Sindh Univ. Res. Jour. (Sci. Ser.) Vol. 47(3):389-394 (2015)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Performance Evaluation of Classification Methods for Heart Disease Dataset

A. SHAIKH[++], N. MAHOTO, F. KHUHAWAR*, M. MEMON

Department of Software Engineering, Mehran UET, Jamshoro, Sindh, Pakistan

Received 12th April 2015 and Revised 24th August 2015

Abstract: Classification techniques are greatly deployed in several application domains for the purpose of classification, estimation and prediction. This paper evaluates the performance of three basic classifiers (i.e., k-NN, Naïve Bayesian and Decision Tree) for the highly sparse dataset (i.e., medical data), which has been acquired from online machine learning repositories. The goal of the study is to evaluate performance of three basic classifiers, since each classifier works at different learning mechanism and to predict possible treatment outcomes based on well-known patterns of the treatments. The experimental results reveal that k-NN classifier produces better prediction results as compared to other two classifiers for the considered heart disease dataset. Decision Tree classifier predicts poorly for the sparse dataset, specially, when dataset comprises of varying attribute values.

Keywords: Classifier, k-NN, Naïve Bayesian, Decision Tree, performance

1. INTRODUCTION

A large amount of data is everyday being stored electronically about patients' records, their clinical test records and diagnostic examinations by healthcare service providers (Tongsiri, 2013), (Baralis *et al.,* 2010), (Mahoto, *et al.,* 2015). This vast amount of data is hard to process and examine with the help of conventional processes, since this data is very much multifaceted (Antonelli *et al.,* 2013). Furthermore, this data is highly diversified in nature; therefore, it becomes a complex application domain (Cios and Moore, 2002), (Mahoto, 2013). Healthcare comes in the complex application domain, because of the fact that medical data is highly dimensional (Mahoto, 2013); each individual may have its own medical history and separate disease complications (Mahoto, *et al.,* 2014). In this domain, getting the meaningful and understandable information is one of the challenging and open issues for the research (Mahoto, 2013).

Data mining has been profitably used for the better services; healthcare management can also get the benefits by exploiting data mining techniques for immense medical data (Mahoto, *et al.,* 2014), (Mahoto, *et al.,* 2015), (Baralis *et al.,* 2012). In healthcare domain, a number of attempts are carried out in terms of research (Koh, *et al.,* 2011), (Mahoto, *et al.,* 2014), (Mahoto, *et al.,* 2014), (Mahoto, *et al.,* 2015), (Baralis *et al.,* 2012), (Antonelli *et al.,* 2012). Data mining transforms large collection of data into useful and non-trivial information to make better decisions. Data mining techniques, for instance clustering, classification, association analysis and sequential pattern mining for pattern discoveries are widely used in several domains including healthcare (Gupta, *et al.,* 2011), (Mahoto, *et al.,* 2014). The classification technique – also known as supervised learning is well-established and popular data mining technique that is greatly used in text classification, estimations and predictions. The aim of classification technique is to assign class labels to new cases (i.e., unknown data points) based on already known records (i.e., training set) for the given data (Aggarwal and Philip, 2008). The k-Nearest Neighbors (k-NN), Naïve Bayesian, CART, Decision Tree, Support Vector Machine (SVM) and Neural Networks are among the well-known classification methods. The performance of each classification technique relies on its accuracy rate.

This study uses three classification methods namely k-NN, Naïve Bayesian and Decision Tree for evaluating their performance for the medical data. In particular, classifier learns using three algorithms separately and predicts possible treatment outcomes based on well-known patterns of the treatments. The considered learning algorithms (i.e., k-NN, Naïve Bayesian and Decision Tree) cover different learning methods of classification. For instance, Decision Tree algorithm classifies given data points based on inductive learning, Naïve Bayesian algorithm works on probabilities basis, and k-NN algorithm uses distance measures to train the classification model. Thus, considered algorithm represent the broader domain of classification methods. The prediction of the treatment patterns may help the healthcare experts as well as management to get benefits the extracted information. The healthcare management can also effectively utilize human resources and medical equipment's.

