



Performance Comparison of Classification Algorithms for Medical Diagnosis

Anju Jain*, Saroj Ratnoo and Dinesh Kumar

Department of Computer Science and Engineering, Guru Jambheshwar University of Science and Technology, Hisar-125001, India

ABSTRACT

Knowledge extraction from medical datasets is a challenging task. Medical datasets are known for their complexity in terms of noise, missing values and imbalanced class distribution. Classification algorithms can assist medical experts in disease diagnosis provided that a rigorous and methodological evaluation of classification models is applied by selecting appropriate sampling techniques, performance metrics and statistical tests. An ad hoc approach in this regard can result in unexpectedly high misclassification rates, which may prove very costly in terms of people's health and lives. In this paper, we illustrate a methodology to evaluate and compare multiple classification algorithms on multiple medical datasets. The example experiment is conducted by applying five well-known machine learning algorithms i.e. the Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Ant Colony Optimisation (ACO) and Genetic Algorithm (GA) for discovering classification models for disease diagnosis from 11 publicly available medical datasets from the UCI machine learning laboratory. We conclude through a stepwise evaluation process that the performance of the Random Forest classifier was significantly better in diagnosing various diseases. The paper also addresses the issue of class imbalance and non-uniform misclassification costs, usually prevalent in datasets for disease diagnosis.

Keywords: Classification algorithms, machine learning, medical diagnosis, performance evaluation

INTRODUCTION

Machine learning, a sub-discipline in the field of Artificial Intelligence, involves the study and design of algorithms that can learn from data (Brink & Richards, 2013). Machine

learning is mainly useful in cases where deterministic solutions are not available. Medical diagnosis is one of the important activities in the field of medicine because early and accurate diagnosis helps people recover faster and saves the cost of prolonged treatment. Research on automated knowledge

Article history:

Received: 1 Jun 2017

Accepted: 24 August 2017

E-mail addresses:

anju_jain7@rediffmail.com (Anju Jain)

ratnoo.saroj@gmail.com (Saroj Ratnoo)

dinesh_chutani@yahoo.co.in (Dinesh Kumar)

*Corresponding Author

extraction from medical data has been growing rapidly (Esfandiari, Babavalian, Moghadam, & Tabar, 2014). However, extracting accurate and comprehensive knowledge from enormous medical data is a challenging task. Medical datasets essentially have missing values and class imbalance that consequently influence the accuracy and comprehensibility of classification algorithms used for disease diagnosis.

Classification algorithms are being widely used in various medical applications related to the predictive domain (Çınar, Engin, Engin, & Ateşçi, 2009; Azar & El-Metwally, 2012). A classification algorithm works in two phases: i). The training phase learns the classifier from the training data set; and ii) The classification/prediction phase uses the model to predict the instances unseen during the training phase. The performance of a classification algorithm is analysed on a test set.

A proper way of evaluating classification algorithms for medical diagnosis is of crucial importance. Though plenty of research has been done on evaluating the performance of classifiers, yet it has been observed that most researchers do not follow a rigorous approach for evaluating classification algorithms (Japkowicz & Shah, 2014; Vanaja & Rameshkumar, 2015). A common approach for performance evaluation of classification algorithms is based on their accuracy or error rates over a large number of datasets using k-fold cross validation (CV). Subsequently, the results are averaged and the paired t-test is applied to test statistical significance in the observed performances of different classification algorithms. There are several issues with this approach such as error estimation, aggregation of the results, selection of evaluation metrics and statistical significance testing. The most widely used evaluation metric for classification algorithm is accuracy/error rate. However, it may be misleading to use accuracy as the lone performance criterion, particularly, in the presence of class imbalance and non-uniform misclassification costs; this is most often the case with medical datasets for disease diagnosis. At the same time, the most widely used test of statistical significance, the t-test, is certainly not suitable in cases where such testing is to be performed over multiple classifiers and datasets.

The aim of this research was to illustrate proper evaluation of classification algorithms on datasets related to disease diagnosis in terms of choice of evaluation methods, performance metrics and appropriate statistical analysis. A comparison of two classification algorithms over a single dataset is simple and can be carried over by applying t-test for performance metric values obtained over the tenfold cross validation sampling technique. Comparing multiple classifiers on multiple datasets is relatively a more complex problem that needs the attention of the people working in the field of computational medical diagnosis. In this paper, we demonstrate the complete process of the evaluation of multiple classifiers on multiple medical datasets. We have divided the whole process into five basic components: i) Selection of datasets and handling missing values in datasets ii) Choice of learning algorithms; iii) Performance measures of interest; iv) Error estimation/Sampling method; and v) Application of statistical tests and Dealing with non-uniform misclassification costs.

The rest of the paper is organised as follows: The next section describes the related work. Section 3 depicts the proposed classifier evaluation framework for comparing multiple classification algorithms applied to multiple datasets for medical diagnosis and datasets used

in the proposed work. We report the experimental setup with the results of our experimentation and address the issue of non-uniform misclassification cost in Section 4. Finally, we conclude the paper in the last section.

RELATED WORK

It is often emphasised that researchers should use appropriate evaluation methods, performance metrics and statistical analysis for evaluating classifiers. However, many a time, classification algorithms are evaluated and compared solely on the basis of point estimates of accuracy or error rates. This lack of proper understanding and appreciation of the context in which the different components of the evaluation framework operate may lead to misinterpretation of the evaluation outcomes (Japkowicz & Shah, 2014) and may have adverse influence in fields like medical diagnosis due to high misclassification costs. In this direction, researchers Tanwani, Afridi, Shafiq and Farooq (2009) have shown that the nature of the given dataset plays an important role in the classification accuracy of algorithms. They have provided 10 generic guidelines that can help researchers of the biomedical classification community to select an appropriate classifier for a given problem. Danjuma and Osofisan (2014) have underlined that since clinical data are noisy and imbalanced, appropriate sampling techniques and evaluation metrics should be used to estimate the performance of classifiers and subsequently, the results should further be subjected to rigorous assessment before applying them to real life clinical decisions. Further, it has also been highlighted that the quality of clinical decision support systems depends on the ability of researchers to carefully choose, tune and apply machine learning algorithms to clinical data (Danjuma, 2015; Bal et al., 2014).

The accuracy estimates of many classification algorithms lose their credibility in the presence of missing values in the datasets. The classification accuracy is predominantly affected by the presence of missing feature values because most of the classification algorithms cannot deal with these automatically (Marlin, 2008; Howell, 2007). Therefore, handling missing attribute values is a common concern in the field of data mining and knowledge discovery from medical datasets.

