European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 08, Issue 1, 2021

Predicting the possibility of cancer with supervised Learning Algorithms

I.Jeena Jacob[1], Archana S.Nadhan[2], Swasthika Jain[3], Spandan Gunti[4], D.Sathya[5]
[1,2,3,4]*Department of CSE, GITAM University, Bengaluru Campus*
[5]*Department of CSE, Kumaraguru College of Technology, Coimbatore*

Abstract
According to the statistics, the disease with which most of the women die is breast cancer. Lot of new cases and deaths because of this cancer places this disease as a major public health issue. The diagnosis of this disease at the starting stages will help the treatment by which the mortality can be reduced. This leads to extensive research in the diagnosis and classification of patients based on its malignancy. Lot of machine learning algorithms was used to diagnose this disease. This work analyses the various works done in this area. Also it shows the comparative study of those algorithms.

Keywords- Wisconsin breast cancer dataset, ***Nearest Neighbor, Support Vector Machines, Naïve Bayes and Decision Tree Algorithm***

I. INTRODUCTION

Many deadly diseases can be cured, if they can be diagnosed earlier. Such a disease is breast cancer [11-12], by which many middle aged women are affected. This has become a main source of mortality in women nowadays. If the disease can be diagnosed or if it can be predicted in early stage, this disease can be easily cured. Because of our easy accessibility to relevant datasets, many of the prediction algorithms give better results. This work helps to formulate a model which will predict the possibility of breast cancer based on the other conditions. The experimental analysis was done in the Wisconsin breast cancer dataset (WBCD). Majority of the cancer cells are being diagnosed with the help of efficient and latest methods. The classifications based on the various conditions will aid the physicians to have a prediction on the disease. This will help us to avoid the errors in physician's diagnosis since this information will give a caution note to the physician in a very lesser time. In turn, it will help the physicians to make an accurate decision.

II. RELATED WORK

The prediction of the value of the variable can be done based on the training data which has the known samples of data. Algorithm which is being used for the classification system plays the vital role in all the classification and prediction system. Many algorithms like Linear Regression (LR) [9], Multi-Linear Regression (MLR), Support Vector Machine (SVM) [1-2], K-Nearest Neighbor (KNN) classifiers [6], Neural Network [4-5],Random Forest [7], Bayes Classifier [8], Adaboost [10], etcare used in the classification works. KNN classifier is aninstance-based supervised learning classifierwhich uses the training data for learning.K level of validation in cross-level is done for training the data. First, the data are partitioned into k groups with same count. Validation of the algorithm is done with every individual group and other k-1 groups are used for training.

III. PROPOSED WORK

The classification or prediction efficiency of any variable depends on the other values and also on the algorithms with which the work is defined. Learning may be supervised, unsupervised or semisupervised. Guidance will be there in supervised learning and guidance will not be there in unsupervised learning. The dependent variable of this work will help us to identify M (Malign) or B(Benign). We used four different supervised algorithms, Nearest Neighbor, Support Vector Machines, Naïve Bayes and Decision Tree Algorithm

The data preprocessing should be done first which will help us to eradicate the noises from the data. The null or missing data should be found out and they should be deleted. The categorical data like country, age, etc need to be replaced with some numbers. For this encoding scheme is used. Thus the data will be prepared for applying the algorithm. These works need thorough data observation. The preprocessed data should be split into testing and training set. The training set is fed to the algorithm and they will help the algorithm to learn. The test set is used for

testing the accuracy.Scaling of the features should be done to aid the comparison easier. The normalization of the data to 0 to 1 or 0 to 100 should be done. The models are applied with the data.Fig.1 shows the Proposed Framework