[++]Corresponding Author Email: anoud.shaikh@faculty.muet.edu.pk ; Tel.: +92-22-2771351 ext. 6907

* Institute of Advanced Research Studies in Chemical Sciences, University of Sindh, Jamshoro, Sindh, Pakistan

The rest of the paper is structured as: Section 2 reports classification techniques carried out in medical domain, Section 3 presents classifier designing and Section 4 discusses several experiments performed using three algorithms on a heart disease dataset. Finally conclusions are reported in Section 5.

2. LITERATURE REVIEW

In medical domain, a vast number of researches are being carried out due to the fact that it is inherently complex application domain. For instance, data mining techniques have been applied on breast cancer (Seera and Lim, 2014), colon cancer (Antonelli *et al.,* 2012), heart disease (Shouman, *et al.,* 2011), (Shouman, *et al.,* 2012), diabetes (Antonelli *et al.,* 2013), (Baralis *et al.,* 2010), (Mahoto, *et al.,* 2014), and eye patients (Mahoto, *et al.,* 2014). Especially several data mining techniques are exploited, for instance, classification technique (Shouman, *et al.,* 2011), sequential pattern mining (Baralis *et al.,* 2010), (Mahoto, *et al.,* 2014), clustering (Antonelli *et al.,* 2013), (Mahoto, *et al.,* 2014).

Classification technique, particularly, is also widely used in medical domain. For example, improving Naïve Bayes classifier's accuracy (Abraham, *et al.,* 2006), hybrid intelligent model (Seera and Lim, 2014), and hybrid classification approach (HCA) (Chen, *et al.,* 2014) for medical dataset. The study in (Soni *et al.,* 2011) presents data mining techniques for heart disease prediction. In particular, decision tree, k-NN, neural networks and Bayesian classification are discussed.

The study in (Shouman, Turner and Stocker, 2012) applies k-NN for heart disease dataset and showed its improved results. However, this paper investigates the other classification algorithms on the heart disease dataset. The focus of this study is to investigate the different aspects of the medical data, when classification methods (i.e., k-NN, Naïve Bayesian and Decision tree) are applied. For example, what would be the impact of number of attributes in dataset on the performance of classifier? How does training set affect the accuracy of the classifier? What is the behavior of the considered algorithms when applied on sparse dataset (i.e., medical dataset)?

3. PREDICTING MEDICAL DATA

Classification technique for prediction relies on the accuracy of the classification model. This study implies supervised learning methods for predicting medical dataset (i.e., heart disease). The heart disease dataset has been acquired from online UCI Repository[1]. The online repository dataset are freely available for research purposes. The heart disease dataset has been considered for the study, because this dataset comprises of large number of attributes (i.e., 14 attributes) with varying number of values. The processed dataset comprises of 294 records with 14 attributes (see details in (Table-1).

The **(Fig. 1)** illustrates building a classification model. This study applies supervised learning algorithms namely - k-NN, Naïve Bayes and Decision Tree. Once a classification model is built, unseen/ unknown data are provided as an input for the classification/prediction. The predicted results can be evaluated using evaluation methods. This study exploits confusion metrics to evaluate the classifier performance.

3.1 Building Classification Model

The details of each component of Classification model are described in the following:

Training Set - The dataset, which is used to train the classifier during its learning process. The training set used in building classification model is subset of the entire dataset. The records, which satisfy the following criterion, are included in the training set:

- Dissimilar attributes values with respect to rest of the dataset
- Containing all different attributes values available in the dataset

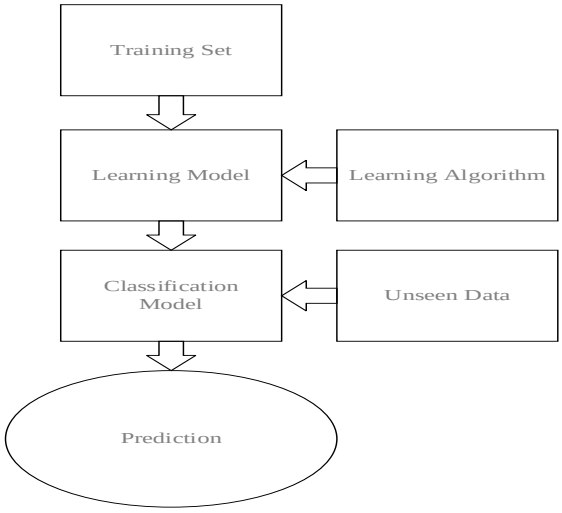


Fig. 1, Classification Modeling

Learning Model - The learning model is the classifier that learns known patterns from the training set based on learning method/algorithm.