Ferri et al. (2009) have expressed concerns about what aspects of classifiers' performance need to be evaluated (Ferri, Hernández-Orallo, & Modroi, 2009). The authors have emphasised that accuracy as a single performance evaluation metric is simple in its implementation, but less efficient in covering various aspects of the evaluation process and it should be supplemented with a Receiver Operating Curve (ROC). The ROC curve has a complex implementation, but it is more efficient and gives meaningful interpretations in medical domain (Esfandiari et al., 2014; Fawcett, 2006; Yousef, Wagner, & Loew, 2006; Prati, Batista, & Monard, 2011).

Some authors have raised objections to using the Null Hypothesis Significance Testing (NHST) in the context of machine learning algorithms (Demsar, 2008; Drummond, 2006; Ioannidis, 2005). Demsar (2008) propagated that NHST does not tell us all that we need to know and what many researchers believe it conveys. The test computes the (conditional) probability of certain statistics and says nothing about the hypothesis. Drummond (2006) also argued that rather than being confirmatory, the experiments should remain exploratory. Nevertheless, statistical testing does remain relevant for comparing classification algorithms

provided that appropriate statistical tests are applied. Many researchers use repeated paired t-tests for comparing more than two classification algorithms on multiple datasets. However, the t-test is only appropriate for comparing two classifiers on a single dataset. Further, Salzberg (1997) has suggested that because of the large number of experimental runs, a significance level of 0.05 most often used for t-tests is not stringent enough and the inferences drawn may have been obtained by chance (Hubbard & Lindsay, 2008). Such problems are well addressed by using the ANOVA and the Friedman tests. In addition, Demsar (2008) has compared the ANOVA and Friedman tests and he concluded that the Friedman test, being non-parametric, is safer than ANOVA for comparing multiple classifiers over multiple datasets since it does not assume normal distribution for the values of attributes and there is less chance of committing a type I error. Such non-parametric tests can be applied to classification accuracies, error ratios or any other performance metric of classification models.

Most of the medical datasets suffer from the problems of class imbalance and asymmetric misclassification cost. Sampling and cost-sensitive learning are the two commonly used methods to address these problems (Kotsiantis, Kanellopoulos, & Pintelas, 2006; Weiss, McCarthy, & Zabar, 2007; Ling & Sheng, 2011). Sampling includes under-sampling, over-sampling and a combination of the two. Cost-sensitive learning includes methods like MetaCost, CostSensitiveClassifier and Empirical Thresholding. Cost-sensitive learning has many times outperformed sampling methods (Japkowicz & Stephen, 2002). MetaCost, a general method for making a classifier cost sensitive, produces large execution cost reduction compared to the sampling techniques (Domingos, 1999; Kim, Choi, Kim, & Suh, 2012).

MATERIALS AND METHOD

Proposed Classifier Evaluation Framework for Medical Diagnosis

There are a few important steps for evaluating the performance of data mining algorithms. These steps depend on the type of the data mining task under consideration, underlying datasets and the purpose of the evaluation. Figure 1 shows the proposed classifier evaluation framework. This section describes the proposed framework for evaluating multiple classification algorithms on multiple datasets for the purpose of medical diagnosis.

1. Selection of datasets: This is not independent of other components of the proposed model since it affects the other components like sampling method, selection of performance measures and statistical tests. Therefore, keeping the purpose of the research in mind, a variety of medical datasets are chosen from the UCI repository. These cover a broad spectrum of variability in data characteristics such as dimensionality and class distribution. The class distribution is specifically important for the field of disease diagnosis.
2. Handling missing values in datasets: Accuracy achieved by any classification algorithm is inversely proportional to the amount of noise and missing values present in medical datasets. The removal of the examples with missing values belonging to the minority class can significantly affect the performance of a classification algorithm. Therefore, replacing missing values instead of removing the examples that contain missing values

may be more beneficial. In this work, a non-parametric method, called ‘missForest’, was used to replace missing values. This method was found to be suitable for replacing missing values because it can cope with mixed-type variables, nonlinear relations and high dimensionality of datasets (Stekhoven & Bühlmann, 2012). The algorithm is based on the Random Forest technique. For each variable, the ‘missForest’ algorithm fits a random forest on the observed values and then predicts the missing values.

3. Choice of Learning Algorithms: Among various algorithms in data modelling for disease diagnosis, a decision tree is well known for its simplicity and interpretability (Chang & Chen, 2009; Azar & El-Metwally, 2012). Other efficient algorithms such as the Support Vector Machine (SVM) and Random Forests have also become popular (Esfandiari et al., 2014; Hassanien & Kim, 2012; Jin, Tang, & Zhang, 2007) in the medical domain. The use of evolutionary and swarm intelligence algorithms have also gained popularity for discovering knowledge for medical diagnosis (Peña-Reyes & Sipper, 2000; Parpinelli, Lopes, & Freitas, 2002; Chen et al., 2011; Fidelis, Lopes, & Freitas, 2000; Freitas, 2013; Kumar, 2011). Five algorithms were selected for the evaluation exercise. These algorithms included SVM, the Decision Tree (DT) and the Random Forest (RF) as decision-based classifiers. The other two algorithms, General Algorithm (GA) and Ant Colony Optimisation (ACO), are nature-inspired classifiers. An algorithm known as the Pittsburgh Genetic Interval Rule Learning Algorithm (PGIRLA) was included in the GA-based classification algorithm (Corcoran & Sen, 1994). This GA follows the Pittsburgh approach, which is considered better than the Michigan approach because it evaluates the classification model as a whole. The most preliminary ACO algorithm for classification rule mining is the Ant Miner (Parpinelli et al., 2002). Several variants of this algorithm have been proposed with advanced features. An algorithm, Advanced Ant Miner Plus-C, was also included in this work. This algorithm is known for having discovered simple and small rule lists. This improves comprehensibility of the knowledge discovered (Parpinelli et al., 2002). The choice of classification algorithms is not an issue of this study. Researchers can choose classification algorithms of their interest.
4. Performance measures of interest: The selection of appropriate performance measures is essential in the domain of medical diagnosis. In case of imbalanced datasets, the use of accuracy as the lone performance evaluation metric does not give the real picture of how good a classifier is in discerning one class from the other (Chawla, 2005). For example, let us consider a dataset related to diagnosing cancer. Assume that this data set contains 90% of the examples with the negative occurrence of cancer and only 10% of the examples with positive occurrence of cancer. An inconsequential classifier can decide to always predict the majority class (non-cancerous class) for all data examples in the test set. Such a trivial classifier is expected to achieve an overall accuracy of 90 percent. This high accuracy figure may look impressive, but it is misleading because this classifier achieves perfect predictions for the majority class only (non-cancerous class) and no correct predictions at all for the minority class (cancerous class). All of the 10% of people who actually have cancer will be predicted as having no cancer, and this is a bad scenario for medical diagnosis. This example clearly indicates that

accuracy in itself is not an adequate measure to use for the datasets with a highly skewed class distribution. In medical applications, the sensitivity and specificity of a classifier matter more than its overall accuracy. Sensitivity measures the capability of a classifier to predict a diseased person correctly and specificity measures the capability of a classifier to diagnose a healthy person correctly. Therefore, we should use a performance metric that accounts for sensitivity as well as specificity (SE*SP). Another metric used in the medical domain is Youden's index (J), which is computed as (sensitivity+specificity-1) (Newby, Freitas & Ghafourian, 2013). One more simple measure of interest could be F-measure that takes account of precision and recall. The formula for F-measure is given below.

