International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 10 • Issue 2 pp. 794-800 July 2018 www.csjournals.com

Primary Tumor Disease Detection Using Ensemble

[1]Sanjeev Kumar, [2]SunilaGodara
[1,2]Department of CSE, Guru Jambheshwar University of Science & Technology, Hisar, India

Abstract: Medical science diligence has enormous amount of data, but the majority of this data is not mined to find out concealed information in data. Diagnosing of Primary Tumor is one of important issue in medical decision support system. In this paper an attempt is made to analyze Decision Tree, Support Vector Machine (SVM), ensemble of Decision Tree and ensemble of SVM on Primary Tumor dataset. Performance of these techniques is measured through on basis of Precision, Recall, F measure, ROC, RMS Error and accuracy. Our analysis shows that SVM is more stable than Decision Treeclassifier to detect primary tumor with high accuracy and lowest error rate.Further ensemble of SVM outperforms other.

I.Introduction

A primary tumor is a tumor developing at the anatomical place where tumor progression began and proceeded to yield acancerous mass. The majority of cancers cell develop at their primary site but then go on to large size and spread to other parts of the body. These large tumors are secondary tumor. A primary tumor [1]begins in the brain or spinal cord. They often do not spread, even if they grow quickly. Some of the brain and spine tumors that do metastasize will usually spread only within the brain and spinal cord area and not to the rest of the body.Human brain represents only 2% mass of total body but uses 20% body's energy. Brain controls all the activities of the human body. So the brain needs to operate with its maximum efficiency. Now-a-days, a lot of people are suffering from brain tumor which causes even death, if not treated at time. Brain Tumor is a bunch of abnormal cells growing rapidly in the brain. It may happen to any human being at any age and appear at any location in the brain. Tumor is further categorized in two types: malign and benignant. Benignant tumor is the tumor which has homogeneous structure and don't have cancer cells while malign tumor has heterogeneous structure and contain cancer cells. Benign tumors are treated by either radio-logically or surgically and have rare chances of grow back. Malign tumor are life threatening and can be treated by chemotherapy,radiotherapy and their combination. So, need to diagnose the tumor at an early stage is essential for future treatments.Doctors utilize tumor grade and other factors, such as cancer phase and a patient's age and common health, to build up an action map and to decide a patient's prognosis (the probable result or path of a disease; the possibility of revival or repetition). Usually, a junior grade indicates aimproved prognosis. A higher-grade cancer may grow up and spread more rapidly and may require immediate or more hostile action.

ML has been known for evolving out some important features from large amount of data. Due to this specialty of ML, this field is used in combination with medical science for the accurate diagnosis of the patient disease. A number of ML techniqueshave been evolved and in order to achieve best accuracy of a model ensembles are widely used.P. Kalaiselvi et .al, [9] compared the performance of the different classifiers like Bagging, Dagging, Decorate, MultiBoostAB and Multi ClassClassifierand concluded that Bagging is best algorithm to finding the accuracy. YosvanyLlerena Rodriguez [10] worked on the essential detection of the pertinent segments using SVM, Naive Credal Classifier2 and MultiBoostAB. The experimental outcome showed that MultiBoostAB is the best classifier and gives better accuracy than two. Geoffrey I.Webb [17] used 36 delegate datasets from the UCI repository for their work and MultiBoostAB and AdaBoost are used. Between these Multi Boost gives lower error rate than AdaBoost. Divya Sravaniet.al, [18] had done work on on protein fold classification using different machine learning algorithms like SVM, MultiBoostAB, AdaBoost, K-NN and NN . In this paper we will compare the accuracy determined by the Decision Tree ,SVM ,Ensemble of Decision Tree and ensemble of SVM on primary tumor dataset obtained from the UCI Web Repository.