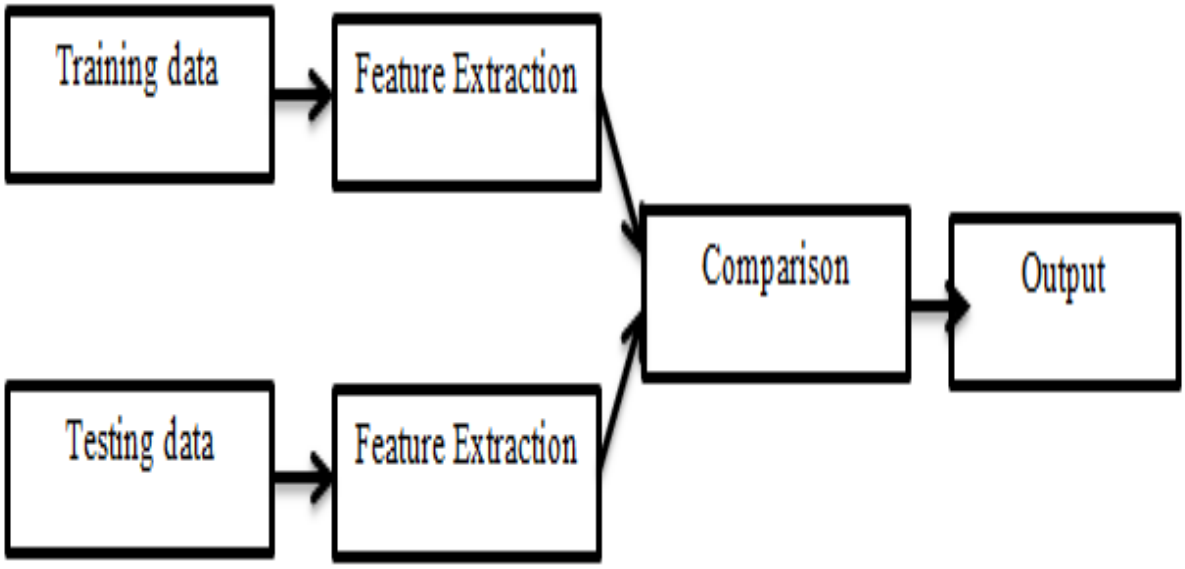


Fig.1: Proposed Framework

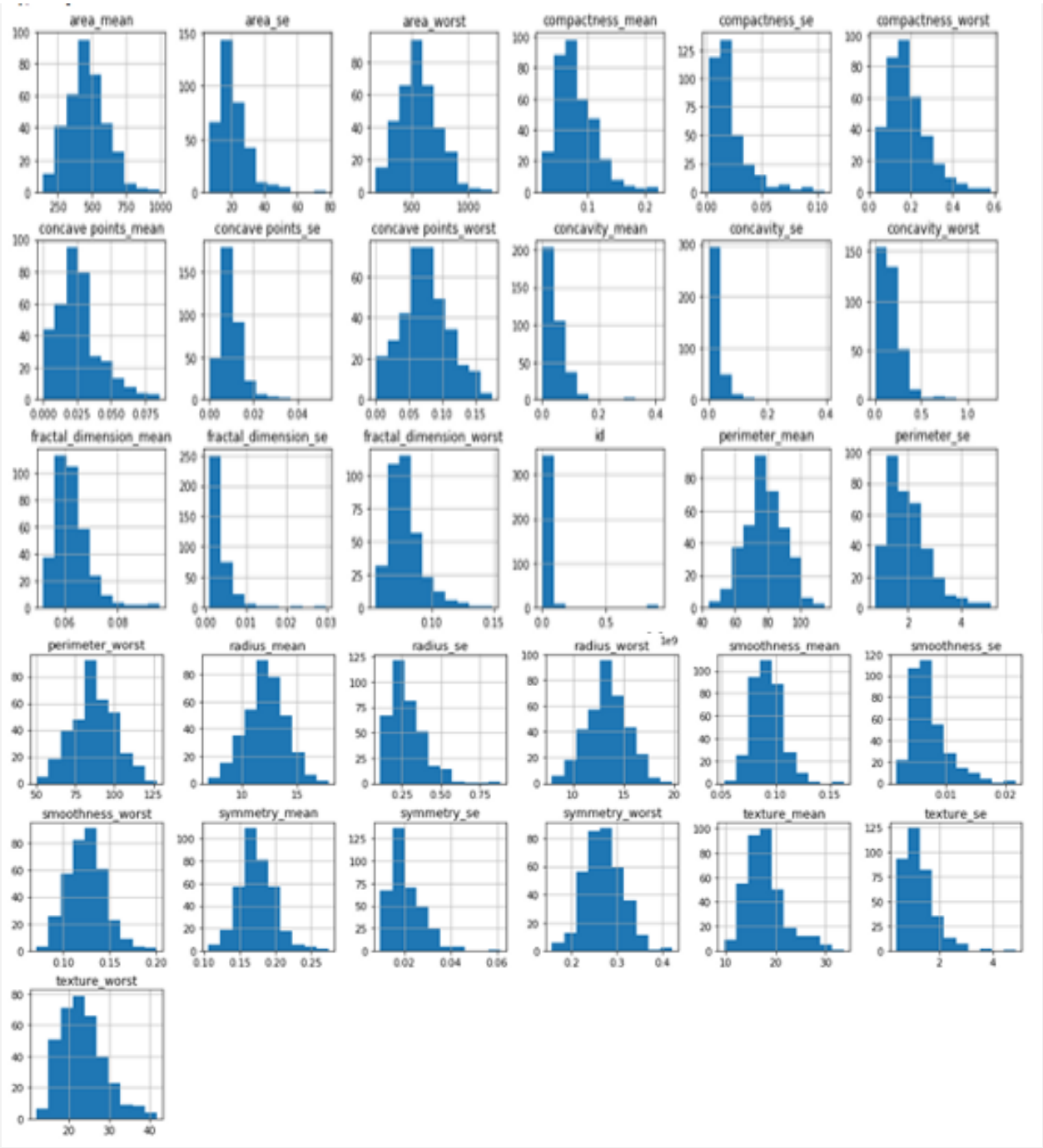
Algorithm:

