Efficient Feature Selection and ML Algorithm for Accurate Diagnostics

Abstract - Machine learning algorithms (MLs) can potentially improve disease diagnostic, leading to early detection and treatment of these diseases. Although conventional ML techniques such as classification attain good classification accuracies in medical diagnoses, their performance diminishes when presented with imbalanced dataset more so in detection of minority category. In addition, numerous factors negatively impact on the performance of current classification models when applied to real data, such as class imbalance of the training dataset. Consequently, these models are often biased towards majority class and hence unable to generalize the learning. Ensemble learning which involves the utilization of a group of decision making systems that apply various strategies to combine classifiers may be helpful here to boost prediction on new data. However, current ensemble ML techniques rarely consider comprehensive evaluation metrics to evaluate the performance of individual classifiers. This comprehensive evaluation is necessary so as to deploy ML algorithms that are not only accurate but also efficient in terms of computation costs involved. In this paper, an ensemble machine learning algorithm is developed based on Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes (NB) and K-nearest neighbor (KNN). This algorithm is then executed on Breast cancer data and evaluated using execution time, correctly classified instances (CCI), incorrectly classified instances (ICI), FP rate (FPR), recall(R), precision (P) and F-measure (F-M). The results indicated. Experimental results show that SVM is the best classifier, in which the probability of having best classification is 0.9652% at lowest error rate of 0.0206. On the other hand, NB had the worst performance of 0.8475% classification at 0.0738 error rate.

Keywords: Accuracy, classification, precision, recall, F-measure, NB, RF, SVM, KNN.

1 Introduction

The process of automated prediction of disease is key for better treatment and life saving. As such, many machine learning (ML) based methods have been developed for various diseases. Breast cancer (BC) is a fatal disease that arises from human breast tissue cells and it accounts for 13.7% of deaths in women. As such, early diagnosis of BC is a rich application domain for data mining algorithms. The growing utilization of machine learning algorithms is attributed to the huge surge in digital storage of health records, where ML algorithms help in uncovering patterns existing in these health records. By doing so, interesting insights are gained that assist in diagnosis of various ailments. Authors in [1] explain that data mining models such as artificial neural networks (ANNs), decision tree (DT) analysis, support vector machines (SVMs), Naïve Bayes (NB), and K-Nearest neighbor (KNN) have been deployed for medical diagnosis.

As explained in [2], the development of newer technologies such as analytics, artificial intelligence and machine learning have influenced a number of sectors including health care. Here, these schemes are deployed for improving patient wellness, clinical decision support, and better care coordination. Authors in [3] note

that there is a growing literature on the deployment of machine learning techniques for the development of psychopathology risk algorithms that inform preventive intervention. For instance, supervised machine learning methods can serve as an alternative to conventional techniques for internalizing disorder (ID). Here, these ML algorithms are critical for the optimization of early detection.

World health organization (WHO) reports indicate that many cancer cases are diagnosed too late [4]. However, if accurate diagnosis could be done early, more than 30% of these patients can service the disease. This calls for the design of effective techniques for early detection of diseases so as to improve societal healthcare. The complex nature of actual medical dataset needs careful management due to serious consequences of prediction errors [5].

Machine learning (ML) techniques can effectively extract useful knowledge from large, complex, heterogeneous and hierarchical time series clinical data [6]. As such, many machine learning algorithms have been proposed for deployment in medical diagnosis. As explained in [7], data mining and machine learning techniques present new and powerful solutions for discovering hidden relationships in complex datasets. In most cases, raw datasets available from different medical science sources have useful information which traditional data classification approaches cannot unravel. In addition, although these manual classification schemes may unravel some latent information, they require longer durations and are prone to human mistakes. Consequently, the provision of reliable and trustworthy predictive models with the highest precision and accuracy is the main goal of data mining and machine learning approaches [7]. It is also important for the predictive models to have negligible error rates for effective diagnosis and treatment.

Although machine learning-based techniques have been successful in many areas of medical science, there is need to optimize and improve these methods [8]. Ensemble learning is one such improvement that has enhanced machine learning tasks. Here, a classifier consists of a set of individual classifiers coupled with a mechanism, such as majority voting that combines the predictions of the individual classifiers. Authors in [9] discuss that ensemble classifiers exhibit better performance compared to conventional classifiers. This superiority results from the utilization of a group of decision making systems that apply various strategies to combine classifiers to boost prediction on new data. Authors in [10] concur that ensemble learning can yield more accurate classification results than a single classifier due to incorporation of benefits from both the performance of different classifiers and the diversity of the errors. The contributions of this paper include the following:

- An ensemble classifier leveraging on RF, KNN, BN and SVM is developed to boost breast cancer detection accuracies.
- A comprehensive mathematical modeling of the proposed classifier is carried out to unravel its technical structures and operations.
- Principal component analysis is deployed to reduce the feature space for enhanced classification accuracies.
- Performance evaluations shows that SVM had the best classification accuracies among all the other classifiers.

