MEDWIN PUBLISHERS
Committed to Create Value for researchers

Bioinformatics & Proteomics Open Access Journal

Application of Machine Learning Algorithms in Bioinformatics

Siraj-Ud-Doulah*

Department of Statistics, Begum Rokeya University, Bangladesh

***Corresponding author:** Md Siraj-Ud-Doulah, Department of Statistics, Begum Rokeya University, Rangpur, Bangladesh, Email: sdoulah_brur@yahoo.com

Research Article

Volume 3 Issue 1

Received Date: February 25, 2019

Published Date: March 16, 2019

Abstract

Bioinformatics is a developing domain that discourses the necessity to achieve and understand the data that in the past era was enormously produced by genomic investigation. This domain signifies the convergence of genomics, biotechnology and involves examination and understanding of data, displaying of biological occurrences, which is one of the most significant field for the study of biological schemes today and expansion of algorithms and statistics. The paper objectives at a relative study of Machine learning algorithms on a breast cancer dataset. The algorithms used for comparison of ANN, ANFIS, ADT, BN, BFT, C4.5, CART, CBR, DT, DESVM, EP, GP, GA, GRU-SVM, IBk, J48, kNN, K*, KMSVM, kNN+ICA, kNNGA, LB, LMT, LR, MDR, MLP, NNge, NB+kNN, NNANF, NB, PART, RBFNN, RF, SVM, SMO, SL, SOM, SR, SMO-SVM, ZeroR– are supervised learning algorithms used extensively for classification purposes and are chosen for their variety. The performance of those algorithms are compared by using the different performance criterions based on confusion matrix such as Recall /Sensitivity, Specificity, Precision, Accuracy, F-Score and Mathews Correlation Coefficient. Based on analysis of this data, Artificial Neural Networks (ANN) is better at classification with 99.5% accuracy and 0.9901 Mathews correlation coefficient among others classifiers.

Keywords: Bioinformatics; Machine Learning; Classification; Performance Evaluation; Breast Cancer Data

Abbreviations: ANN: Artificial Neural Networks; ANFIS: Adaptive Neuro-Fuzzy Inference System; ADT: Alternating Decision Tree; BN: Bayes Net; BFT: Best First Tree; CART: Classification and Regression Trees; CBR: Case Based Reasoning; DT: Decision Tree; EP: Emerging Pattern; GP: Genetic programming; LMT: Logistic Model Tree; LR: Linear Regression; MDR: Multifactor Dimensionality Reduction; NB: Naive Bayesian; RBFNN: Radial Basis Function Neural Network; DESVM: Differential Evolution Support Vector Machine; RF: Random Forest; SMO: Sequential Minimal Optimization; SL: Simple Logistics; SOM: Self Organizing Maps; SR: Softmax Regression.

Introduction

Bioinformatics, or computational biology, is a versatile domain of study for understanding biological data using computer science. The significance of this new arena of investigation will develop as we endure to produce and incorporate huge amounts of genomic, proteomic and other data. A fascinating field of study in bioinformatics is the application and progress of machine learning algorithms to explain biological difficulties. Numerous labors are being made by the computer scientists and statisticians to policy and gadget algorithms and methods for well-organized storage, management, treating and investigation of biological databases. In this study we analysis the application of some of the frequently used and appropriate machine learning techniques in the arena of bioinformatics. We have used forty different machine learning techniques such as Artificial Neural Networks (ANN), Adaptive Neuro- Fuzzy Inference system (ANFIS), Alternating Decision Tree (ADT), Bayes Net (BN), Best First Tree (BFT), C4.5, Classification and Regression Trees (CART), Case-Based Reasoning (CBR), Decision Tree (DT), Differential Evolution & SVM (DESVM), Emerging Pattern (EP), Genetic programming (GP), Genetic Algorithm(GA), GRU-SVM, IBk, J48, k-Nearest Neighbors (kNN), KStar (K*), K-means and Support Vector Machine (KMSVM), kNN+ICA, k Nearest neighbor algorithm and Genetic algorithm (kNNGA), Logit Boost (LB), Logistic

