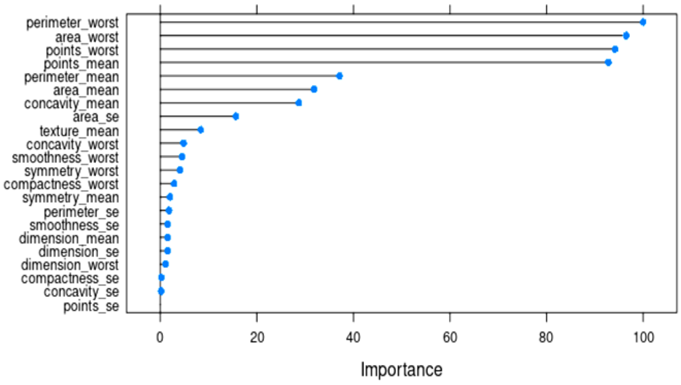


Fig 4: importance of features

The figure above shows the importance of individual features towards the discrimination of sick patients from non sick patients. We noted that perimeter_worst, area_worst, points_worst and points mean carry the most weight towards the breast cancer prediction as compared to other features. On the other hand, points_se, concavity_se and compactness_se carry the least weight towards breast cancer prediction.

However, we must state that despite the difference in weights, all features are significantly important for the prediction hence less weight does not equate to useless.

Specificity versus Sensitivity

plot(rocCurve.gbm,add=TRUE,col=c(3)) # color green is g
plot(rocCurve.rf,add = TRUE, col=c(6)) #color is purple
plot(rocCurve.bagg,add=TRUE,col=c(2)) # color is red

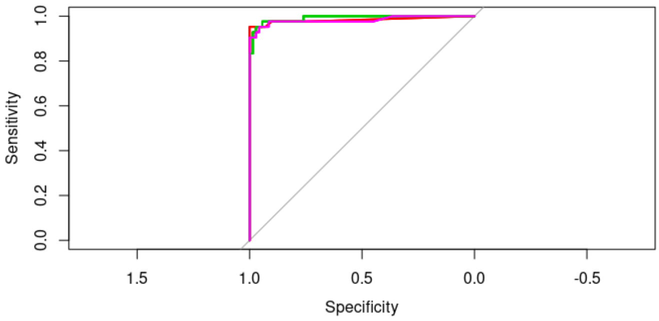


Fig 5: ROC curves for the ensembles

the figure above shows the receiver operating characteristic(ROC) curves with the Area Under the Curve showing the bias variance trade-off among the ensemble models of stochastic gradient boosting(gbm) with random forests, bagging with Decision trees(bagg) and stacking with random forests(rf). The gbm emerged a winner followed by rf and then bagg.

4.0 DISCUSSION, CONCLUSION AND RECOMMENDATIONS

We discuss the different models, their performance, strengths and weaknesses towards breast cancer diagnosis.

The researcher designed different models including logistic regression, decision trees, random forests and partial least squares discriminant analysis model. We further combined different models to produce ensemble models with different combination criteria of bagging, boosting and stacking.

We applied 10-fold cross-validation where we would repeat the construction of a model only on data not seen during training which would allow us to use each and every example in both training and evaluating models (Perlich & Świrszcz, 2011) & (Mount & Thomas, 2020)

The developed models were then subjected to testing for how well they discriminate unknown data and evaluated using a confusion matrix to determine the false positives and false negatives which were used to calculate the accuracy, sensitivity and specificity. We used Sensitivity to measure the ability of a model to measure the proportion of all malignant patients and specificity to measure the proportion of all benign patients captured by our model(Kantardzic, 2020).

Our applied bootstrap aggregating method helped to optimize the size of the tree while tuning the complexity parameter. With the bagged tree model, we achieved better accuracy rates at 0.9646% and 0.99% reliability rate of 0.92 as compared to the original DT model. The sensitivity and specificity of the bagged tree was well balanced. However, the performance was not optimal and we further endeavored to apply boosting techniques to RF model for a better performance.