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

An additional important performance metric for classification problems related to medical diagnosis is the area under the Receiver Operating Curve (ROC). This measure can characterise the behaviour of a classifier over the full operating range and is more robust than accuracy in class imbalanced situations. Finally, we propose to use Sensitivity (SE), Specificity (SP), SE*SP, F-measure and ROC as the performance metrics to compare classification algorithms for medical diagnosis.

5. Error estimation/sampling method: Selecting the appropriate error estimation and re-sampling method is important while working with medical datasets. A stratified tenfold cross validation is suitable for medical diagnosis because of the presence of skewed class distribution. It ensures that the data distribution is respected in the training and testing sets created at every fold.
6. Application of statistical tests: The last step is to select appropriate tests for assessing the statistical significance of the difference in performance of the classification algorithms. For comparing the performances of multiple classifiers on multiple domains, the one-way repeated ANOVA measure and Friedman tests fit the bill. Due to lack of strong assumption about the distribution of underlying medical data, the non-parametric Friedman test is more appropriate. The null hypothesis of the Friedman test is that there is no significant difference in the performance of the classification algorithms (Japkowicz & Shah, 2014). Further, in case the Friedman test returns affirmative results in terms of statistical significance i.e. it rejects the NULL hypothesis that there is no significant difference in the classification algorithms, the post-hoc Nemenyi test is applied to identify the significantly different pairs of classifiers.
7. Dealing with non-uniform misclassification costs: The datasets for disease diagnosis often have a class imbalance. For example, people who are diagnosed positive for cancer will be far fewer than those who are diagnosed negative for the disease. This introduces a learning bias for the majority class and will have a poor prediction rate for diagnosing the disease. The additional issue is of non-uniform misclassification costs, which means that the type of misclassifications have unequal costs associated with them. In the domain of disease diagnosis, we consider the occurrence of cancer as the positive class (minority class in the data) and non-occurrence as the negative class

(majority class in the data). Although false positive and false negative, both types of errors have some cost associated with them, but false negative errors are considered costlier than false positive errors in disease diagnosis i.e. it is costlier to send those patients home who actually had the cancer but were diagnosed otherwise. The false negative error rate tends to increase in the presence of the class imbalance scenario. The performance metric, accuracy, does not account for unequal misclassification costs. We can address the problem either by using sampling or cost-sensitive learning techniques. The cost-sensitive learning techniques have been proven more effective in dealing with non-uniform misclassification costs than the sampling methods (Ling & Sheng, 2011). Therefore, we used a cost-sensitive learning method known as MetaCost to deal with non-uniform classification costs.

Datasets Used

This study used five classification algorithms on 11 medical datasets (Table 1) from the UCI Machine Learning repository. All the datasets were two-class datasets, except for the heart dataset, and are numeric in nature. The heart dataset, a multi-class dataset, is modified to obtain a two-class problem by defining the joint of two or more classes as negative and one class as positive. The datasets considered are partitioned using the tenfold cross validation (10-fcv) procedure. The main characteristics of these datasets are summarised in Table 1.

RESULTS AND DISCUSSION

This study applied the Decision Tree (J48), Random Forest and Support Vector Machine in WEKA (Bouckaert et al., 2002) and PGIRLA and Advanced_Ant_Miner_Plus-' in KEEL (Alcalá-Fdez et al., 2008) library. The entire process of applying classification algorithms on datasets is shown in Figure 2. Missing values in the dataset, if any, were replaced prior to running classification algorithms using the 'missForest' package in statistical software R, which was also used for applying the statistical tests (Venables & Smith, 2008). Since the purpose of this experiment was not to tune the parameters for optimised performance, all the classification algorithms were run with their default parameters as such. The parameters of these algorithms are presented in Table 2.

Table 3 shows the results obtained for the accuracy of the chosen algorithms and a classifier known as ZeroR. The ZeroR classifier has no predictive power and is useful for determining a baseline performance as a benchmark for other classification methods. It ignores all the predictors and predicts every instance in the test data to belong to the majority class (Witten, Frank, & Hall, 2011). It is noteworthy that in case of four datasets highlighted in bold font, ZeroR is comparable with the rest of the classifiers. The accuracy achieved by SVM classifier is almost the same as that of ZeroR classifier except for the Mammographic Mass dataset. This shows that accuracy does not necessarily suffice in its classical form for evaluating classifiers.

Table 4 shows the results obtained for the ROC, Specificity (SP), Sensitivity (SE), $SP \cdot SE$ and F-measure for classifiers J48, RF, SVM, GA and ACO. The values of specificity are significantly greater than the values of sensitivity for almost all the classifiers and datasets. This shows the bias of classification algorithms towards the majority class. This difference is more visible for the datasets (heart, hepatitis, WBCD, ILPD and thoracic) with high class imbalance. This fact highlights the need for addressing the issue of class imbalance.

A comparison of values of sensitivity for different datasets with all the five classifiers is shown in Figure 3. Figure 3 shows that the performance of SVM is the worst when the value of sensitivity is zero or very low for all the datasets except for the mammographic mass dataset. This means that the SVM fails to make correct predictions for diseased people. The figure also depicts that heart and thoracic are the most difficult datasets to classify. None of the algorithms work well on these datasets.