Model Tree (LMT), Linear Regression (LR), Multifactor Dimensionality Reduction (MDR), Multilayer Perceptron (MLP), NNge, NB+kNN, NNANF, Naive Bayesian (NB), PART, Radial Basis Function Neural Network (RBFNN), Random Forest (RF), Support vector machine (SVM), Sequential Minimal Optimization (SMO), Simple Logistics (SL), Self Organising Maps (SOM), Softmax Regression (SR), SMO-SVM, ZeroR to discourse some difficulties challenged in computational biology. A lot of study is being done on breast cancer diagnosis. Many methods and procedures are constantly established attain correct and well-organized diagnosis outcomes. SVM has been used for knowledge based gene expression analysis, recognition of functional classes of genes gene selection [1,2]. kNN was successfully used for making a model which was capable of predicting the identity of unknown cancer samples [3]. The problem of gene reduction from huge microarray data set was solved by neural network moreover; it was able to identify the genes responsible for particular type of cancer occurrence [4]. Genetic Algorithm was used for building selectors where the state of allele denotes whether it (gene) is selected or not [4]. Genetic Programming has been shown to work excellent in case for recognition of structures for large data sets [5]. It was also applied to microarray data to generate programs that predict the malignant states of cancerous tissue, as well as classify different types of tissues [5]. Bayesian Network was well applied for identification of gene regulatory network from time course microarray data [5]. Sequential Minimum Optimization (SMO) shows good result for gene clustering [6]. Naive Bayes has been used by several researchers for gene selection and classification [6]. In the authors carried out a comparative study on diabetes disease diagnosis using neural networks on a pima-diabetes dataset and show that multilayer neural networks with Levenberg–Marquardt (LM) algorithm outperformed other neural network based classifiers [6]. In the authors performed a comparative study on Parkinson's disease diagnosis and compared the results of DMNeural, Neural Network, Regression, and Decision tree classification models on a data of 197 Parkinson's disease patients [7]. A total of 22 factors were compared and neural networks outperformed rest of the classification techniques with an accuracy of 92.9%. In Chen MS, et al. [7], the authors used Na.ve Bayes , Support vector machines Radial Basis Function (RBF) kernel, Radial basis neural networks, Decision trees J48 and simple CART for disease detection and found that SVM RBF kernel method outperformed other classifier techniques. Different data mining technique that were discovered and developed include the artificial neural network (ANN) as the commonest along with support vector machine (SVM), adaptive neuro-fuzzy inference system (ANFIS), k nearest neighbor decision tree (DT) case-based reasoning (CBR) and rough set theory (RST) algorithms on 4 different biological datasets. A hybrid combination of SVM, particle swarm optimization (PSO), and cuckoo search (CS) in a novel machine method was proposed in kumar Y, et al. [8]. The novel method is comprised of two phases: phase one makes use of CS for optimizing the SVM parameter which is specially developed to identify the kernel function best initial parameters, and the second on is the PSO application for SVM training continuation and finding of the best SVM parameters. Least Squares Support Vector Machine (LS-SVM) and Differential Evolution (DE) were applied for BC diagnosis in kumar Y, et al. [8]. In a two feature ranking algorithms of the Case-Based Reasoning (CBR) system, Adaptive Neuro-Fuzzy Inference system (ANFIS) and ANN based on PSO were proposed for the classification of breast cancer application [9]. SVM are utilized in a proposed method in with its combination with feature selection for the classification of breast cancer [9]. In uses a system that utilizes (ReliefF) algorithm for reducing the dimension data and Bayesian network for breast cancer classification. Emerging Pattern is markedly good for microarray data analysis [9]. Moreover it has an advantage of designing interaction among genes which enhances classification accuracy [9]. The need for powerful ML models is more pressing than ever. It is evident that current methods require further development before successful application to these enormous data sets can be claimed and their outputs enhance understanding of the genetic epidemiology of disease or become useful in a clinical disease risk predictive setting. It worth mentioning that some other similar researches have been done on application of Machine learning methods in some specific field of biology. But still which method one should apply for the classification and prediction of a particular cancer with high accuracy remains a challenge.

The objective of this study is to provide some contextual information of Machine Learning methods in Bioinformatics. This paper is planned as follows. Section 1 presents a brief summary of machine learning. Section 2 outlines the materials and methods used in this study. Section 3 presents performance evaluation, dataset and software used in this study. Section 4 presents the results and discussion, and the final section summarizes this work.

Materials and Methods

There are a number of classification algorithms that has been proposed by several researchers in the field of classification applications and investigated in breast cancer data. The working process is shown in Figure 1. We selected classification algorithm to find the most suitable one for predicting breast cancer.

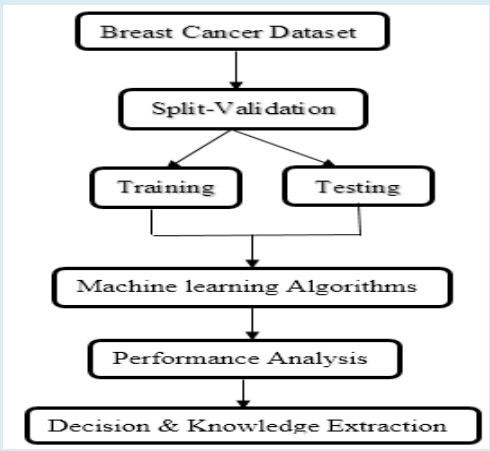

Figure 1: Working Process.

In this research work, the breast cancer data is to be classified using frequently used forty machine learning algorithms are given below-

Artificial Neural Networks (ANN)

Artificial Neural Networks (ANN) is an interconnected group of nodes that uses a computational model for information processing. It changes its structure based on external or internal information that flows through the network. ANN can be used to model a complex relationship between inputs and outputs and find patterns in data. The output of ANN is determined by characteristics of the features and the weights associated with the interconnections among them. The connections between nodes are modified in the training process to adapt the network to desired outputs [10].