II. Primary Tumor Detection Models

Under this section we will discuss following data mining classification models to detect brain tumor:

A. Decision Tree

Decision tree is the powerful and greedy classification technique. The most accepted are Quinlan's ID3, C4.5 and CART and J48 algorithm. As the name implies, a tree is made in a top-down recursive divide and conquer manner. At begin, all the observations are at the root.Then the test attributes are chosen on the basis of some heuristic or statistical measure, (e.g. information gain). It splits the input observations into two or more parts. This process continues recursively until the complete tree is formed. The main objective is to obtain the variable-threshold pair which best splits the observations into subparts. The majorityof mathematical algorithms used for splitting includes Entropy based information gain.

B. Support Vector Machine

SVM is used for the classification of both linear and non-linear data. This technique is derived from statistical learning theory given by Vipnik in 1992. SVM technique solvesthe problem by finding out the hyper-plane with maximum margin.For nonlinearlyseparable data, it transformsthe training data into a higher dimension space by doing non linear mapping. By transforming it into high dimensional space, it searches for linear optimal separating hyper-plane. This transformation technique into high dimension always helps in searching for an optimal hyper-plane using support vectors and margins[13]. SVM achieved classification by finding optimal MMHand minimizing the classification errors.

C. Bagging

Bagging[9] is a machine learning ensemble meta-algorithm designed to improve the strength and accuracy of machine learning algorithms. It is used in statistical classification and regression. It also minimizes the variance and avoid overfitting. For example, M different trees or SVM can be trained on different subsets of the data (chosen randomly with substitute) and calculate the ensemble. Figure 1 shows working of bagging approach.

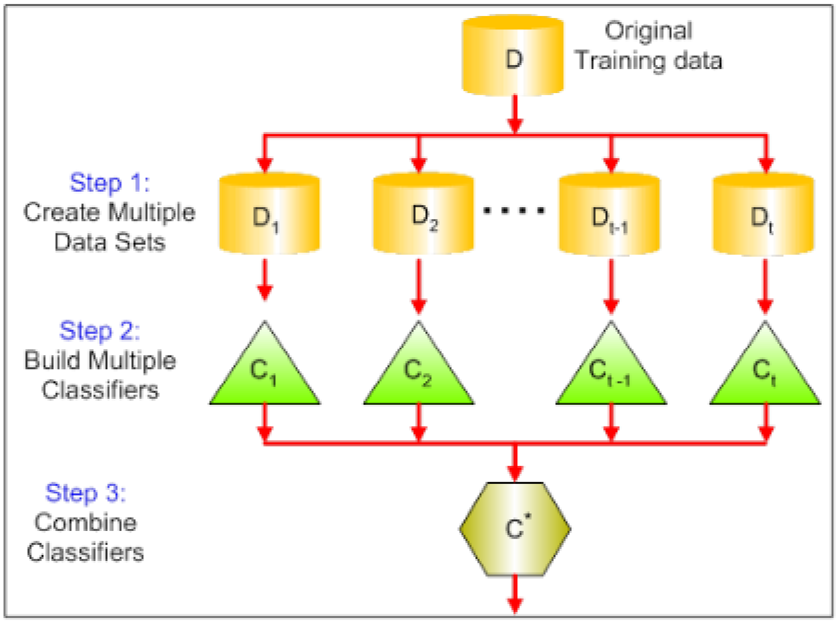


Figure 1: Bagging Approach Ensemble is performed using formula given below:

III. $$f(x) = 1/M \sum_{m=1}^{M} f_m(x)$$

Data Source

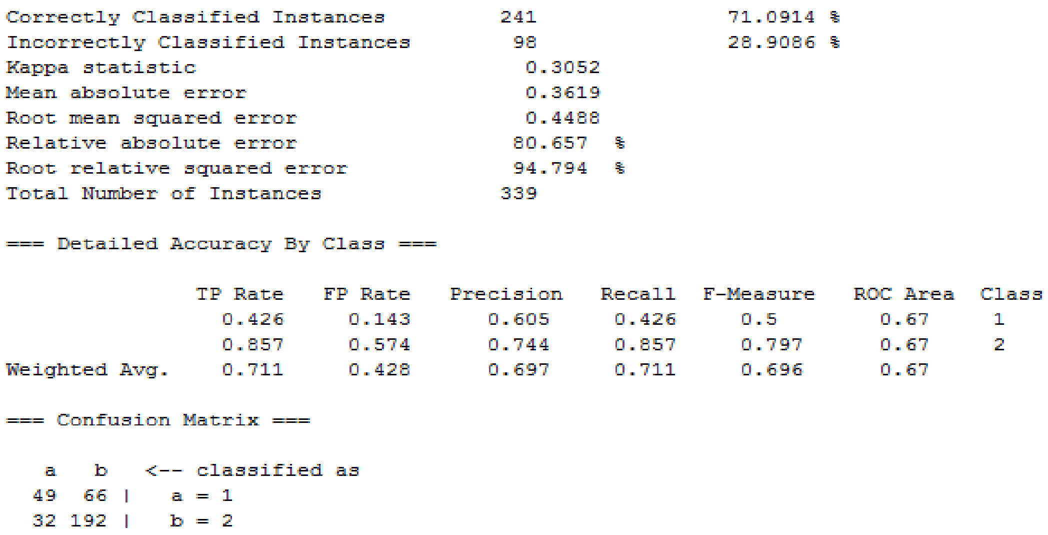
To compare these Machine Learning techniques,Primary Tumor dataset wastaken from UCI repository. The Primary Tumor dataset has 17 attributes and 339. Table 1 below lists these attributes:

Table1: Primary Tumor Data Set Description

No.	Name	Description	No.	Name	Description
1	Class	lung, head & neck, esophasus, thyroid, stomach	10	Penitoneum	Yes, no
2	Age	<30,30-59,>=60	11	Lever	Yes, no
3	Sex	Male, female	12	Brain	Yes, no
4	Histologic-type	Epidermoid, Adeno, Anaplastic	13	Skin	Yes, no
5	Degree-of-diffe	Well, fairly, poorly	14	Neck	Yes, no
6	Bone	Yes, no	15	Superclavicular	Yes, no
7	Bone-marrow	Yes, no	16	Axillar	Yes, no
8	Lung	Yes, no	17	Mediastinum	Yes, no
9	Pleura	Yes, no	18	Abdominal	Yes, no