Learning Algorithm- The algorithms that make classifier capable to classify/predict are applied. This study applies supervised learning algorithms for performance evaluation of these algorithms in medical

[1] URL: http://archive.ics.uci.edu/ml/datasets/Heart+Disease

domain (i.e., heart disease dataset). The details of each algorithm are reported in the following.

1. **Decision Tree:** Decision Tree builds tree structure for the classification problem. It incrementally develops tree structure, in parallel to breaking down given dataset into small subsets. The final nodes in the tree are leaf nodes or decision nodes. Decision Tree classification works on the core ID3 algorithm that implies entropy or information gain to build decision tree.

2. **Naïve Bayes algorithm:** Naïve Bayes algorithm is greatly applied in classification and prediction problems. This algorithm is also widely used for text classification, because of its nature of probabilistic conditioning. Naïve Bayesian algorithm works on the basis of Bayes theorem, which is easy and free from iterative estimation parameters such as found in k-NN algorithm. Due to its simplicity, this algorithm is very useful for large datasets and produces promising outcomes. Bayes theorem considers the value of unknown data point for a given dataset is independent of other unknown data points. Naïve Bayesian classifier considers each individual data point has independent distribution, therefore, it estimates the probabilities of each unknown data point, as given below:

$$P\left(\frac{c_j}{d}\right) = \frac{P\left(\frac{d}{c_j}\right)P(c_j)}{P(d)}$$

Where $P(c_j/d)$ is the probability of instance d (i.e., unknown data point) to be found in the class c_j (i.e., known data points), $P(d/c_j)$ shows the probability of instance d in a given class c_j, $P(c_j)$ is the probability of occurrence of class c_j and $P(d)$ is the probability of occurrence of instance d.

3. **k-NN algorithm:** k-nearest neighbor (k-NN) algorithm has been widely used for classification, estimation and prediction (Moreno-Seco, *et al.*,, 2003), (Larose, 2005). This algorithm, being instance-based learning, compares the unknown records with the most similar stored known trainee data. The k-NN algorithm based on training set selects the most near known data point and label/predict the unknown data point. The k represents the number of nearest neighbors for the unknown data point. For instance, if k=2 then k-NN would choose the two closest known data points from the known data points to classify the new unknown data point. The similarity between data points is measured by various distance measures. For example, Euclidean distance function can be used for similarity measurement (Larose, 2005).

Classification Model - The classification model is the actual classifier, which is capable to predict the unknown dataset based on learned patterns.

Unseen data- The dataset, which undergoes for the prediction/classification.

Prediction - The outcomes/predictions of the classifier.

3.2 Performance Evaluation Metrics

The performance of classifier/prediction model is measured using evaluation metrics such as Precision, Recall, Accuracy and F-measure. In particular, the performance of the classification model is measured to distinguish between actual and predicted class/label. For example, consider two classes/labels {Positive, Negative} for the simplicity.

Given a classification model and an instance (i.e., unknown data point), there are four possible classification model outcomes. If the instance is POSITIVE and it has been predicted as POSITIVE, it means it is considered as TRUE POSITIVE (TP). In case it has been predicted as NEGATIVE, it means it is considered as FALSE NEGATIVE (FN).

Likewise, if the instance is NEGATIVE and it has been predicted as NEGATIVE, it means it is considered as TRUE NEGATIVE (TN). In case the instance has been predicted as POSITIVE, it means it is assumed as FALSE POSITIVE (FP). These predicted and actual values form confusion matrix (also contingency table) (Fawcett, 2004).

The performance of the classification model is computed based on confusion matrix. This matrix is the base for the common evaluation metrics such as precision, recall, accuracy and F-measure (Fawcett, 2004).