Adaptive Neuro- Fuzzy Inference System (ANFIS)

In a two feature ranking algorithms of the Case-Based Reasoning (CBR) system, Adaptive Neuro- Fuzzy Inference system (ANFIS) and ANN based on PSO were proposed for the classification of breast cancer application [11].

Alternating Decision Tree (ADT*)*

Alternating Decision Tree is one of the classification method used in Machine learning which consists of decision nodes and prediction nodes. An instance is classified by an ADTree for which all decision nodes are true and summing any prediction nodes that are traversed. This makes it different from basic classification tree models that follow only one path through the tree [11].

Bayes Net (BN)

Bayes Nets or Bayesian networks are graphical representation for probabilistic relationships among a set of random variables. A Bayesian network is an annotated Directed Acyclic Graph (DAG) that encodes a joint probability distribution [11].

Best First Tree (BFT)

Judeay described the best-first search which searches up to the collected point and additional knowledge about the problem domain. It expands the most promising node chosen based on specified rules. The following extended algorithm is to use an additional list, consists of all nodes that have been evaluate. It will avoid the node which evaluated twice and it is not have the concept of infinite loops [11].

C4.5

It is an extension of ID3 algorithm proposed by Ross Quinlan which is used to generate a decision tree for classification. C4.5 is also known as statistical classifier [12]. The C4.5 constructs decision tree from a set of training dataset using information entropy concepts.

Classification and Regression Trees (CART)

CART is a recursive and gradual refinement algorithm of building a decision tree, to predict the classification situation of new samples of known input variable value. Breiman, et al. provided this algorithm and is based on Classification and Regression Trees (CART) [13].

Case-Based Reasoning (CBR)

In a two feature ranking algorithm of the Case-Based Reasoning (CBR) system was proposed for the classification of breast cancer application [13].

Decision Tree (DT)

It is an easy way to model complicated logics. These are flowcharts based on if, then, else, switch cases statements and associate conditions with actions to perform. Each decision is related to a variable, relation, condition alternatives dependencies. Operations to be performed are actions which correspond to specific entry. Each entry specifies whether or in what order the action is to be performed for the given set of condition alternatives the entry corresponds to [13].

Differential Evolution & SVM (DESVM)

A hybrid combination of SVM, particle swarm optimization (PSO), and cuckoo search (CS) in a novel machine method was proposed in Chaudhary A, et al. [14]. The novel method is comprised of two phases: phase one makes use of CS for optimizing the SVM parameter which

is specially developed to identify the kernel function best initial parameters, and the second on is the PSO application for SVM training continuation and finding of the best SVM parameters.

Emerging Pattern (EP)

Emerging Pattern is markedly good for microarray data analysis. Moreover it has an advantage of designing interaction among genes which enhances classification accuracy [14].

Genetic programming (GP)

Genetic programming aims to 'evolve' computer programs to solve complex problems [15]. First, an initial population of randomly generated computer programs is produced. Each program is run on a problem and assigned a fitness value based on its performance.

Genetic Algorithm (GA)

Genetic algorithms are a type of meta-heuristic search which simulates the natural selection process [15]. Genetic algorithms are related to the Evolutionary algorithms (EA) larger class which is primarily used for optimization problem/ solution generation. The Genetic Algorithm (GA) is moved by the genetics of the population (including gene and heredity frequencies), along with population levelled evolution, as well as the inspiration by the Mendelian structure understanding (such as alleles, genes and chromosomes) along with the related mechanisms (including mutation and recombination).

GRU-SVM

A neural network architecture [16] combining the gated recurrent unit (GRU) variant of recurrent neural network (RNN) and the support vector machine (SVM), for the purpose of binary classification.

IBk

IBk is a *k*-nearest-neighbor classifier that uses the same distance metric. k-NN is a type of instance based learning or lazy learning where the function is only approximated locally and all computation is deferred until classification. In this algorithm an object is classified by a majority vote of its neighbors [16].

J48

The J48 algorithm is WEKA's implementation of the C4.5 decision tree learner. The algorithm uses a greedy technique to induce decision trees for classification and uses reduced-error pruning [17].

k-Nearest Neighbors (kNN)

kNN is a lazy learning algorithm for instance based learning used to classify objects based on their closest training examples in the feature space [17]. An object is classified in a class to which its k-nearest neighbors belong. In the kNN algorithm, the classification of a new test feature vector is determined by the classes of its k-nearest neighbors.

KStar (K*)

Aha, Kibler & Albert describe three instance-based learners of increasing sophistication. IB1 is an implementation of a nearest neighbor algorithm with a specific distance function. IB3 is a further extension to improve tolerance to noisy data. Instances that have a sufficiently bad classification history are forgotten and only instances that have a good classification history are used for classification. Aha described IB4 and IB5, which handle irrelevant and novel attributes [18].