The random forest model changes the algorithm for the way that the sub-trees are learned so that the resulting predictions from all of the subtrees have less correlation. Therefore, we hoped to reduce the bias through using random forests which is advanced DTs and further applied boosting with stochastic gradient boosting for the RFs. Boosting algorithms tend to be prone to overfitting and hence, we included parameter tuning as a crucial part of boosting algorithms to make them avoid overfitting.

With the boosted random forest model, we achieved a highest model accuracy of 0.9823 and reliability of 0.9617 with least error rate. The model also achieved the highest level of sensitivity and classified best for benign tumors with no false negatives. Also, the model was highly reliable with kappa of 0.96. With only 0.03 specificity error, the model would work well on unknown data in the real-world cases.

Even though a good prediction accuracy rate was reached with boosted RF model, it is believed that combining different models is far better than boosting or bagging single models. Stacking heterogeneous models would produce a more robust model that would generalize better on new data as compared to homogenous ensemble as seen with bagging DTs and boosting RFs. The difference between stacking and bagging is the algorithm used by the classifiers in the base learning pool. In bagging, each classifier uses the same classification learning algorithm (such as a decision tree), while stacking uses different algorithms to train different classifiers (such as decision trees, random forests, SVMs, and neural networks, etc.). the latter being a model based on homogenous learners. The rationale behind heterogeneous methods is that different models may have different views about the data as they're built on different mathematical paradigms(Narassiguin, 2019).

Here we built a meta learner to combine predictions from multiple models. We tested for correlation in predictions from the base learners since a heterogeneous based ensemble works better if the predictions from the sub-models are uncorrelated or at best weakly correlated. We then built a two layered ensemble with six models. We employed base learners of DTs, RFs, Bagged Trees, Boosted RFs, Logistic regression and PLS-DA produced predictions that suggested that the models are skillful but in different ways, allowing the meta learner to figure out how to get the best from each model for an improved score. This was validated by the low correlations below 0.75. If the predictions for the sub-models were highly correlated (>0.75) then they would be making the same or very similar predictions most of the time reducing the benefit of combining the predictions.

We stacked our models using logistic regression (glm) as a meta learner first and then also used random forests (rf) as a meta learner in the second simulation. A model produced from using glm achieved an accuracy of approximately 98% same as one with rf meta leaner but the later performed better and with greater reliability. Stacking with rf meta learner achieved same prediction accuracy as boosted RFs. Even though the later had the least error, due to their proneness to over fitting, we generally determine that stacked models are better since they combine different heterogeneous learners and can work well in real world situations. An optimum level of bias and variance are always an aim of our study with ensembles.

However, a race to reduce one normally leads to an increase in the other and hence a tradeoff must be reached by a practitioner on which model to deploy. A question of whether it is better to have a model that classifies better for sick patients or one that classifies better for non-sick patients creates a dilemma for medical practitioners and the patient as well. But we believe that with the help on such models whose prediction accuracy is up to 98%, an informed decision is bound to be made.

CONCLUSION

On the current medical environment, computer-based tools to assist in decision making has changed the fabric of health and diagnostic systems. The avenue of applying machine learning tools and ensemble-based decision is fundamental since humans use history, memory and their inherent experience to make judgement that is prone to errors. And as proposed by

(Lancia & Serafini, 2021), we employed algorithms to build models that could first of all tackle the problem of insufficient memory with respect to the size of the data set and secondly, that could run be deployed fast with less computational complexity given the limited and inadequate data acquisition tools and lacking computer machinery in our health institutions.

In this study we attempted to study and design models based on machine learning and predictive analytics to solve a problem of miss-classification in the diagnosis of breast cancer in women. Our major objective was to help reduce errors in a final judgement as to whether a breast tumor is cancerous or non-cancerous. Our models would therefore aim to maximize accuracy with high sensitivity and specificity.