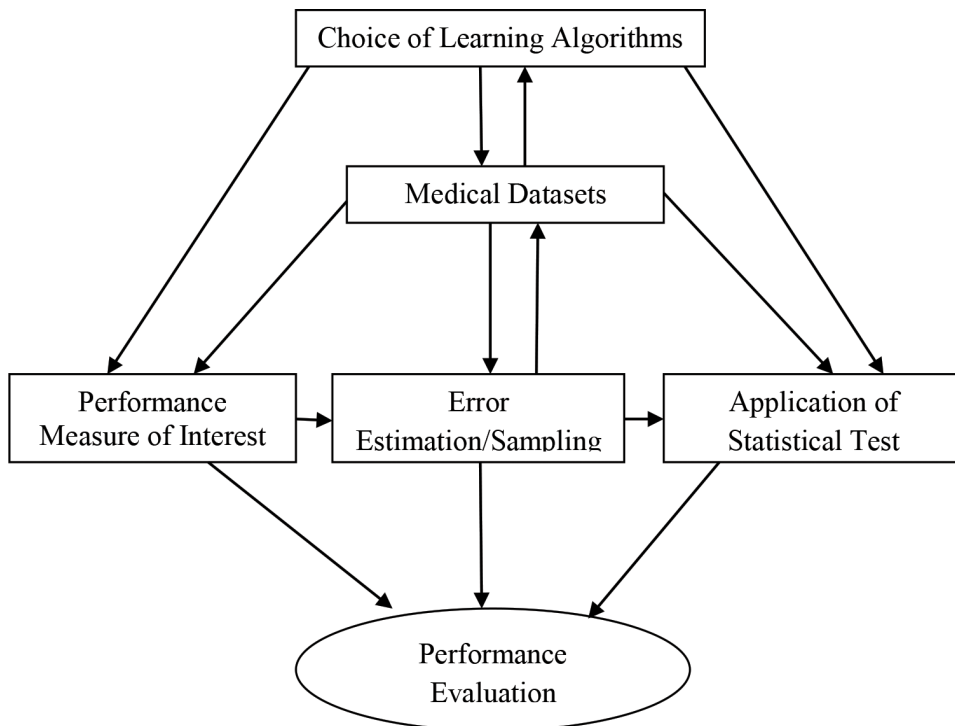


Figure 1. Proposed classifier evaluation framework for medical diagnosis

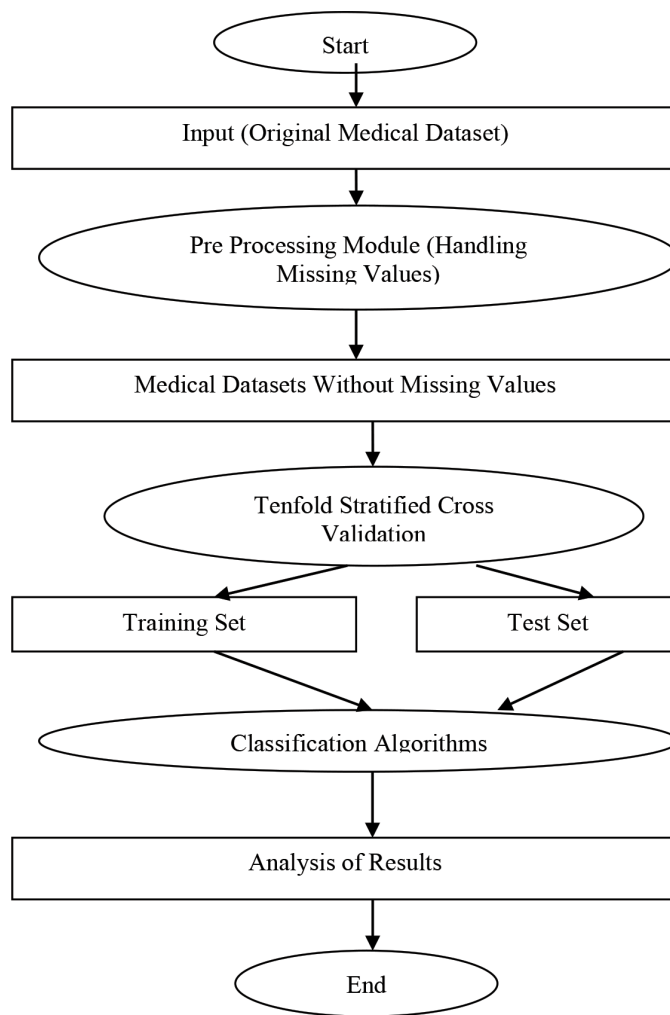


Figure 2. Flowchart of applying classification algorithms on medical datasets

Table 1

Datasets used in the experiment

Datasets	#Instances	#Attributes	#Classes	Class distribution ratio Majority: Minority
Heart (HE)	303	14	5	Remainder:19
Hepatitis (HEP)	155	20	2	79:21
Chronic Kidney Disease (CKD)	400	25	2	63:37
WBCD	699	11	2	66:34
Bupa Liver Disorder (BLD)	345	7	2	58:42
Pima Indian Diabetes (PID)	768	9	2	65:35
ILPD	583	11	2	72:28
Thoracic Surgery (TS)	470	17	2	85:15
Statlog Heart (SH)	270	13	2	56:44
Mammographic Mass Data (MMD)	961	6	2	54:46
WPBC	198	34	2	76:24

Table 2

Parameters considered for the classifiers

Algorithms	Parameters
J48	Confidence factor = 0.25, numFolds = 3
RF	numTrees = 100, Number of random features considered = 5
SVM	KernelType = radial basis function
GA	No. of generations = 1000, Pop-size = 61, Crossover probability = 0.7, Mutation probability = 0.5
ACO	No. of ants = 1000, Max_Pheromone = 0.99, Min_Pheromone = 0.1

Table 3

Results obtained for accuracy

Datasets	J48	RF	SVM	GA	ACO	ZeroR
HE	52.8	58.41	54.12	53.97	56.0	54.12
HEP	89.03	89.67	79.35	79.41	85.3	79.35
CKD	99.5	99.75	62.5	93.25	98.75	62.5
WBCD	94.27	96.7	66.38	92.69	94.54	65.5
BLD	68.69	73.33	59.42	64.9	43.48	57.97
PID	73.82	74.34	65.1	73.31	75.0	65.10
ILPD	68.78	70.15	72.38	69.44	68.45	71.35
TS	84.46	84.04	85.1	84.89	84.0	85.10
SH	76.66	81.85	55.92	70	81.48	55.55
MMD	84.7	80.12	81.37	80.41	82.75	53.6
WPBC	76.26	80.81	76.26	67.9	63.68	76.26
Average Accuracy	78.99727	80.83364	68.9	75.47	75.76636	66.03

Table 4

*Result obtained for ROC, SP, SE, SP*SE, F-measure for J48,RF,SVM,GA and ACO classifiers*

DATASETS	Results for various performance measures for Decision Tree classifier (J48)					Results for various performance measures for Random Forest classifier				
	ROC	SP	SE	SP * SE	F-measure	ROC	SP	SE	SP * SE	F-measure
HE	0.773	0.85	0.11	0.09	0.1	0.9	0.9	0.135	0.122	0.15
HEP	0.797	0.943	0.688	0.64	0.721	0.897	0.951	0.688	0.654	0.733
CKD	0.999	0.992	1	0.99	0.993	1	0.996	1	0.996	0.997
WBCD	0.933	0.956	0.917	0.87	0.917	0.992	0.965	0.971	0.93	0.953
BLD	0.665	0.8	0.531	0.424	0.588	0.764	0.815	0.621	0.5	0.662
PID	0.751	0.814	0.597	0.48	0.614	0.815	0.836	0.571	0.47	0.608
ILPD	0.674	0.829	0.335	0.27	0.674	0.742	0.861	0.305	0.26	0.37
TS	0.502	0.99	0.014	0.013	0.027	0.642	0.983	0.029	0.02	0.051