K-means and Support Vector Machine (KMSVM)

In Vanneschi L, et al. [18], a new system was propose for breast cancer classification. The new system uses a hybrid of K-means and Support Vector Machine (SVM).

kNN+ICA

Breast Cancer Detection with Reduced Feature Set in have been discussed using various data mining technique such as k- Nearest neighbor (k-NN) which is utilized for diagnosis with feature reduction properties using Independent Component Analysis (ICA) for reducing the one-dimensional feature vector that is involved in the computation of an independent component (IC) [19].

kNNGA

In Friedman N, et al. [19] new system was proposed using k Nearest neighbor algorithm (kNN) and Genetic algorithm (GA) with imputation techniques which is used instead of removing the values that are missing from the Mammographic Mass data.

Logit Boost (LB)

Boosting was well describe by "Freund and Schapire" that it is a classification which works by sequential implementation of a classification algorithm to reweighted training data and then taking the sequence classifiers produce by the weighted majority vote [20].

Logistic Model Tree (LM)

This is the combined version of linear logistic regression and tree induction. The former produces low variance high bias and the later produces high variance low bias. These two techniques were combined into learner which depends upon simple regression models if only little and/or noisy data is present. It adds more complex tree structures if enough data is available to such structures. Thus logistic model trees are the decision trees having linear regression model at leaves [21].

Linear Regression (LR)

Despite an algorithm for regression problem, linear regression was used as a classifier for cancer data analysis [21].

Multifactor Dimensionality Reduction (MDR)

MDR was one of the first ML methods developed to detect and characterize gene-gene interactions [22]. In the first stage of MDR, n genetic factors (e.g. SNPs) are selected from the entire set of factors. All possible multifactor (SNP genotype) combinations are represented in cells in n-dimensional space and each cell is assigned a case-control ratio.

Multilayer Perceptron (MLP)

Multilayer Perceptron is a nonlinear classifier based on the Perceptron. A Multilayer Perceptron (MLP) is a back propagation neural network with one or more layers between input and output layer. The following diagram illustrates a perceptron network with three layers [22].

NNge

Instance-based learners are "lazy" in the sense that they perform little work when learning from the data set, but expend more effort classifying new examples. The simplest method, nearest neighbor, performs no work at all when learning. NNge does not attempt to out-perform all other machine learning classifiers. Rather, it examines generalized exemplars as a method of improving the classification performance of instance-based learners [23].

NB+kNN

In Medjahed S, et al. [23] new system was proposed using k Nearest neighbor algorithm (kNN) and Naïve Bayes with imputation techniques which is used instead of removing the values that are missing from the Mammographic Mass data.

NNANF

In a two feature ranking algorithm of the Adaptive Neuro Fuzzy Inference system (ANFIS) and ANN based on PS were proposed for the classification of breast cancer application [24].

Naive Bayesian (NB)

Naive Bayesian classifier is developed on bayes conditional probability rule used for performing classification tasks, assuming attributes as statistically independent; the word Naive means strong. All attributes of the data set are considered as independent and strong of each other [24].

PART

PART uses the separate-and-conquer strategy, where it builds a rule in that manner and removes the instances it covers, and continues creating rules recursively for the remaining instances. Where C4.5 and RIPPER doe's global optimization to produce accurate rule sets, this added simplicity is the main advantage of PART [25].

Radial Basis Function Neural Network (RBFNN)

Breast Cancer Detection with Reduced Feature Set in Akay M [25] have been discussed using data mining technique such as radial basis function neural network (RBFNN).

Random Forest (RF)

Random forest is an ensemble classifier which consists of many decision tree and gives class as outputs i.e., the mode of the class's output by individual trees. Random Forests gives many classification trees without pruning [26].

Support Vector Machine (SVM)

Support vector machine (SVM) belongs to the same field as the neural networks. In their simplest form, SVMs are based on hyperplanes that separate the training data by a maximal margin. All vectors lying on one side of the hyperplane are labeled as -1, and all vectors lying on the other side are labeled as 1. The training instances that lie closest to the hyperplane are called support vectors [26].

Sequential Minimal Optimization (SMO)

Sequential Minimal Optimization (SMO) is used for training a support vector classifier using polynomial or RBF kernels. It replaces all missing the values and transforms nominal attributes into binary ones [27]. A single hidden layer neural network uses exactly the same form of model as an SVM.

Simple Logistics (SL)

It is a classifier used for building linear logistic regression

models. LogitBoost with simple regression functions are base learners used for fitting the logistic models. The optimal number of LogitBoost iterations to perform is cross-validated, which leads to automatic attribute selection [27].

Self Organising Maps (SOM)

Self Organising Maps [28], have been used for feature selection, attribute reduction, class prediction and classification using gene expression data.

Softmax Regression (SR)

This is a classification model generalizing logistic regression to multinomial problems. But unlike linear regression that produces raw scores for the classes, softmax regression produces a probability distribution for the classes [29].