The rest of this paper is organized as follows: Section 2 discusses related work while section 3 elaborates the system model employed to achieve the paper objectives. On the

other hand, section 4 presents results, discusses them and evaluates the developed protocol. Lastly, section 5 concludes this paper and gives future direction in this research area.

2 Related Work

The field of disease diagnostics has attracted a lot of research efforts from both the industry and academia. This can be attributed to the ease with which diseases such as cancer, diabetes cardiovascular diseases (CVDs), and Rheumatoid arthritis (RA) can be treated if they are detected early. According to [1], there is need to identify the causes of such diseases and be able to diagnosis them early enough. Artificial intelligence based algorithms have been deployed for this early diagnosis for a number of diseases. For instance, authors in [11] have applied KNN, ANN, radial basis function (RBF) neural network (RBFNN), and SVM techniques for BC data classification. In addition, Genetic Algorithm (GA) and Random Forest (RF) algorithms have been deployed for BC detection in [12].

A data mining method for accurate breast cancer (BC) prediction has been developed in [13], by combining SVMs and ANNs for BC data analysis. The results showed that this approach improved the performance of the conventional machine learning algorithms, attaining an accuracy of 100%. A probabilistic neural network (PNN), convolutional neural network (CNN), multilayer perceptron neural network (MLPNN), recurrent neural network (RNN) and SVM have been utilized in [14] for BC prediction. The results showed SVM achieved the best prediction accuracy of 99.54%. On the other hand, Association Rules (AR) and neural network (NN) techniques have been applied in [15] for BC detection, attaining a classification accuracy of 97.40%. Separately, NB technique in combination with a weighting approach have been deployed in [16], yielding a BC prediction accuracy of 98.54%.

On the other hand, ANFIS technique coupled with GA have been applied in [17] for BC prediction, reporting a prediction accuracy of 71%. Authors in [18] have evaluated SVM, DT (C4.5), NB and KNN algorithms for BC classification, with SVM reporting the highest accuracy of 97.13% among other classifiers. Similarly, authors in [19] have applied multilayer perceptron (MLP), RF, Random RT and Ensemble Classifier (EC) for BC detection, with EC yielding the highest accuracy of 83.50% among all classifiers. On the other hand, the scheme proposed in [20] exhibited an accuracy of 99.68%.

Authors in [21] have used lazy association classification algorithm on heart disease data set and recorded 10.26% improvement over J4.8 and 8.6% improvement against NB classification algorithm. On the other hand, a hybrid model of neural network tools and genetic algorithms for prediction of heart disease have been presented in [22], yielding trained data accuracy of 96% and validation accuracy of 89%. Similarly, NB algorithm have been deployed in [23] for heart disease prediction, obtaining an accuracy of 86.29%. On the other hand, authors in [24] have employed AdaBoost and feature subset selection method principal component analysis (PCA) for heart disease data analysis, which improved prediction rates by 2.11% over classification accuracy of J4.8 and 7.33% over 10 cross validations. Authors in [25] have employed decision tree algorithm C4.5 for heart disease prediction, yielding the highest accuracy of 75% while a combination of KNN and NB classifier have been presented in [26] for heart disease prediction, achieving an accuracy of 82.6% on

heart disease data set. The scheme developed by in [27] was shown to improve prediction accuracy by more than 4% compared with other schemes.

3 System Model

In this section, the mathematical basis for the deployed machine learning algorithms is provided. This is followed by data set description, data pre-processing, PCA, and experimentations as explained in the sib-sections that follow.

3.1 Mathematical Modeling of ML Algorithms

In this sub-section, the mathematical formulations for K-nearest neighbor, Naïve Bayes, Random Forest and support vector machine are presented.

3.1.1. K-Nearest Neighbours

Taking α_i as an M-dimensional training vector and β_i as the consequent class label, then the training set is formualted as in (1):

$$\{(\alpha_i, \beta_i)\}_{i-1}^{N} \in G \quad (1)$$

Suppose that α' is a particular query from some test set (α', β'). Based on this, the unknown class label β' is derived as shown in steps 2 to 5.

Step 1: Calculate Euclidean distance $\mathbb{Z}$ between α' and each training set (α_i, β_i):

$$\mathbb{Z}(\alpha', \alpha_i) = ||\alpha' - \alpha_i||_{\Re_2} \quad (2)$$

Equation (2) can also be expressed as follows: suppose that ω is the number of training samples , and Ψ is the number of feature vectors. Then for a particular test feature set $(\beta_1, \beta_2, \beta_3 \ldots . \beta_n)$ and training feature set $(\alpha_1, \alpha_2, \alpha_3 \ldots . \alpha_n)$, $\mathbb{Z}_j$ is derived as in (3):

$$\mathbb{Z}_j = \sqrt{\sum_{i=1}^{\Psi}(Test_i - Train_j)} \quad (3)$$

Step 2: Organize the obtained Euclidean distance $\mathbb{Z}$s in ascending order

Step 3: Designate some weight γ_i to ith nearest neighbour as in (4):

$$\gamma_i = \frac{1}{\mathbb{Z}(\alpha', \alpha_i)^2} \quad (4)$$

Step 4: For equally-weighted KNN rules, designate $\gamma_i = 1$

Step 5: Suppose $\mathcal{F}(.)$ is the Dirac-delta function, η is the class label, and β' is the class label for ith nearest neighbour among its K-nearest neighbours. Then depending on the majority vote of its nearest neighbours, the class label for α' is assigned as in (5):

$$\beta' = arg\,{}_{\eta}^{max} \sum_{(\alpha_i, \beta_i) \in G'} \gamma_i \mathcal{F}(\eta = \beta') \quad (5)$$

Here, $\mathcal{F}(.)$ assumes the value of unity (1) when its argument is true and zero otherwise.