Table 4 (*continue*)

DATASETS	Results for various performance measures for Decision Tree classifier (J48)					Results for various performance measures for Random Forest classifier				
	ROC	SP	SE	SP * SE	F-measure	ROC	SP	SE	SP * SE	F-measure
SH	0.744	0.793	0.733	0.58	0.736	0.891	0.867	0.758	0.65	0.788
MMD	0.864	0.849	0.845	0.71	0.836	0.877	0.812	0.789	0.64	0.786
WPBC	0.58	0.861	0.447	0.38	0.472	0.691	0.993	0.213	0.21	0.345
AVERAGE	0.752	0.88	0.565	0.495	0.607	0.8373	0.907	0.552	0.495	0.585

DATASETS	Results for various performance measures for SVM Classifier					Results for various performance measures for GA classifier (PGIRLA)				
	ROC	SP	SE	SP * SE	F-measure	ROC	SP	SE	SP * SE	F-measure
HE	0.5	1	0	0	0	0.5	0.925	0.07	0.06	0.09
HEP	0.5	1	0	0	0	0.5	1	0	0	0
CKD	0.5	1	0	0	0	0.904	0.924	0.93	0.85	0.904
WBCD	0.512	1	0.025	0.025	0.049	0.95	0.98	0.28	0.27	0.42
BLD	0.518	0.995	0.041	0.04	0.079	0.59	0.805	0.29	0.23	0.372
PID	0.5	1	0	0	0	0.7	0.83	0.54	0.44	0.58
ILPD	0.518	1	0.036	0.036	0.069	0.5	0.95	0.03	0.02	0.05
TS	0.5	1	0	0	0	0.4	0.99	0.02	0.019	0.03
SH	0.504	1	0.008	0.008	0.017	0.67	0.84	0.51	0.428	0.59
MMD	0.813	0.822	0.804	0.66	0.8	0.81	0.85	0.79	0.67	0.8
WPBC	0.5	1	0	0	0	0.3	1	0	0	0
AVERAGE	0.533	0.983	0.083	0.069	0.09	0.6313	0.918	0.314	0.2715	0.348

DATASETS	Results for various performance measures for ACO Classifier (Advanced Ant Miner Plus-C)				
	ROC	SP	SE	SP * SE	F-measure
HE	0.56	0.9	0.05	0.045	0.05
HEP	0.65	0.91	0.62	0.5642	0.62
CKD	0.98	1	0.96	0.96	0.97
WBCD	0.93	0.96	0.9	0.864	0.91
BLD	0.5	0.095	0.9	0.085	0.56
PID	0.7	0.85	0.56	0.476	0.61
ILPD	0.64	0.72	0.58	0.4176	0.5
TS	0.5	0.98	0.05	0.049	0.08
SH	0.76	0.89	0.71	0.63	0.76
MMD	0.82	0.89	0.81	0.72	0.81
WPBC	0.59	0.6	0.72	0.43	0.48
AVERAGE	0.693	0.8	0.623	0.476	0.577

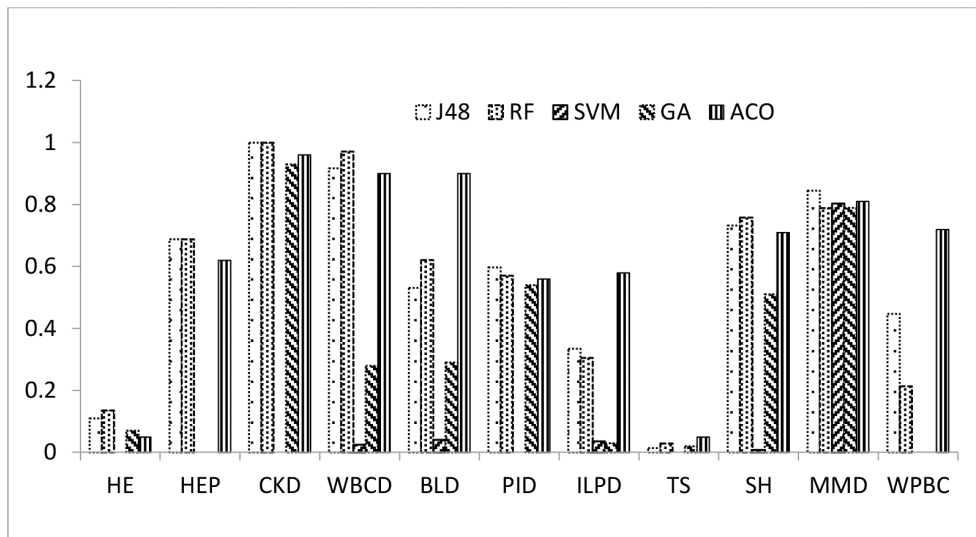


Figure 3. Comparison of sensitivity (SE) of five classifiers

One of the common techniques practised for an overall comparison of classification algorithms is to aggregate their performance over several datasets. Figure 4 shows the average values for ROC, SP*SE, F-measure and SE. The graph shows that the Random Forest classification algorithm is superior to all other classifiers in terms of the aggregated performance over all the datasets. Out of the selected classifiers, Random Forest has the highest average ROC (83%), SP*SE (49%), F-measure (59%) and SE (55%), followed by the Decision Tree and ACO classifiers. However, this is not sufficient and does not rule out the possibility that the result may have been obtained by chance. Therefore, we need to apply appropriate statistical tests.

This study applied the Friedman test, a non-parametric statistical test for comparing multiple algorithms on multiple domains, to see if indeed the performance of the Random Forest classifier on different domains is significantly better than that of the other classifiers. The test results are summarised in Table 5. The results show that the Random Forest Algorithm is significantly better than the rest of the classification algorithms on the performance metrics of ROC area, SP*SE, F-measure and SE at a significance level of 5% and degree of freedom, which equals 4. The results of the Friedman test agree on all metrics of the performance measures in this study but it may not be true for similar experimental studies with different combinations of classification algorithms and datasets.

This study also applied the post-hoc Nemenyi test to make a pair-wise comparison by using the Random Forest as the control classifier. The resulting Q values computed for the Nemenyi test for discovering the pair-wise differences between the performance of RF and other classifiers with respect to the metrics of ROC, SP*SE, F-measure and SE are given in Table 6.