SMO-SVM

Sequential minimal optimization (SMO) is an efficient algorithm which solves the optimization problem of Support Vector Machine (SVM) which arises during training. It breaks the optimization problem into several sub-problems, which are then solved analytically. The larger multiplier of the problem has linear equality constraint and therefore the smallest possible problem has two such multipliers [29].

ZeroR

ZeroR is the simplest classification method which depends on the target and ignores all predictors. ZeroR classifier simply predicts the majority category (class). Although there is no predictability power in ZeroR, it is useful for determining a baseline performance as a benchmark for other classification methods [29].

Performance Evaluation

We created a confusion matrix (contingency table) to assess the classifier's outcomes in Table 1 [30]. True positives (TP) denote the correct classifications of positive examples. True negatives (TN) are the correct classifications of negative examples. False positives (FP) represent the incorrect classifications of negative examples into class positive and false negatives (FN) are the positive examples incorrectly classified into class negative.

		Predicted	
		Positive	**Negative**
	Positive	TP	FN
Actual	Negative	FP	TN

Table 1: Confusion Matrix.

The classifier evaluation measures presented in this section are summarized in Table 2.

Tools	Statistic
Recall /Sensitivity	$R = TP/TP + FN$
Specificity	$S = TN/TN + FP$
Precision	$P = TP/TP + FP$
F-Score	$F = (2 \times \mathrm{Pr}ecision \times \mathrm{Re}call)/(\mathrm{Pr}ecision + \mathrm{Re}call)$
Mathews Correlation Coefficient	$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$

Table 2: Classifier Evaluation Measures.

Based on the contingency table, numerous measurements can be carried out to assess the performance of the induced classifier. The frequently used performance assessment techniques applied in prediction or classification knowledge is classifier accuracy which measures the proportion of correctly classified instances.

Datasets

Breast cancer (BC) is one of the main anxieties nowadays and is one of the maximum chief causes of death among women as it is extremely dominant cancer type later lungs cancer [31]. Early recognition of breast cancer aids to

shrinkage death and recover the superiority of life among patients. Here, the breast cancer data used for this paper was achieved from [32]. The breast cancer data contains 107 instances, 10 attributes: 9 features+1class attribute, Six classes of freshly excised tissue were studied using electrical impedance measurements: Carcinoma (Car), Fibro-adenoma (Fad), Mastopathy (Mas), Glandular (Gla), Connective (Con), Adipose (Adi).

Software Used

The Machine learning classification algorithms in this investigation are engaged based on R 3.2.4 version & Weka 3.6.6 version open source software for this analysis, an open gathering of machine learning algorithms endures the investigator to dig his own data for methods and plans [33,34]. The algorithms can either be useful conventional to a dataset or submitted from the researchers own code. R & Weka grasp tools for data pre-processing, classification, regression, clustering, association rules, and visualization. In our research we completed different supervised ML classification algorithms.

Results & Discussion

A specific vigorous zone of investigation in bioinformatics is the application and expansion of artificial intelligence techniques to answer biological difficulties. Investigating huge biological data sets necessitates creating sense of the data by concluding construction or simplifications from the data. Here, we have taken breast cancer data and the analysis is given below-

From Figure 2 we observed that ANN was found to be the best technique for classifying cancer class. From Figure 3 we showed that ANN and J48 based techniques shows the highest Recall/ Sensitivity and Specificity over the others. From Figure 4 we observed that ANN, DESVM and SVM gave the largest precision and F-score among all other methods. From Figure 5 we observed that the Mathews Correlation Coefficient (MCC) value is 0.9901 of ANN that indicates the perfect prediction. Alternatively, ANN outperforms the other techniques with an accuracy level of 99.5% followed by rest of the methods. The confusion matrix helps us to find the various evaluation measures like recall/sensitivity, specificity, precision, accuracy, F-Measure and Mathew's correlation coefficient (MCC). The performance of forty machine learning algorithms is given the Table 3. The results from Table 1 have been analyzed and it indicates that the classifiers ANN and SVM have performed better. Therefore, we see a great potential to increase the interaction between machine learning and bioinformatics. As far as application of machine learning techniques in bioinformatics is concerned, there is no perfect method to solve a biological problem; however, most of the times we better compare them with each other and then apply them in our problem.

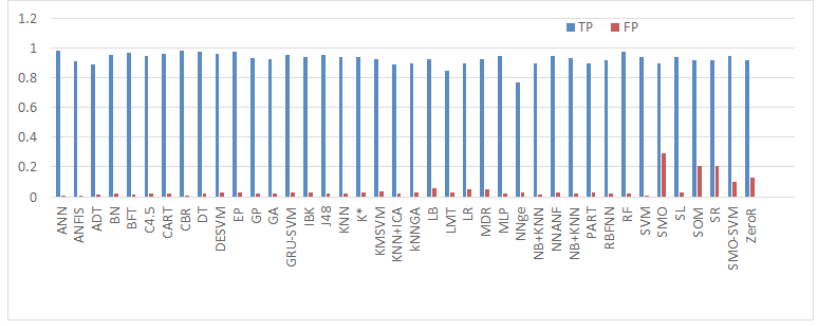

Figure 2: True Positive (TP) Versus False Positive (FP).