It was still impossible to achieve the accuracy of each decision maker's decision with a nonzero variability and we noted that any classification error encountered by any model was composed of two components that we could control: bias, the accuracy of the classifier; and variance, the precision of the classifier. A low bias and a low variance, although they most often vary in opposite directions, are the two most fundamental features expected for a model. Indeed, to be able to "solve" a problem, we aimed to achieve a model that could attain enough degrees of freedom to resolve the underlying complexity of the breast cancer data, but we also required that it would not have too much degrees of freedom to avoid high variance and be more robust.

We therefore applied ensembling methods where we also noted that averaging through bootstrap aggregating, boosting and stacking the model can have a smoothing (variance-reducing) effect. Model stacking also improves both on bias and variance and produced an acceptable balance between bias and variance for our more robust model that can generalize better for unknown data.

Our best model was a stochastic gradient boosted random forest with random forests with an error of 0.0177. We also noted that stacked ensembles with logistic regression and also with random forests attained same accuracy of 98% as the boosted random forests and sensitivity 0.1 and zero negative predictions.

We however noted that ensembling through stacking reduces the model interpretability and makes it very difficult to draw any crucial insights at the end and also selecting the base learners for a stacked generalization required great expertise and a lot of time for simulation to determine which base learners produce the best generalization.

RECOMMENDATIONS AND FUTURE WORK

Given our findings about ensembles and classification models for breast cancer diagnosis, the tradeoff between variance and bias is still an issue for further study. We recommend advanced methodologies of boosting especially xgboost for further study to improve the diagnosis results. We also propose more complex ensemble for heterogeneous learners with more layer of meta learners. Since the machine learning and automation of the diagnostics is a paramount issue, we would love to carry out further studies with deep learning with Neural networks for an advanced scene in the medical field given the invention of more powerful and yet cheaper technology embedded in latest computers and the use of cloud computing for exhaustive analytics with big data. This is because a shared environment could present enough resources for better simulation and experimentation as well as feedback from stakeholders for a more accurate process and results.

REFERENCES

[1] AcadRadiol, G. .. (2002). *Computer-aided diagnosis in radiology.*

[2] AdnanO.M.Abuassba. (2017). *Improving Classificatino Performance through an Advanced Ensemble Based Heterogeneous Extreme Learning Machines.*

[3] American Cancer Society. (2020). *Cancer Facts and Figures, Atlanta, Ga: American Cancer Society.*

[4] American Cancer Society. (2020). Cancer Facts and Figures. Atlanta, Ga: American Cancer Society. .

[5] American Joint Committee, o. C. (2017). *Breast. In: AJCC Cancer Staging Manual. 8th ed. New York, NY: Springer;. .*

[6] Animesh Hazra, E. (2016). "Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms", International Journal of Computer Applications.

[7] Arno D.B., M. (2016). *Introducing Data Science: Big data, machine learning, and more, using Python tools.*

[8] Arunachalam, A. (2017). Combining Heterogeneous Ensemble Learners Into a Single Meta-Learner in an Amateur Way.

[9] Asri, H. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis.

[10] Ayer., T. (2011). Breast Cancer Risk Estimation with Artificial Neural Networks Revisited: Discrimination and Calibration.

[11] Balogh EP, M. (2015). Improving Diagnosis in Health Care.

[12] Bashir, S. Q. (2015). Heterogeneous classifiers fusion for dynamic breast cancer diagnosis using weighted vote-based ensemble.

[13] Bowles., M. (2015). *Machine Learning in Python Essential Techniques for Predictive Analysis.*

[14] Breit, C. A. (2019). Breast cancer risk assessment in patients who test negative for a hereditary cancer syndrome. The American Journal of Surgery.

[15] Chan, Y.-T. (2020). *An introduction to approaches and modern applications with ensemble learning. .*

[16] Chaurasia, V. (2007). Data mining techniques: To predict and resolve breast cancer survivability. .