1. Training data is fed for extracting the relevant features and indexing
2. Test data also fed into corresponding model and relevant features are extracted
3. Similarity measures are used to compare both the features of query data and training data
4. Test image is classified into that of training to which the difference is minimum

IV. EXPERIMENTAL ANALYSIS

4.1.Dataset

Breast Cancer Wisconsin (Diagnostic) Data Set [3] is used in this work. The dataset consists of two categories of cancers, benign and malignant with ten features of cell nuclei of concerned portion. The characteristics of cell nuclei are texture, radius, area, perimeter, smoothness, concavity, symmetry, compactness, concave points and fractal dimension. For these features, average of largest three numbers, standard error and mean are calculated for all the images. Thus total 30 features will be there in the dataset. Some features will be helping the physicians in prediction. One among them is mammography which will be helping doctors to identify cancer before 2 years. The women of age 40 to 45 are very much prone to this disease. Another feature is the women who have family history of breast cancer are prone to this disease. Figure 1 shows the visualization of the data.


area_mean
area_se
area_worst
compactness_mean
compactness_se
compactness_worst
concave points_mean
concave points_se
concave points_worst
concavity_mean
concavity_se
concavity_worst
fractal_dimension_mean
fractal_dimension_se
fractal_dimension_worst
id
perimeter_mean
perimeter_se
perimeter_worst
radius_mean
radius_se
radius_worst
1e9
smoothness_mean
smoothness_se
smoothness_worst
symmetry_mean
symmetry_se
symmetry_worst
texture_mean
texture_se
texture_worst