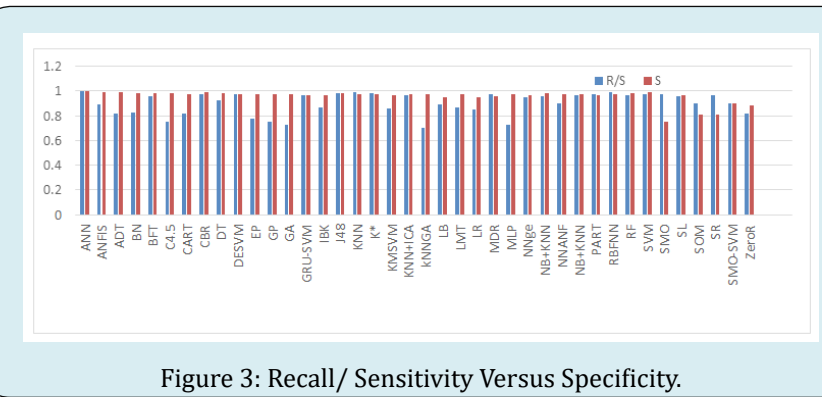

Figure 3: Recall/ Sensitivity Versus Specificity.

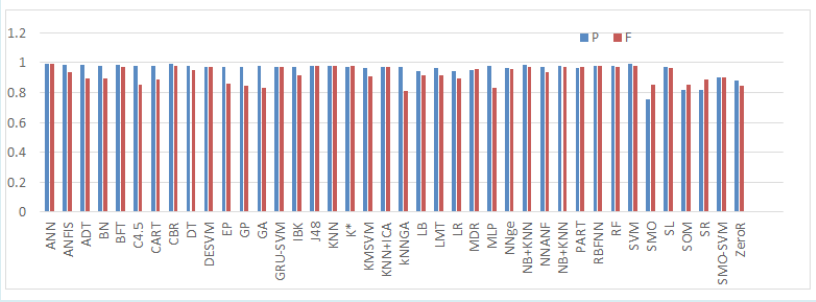

Figure 4: Precision Versus F-score.

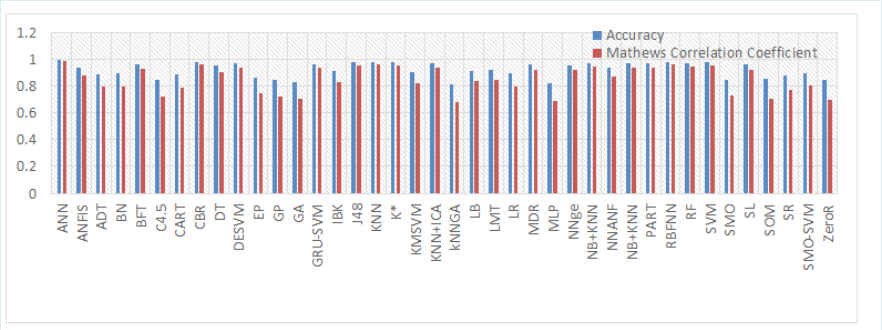

Figure 5: Accuracy versus Mathews Correlation Coefficient.

ML Algorithms	TP	TN	FP	FN	R/S	S	P	A	F	MCC
ANN	0.9812	0.9611	0.0051	0.0045	0.9954	0.9947	0.9948	0.995	0.9951	0.9901
ANFIS	0.9135	0.8617	0.0107	0.10781	0.8944	0.9877	0.9884	0.9374	0.939	0.8796
ADT	0.8934	0.8713	0.01137	0.2018	0.8157	0.9871	0.9874	0.8922	0.8934	0.8011
BN	0.9537	0.9125	0.02076	0.20154	0.8255	0.9777	0.9786	0.8936	0.8956	0.8005
BFT	0.9687	0.8459	0.01435	0.0475	0.9532	0.9833	0.9854	0.967	0.969	0.9344
C4.5	0.9456	0.8841	0.01865	0.3154	0.7498	0.9793	0.9806	0.8456	0.8498	0.7234
CART	0.9604	0.8613	0.02057	0.2145	0.8174	0.9766	0.979	0.8857	0.8909	0.7868
CBR	0.9845	0.8931	0.01036	0.0259	0.9743	0.9885	0.9895	0.981	0.9819	0.9621
DT	0.9741	0.9244	0.01976	0.0758	0.9278	0.979	0.9801	0.952	0.9532	0.9056
DESVM	0.9632	0.9431	0.02871	0.029	0.9707	0.9704	0.971	0.9706	0.9709	0.9412
EP	0.9744	0.9754	0.02837	0.287	0.7724	0.9717	0.9717	0.8607	0.8607	0.7442
GP	0.9345	0.9459	0.02514	0.3143	0.7483	0.9741	0.9738	0.847	0.8462	0.7234
GA	0.9254	0.9702	0.02219	0.3546	0.7229	0.9776	0.9765	0.8341	0.8308	0.7047
GRU-SVM	0.9582	0.9031	0.02999	0.03246	0.9672	0.9678	0.9696	0.9675	0.9684	0.935
IBK	0.9381	0.8154	0.02812	0.1419	0.8686	0.9666	0.97087	0.9116	0.9169	0.8289
J48	0.9568	0.9727	0.02156	0.02178	0.9777	0.9783	0.9779	0.978	0.9778	0.956
KNN	0.9382	0.9016	0.02384	0.01287	0.9864	0.9742	0.9752	0.9804	0.9808	0.9609