[17] Cortes, C. (n.d.). *Support-vector networks. Machine Learning.*

[18] Dietterich., T. (2020). Ensemble Methods in Machine Learning. . *Springer Berlin Heidelberg. Berlin, Heidelberg:.*

[19] Eleanor Black, &. R. (n.d.). Improving early detection of breast cancer in sub-Saharan Africa: why mammography may not be the way forward. *2019*.

[20] Faure, C. A. (n.d.). Empirical and fully Bayesian approaches for the identification of vibration sources from transverse displacement measurements. Mechanical Systems and Signal Processing. . *2017*.

[21] Forysth A., B. R. (n.d.). Machine learning methods to extract documentation of breast cancer symptoms from electronic health records. . *2018*.

[22] Frankenfield., J. (2020). *An Introduction to Machine Learning.*

[23] Fred Nwanganga, M. M. (2020). *Practical machine learning in R.*

[24] H Hasan, &. N. (2019). Feature selection of breast cancer based on principal component Analyis.

[25] Habib Dhahri, E. A. (2019). Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms.

[26] Harris JR, L. M. (2014). *Physical Exam of the Breast. Diseases of the Breast. 5th ed. Wolters Kluwer Health.*

[27] Henry NL, S. P. (2020). *Cancer of the Breast.*

[28] Ian H. Witten, E. F. (2011). *Data Mining Practical Machine Learning Tools and Techniques.*

[29] Jagpreet Chhatwal, O. A. (2010). Optimal Breast Biopsy Decision-Making Based on Mammographic Features and Demographic Factors.

[30] Johannes Uhlig, M. M. (2019). Discriminating malignant and benign clinical T1 renal masses on computed tomography, A pragmatic radiomics and machine learning approach.

[31] Kantardzic, M. (2020). *Data Mining. Concepts, Models, Methods, and Algorithms 3ed.* .

[32] Khairunnahar, L. H. (2019). Classification of malignant and benign tissue with logistic regression.

[33] L.G.Ahmad, A. E. (2015). Using three machine learning techniques for predicting.

[34] Latrach AfefRania, T. T. (2018). Comparison Study for Computer Assisted Detection and Diagnosis 'CAD' systems Dedicated to Prostate Cancer Detection Using MRImp Modalities.

[35] Li. W., V. V. (2017). Extraction of modal parameters for identification of time-varying systems using data-driven stochastic subspace identification. Journal of Vibration and Control.

[36] Liu. N., Q. E. (2019). A novel intelligent classification model for breast cancer diagnosis.

[37] M. I. Jordan, T. M. (2015). *Machine learning: Trends, perspectives, and prospects.*

[38] M., E. (2007). The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process. .

[39] Madeh, P. S.-D. (2020). Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems.

[40] Maldonado, S. P. (2014). Feature selection for support vector machines via mixed integer linear programming. .

[41] Mark Menagie. (2018). *A comparison of machine learning.*

[42] Mohamed Hosni, M. H.-A.-A. (2019). Reviewing Ensemble Classification Methods in Breast Cancer.

[43] Mustafa. M., N. A. (2016). Breast cancer: Detection markers, prognosis, and prevention. *IOSR Journal of Dental and Medical sciences.*

[44] Nasser. F., L. Z. (2016). An automatic approach towards modal parameter estimation for high-rise buildings.

[45] Nina Zumel, J. M. (2020). *Practical Data Science with R.*

[46] Noske, A. A. (2020). Risk stratification in luminal-type breast cancer: Comparison of Ki-67 with EndoPredict test results.

[47] Quinlan, A. A. (1996). Improved Use of Continuous Attributes in C4.5. . *Journal of Artifitial Intelligence Research.* .

[48] R, S. (2000). Generating concise and accurate classification rules for breast cancer diagnosis. Artificial Intelligence Medicine.

[49] Ricvan, D. N. (2018). Diagnostic Accuracy of Different Machine Learning Algorithms for Breast Cancer Risk Calculation: a Meta-Analysis.