The value of (q_α) (for a degree of freedom that equals 40) computes to 2.68 for a significance level of $\alpha = 0.05$. The null hypothesis are rejected only if the respective Q value for J48, SVM, GA and ACO exceed the value of q_α . One can conclude from the Q values given above that the RF classification algorithm was significantly better than SVM, GA and ACO classifiers

individually. Further, RF was significantly better than J48 on the basis of the metrics SE*SP and the ROC area. However, the NULL hypothesis could not be rejected for the RF and J48 classifiers on the basis of the F-measure and SE.

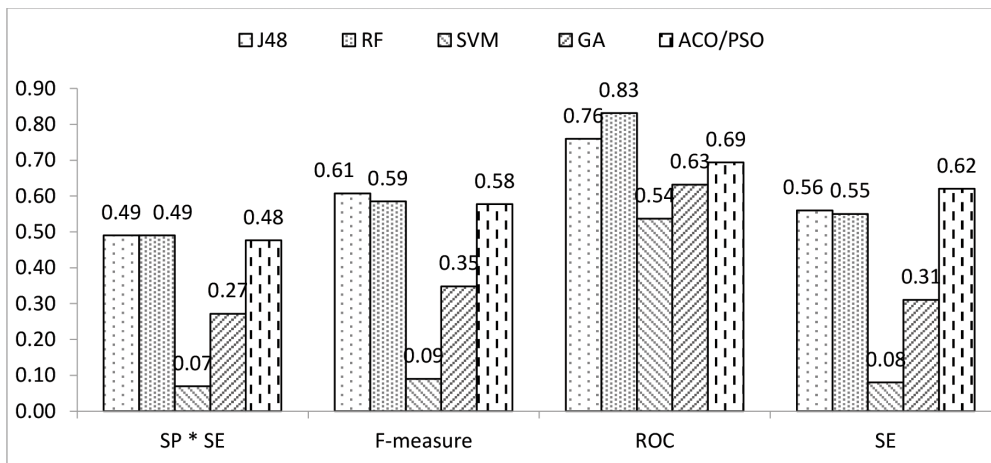


Figure 4. A comparison of aggregate performances of five classifiers of 11 medical datasets

Table 5

Summary of the results of the Friedman test

Domains	Friedman chi-squared value	Degree of freedom	p-value	Null hypothesis rejected/accepted
ROC	34.930	4	0.000000479	Rejected
SP*SE	25.981	4	0.00007695	Rejected
F-measure	27.410	4	0.000016422	Rejected
SE	27.037	4	0.000019539	Rejected

Table 6

Nemenyi test results

	ROC Area	SP*SE	F-measure	SE	Turkey's critical value (q_α)	Null hypothesis rejected/accepted
$Q_{J48, RF}$	21.64	3.73	0	-1.49	2.68	Rejected for ROC and SP*SE, Not rejected for F-measure and SE
$Q_{SVM, RF}$	56.71	45.52	44.02	41.79		Rejected
$Q_{GA, RF}$	50.74	32.08	32.08	31.34		Rejected
$Q_{ACO, RF}$	23.5	4.47	4.47	2.98		Rejected

Cost Sensitive Analysis

The issue of class imbalance and unequal misclassification costs is itself a very wide topic and needs a separate study. Thus, we have left a detailed study on the topic as a future scope and addressed the issue in a very limited way in this paper.

One of the important techniques to deal with unbalanced class medical datasets is to increase the cost of misclassification of minority class. There are two misclassification errors: False Positive (FP) errors and False Negative (FN) errors. In case of medical datasets, an FP error amounts to diagnosing a healthy person as diseased, whereas an FN error pertains to diagnosing a diseased person as healthy. Often, the datasets for disease diagnosis are biased towards the non-diseased class. Therefore, of these two misclassification errors, more cost needs to be assigned to FN errors because the class representing diseased people is the minority class.

The objective of a cost-sensitive approach is to build a classifier with the lowest misclassification cost among all the examples in a test dataset. We can reduce certain types of error (either FP or FN) by assigning a higher cost for such misclassifications (FN in our case). We have used a cost-sensitive learning method known as MetaCost to demonstrate how to take account of non-uniform misclassification costs. The MetaCost algorithm takes in misclassification costs usually determined by the subject experts. Assigning a high misclassification cost to FN errors will result in reducing the FN errors and increasing the sensitivity of a classification model. Table 6 depicts the results of using MetaCost in combination with the Decision Tree (J48) and Random Forest on some of the datasets with highly imbalanced or skewed class distributions.

It is clear from Table 6 that sensitivity increases and specificity decreases with the increasing misclassification cost for FN errors. These findings show that using a cost-sensitive technique can overcome the bias of a classifier towards the majority class up to some extent i.e. the tradeoff between specificity and sensitivity can be adjusted according to requirements. It is pertinent to observe from Table 7 that even after applying MetaCost only 20% and 15% of the diseased class instances of the thoracic surgery dataset (a highly class-imbalanced dataset) were classified correctly by the J48 and RF classification algorithms, respectively. This indicates the need for more research to correct the problem of class imbalance.

Figure 5 depicts the trend for sensitivity, specificity and accuracy with the increasing misclassification cost of FN errors for J48 and RF classifiers for two datasets with high class imbalance. The figure shows that sensitivity increases and specificity and accuracy decrease with the increasing misclassification costs for FN errors. The important conclusion is that a classifier with slightly low accuracy is not essentially a bad classifier for medical diagnosis because we are interested in the tradeoff between sensitivity and specificity. A classifier with a slightly low accuracy and high sensitivity rate may be considered better for medical diagnosis. The balance of tradeoff between FN and FP errors is finally an expert's decision.

CONCLUSION

The application of machine learning algorithms can only succeed and be useful in fields like medical or fault diagnosis if the domain is well understood and a proper evaluation framework is practised. The common approach for the performance evaluation of

classification algorithms for medical diagnosis has many issues such as selection of re-sampling techniques, choice of appropriate evaluation metrics and statistical significance testing. Since as a part of experimental design these choices influence the performance of classifiers, we have to be extremely careful in designing experiments for selecting classification algorithms for a sensitive field like medical diagnosis. In this study, we proposed and demonstrated a performance evaluation framework for comparing multiple classification algorithms across multiple datasets for the purpose of medical diagnosis. This approach will help researchers in sensitive fields like medical diagnosis, where misclassification costs are high, to adopt an appropriate approach for evaluating classification algorithms.

The nature of medical data makes it difficult to quickly select, tune and apply classification algorithms to disease diagnosis. The classification algorithms for disease diagnosis must specifically take care of missing values and skewed class distributions in medical datasets. Some techniques for replacing missing values need to be used and classifiers' bias towards the majority class needs to be taken care of either through sampling techniques or cost-sensitive analysis.