3.1.2 Support Vector Machine

This classifier takes in an input feature vector and establishes the class to which this vecor belongs to. Suppose that $\alpha_i, i = 1,2,3, \ldots N$ are the feature vectors for training set Ť. Here, Ť may belong to either $\ddot{Y}_1$ or $\ddot{Y}_2$. Based on this training data, the hyperplane is mathematicaly represented as in (6):

$$\mathbb{H}(\alpha)= \gamma^{d}\alpha_i+ \mathcal{L} =0 \quad (6)$$

Where γ=[γ_1, γ_2… γ_q] represents the weight vector and $\mathcal{L}$ is the bias. Here, the binary classification degenerates into the solution of decision function in (7):

$$\wp(\alpha)=\text{sign}\ (\gamma^{d}\alpha_i+ \mathcal{L}) \quad (7)$$

Due to the possibility of many hyperplanes that seprate the feature vectors, the role of SVM is to find the one with the largest margin. For non-linearly separable feature vectors, the input space is mapped into high dimensional feature space using kernel functions that transforms it into linear separable. In essense, kernel functions serve to transform feature vectors from finite to infinite dimensional space. As such, the performance of SVM is influenced immensely by the underlying kernel function. The five most prominent kernel functions include linear, Mahalanobis, radial basis function(RBF), polynomial and sigmoid (also known as hyperbolic tanget or multi-layer perception kernel) whose mathematical formulation are derived in (8) to (12).

In these formulations, $\mathcal{M}(>0)$ is the scaling factor, D is the dimension of the data set, V is the covariance matrix, and $\mathcal{L}$ denotes polynomial kernel degree, which is adjustable just like parameters $\wp$ and $\mathcal{E}$ based on the underlying data.

$$\mathrm{K}(\alpha_i, \alpha_j) = (1+\alpha_i^{d}\alpha_j)\ , \text{(linear)} \quad (8)$$

$$\mathrm{K}(\alpha_i, \alpha_j) = (\wp\ \alpha_i^{d}\alpha_j + 1)^{\mathcal{L}}\ , \wp > 0 \text{ (Polynomial)} \quad (9)$$

$$\mathrm{K}(\alpha_i, \alpha_j) = e^{(-\wp||\alpha_i-\alpha_j||^2)}\ , \wp > 0 \text{ (RBF or Gaussian)} \quad (10)$$

$$\mathrm{K}(\alpha_i, \alpha_j) = \tanh(\wp\ \alpha_i^{d}\alpha_j + \mathcal{E}) \quad \text{(Sigmoid)} \quad (11)$$

$$\mathrm{K}(\alpha_i, \alpha_j) = -\frac{\mathcal{M}}{D}\left(\alpha_i - \alpha_j\right)^{d} V^{-1}(\alpha_i - \alpha_j) \quad \text{(Mahalanobis)} \quad (12)$$

In equation (10), $\mathcal{M}$ serves to control the Mahalanobis distance.

Consdering a set of q data samples that belong to two classes (α_1, β_1), (α_2, β_2)…., $\left(\alpha_q, \beta_q\right)$ that are mapped to a higher dimensional space, where $\beta_i \in \{-1,1\}$. For the correct classification process, the separating hyperplane should be optimized. Taking γ as some weight vector and $\mathcal{L}$ as bias weight, the optimization problem in SVM degenerates to the determination of the hyperplane that segrates positive and negative classes given in (14) and (15):

$$\mathbb{H}(\alpha)= \gamma\alpha_i+ \mathcal{L} =0 \quad (13)$$

$$(\gamma\alpha_i+ \mathcal{L}) \geq 1, \text{ for } \beta_i=1 \quad (14)$$

$$(\gamma\alpha_i+ \mathcal{L}) \leq -1, \text{ for } \beta_i= -1 \quad (15)$$

To accomplish this, the margin between the two classes is maximized by determining γ and $\mathcal{L}$ that maximizes (16):

$$\frac{1}{2}||\gamma||^2 \quad (16)$$

In essence, an optimal hyperplane denotes an error-free plane with largest possible separation margin. Ideally, this is the hyperplane that minimizes the cost function in (17):

$$\mathbb{C}(\gamma) = \frac{1}{2}\gamma^{d}.\gamma \quad (17)$$

The optimization in (17) is subject to some constant in (18):

$$\beta_i(\gamma^{d}.\alpha_i + \mathcal{L}) \geq 1, i = 1{:}\,q \quad (18)$$

Due to the convex nature of $\mathbb{C}(\gamma)$, Lagrange multipliers (ℓ_1, ℓ_2…. ℓ_q) are employed to reduce this constrained optimization problem. This is achieved through the process

of weighing each data point based on its criticality in the determination of the segregating information of the two classes. Mathematically, this is derived as in (19):

$$\max \mathrm{L}(\gamma, \mathcal{L}, \ell) = \sum_{l=1}^{q} \ell_i - \frac{1}{2}\sum_{i=1}^{q}\sum_{j=1}^{q} \ell_i \ell_j \beta_i \beta_j (\alpha_i . \alpha_j) \quad (19)$$