K*	0.9381	0.9012	0.02561	0.0178	0.98137	0.9723	0.9734	0.9769	0.9773	0.9539
KMSVM	0.9258	0.9451	0.03501	0.1556	0.8561	0.964	0.9635	0.9075	0.9066	0.8212
KNN+ICA	0.8897	0.9453	0.02488	0.03184	0.9654	0.9743	0.9727	0.97	0.9691	0.94
kNNGA	0.8973	0.9512	0.02562	0.3875	0.6983	0.9737	0.972	0.8173	0.8128	0.6774
LB	0.9256	0.9591	0.05431	0.1153	0.8893	0.9464	0.9445	0.9174	0.916	0.836
LMT	0.8469	0.9628	0.02828	0.1289	0.8679	0.9714	0.9676	0.92	0.915	0.8444
LR	0.8964	0.8919	0.05146	0.1571	0.8508	0.9454	0.9457	0.8955	0.8957	0.7961
MDR	0.9253	0.9702	0.04691	0.02855	0.97	0.9538	0.9517	0.9617	0.9608	0.9235
MLP	0.9465	0.8518	0.02013	0.357	0.7261	0.9769	0.9791	0.8266	0.8338	0.6933
NNge	0.7683	0.9364	0.03056	0.04124	0.949	0.9683	0.9617	0.9595	0.9553	0.9185
NB+KNN	0.8975	0.9651	0.01502	0.03769	0.9596	0.9846	0.9835	0.9724	0.9714	0.9451
NNANF	0.94697	0.9615	0.02993	0.10247	0.9023	0.9698	0.9693	0.9351	0.9346	0.8726
NB+KNN	0.9364	0.8341	0.02109	0.03571	0.9632	0.9753	0.9779	0.9689	0.9705	0.9377
PART	0.8975	0.8655	0.03148	0.0213	0.9768	0.9649	0.9661	0.9709	0.9714	0.9419
RBFNN	0.9217	0.9314	0.02287	0.01254	0.9865	0.976	0.9757	0.9812	0.9811	0.9625
RF	0.9744	0.9616	0.02044	0.0325	0.9677	0.9791	0.9794	0.9733	0.9735	0.9468
SVM	0.9411	0.9317	0.0101	0.0287	0.9704	0.9892	0.9893	0.9797	0.9798	0.9595
SMO	0.8975	0.8846	0.2897	0.0251	0.9727	0.7533	0.755	0.8498	0.8507	0.7272
SL	0.9411	0.9262	0.02992	0.0425	0.9567	0.9687	0.9691	0.9626	0.9629	0.9254
SOM	0.9218	0.8886	0.20646	0.10578	0.897	0.8114	0.817	0.8529	0.8551	0.7095
SR	0.91995	0.8651	0.20541	0.0342	0.9641	0.8081	0.8174	0.8816	0.8847	0.7758
SMO-SVM	0.9491	0.9527	0.1034	0.1054	0.9	0.902	0.9017	0.901	0.9009	0.8021
ZeroR	0.9197	0.9436	0.1256	0.2074	0.8159	0.8825	0.8798	0.8483	0.8467	0.699

Table 3: Performance Assessment Indicators of Machine Learning Algorithms.

Conclusion

Initial identification of Breast Cancer is meaningfully essential to treat the disease simply therefore it is essential to improve methods that can aid physicians to get perfect diagnosis. In this paper, we assessed the classification accuracy of forty Machine Learning algorithms on breast Cancer dataset. The goal of this relative study was to find the most perfect machine learning algorithm that can performance as a device for diagnosis of breast cancer. According to the prediction results, ANN and RBFNN have uppermost accuracy as well as mathews correlation coefficient for the given dataset. This shows that ANN and RBFNN can be used for prediction of breast cancer as compared to the rest of the methods. We want to develop network grounded software for performance assessment of numerous classifiers where the researchers can just call their data set and assess the outcomes on the prompt.

References

1. Xiaoyong L, Fu H (2014) PSO-Based Support Vector Machine with Cuckoo Search Technique for Clinical Disease Diagnoses. Scientific World Journal 2014: 548483.

2. Dubey A (2015) Machine Learning Classification for HIV Biomarkers. Online Journal of Bioinformatics 16(3): 344-356.

3. Syed SS, Shanthi S, Chitra VM (2013) Application of Data Mining techniques to model breast cancer data. International Journal of Emerging Technology and Advanced Engineering 3(11): 362-369.