Accuracy, as the sole criteria for measuring the performance of classification algorithms in the domain of disease diagnosis, can be misleading in the presence of skewed class distributions. Keeping in view such circumstances, we need to focus on other criteria like sensitivity (rate of diseased persons classified correctly), specificity (rate of healthy persons classified correctly) and ROC. In domains like medical diagnosis, it is important for a classifier to perform well at discovering positive instances (belonging to the diseased class, which is actually the minority class) than to have an overall high accuracy rate. A classifier for medical diagnosis should keep the false negative rate (number of instances of diseased persons misclassified as healthy) to be as close to zero as reasonably possible. An expert's opinion is important for fixing the tradeoff between sensitivity and specificity. An expert may very well decide to have minimal FN errors. This may increase FP errors (healthy persons misclassified as diseased); however, these may be eliminated by conducting further tests on these instances.

Table 7

A cost-sensitive analysis using metacost with j48 and rf classifiers on four datasets

Cost FP:FN	Datasets	J48				RF			
		SE	SP	SE*SP	Accuracy	SE	SP	SE*SP	Accuracy
1:1	Hepatitis	0.781	0.951	0.742	91.61	0.656	0.959	0.629	89.67
	ILPD	0.347	0.839	0.291	69.81	0.275	0.858	0.235	69.12
	Thoracic Surgery	0.014	0.983	0.013	83.82	0	0.988	0	84.04
	WPBC	0.383	0.848	0.324	73.73	0.191	0.993	0.189	80.30
	Hepatitis	0.75	0.951	0.713	90.96	0.719	0.902	0.648	86.45
1:2	ILPD	0.467	0.755	0.352	67.23	0.443	0.781	0.345	68.43
	Thoracic Surgery	0.057	0.94	0.053	80.85	0.029	0.978	0.028	83.61
	WPBC	0.447	0.795	0.355	71.21	0.255	0.993	0.253	81.81

Table 7 (continue)

Cost FP:FN	Datasets	J48				RF			
		SE	SP	SE*SP	Accuracy	SE	SP	SE*SP	Accuracy
1:3	Hepatitis	0.813	0.935	0.76	90.96	0.844	0.902	0.761	89.03
	ILPD	0.563	0.661	0.372	63.29	0.665	0.704	0.461	69.29
	Thoracic Surgery	0.071	0.903	0.064	77.87	0.043	0.95	0.04	81.48
	WPBC	0.468	0.788	0.368	71.21	0.277	0.947	0.262	78.78
	Hepatitis	0.844	0.894	0.754	88.38	0.844	0.878	0.741	87.09
1:4	ILPD	0.689	0.647	0.445	65.86	0.772	0.637	0.491	67.58
	Thoracic Surgery	0.129	0.845	0.109	73.82	0.114	0.93	0.106	80.85
	WPBC	0.426	0.662	0.282	60.60	0.362	0.834	0.301	72.22
	Hepatitis	0.813	0.821	0.667	81.93	0.875	0.846	0.740	85.16
1:5	ILPD	0.671	0.635	0.426	64.49	0.832	0.567	0.471	64.32
	Thoracic Surgery	0.2	0.783	0.156	69.57	0.157	0.895	0.140	78.51
	WPBC	0.596	0.629	0.374	62.12	0.426	0.722	0.307	65.15

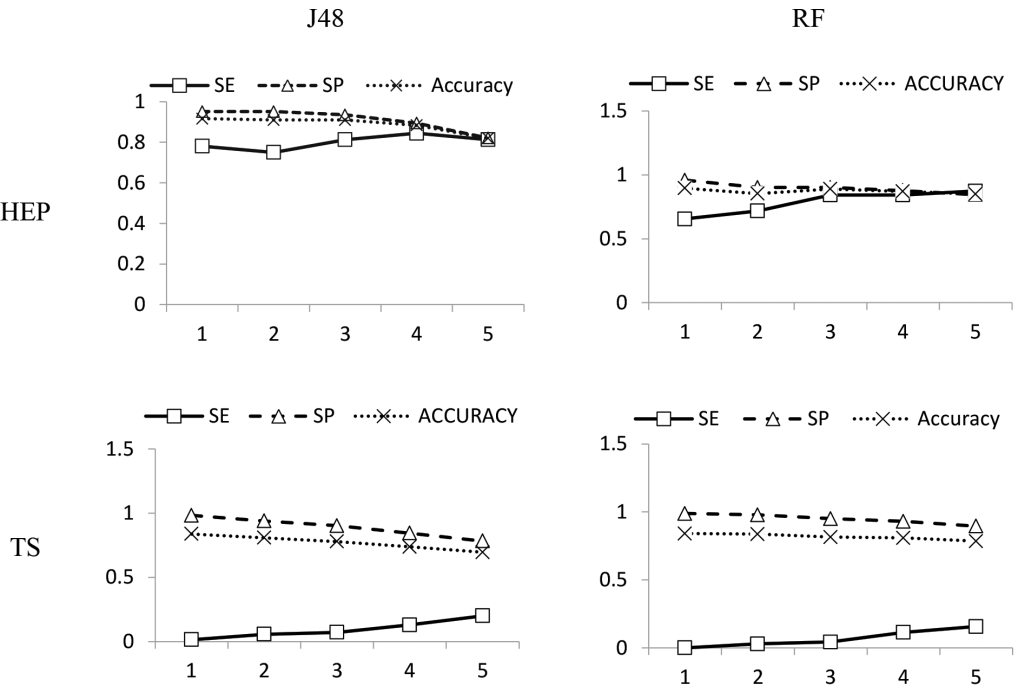


Figure 5. Change in sensitivity, specificity and accuracy with increasing misclassification cost for datasets using J48 and RF classifiers