Precision - This is the positive predictive value (PPV).

$$precision = \frac{TP}{TP + FP}$$

Recall - This is the sensitivity. In the expression of recall, P is the actual total POSITIVE data points.

$$recall = \frac{TP}{P}$$

Accuracy - This value presents the correctness of the classification model in predicting unknown data points. In the expression of accuracy, N is the actual total NEGATIVE data points, and P is actual total POSITIVE data points.

$$accuracy = \frac{TP + TN}{P + N}$$

F-measure - It is the harmonic mean of precision and recall (Sasaki, 2007).

$$F - measure = \frac{2}{1/precision + 1/recall}$$

4. EXPERIMENTAL RESULTS

Several experiments have been performed on the heart disease dataset. The entire dataset comprises of 14 different attributes/features, the details are given in **(Table-1)**

Table 1: Attributes of Heart Disease Dataset

Serial Number	Attribute Name	Distinct Possible Values
1	Age	38
2	Sex	02
3	Cp	04
4	Trestbps	32
5	Chol	144
6	Fbs	02
7	Restecg	02
8	Thalach	73
9	Exang	07
10	Oldpeak	06
11	Slope	11
12	Ca	5
13	Thal	5
14	Num	5

To evaluate the performance of each considered classification algorithm, the dataset has been segmented in different cases. Each case represents the distribution of dataset. The different cases considered in this study allows to evaluate the performance of the classification model with respect to varying number of attributes and their varying values in the dataset. For example, the attribute 'Chol' may have 144 different values and attribute 'Fbs' has only 2 possible values. In all cases, NUM attribute is used for the prediction, since NUM represent the complications for heart disease. Different cases used in this study are described in the following

Case 1: In this case, all the attributes (i.e., 14 attributes) of the dataset are used. Attributes: Age, Sex, CP, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, num

Case 2: Subset (i.e., 8 attributes) of the dataset is used in this case. Attributes: Age, Sex, num, trestbps, chol, restecg, thalach, ca

Case 3: Subset (i.e., 7 attributes) of the dataset is used in this case. Attributes: Ca, fbs, exang, oldpeak, slope, thal, num

The classifier results of each classifier considered in this study are reported in the following.

Fig. 2. reports the accuracy results achieved using k-NN algorithm for the three cases of the heart disease dataset, when training set is varying. When training set is 17% of the total dataset, k-NN classifies up to 78% accurate results. Likewise with 50% of training set, it produces 90% accurate results. It is observed that k-NN algorithm produces almost similar results in each case of the dataset. However, case 3 is having slightly more accuracy comparing with other two cases.

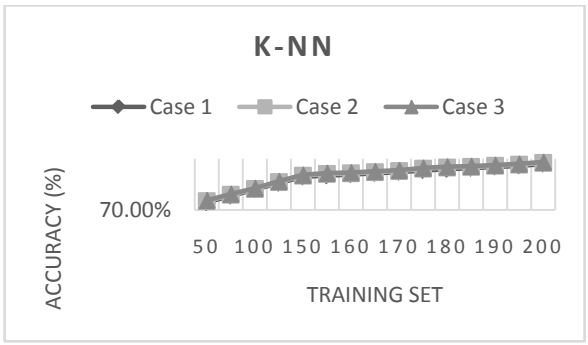


Fig. 2. k-NN Classification/Prediction Results

The classification/prediction results obtained using Naïve Bayesian and Decision Tree classifiers are presented in **(Fig. 3 4 and 5)** respectively. It is quite visible that case 3 in Naïve Bayesian is having higher accuracy that that of other two cases. Similarly, using Decision Tree, Case 3 results are more accurate than that of Case 1 and Case 2. However, among all three classifiers k-NN outperformed in performance in all three cases, then Naïve Bayesian and lastly Decision Tree.

The promising results are obtained in Case 3, this shows that classifiers better perform in case of less attributes as well as each individual attribute value ranges at smaller range.