4. Vaidehi K, Subashini TS (2015) Breast tissue characterization using combined K-NN classifier. Indian Journal of Science and Technology 8(1).

5. Arora R, Suman S (2012) Comparative analysis of classification algorithms on different datasets using WEKA. International Journal of Computer Applications 54(13): 21-25.

6. Poomani N, Porkodi RA (2015) Comparison of Data Mining classification algorithms using breast cancer microarray dataset: A study. International Journal for Scientific Research and Development 2(12): 543-547.

7. Chen MS, Han J, Yu PS (2002) Data mining: an overview from a database perspective. IEEE Transactions on Knowledge and Data Engineering 8(6): 866-883.

8. kumar Y, Sahoo G (2012) Analysis of Bayes, Neural Network and Tree Classifier of Classification Technique in Data Mining using WEKA. Computer Science & Information Technology 359-369.

9. Baldi P, Brunak S (2001) Bioinformatics: The Machine Learning Approach. 2nd (Edn.), MIT Press.

10. Doulah MSU (2019) Time Series Forecasting: A Comparative Study of VAR ANN and SVM Models**.** Journal of Statistical and Econometric Methods.

11. Palaniappan R, Sunderaj K, Sundaraj S (2014) A comparative study of the svm and k-nn machine learning algorithms for the diagnosis of respiratory pathologies using pulmonary acoustic signals. BMC Bioinformatics 15: 223.

12. Temurtas H, Yumusak N, Temurtas E (2009) A comparative study on diabetes disease diagnosis using neural networks. Expert Systems with applications 36(4): 8610-8615.

13. Das R (2010) A comparison of multiple classification methods for diagnosis of Parkinson disease. Expert Systems with Applications 37(2): 1568-1572.

14. Chaudhary A, Kolhe S, Kamal R (2013) Machine Learning Classification Techniques: A Comparative Study. International Journal on Advanced Computer Theory and Engineering 2(4): 2319-2526.

15. Lu Y, Han J (2003) Cancer classification using gene expression data. Information Systems 243-268.

16. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning 46(1-3): 389-422.

17. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, et al. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature medicine 7(6): 673-679.

18. Vanneschi L, Farinaccio A, Mauri G, Antoniotti M, Provero P, et al. (2011) A comparison of machine learning techniques for survival prediction in breast cancer. Bio Data Min 4: 12.

19. Friedman N, Linial M, Nachman I, Peer D (2000) Using Bayesian networks to analyze expression data. Journal of computational biology: a journal of computational molecular cell biology 7: 601-620.

20. Mitchell T (1997) Machine Learning. McGraw-Hill 414.

21. Wang Y, Tetko IV, Hall MA, Frank E, Facius A, et al. (2005) Gene selection from microarray data for cancer classification-a machine learning approach. Comput Biol Chem 29(1): 37-46.

22. Brown MPS, et al (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. PNAS 97(1): 262-267.

23. Medjahed S, Saadi TA, Benyettou A (2013) Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules. International Journal of Computer Applications 62(1): 1-5.

24. Liu JJ, Cutler G, Li W, Pan Z, Peng S, et al. (2005) Multiclass cancer classification and biomarker discovery using GA-based algorithms. Bioinformatics 21(11): 2691-2697.

25. Akay M (2009) Support vector machines combined with feature selection for breast cancer diagnosis. Expert Systems with Applications 36: 3240-3247.

26. Bichen Z, Yoon SW, Lam SS (2014) Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Systems with Applications 41: 1476-1482.

27. Abed BM (2016) A Hybrid Classification Algorithm Approach for Breast Cancer Diagnosis. IEEE: 264-269.

28. Doulah MSU, Islam MN (2019) Defining Homogenous Climate zones of Bangladesh using Cluster Analysis. International Journal of Statistics and Mathematics 6(1): 119-129.

29. KJlJ NA, Bilgili E, Akan A (2015) Breast Cancer Detection with Reduced Feature Set. Computational and Mathematical Methods in Medicine 11: 55-68.

30. Han J, Kamber M (2006) Data Mining: concepts and techniques. Morgan Kaufmann, San Francisco.

31. Soliman S, ElHamd E (2014) Classification of Breast Cancer using Differential Evolution and Least Squares Support Vector Machine. International Journal of Emerging Trends & Technology in Computer Science 3(2): 119-129.

32. Silva JE, Marques, Jossinet J (2000) Classification of Breast Tissue by Electrical Impedance Spectroscopy.

Med & Bio Eng & Computing 38(1): 26-30.

33. Crawley MJ (2007) The R Book 1st (Edn.), John Wiley & Sons Ltd England: 811-827.

34. Mark H, Frank E, Holmes G, Reutemann BPP, Ian H, et al. (2009) The WEKA data mining software: an update. ACM SIGKDD Exploration Newsletters 11(1): 10-18.