REFERENCES

- Alcalá-Fdez, J., Sánchez, L., García, S., del Jesus, M. J., Ventura, S., Garrell, J. M., ... & Herrera, F. (2008). KEEL: A software tool to assess evolutionary algorithms for data mining problems. *Soft Computing*, 13(3), 307–318.
- Azar, A. T., & El-Metwally, S. M. (2012). Decision tree classifiers for automated medical diagnosis. *Neural Computing and Applications*, 23(7-8), 2387–2403.
- Bal, M., Amasyali, M. F., Sever, H., Kose, G., Demirhan, A., Bal, M., ... & Demirhan, A. (2014). Performance evaluation of the machine learning algorithms used in inference mechanism of a medical decision support system. *The Scientific World Journal*, 2014, 1-15.
- Bouckaert, R. R., Frank, E., Kirkby, R., Reutemann, P., Seewald, A., & Scuse, D. (2002). WEKA Manual for Version 3-7-8. *University of Waikato, New Zealand*, 588-595.
- Brink, H., & Richards, J. W. (2013). *Real-World machine learning* (7th Ed.). United States of America, USA: Manning Publications.
- Chang, C. L., & Chen, C. H. (2009). Applying decision tree and neural network to increase quality of dermatologic diagnosis. *Expert Systems with Applications*, 36(2, Part 2), 4035–4041.
- Chawla, N. V. (2005). Data mining for imbalanced datasets: An overview. In O. Maimon & L. Rokach (Eds.), *Data mining and knowledge discovery handbook* (pp. 853–867). Springer US.
- Chen, H. L., Yang, B., Wang, G., Wang, S. J., Liu, J., & Liu, D. Y. (2011). Support vector machine based diagnostic system for breast cancer using swarm intelligence. *Journal of Medical Systems*, 36(4), 2505–2519.
- Çınar, M., Engin, M., Engin, E. Z., & Ateşçi, Y. (2009). Early prostate cancer diagnosis by using artificial neural networks and support vector machines. *Expert Systems with Applications*, 36(3), 6357–6361.
- Corcoran, A. L., & Sen, S. (1994). Using real-valued genetic algorithms to evolve rule sets for classification. In *IEEE Conference on Evolutionary Computation Proceedings* (pp. 120–124). IEEE.
- Danjuma, K. J., & Osofisan, A. O. (2014). Evaluation of predictive data mining algorithms in erythematous disease diagnosis. *International Journal of Computer Science Issues*, 11(6), 85–94.
- Danjuma, K. J. (2015). Performance evaluation of machine learning algorithms in post-operative life expectancy in the lung cancer patients. *International Journal of Computer Science Issues*, 12(2), 1-11.
- Demsar, J. (2008). On the appropriateness of statistical tests in machine learning. In *Proceedings of ICML'08*. Association for Computing Machinery.
- Domingos, P. (1999). MetaCost: A general method for making classifiers cost-sensitive. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 155–164). ACM.
- Drummond, C. (2006). Machine learning an experimental science (revisited). In *AAAI'06 Workshop on Evaluation Methods for Machine Learning* (pp. 1–5).
- Esfandiari, N., Babavalian, M. R., Moghadam, A.-M. E., & Tabar, V. K. (2014). Knowledge discovery in medicine: Current issues and future trend. *Expert Systems with Applications*, 41(9), 4434–4463.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, 27(8), 861–874.

- Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 30(1), 27–38.
- Fidelis, M. V., Lopes, H. S., & Freitas, A. A. (2000). Discovering comprehensible classification rules with a genetic algorithm. In *Evolutionary Computation, 2000. Proceedings of the 2000 Congress on* (Vol. 1, pp. 805–810). IEEE.
- Freitas, A. A. (2013). *Data mining and knowledge discovery with evolutionary algorithms*. New York, NY: Springer Science and Business Media.
- Hassanien, A. E., & Kim, T. (2012). Breast cancer MRI diagnosis approach using support vector machine and pulse coupled neural networks. *Journal of Applied Logic*, 10(4), 277–284.
- Howell, D. C. (2007). The treatment of missing data. *The SAGE handbook of social science methodology* (pp. 212–226). Great Britain: SAGE Publications Ltd.
- Hubbard, R., & Lindsay, R. M. (2008). Why P values are not a useful measure of evidence in statistical significance testing. *Theory and Psychology*, 18(1), 69–88.
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Med*, 2(8), e124.
- Japkowicz, N., & Shah, M. (2014). *Evaluating learning algorithms: A classification perspective* (1st Ed.). United States of America, USA: Cambridge University Press.
- Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5), 429–449.
- Jin, B., Tang, Y. C., & Zhang, Y.-Q. (2007). Support vector machines with genetic fuzzy feature transformation for biomedical data classification. *Information Sciences*, 177(2), 476–489.
- Kim, J., Choi, K., Kim, G., & Suh, Y. (2012). Classification cost: An empirical comparison among traditional classifier, cost-sensitive classifier, and metacost. *Expert Systems with Applications*, 39(4), 4013–4019.
- Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 30(1), 25–36.
- Kumar, D. (2011). A Genetic algorithm with entropy based probabilistic initialization and memory for automated rule mining. In *Advances in Computer science and information technology* (pp. 604–613). Springer, Berlin, Heidelberg.
- Ling, C. X., & Sheng, V. S. (2011). Cost-sensitive learning and the class imbalance problem. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 231–235). Springer US.
- Marlin, B. (2008). Missing data problems in machine learning. (Doctor of Philosophy in Dept of Computer Science). University of Toronto, Canada.
- Newby, D., Freitas, A. A., & Ghafourian, T. (2013). Coping with unbalanced class data sets in oral absorption models. *Journal of Chemical Information and Modeling*, 53(2), 461–474.
- Parpinelli, R. S., Lopes, H. S., & Freitas, A. A. (2002). Data mining with an ant colony optimization algorithm. *IEEE Transactions on Evolutionary Computation*, 6(4), 321–332.
- Peña-Reyes, C. A., & Sipper, M. (2000). Evolutionary computation in medicine: An overview. *Artificial Intelligence in Medicine*, 19(1), 1–23.

- Prati, R. C., Batista, G. E. A. P. A., & Monard, M. C. (2011). A survey on graphical methods for classification predictive performance evaluation. *IEEE Transactions on Knowledge and Data Engineering*, 23(11), 1601–1618.
- Salzberg, S. L. (1997). On comparing classifiers: pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1(3), 317–328.
- Stekhoven, D. J., & Bühlmann, P. (2012). MissForest - Nonparametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), 112–118.
- Tanwani, A. K., Afridi, J., Shafiq, M. Z., & Farooq, M. (2009). Guidelines to select machine learning scheme for classification of biomedical datasets. In *Proceedings of the 7th European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics* (pp. 128–139). Berlin, Heidelberg: Springer-Verlag.
- Vanaja, S., & Rameshkumar, K. (2015). Performance analysis of classification algorithms on medical diagnoses-a survey. *Journal of Computer Science*, 11(1), 30–52.
- Venables, W. N., Smith, D. M., & R Development Core Team. (2008). *An Introduction to R*. United Kingdom, UK: Network Theory Ltd.
- Weiss, G. M., McCarthy, K., & Zabar, B. (2007). Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs? In R. Stahlbock, S. F. Crone, & S. Lessmann (Eds.), *Proceedings of the 2007 International Conference on Data Mining, DMIN 2007* (pp. 35–41). CSREA Press.
- Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques* (3rd Ed.). Burlington, MA: Morgan Kaufmann.
- Yousef, W. A., Wagner, R. F., & Loew, M. H. (2006). Assessing classifiers from two independent data sets using ROC analysis: A nonparametric approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11), 1809–1817.