The optimization in (19) is subject to the conditions in (20):

$$\ell_i \geq 0 \ \& \ \sum_{i=1}^{q} \ell_i \beta_i = 0 \quad (20)$$

Incorporating Lagrange multipliers to the decision function in (7) results in (21):

$$\wp(\alpha) = \text{sign}\ (\sum_{i=1}^{q} \ell_i \beta_i (\alpha_i . \alpha) + \mathcal{L}) \quad (21)$$

Taking $\mathbb{C}(\alpha)$ as the transformation function that maps lower dimension feature vectors to higher dimensional feature space, then the Kernel function in (22) is deployed for these transformations:

$$\mathrm{K}(\alpha, \beta) = \mathbb{C}(\alpha)\mathbb{C}(\beta) \quad (22)$$

Based on (22), the decision function is modified as in (23):

$$\wp(\alpha) = \text{sign}\ (\sum_{i=1}^{q} \ell_i \beta_i \mathrm{K}(\alpha_i, \alpha) + \mathcal{L}) \quad (23)$$

3.1.3 Random Forest

This classifier comprises of classification tree $T(J, K_i), i = 1,2..q$. Here, K_i represents a vector that is identically and indepently disributed (IID) with each tree vote at its input J. Basically, a random forest comines several decision trees so as to minimize over-fitting. Suppose that $T_1(S), T_2(S), \ldots T_q(S)$ is an ensemble classifier with arbitrary training data gotten from vector S and Q (the prediction class), f is the indicator function, Ă is the mean, the margin function is formulated as in (24):

$$\text{mg (S,Q)} = \check{\mathrm{A}} f(T_1(S) = Q) - \max_{R \neq Q} f(T_1(S) = R) \quad (24)$$

In (24), $T_i(S) = Q$ denote classification result while $T_i(S) = R$ is classification result with R. In RF, the margin is utilized to establish the mean value of votes S and Q, such that the greater the margin, the more accurate is the classification. Here, the generalization error $\hat{\mathrm{G}}$ is derived as in (25):

$$\hat{\mathrm{G}} = \mathrm{W}_{\mathrm{S,Q}}(\text{mg(S,Q)} < 0) \quad (25)$$

In (25), $\mathrm{W}_{\mathrm{S,Q}}$ signifies that the probability is more than S,Q dimension. Considering training sample $\mathfrak{T}_{\mathrm{p}} = \{(\alpha_1, \beta_1), \ldots, (\alpha_p, \beta_p)\}$ of IID $[0,1]^l$. Using $\mathfrak{T}_{\mathrm{p}}$, the objective is to estimate the regression function $\mathrm{R_F}(\alpha) = \mathbb{E}[\beta[\alpha = \mathfrak{g}]]$for some fixed $\mathfrak{g} \in [0,1]^l$. Generally, RF classifier consists of a set of stochastic regression tree $\{\mathrm{R_T}(\mathfrak{g}, \mathrm{h_q}, \mathfrak{T}_{\mathrm{p}}), q \geq 1\}$. Here, $\mathrm{h_1}$, $\mathrm{h_2}$,…. denote IID outputs of a randomization construct h. By combining these random trees ($\mathrm{R_T}$s), an amalgamated regression estimate is obtained as in (26):

$$\overline{R_T}(\alpha, \mathfrak{T}_{\mathrm{p}}) = \mathbb{E}_h[\mathrm{R_T}(\mathfrak{g}, \mathrm{h}, \mathfrak{T}_{\mathrm{p}}) \quad (26)$$

In (26), $\mathbb{E}_h$ is conditionally associated with random constructs on $\mathfrak{g}$ and $\mathfrak{T}_{\mathrm{p}}$. Here, the dependency of sample estimates is denoted as $\overline{R}_T(\mathfrak{g})$ and h is utilized to establish how successive divisions are executed when building individual trees.