[50] Rokach, L. (2010). *Ensemble-based classifiers.*

[51] Runjie Shen, Y. Y. (2015). Intelligent breast cancer prediction model and clinical features: A comparative investigation in machine learning paradigm.

[52] Sabyasachi, D. A. (2019). Big data in healthcare: management, analysis and future prospects.

[53] Salama, G. A. (2012). Breast cancer diagnosis on three different datasets using multi-classifiers.

[54] Sidey-Gibbons, J. A.-G. (2019). *Machine Learning in Medicine: A Practical Introduction.*

[55] Siegel, R. M. (2015). Cancer statistics, 2015. *Ca A Cancer Journal for Clinicians.*

[56] Singh, .. (2019). Determining relevant biomarkers for prediction of breast cancer using anthropometric.

[57] Timothy Masters. (2020). *Modern Data Mining Algorithms in C++ and CUDA C: Recent Developments in Feature Extraction and Selection Algorithms for Data Science.* .

[58] Ting. F., T. Y. (2019). Convolutional neural network improvement for breast cancer classification. Expert Systems with Applications,. .

[59] Toğaçar. M., E. B. (2020). Application of breast cancer diagnosis based on a combination of convolutional neural networks, ridge regression and linear discriminant analysis using invasive breast cancer images processed with autoencoders.

[60] Trieu, P. (2019). Improvement of cancer detection on mammograms via BREAST test sets.

[61] Verboven, P. C. (2005). Improved total least squares estimators for modal analysis. Computer & Structure.

[62] Vrigazagova, B. (2019). Optimization of the ANOVA procedure for support vector machines. . International Journal of Recent Technology and Engineering. .

[63] Wang P., s. Q. (2020). Cross-task extreme learning machine for breast cancer image classification with deep convolutional features.

[64] Wang, H. Z. (2018). A support vector machine-based ensemble algorithm for breast cancer diagnosis.

[65] Wang, S. W. (2019). An improved random forest based rule extraction method for breast cancer diagnosis.

[66] Wu M., Z. X. (2019). Prediction of molecular subtypes of breast cancer using BI-RADS features based on a "white box" machine learning approach in a multi-modal imaging setting. .

[67] Xin Yu Liew, N. (2021). A Review of Computer-Aided Expert Systems for Breast Cancer Diagnosis.

[68] Yaghoubi V., V. M. (2017). Automated Modal Parameter Estimation Using Correlation Analysis and Bootstrap Sampling. Mechanical Systems and Signal Processing.

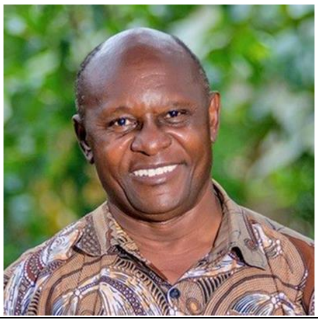
[69] Yan. R., R. F. (2019). Breast cancer histopathological image classification using a hybrid deep neural network.

[70] Zahra. Hematzadeh, R. I. (2015). Comparative studies on breast cancer classifications with k-fold cross validations using machine learning techniques.

[71] Zhang X., Z. Y. (2019). Extracting comprehensive clinical information for breast cancer using deep learning methods.

[72] Zonno G., A. R. (2017). Laboratory evaluation of a fully automatic modal identification algorithm using automatic hierarchical clustering approach. Procedia Engineering. .

AUTHOR'S BIOGRAPHIES

	Mr. AKAMPURURA PAUL A computer Scientist specializing in software engineering and algorithm optimization. An enthusiast in data science and health informatics. Machine learning and deep learning intrigue me.
	PAUL SSEMALUULU (PhD) I am a scholar of information systems with several decades' professional and technical experience. I look forward to contributing my expertise and perspectives to any area in Computer Science, Information Systems, IT work, and teaching.
	ELLY GAMUKAMA (PhD) Research inclination: Modelling and simulation of real world problems with a focus on providing sustainable solutions through the use of ICT systems

****