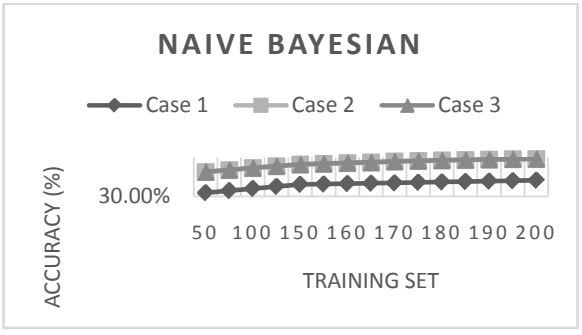


Fig. 3, Naïve Bayesian Classification/Prediction Results

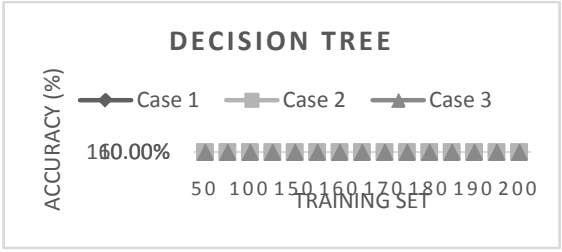


Fig. 4: Decision Tree Classification/Prediction Results

The Case 3 results are reported in **(Fig.5),** which shows the accuracy of each classifier.

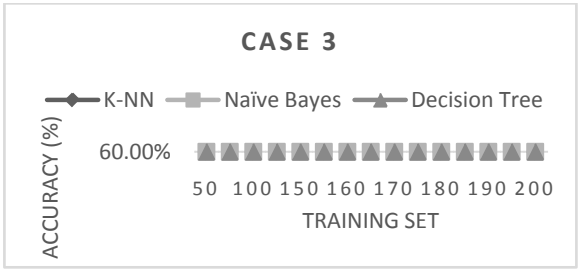


Fig. 5, Case 3 Results of classifiers

The Case 3 is taken as reference for the presentation, since in this case all three considered classifiers (i.e., k-NN, Naïve Bayesian and Decision Tree) produced higher accuracy rate. Therefore, for the Case 3 the performance of each classifier in terms of F-measure and Timing is shown in **(Fig.6, and 7)** respectively.

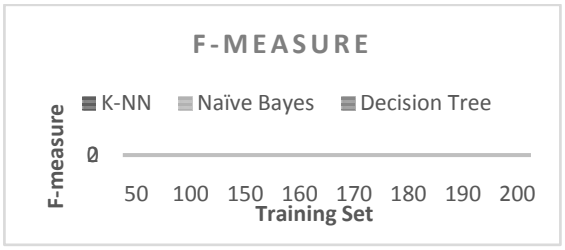


Fig.6, F-measure values of classifiers

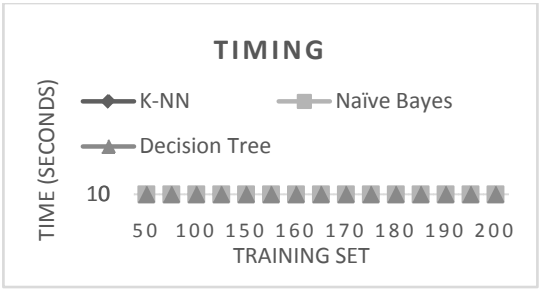


Fig. 7, Timing of each classifier

The performance evaluation of each classifier is shown in **(Fig. 8 and 9).** Particularly, Fig.8 presents the precision and accuracy for each classifier for the Case 3. Similarly, Fig. 9 shows the prediction rate of each classifier for the Case 3 dataset. Prediction rate shows the ratio of prediction of true values against the actual true values. From Fig. 8 and 9. it is observed that k-NN classifier predicted medical dataset with higher accuracy as well as precision. Similarly, k-NN has very less rate for FP (FP represents the false predictions). Therefore, from the experimental results it has been observed k-NN outperformed comparing with Naïve Bayesian and Decision Tree classifiers. Furthermore, k-NN classifier is significantly efficient for the considered dataset. These results are coherent with the study carried out in (Shouman, *et al.,* 2012).