3.1.4 Naïve Bayes

In this algorithm, the probability that attribute $\mathfrak{g}$ takes on particular G when the class is C' is modeled using a real number between 0 and 1. On the other hand, continous attributes are modeled using continous probability distribution over a range of attribute's values. Suppose that $\mathrm{R_V}$ is a random variable representing instance class, and $\mathrm{R_A}$ is a random variable vector representing observed attribute values. Denoting

r_v as a specific class label and r_a as the specific observed attribute value, then if Ɍ is a test case that is to be classified, the probability of each class given the vector of oberved values for the predicitive features is obtained using Bayes' theorem in (27):

$$p(R_V= r_v|R_A= r_a) = \frac{p(RV= rv)p(RA= ra|RV= rv)}{p(RA= ra)} \quad (27)$$

Since an event consists of a juxtaposition of feature values assignments, then using the features conditional independence postulation, equation (27) is written as in (28):

$$p(R_A= r_a |R_V= r_v) = \prod_i p(R_{A_i} = r_{a_i}|RV = rv) \quad (28)$$

Suppose that $\mathcal{P}$ is the training set and $\bar{U}$ is the related class labels. Here, each tuple is denoted by $\ddot{E}$ features, implying that each tuple consists of $\ddot{E}$ values. If there are k class labels $\bar{U}_1$, $\bar{U}_2$,… $\bar{U}k$ for any new tuple Z, the classifier predicts that Z is a member of the class with highest probability state on Z. suppose now that this classifier is presented with a new test set Z that needs to be classified as either benign or malignant. Here, Z can be classified into its respective class $\bar{U}_i$ or $\bar{U}_j$ provided it satisfies the state in (29):

$$P\left(\frac{\bar{U}_i}{Z}\right) > P\left(\frac{\bar{U}_j}{Z}\right) \text{ for } 1 \leq j \leq k \quad (29)$$

In this case, $\bar{U}_i$ becomes the maximum posterior hypothesis since its $\left(\frac{\bar{U}_i}{Z}\right)$ is being maximized. Based on Bayes's theorem:

$$P(\bar{U}_i|Z) = \frac{P(Z|\bar{U}_i)P(\bar{U}_i)}{P(Z)} \quad (30)$$

Since $P(Z)$is unvarying for all the classes, only the values for $P(Z|\bar{U}_i)P(\bar{U}_i)$ needs t be increased. In this case, the formualions reduce to:

$$P\left(\frac{\bar{U}_i}{Z}\right) = P\left(\frac{Z}{\bar{U}_i}\right) * P(\bar{U}_i) \quad (31)$$

During prediction of Z's class label , $P\left(\frac{Z}{\bar{U}_i}\right) * P(\bar{U}_i)$ is evaluated for each class $\bar{U}_i$. In essence, the predictor class label $\bar{U}_i$ for which $P\left(\frac{Z}{\bar{U}_i}\right) * P(\bar{U}_i)$is maximum.

On condition that priori probabilities for class $P(\bar{U}_j)$ is unknown, the assumption made is that the classes are all equally likely and $P(Z|\bar{U}_j)$ needs to be maximized.

During class label or class value Z classification, $P(Z|\bar{U}_i)P(\bar{U}_i)$ is evaluated for both benign and malignant instances in $\bar{U}_i$. In this case, NB classifies Z to class $\bar{U}_i$ on condition that it is the class that maximizes $P(Z|\bar{U}_i)P(\bar{U}_i)$.

3.2 Data Set Description

In this paper, the data set from Wisconsin Diagnostic Breast Cancer (WDBC) repository is deployed. This data set comprises of 699 cases with 11 attributes for each data sample as shown in Table 1.

Table 1: Attribute Space

Attribute	Range
Normal Nucleoli	1-10
Sample code number	1-10
Mitoses	1-10
Clump Thickness	1-10

Bland Chromatin	1-10
Uniformity of Cell Size	1-10
Bare Nuclei	1-10
Uniformity of Cell Shape	1-10
Single Epithelial Cell Size	1-10
Marginal Adhesion	1-10
Class	B or M

The class attribute, which is part of the 11 attributes, has only two values: benign (B) or malignant (M). As such, each instance is either benign or malignant.

3.3 Data Pre-processing

Before the classification process, the data was cleansed and relevance analysis excuted to eliminate redundant attributes from further analysis. Thereafter, data transformation is executed to map the attribute values to a small-scale range of 1 or 0, before the application of PCA for dimensionality reduction. Here, data cleaning involves the removal or reduction of noise and handdling of missing values. The WDBC data set has 16 instances with single missing attribute value. These missing values were replaced by the mean of the particular attribute. In addition, the attribute 'Sample code number' in Table 1 is irrelevant and hence is eliminated. On the other hand, statistical correlations are computed and utilized to eliminate redundant attributes.

3.4 Principal Component Analysis

Based on the deployed data, its input attributes are huge and this may impede classification speed and accuracy. As such, the principal component analyisis (PCA) is utilized for feature selection as one way of dimensionality reduction in the input features. The selection of PCA was informed by the fact that it is a simple and yet widely deployed dimensionality redction technique for the two-class classification problems. In essence, PCA serves to establish peak disparity in the underlying data set. In so doing, the many features in the dataset are reduced to less but crucial features. By applying it to both training and testing samples, patterns in the input dataset are detected based on resemblance and variances among the present attributes.