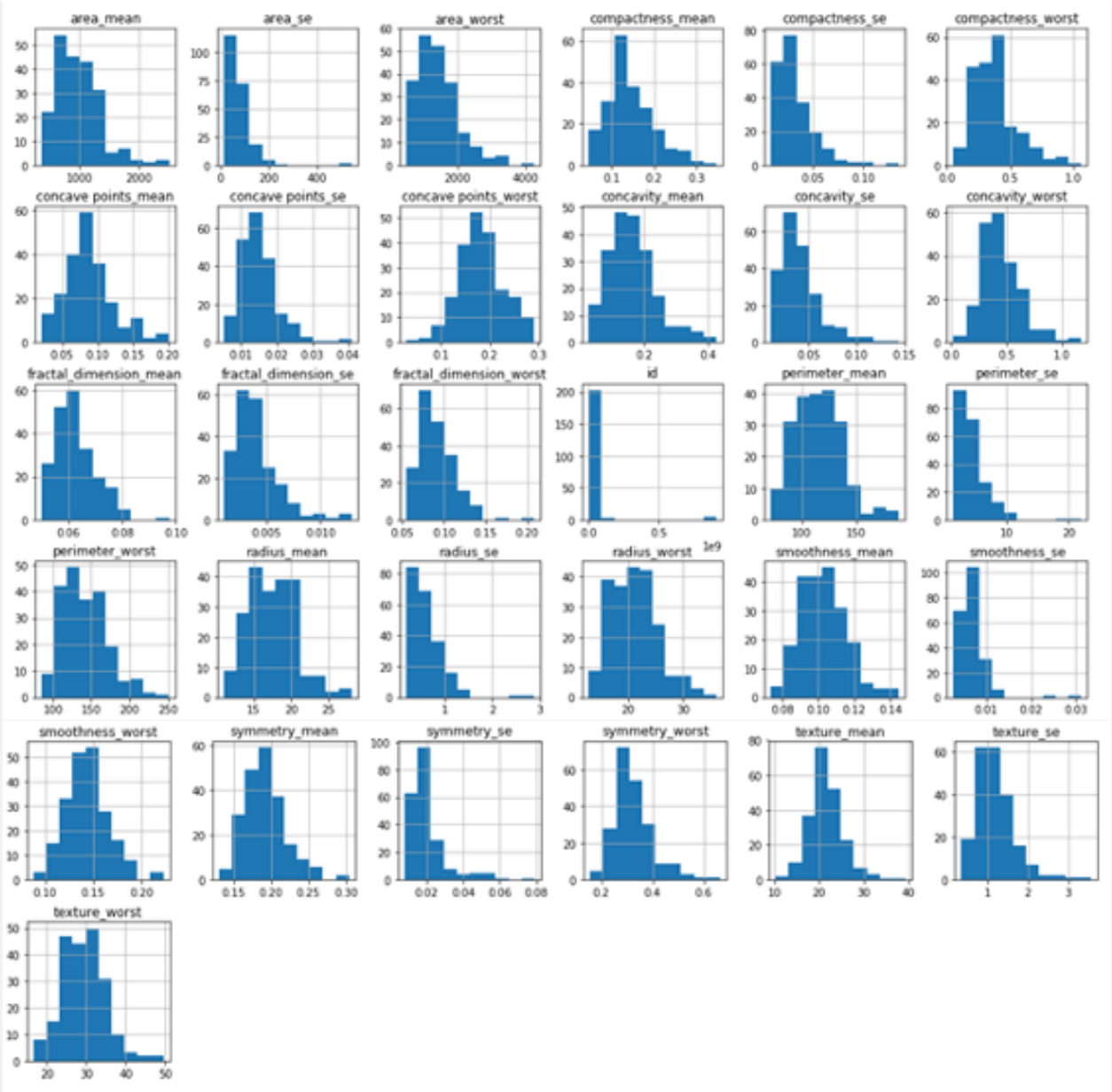


Fig.1 : Visualization of Dataset

The classification is done based on the various models. The accuracy is calculated by finding the ratio of number of correct predictions with total number of input data.To check the correct prediction we have to check confusion matrix object and add the predicted results diagonally which will be number of correct prediction and then divide by total number of predictions.

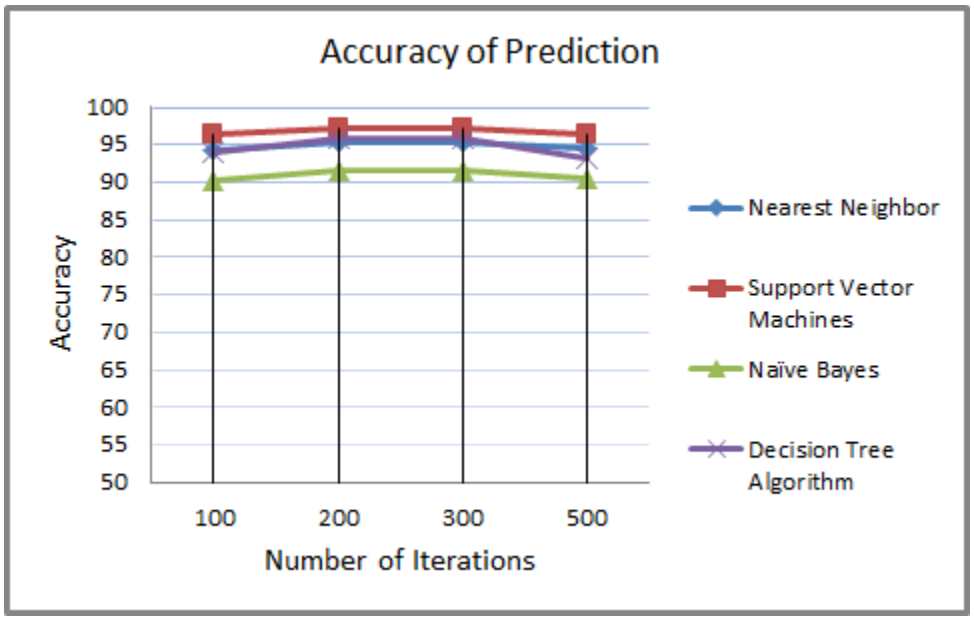


Fig.2: Classification accuracy with number of iterations

The classification is done based on different algorithms like Nearest Neighbor, Support Vector Machines, Naive Bayes and Decision Tree Algorithm. The Nearest Neighbor gives 95.1%, Support Vector Machines gives 97.2%, Naive Bayes gives 91.6% and Decision Tree Algorithm gives 95.8%. Figure 2 gives classification accuracy with number of iterations. Figure 3 gives the classification accuracy with percentage of test to train dataset.