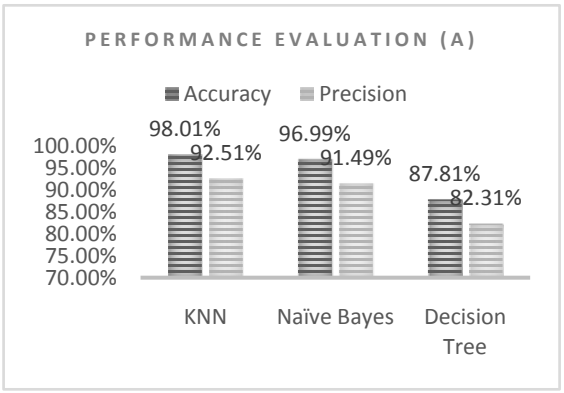


Fig. 8, Accuracy and Precision for Case 3

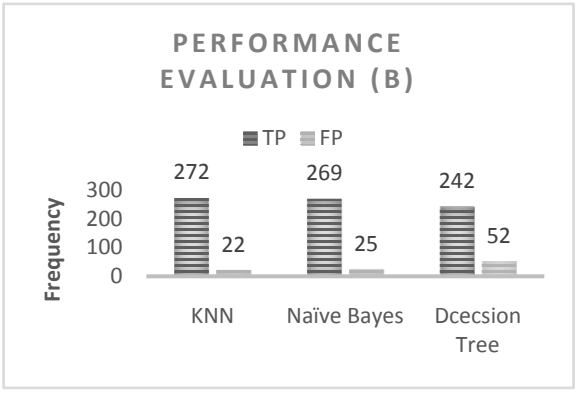


Fig. 9, TP rate and FP rate for Case 3

The possible reason behind better accuracy of k-NN may be the working principle of k-NN algorithm, which is distance measure of data points for the similarity. Decision Tree algorithm needs inductive learning, thus, produced poor prediction results, because of sparse nature of medical dataset. However, Naïve Bayesian resulted in promising results because of its probabilistic approach. Furthermore, k-NN and Naïve Bayesian classifiers take more timing as compared to Decision Tree classifier as shown in Fig.7.

5. CONCLUSION

The study evaluated the performance of three classifiers namely k-NN, Naïve Bayesian and Decision Tree on the highly sparse dataset (i.e., Medical dataset) of Heart Disease patients. It has been observed that k-NN classifier performed better than the other two classifiers (i.e., Decision Tree and Naive Bayes) for the considered datasets. To evaluate the performance of each different classification method (i.e., different working principle of each learning algorithms), the dataset has been portioned into three different cases. These cases represent the diversity of medical dataset. In particular, each case reports the different combination of attributes, and each attribute may have different values. Thus, aim is to evaluate each learning algorithm at different variations related to highly dimensional data. The experimental results based on the considered

dataset represent that k-NN predicted the significantly better results up to 98.01% accuracy rate, when training set is increased higher than 60% of the total dataset.Naïve Bayesian comes next after k-NN in producing higher accuracy in predicting; while Decision Tree produced less accurate results.

REFERENCES:

Abraham, R., J. B. Simha, and S. Iyengar, (2006) 'A Comparative Analysis of Discretization Methods for Medical Datamining with Naive Bayesian Classifier', 9th International Conference on Information Technology, ICIT'06, 235-236.

Aggarwal, C. C. and S.Y. Philip, (2008) *A General Survey of Privacy-Preserving Data Mining Models and Algorithms*, Springer.

Antonelli, D., E. Baralis, G. Bruno, T. Cerquitelli, S. Chiusano, and N. Mahoto, (2013) 'Analysis of Diabetic Patients Through Their Examination History', *Expert Systems with Applications*, vol. 40, 4672-4678, Available: http://www.sciencedirect.com/science/article/pii/S0957417413001206.

Antonelli, D., E. Baralis, G. Bruno, S. Chiusano, N.A. Mahoto, and C. Petrigni, (2012) 'Analysis of Diagnostic Pathways for Colon Cancer', *Flexible Services and Manufacturing Journal*, vol. 24,. 379-399, Available:http://dx.doi.org/10.1007/s10696-011-9095-2.