Suppose that M is the dimension of the data set that has q samples $\{N_i\}_{I=1}^{q}$, in which $N_i \in R^M$. Here, PCA attempts to determine the principle orthogonal directions in which this data set has the highest variances. Provided that majority of these variances occur in one or numerous main directions, these directions form the principal component directions of the data set. These directions are a better representation of the data set with less dimensions. Taking $\tilde{\mathrm{V}}$ as the mean vector of the data samples, the covariance matrix Ω of the sample set is computed as in (32):

$$\Omega=\mathrm{E}[(N_i-\tilde{\mathrm{V}})(N_i-\tilde{\mathrm{V}})^{\mathrm{T}})] \quad (32)$$

Using the eigenvectors of Ω as the basis to span a new coordinate system, the orthogonal coordinate system can be obtained that can eliminate correlations between diverse components of the samples in their initial space. Essentially, the level of Ω's

eigenvalues depict the variance of the samples along the coordinates of the consequent eigenvalues.

Suppose that we have an H X G matrix denoted by $\mathbb{Q}$, in which each row refers to one of *H* trials while each column denotes one of *G* features. We also let $\mathscr{B}_{\mathbb{Q}}$ represent the average of the input, in which case the Eigen values (λ_i) and Eigen vectors (μ_i) of the input correlation matrix are derived as in (33):

$$\sum = \bar{E}^T \bar{E} \tag{33}$$

In which $\bar{E} = \mathbb{Q} - \mathscr{B}_{\mathbb{Q}}$

Taking $\bar{Q}$ as the right singular vector, the principal components are expressed as in (34):

$$P = \bar{E}.\bar{Q}^T \tag{34}$$

Suppose that (λ_1, λ_2,… λ_q) are the eigenvalues of matrix **Ω**, they can be ordered based on their size as: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_q$. Denoting the corresponding eigenvectors as (μ_1, μ_2,…, μ_q), if the first λs are very large compared with the rest, only μs corresponding to these λs are utilized to represent the data set without significant loss to the information. The deployed μs are the principal components axes of the data set while the spanned sub-space by these μs forms the principal component space (PCS). When the first *n* μs are deployed to build the PCS, the resulting representation error ot truncation error *e* is derived as in (35):

$$e = \frac{\sum_{i=n+1}^{M} \lambda_i}{\sum_{i=1}^{M} \lambda_i} \tag{35}$$

This PCA depiction has the minimum error among all the feasible orthogonal n-dimensional represenattion of the sample set.

3.5 Experimentations

Upon data pre-processing, the four machine learning algorithms which included KNN, SVM, RF and NB are applied to the obtained data. To accomplish this, Waikato to Environment for Knowledge Analysis (WEKA) software was utilized. This choice was informed by its ability to implement and facilitate analysis of numerous classification, regression and data mining algorithms. Fig.1 gives the general data flow diagram for the machine leaning algorithm (MLA) classification process.

As shown in Fig.1, the breast cancer classification comprised of a number of steps, starting with the feeding of the WDBC Data set to the MLA upon which data processing was executed. This is followed by training and testing the classifiers. The 10-fold cross validation test is utilized to evaluate the developed predictive models. This technique simply partitions the data set into training and test samples. Here, the training data sample is used to build the model while the test sample evaluates the constructed model.

Here, the classification involved the correct placement of an instance into either the B or M class. The last set of experimentations involved the appraisal of the performance of individual classifiers using the performance metrics in Table 2. Here, TP is the true positive, TN is true negative, FP is false positive, and FN is false negative.

Accuracy represented the overall correctness of the model while precision depicts the ratio of positive cases that were predicted appropriately. On the other hand, the FP-rate is the ratio of negative cases that were incorrectly classified as positive cases. Recall or TP rate represents the ratio of correctly identified positive cases while F-measure is the harmonic mean of precision and recall.

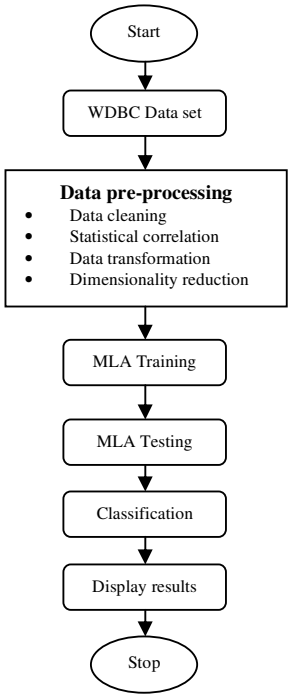

Fig.1 : MLA Classification Process

In terms of error performance, Mean Absolute Error (MAE), Kappa, Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), and Root Relative Squared Error (RRSE) are deployed. Table 3 gives the formulations of these errors.

Table 2: Performance Metrics

Metric	**Formulation**
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
FP-rate	$\frac{FP}{FP+TN}$
Precision	$\frac{TP}{TP+FP}$
Recall / TP-rate	$\frac{TP}{TP+FN}$
F-measure	$\frac{2*Precision*Recall}{Precision+Recall}$

In Table 3, y_i is the predicted value, y_{ij} is the predicted value by individual model i for tuple j out of n tuples, T_j is the target value for tuple j, x_i is the actual value, while n is the number of data points.