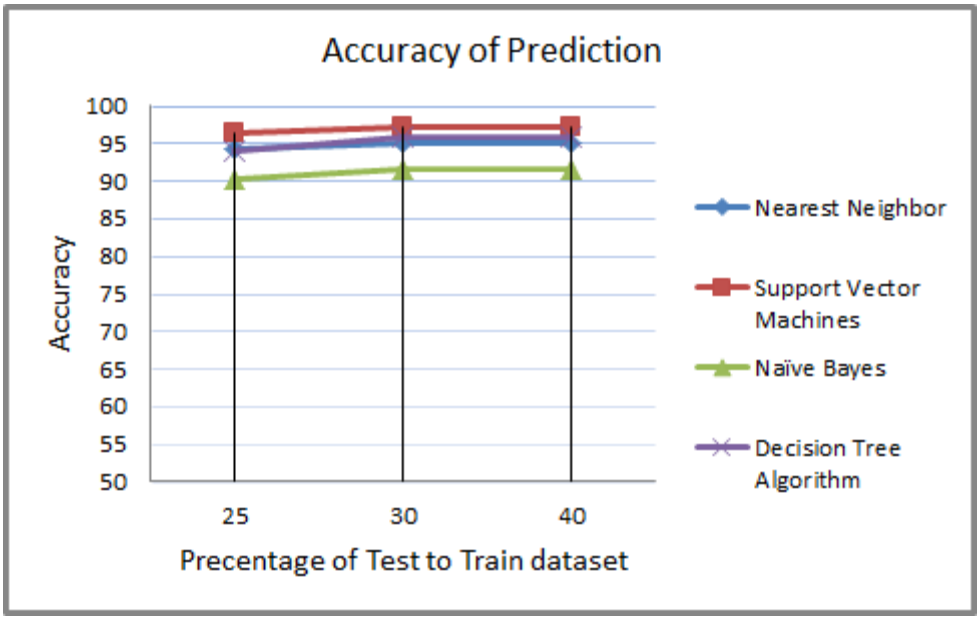


Fig.3: Classification accuracy with percentage of test to train dataset

V. CONCLUSION

Dataset classifications depend upon various properties of the relevant attributes in the dataset. In classification of cancerous cells, they should be categorized either as benign or as malignant. This work compares four different supervised algorithms for predicting the cancerous cells. The classification is done based on different algorithms like Nearest Neighbor, Support Vector Machines, Naive Bayes and Decision Tree Algorithm. Among these algorithms, SVM gives better result.

References

[1] Polat K, Güneş S. Breast cancer diagnosis using least square support vector machine. Digital Signal Processing. 2007 Jul 1;17(4):694- 701.
[2] Akay MF. Support vector machines combined with feature selection for breast cancer diagnosis. Expert systems with applications. 2009 Mar 1;36(2):3240-7.
[3] Marcano-Cedeño A, Quintanilla-Domínguez J, Andina D. WBCD breast cancer database classification applying artificial metaplasticity neural network. Expert Systems with Applications. 2011 Aug 1;38(8):9573-9.
[4] Nahato KB, Harichandran KN, Arputharaj K. Knowledge mining from clinical datasets using rough sets and backpropagation neural network. Computational and mathematical methods in medicine. 2015;2015
[5] Liu L, Deng M. An evolutionary artificial neural network approach for breast cancer diagnosis.InKnowledge Discovery and Data Mining, 2010.WKDD'10. Third International Conference on 2010 Jan 9 (pp. 593-596).
[6] Seyyid Ahmed Medjahed, TamazouztAitSaadi, AbdelkaderBenyettou. Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules. International Journal of Computer Applications (0975 - 8887)
[7] Cuong Nguyen, Yong Wang, Ha Nam Nguyen Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. J. Biomedical Science and Engineering, 2013, 6, 551-560
[8] Diana Dumitru. Prediction of recurrent events in breast cancer using the Naive Bayesian classification.2000 Mathematics Subject Classification.
[9] Turgay Ayer, MS; JagpreetChhatwal, PhD; OguzhanAlagoz, PhD; Charles E. Kahn, Jr, MD, MS; Ryan W. Woods, MD, MPH; Elizabeth S. Burnside, MD, MPH, MS. Comparison of Logistic Regression and Artificial Neural Network Models in Breast Cancer Risk Estimation. RadioGraphics 2010
[10] JarceThongkam, GuandongXu, Yanchun Zhang and Fuchun Huang. Breast Cancer Survivability via Adaboost Algorithm. HDKM '08 Proceedings of the second Australasian workshop on Health data and knowledge management
[11] RasoolFakoor, Faisal Ladhak, Azade Nazi, Manfred Huber. Using deep learning to enhance cancer diagnosis and classification. 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013
[12] Cancer Statistics, 2016. CA: A Cancer Journal for Clinicians