Baralis, E., G. Bruno, S. Chiusano, V.C.Domenici, N.A. Mahoto, and C. Petrigni, (2010) 'Analysis of Medical Pathways by Means of Frequent Closed Sequences', Proceedings of the 14th International Conference on Knowledge-based and Intelligent Information and Engineering Systems:Part 3 Berlin, Heidelberg, 418-425.

Baralis, E. M., S. A. Chiusano, N. A. Mahoto, D. Antonelli, G. Bruno, and C. Petrigni, (2012) 'Extraction of Medical Pathways from Electronic Patient Records', in *Medical Applications of Intelligent Data Analysis: Research Advancements / Rafael Magdalena-Benedito, Emilio Soria, Juan Guerrero Martinez, Juan Gomez-Sanchis, Antonio Jose Serrano-Lopez*, Information Science Reference (IGI Global).

Chen, X.Y., B. Liu, and X. Xia, (2014) 'A Hybrid Approach for Cases Classification of Medical Data', *Applied Mechanics and Materials*, vol. 496, 1965-1970.

Cios, K.J. and G.W.Moore, (2002) 'Uniqueness of Medical Data Mining', *Artificial Intelligence in Medicine*, vol. 26, no. 1, 1-24.

Fawcett, T. (2004) 'ROC graphs: Notes and Practical Considerations for Researchers', *Machine Learning*, vol. 31, 1-38.

Gupta, S., D. Kumar, and A. Sharma, (2011) 'Performance Analysis of Various Data Mining Classification Techniques on Healthcare Data', *International Journal of Computer Science and Information Technology (IJCSIT)*, vol. 3, no. 4, pp. 155-169, Available: DOI: 10.5121/ijcsit.2011.3413.

Koh, H. C., G. Tan, and others (2011) 'Data Mining Applications in Healthcare', *Journal of Healthcare Information Management*, vol. 19, no. 2, 65Pp.

Larose, D.T. (2005) 'k-Nearest Neighbor Algorithm', in *Discovering Knowledge in Data*, John Wiley\& Sons, Inc., Available : http://dx.doi.org/10.1002/0471687545.ch5.

Mahoto, N. A., S F. K.haikh, and A. Q.Ansari, (2014) 'Exploitation of Clustering Techniques in Transactional Healthcare Data', *Mehran University Research Journal of Engineering and Technology*, vol. 33, no. 1, 77-92.

Mahoto, N.A., F. K. Shaikh, and B. Chowdhry, (2015) 'Innovative Architecture to Enhance Quality of Service for Laboratory Management Information Systems', in *Laboratory Management Information Systems: Current Requirements and Future Perspectives*, IGI Global.Mahoto, N.A., Shaikh, A. and Khuhawar, F.Y. (2014) 'CliTA: Discovering Knowledge from Clinical Data', *Sindh University Research Journal (Science Series)*, vol. 46, no. 4, 539-546.

Moreno-Seco, F., L. Mico, and J Oncina, (2003) 'A Modification of the LAESA Algorithm for Approximated k-NN Classification', *Pattern Recognition Letters*, vol. 24, no. 1, 47-53.

Sasaki, Y. (2007) 'The Truth of the F-measure', *Teach Tutor mater*, 1-5.

Seera, M. and C.P. Lim, (2014) 'A Hybrid Intelligent System for Medical Data Classification', *Expert Systems with Applications*, vol. 41, no. 5, pp. 2239-2249.

Shouman, M., T. Turner, and R. Stocker, (2011) 'Using Decision Tree for Diagnosing Heart Disease Patients', Proceedings of the Ninth Australasian Data Mining Conference, 23-30.

Shouman, M., T. Turner, and R. Stocker, (2012) 'Applying k-nearest Neighbour in Diagnosing Heart Disease Patients', *International Journal of Information and Education Technology*, vol. 2, no. 3, pp. 220-223.

Soni, J., U. Ansari, D. Sharma, and S.Soni, (2011) 'Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction', *International Journal of Computer Applications*, vol. 17, no. 8, 43-48.

Tongsiri, S. (2013) 'Electronic Health Records: Benefits and Contribution to Healthcare System', in *Next-Generation Wireless Technologies*, Springer.