Table 3: Error Analysis

Error	Formulation
MAE	$\frac{\sum_{i=1}^{n}\lvert y_i - x_i\rvert}{n}$
Kappa	$\frac{2*((TP*TN)-(FN*FP))}{(TP+FP)*(FP+TN)+(TP+FN)*(FN+TN)}$
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}$
RAE	$\frac{[\sum_{i=1}^{n}(y_i - x_i)^2]^{1/2}}{\left[\sum_{i=1}^{n} x_i{}^2\right]^{1/2}}$
RRSE	$\sqrt{\frac{\sum_{j=1}^{n}(y_{ij} - T_j)^2}{\sum_{j=1}^{n} T_j - \underline{T})^2}}$, where $\underline{T} = \frac{1}{n}\sum_{j=1}^{n} T_j$

4 Results and Discussion

In this section, the developed classifiers are evaluated in terms of their build time, correctly classified instances (CCI), incorrectly classified instances (ICI), FP rate (FPR), recall(R), precision(P) and F-measure(F-M) as shown in Table 4.

Table 4: Performance Comparisons

Classifier	Build time(s)	CCI	ICI	A	FPR	R	P	F-M
RF	0.29	541	28	95.1	0.067	0.951	0.949	0.950
SVM	0.07	561	8	98.6	0.023	0.986	0.982	0.984
NB	0.02	525	44	92.3	0.090	0.923	0.921	0.922
KNN	0.01	5 49	20	96.5	0.042	0.965	0.962	0.963

Based on the values in Table 4, RF takes the longest duration of 0.29 seconds to build the model while KNN took the shortest duration of 0.01 seconds. The explanation for this is that KNN is a lazy learner and hence it does not execute many operations during training. This is unlike other MLAs which need to build models during the training process. In terms of accuracy, SVM had the highest value of 98.6% while NB had the lowest value of 92.3%. This directly follows from SVM's highest values for CCI and lowest value for ICI compared to other classifiers. Table 5 presents the error performance for the various classifiers.

Table 5: Error Performance

Error	Classifiers

	NB	RF	KNN	SVM
MAE	0.0738	0.0749	0.0407	0.0206
Kappa	0.8475	0.9026	0.9238	0.9652
RMSE	0.2637	0.1748	0.1959	0.1462
RAE	15.7832	16.1648	8.6572	4.5142
RRSE	54.7826	35.7492	40.5630	30.0195

It is clear from Table 5 that in SVM, the probability of having best classification is 0.9652% at lowest error rate of 0.0206. On the other hand, NB had the worst performance of 0.8475% classification at 0.0738 error rate. It is evident that both NB and RF have highest error rates which can be accounted by their high ICI. The confusion matrix in Table 6 presents the comparisons of actual class results with the expected results.

Table 6: Confusion Matrix

	M	**B**	
SVM	203	10	M
	2	354	B
NB	189	22	**M**
	23	335	**B**
KNN	203	10	**M**
	12	344	**B**
RF	195	11	**M**
	9	354	**B**

Based on the confusion matrix of Table 6, SVM properly predicts 569 instances out of 699 instances. Out of these 569 correct predictions, 356 are B instances that are actually so, and 213 M instances that are actually so. On the other hand, 12 instances are incorrectly predicted, in which 10 B instances are predicted as M while 2 M instances are predicted as B instances. This explains the high accuracy values for SVM compared with other classifiers. On the other hand, NB had the highest number of incorrectly classified instances of 45(22 B instances incorrectly classified as M, and 23 M instances incorrectly classified as B). This explains why NB has the lowest accuracy of 92.3%.

5 Conclusion and Future Work

Breast cancer is the most common disease among women whose early detection can potentially save lives. However, designing a ML model for the detection of this disease presents some challenges due to its heterogeneous nature. In addition, performance evaluation of breast cancer ML models has been noted to be cumbersome. In this paper, an ensemble MLA is developed based on RF, SVM, NB and KNN. Here, breast cancer classification starts by feeding of the WDBC data set to the MLA upon which data processing is executed. This is followed by training and testing the classifiers, after which 10-fold cross validation test is utilized to evaluate

the developed predictive models. Experimental results show that SVM is the best classifier, while NB is the worst classifier. Based on classification accuracies, SVM was closely followed by KNN, RF and NB in that order. Future work in this research domain will involve building an ensemble classifier encompassing other machine learning algorithms that were not within the scope of the current work. There is also need to evaluate the developed ensemble classifier in other data sets to offer a more comprehensive overview of its performance.

References

[1] Liu, N., Li, X., Qi, E., Xu, M., Li, L., & Gao, B. (2020). A novel Ensemble Learning Paradigm for Medical Diagnosis with Imbalanced Data. IEEE Access, *8*, 171263-171280.

[2] Yekkala, I.; Dixit, S.; Jabbar, M.A. (2017). Prediction of heart disease using ensemble learning and Particle Swarm Optimization. In Proceedings of the 2017 International Conference on Smart Technologies for Smart Nation (SmartTechCon), Bengaluru, India, pp. 691–698.

[3] Rosellini, A. J., Liu, S., Anderson, G. N., Sbi, S., Tung, E. S., & Knyazhanskaya, E. (2020). Developing algorithms to predict adult onset internalizing disorders: An ensemble learning approach. Journal of psychiatric research, 121, 189-196.

[4] Jiang, J., Li, X., Zhao, C., Guan, Y., and Yu, Q. (2017). Learning and inference in knowledge-based probabilistic model for medical diagnosis. Knowl.Based Syst., vol. 138, pp. 58-68.

[5] Baccouche, A., Garcia-Zapirain, B., Castillo Olea, C., & Elmaghraby, A. (2020). Ensemble Deep Learning Models for Heart Disease Classification: A Case Study from Mexico. *Information*, *11*(4), 207.

[6] M. Eshtay, H. Faris, and N. Obeid (2018). Improving extreme learning machine by competitive swarm optimization and its application for medical diagnosis problems. Expert Syst. Appl., vol. 104, pp. 134-152.

[7] A.H. Alkeshuosh, M.Z. Moghadam, I. Al Mansoori, M. Abdar (2017). Using PSO Algorithm for Producing Best Rules in Diagnosis of Heart Disease, in: Computer and Applications (ICCA), 2017 International Conference on, IEEE. pp. 306–311.

[8] Sevakula RK, Verma NK. (2017). Assessing generalization ability of majority vote point classifiers. IEEE Transactions on Neural Networks and Learning Systems. 28(12):2985–97.

[9] Li, H., Cui, Y., Liu, Y., Li, W., Shi, Y., Fang, C., & Lu, Y. (2018). Ensemble learning for overall power conversion efficiency of the all-organic dye-sensitized solar cells. *IEEE Access*, *6*, 34118-34126.

[10] X. Zhang and S. Mahadevan (2019). Ensemble machine learning models for aviation incident risk prediction. Decis. Support Syst., vol. 116, pp. 48-63.

[11] A. Mert, N. Kılıç, E. Bilgili, A. Akan (2015). Breast cancer detection with reduced feature set, Comput. Math. Methods Med. 1–11.

[12] E. Alic˘kovic´, A. Subasi, (2017). Breast cancer diagnosis using GA feature selection and Rotation Forest, Neural Comput. Appl. 28 (4), 753–763.

[13] Abdar, M., & Makarenkov, V. (2019). CWV-BANN-SVM ensemble learning classifier for an accurate diagnosis of breast cancer. Measurement, 146, 557-570.

[14] E.D. Übeyli (2007). Implementing automated diagnostic systems for breast cancer detection, Expert Syst. Appl. 33 (4), 1054–1062.

[15] M. Karabatak, M.C. Ince (2009). An expert system for detection of breast cancer based on association rules and neural network, Expert Syst. Appl. 36 (2), 3465–3469.

[16] M. Karabatak (2015). A new classifier for breast cancer detection based on Naïve Bayesian, Measurement. 72, 32–36.

[17] H. Turabieh, M. Muhanna (2016). GA-based feature selection with ANFIS approach to breast cancer recurrence, Int. Journal of Comput. Sci. Issues (IJCSI) 13 (1), 36.

[18] H. Asri, H. Mousannif, H. Al Moatassime, T. Noel (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Comput. Sci. 83,1064–1069.

[19] D. Kaushik, K. Kaur (2016). Application of Data Mining for high accuracy prediction of breast tissue biopsy results, in: 2016 Third International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC), IEEE, pp. 40–45.

[20] A.M. Abdel-Zaher, A.M. Eldeib (2016). Breast cancer classification using deep belief networks, Expert Syst. Appl. 46, 139–144.

[21] Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2013). Heart disease prediction using lazy associative classification. In 2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s) (pp. 40-46). IEEE.

[22] Amin, S. U., Agarwal, K., & Beg, R. (2013). Genetic neural network based data mining in prediction of heart disease using risk factors. In 2013 IEEE Conference on Information & Communication Technologies (pp. 1227-1231). IEEE.

[23] Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2015). Computational intelligence technique for early diagnosis of heart disease. In 2015 IEEE International Conference on Engineering and Technology (ICETECH) (pp. 1-6). IEEE.

[24] Jabbar, M. A., Deekshatulu, B. L., & Chndra, P. (2014). Alternating decision trees for early diagnosis of heart disease. In International Conference on Circuits, Communication, Control and Computing (pp. 322-328). IEEE.

[25] Karaolis, M., Moutiris, J. A., & Pattichis, C. S. (2008). Assessment of the risk of coronary heart event based on data mining. In 2008 8th IEEE International Conference on BioInformatics and BioEngineering (pp. 1-5). IEEE.

[26] Ferdousy, E. Z., Islam, M. M., & Matin, M. A. (2013). Combination of naive bayes classifier and K-Nearest Neighbor (cNK) in the classification based predictive models. Computer and information science, 6(3), 48-56.

[27] An, N., Ding, H., Yang, J., Au, R., & Ang, T. F. (2020). Deep ensemble learning for Alzheimer's disease classification. Journal of biomedical informatics, 105, 103411.