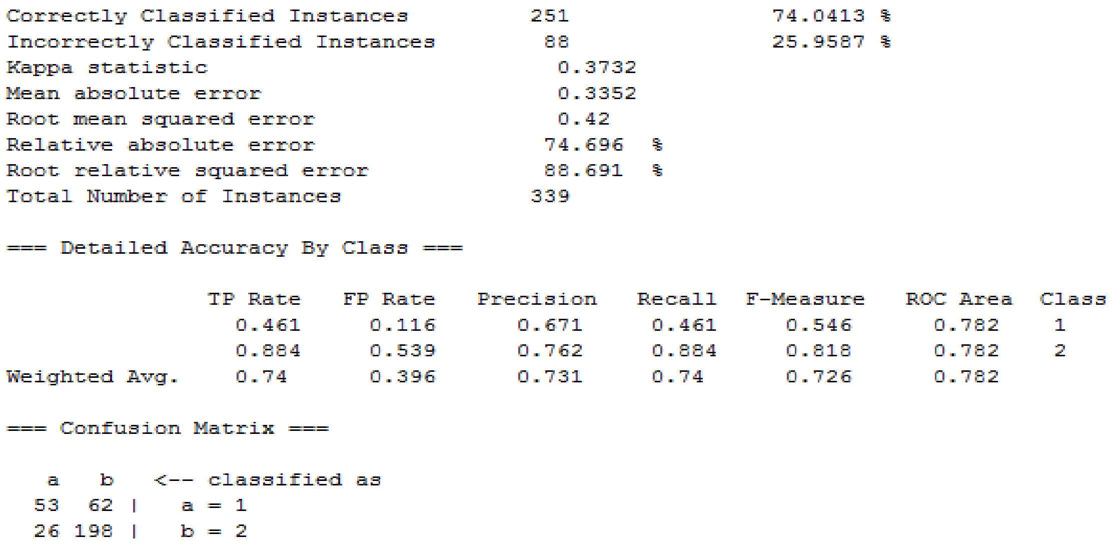
IV. Results

These Machine Learning techniques were implementedusing Weka version 3.6. Initially dataset had 18 attributes and 339 records for Primary Tumor data set. On these records ML techniques: Decision Tree, Support Vector Machine (SVM) ,Bagging of Decision Tree and Bagging of Support Vector Machine are performed and results are compared on basis of Precision,Recall,F measure,ROC and RMS Error.Figure 2 shows results obtained by J48. Figure 3shows results obtained by Bagging of J48.Figure 4 shows results obtained by SVM. Figure 5 shows results obtained by Bagging of SVM.



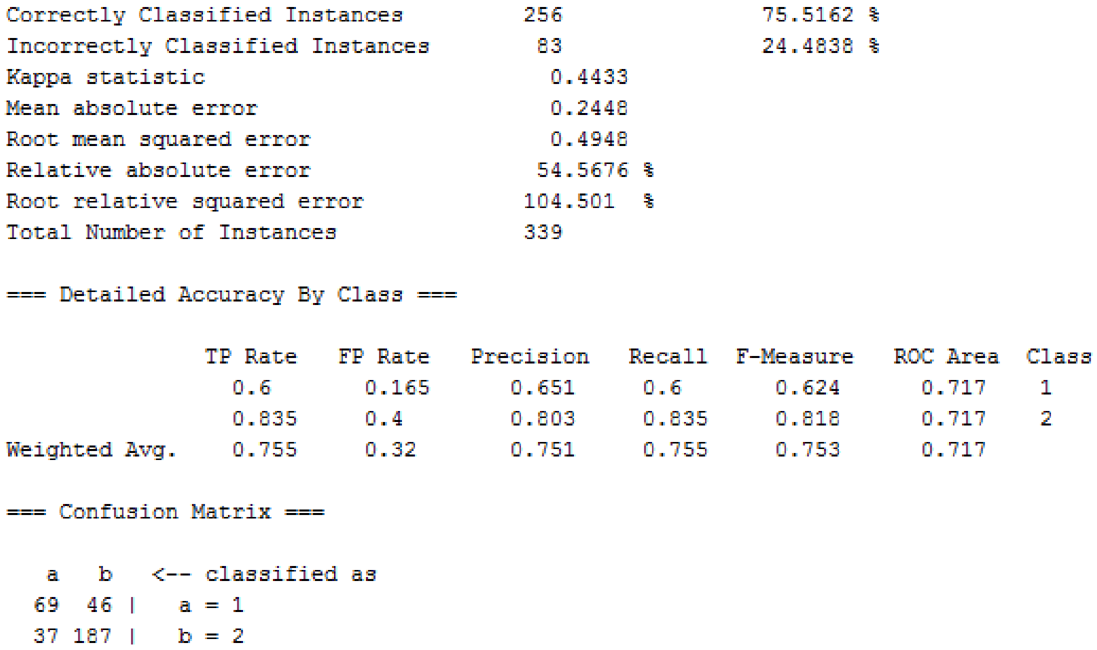
```
Correctly Classified Instances         241               71.0914 %
Incorrectly Classified Instances        98               28.9086 %
Kappa statistic                          0.3052
Mean absolute error                      0.3619
Root mean squared error                  0.4488
Relative absolute error                 80.657  %
Root relative squared error             94.794  %
Total Number of Instances              339

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.426     0.143      0.605     0.426     0.5        0.67     1
                 0.857     0.574      0.744     0.857     0.797      0.67     2
Weighted Avg.    0.711     0.428      0.697     0.711     0.696      0.67

=== Confusion Matrix ===

   a   b   <-- classified as
  49  66 |   a = 1
  32 192 |   b = 2
```

Figure 2: Results obtained by J48



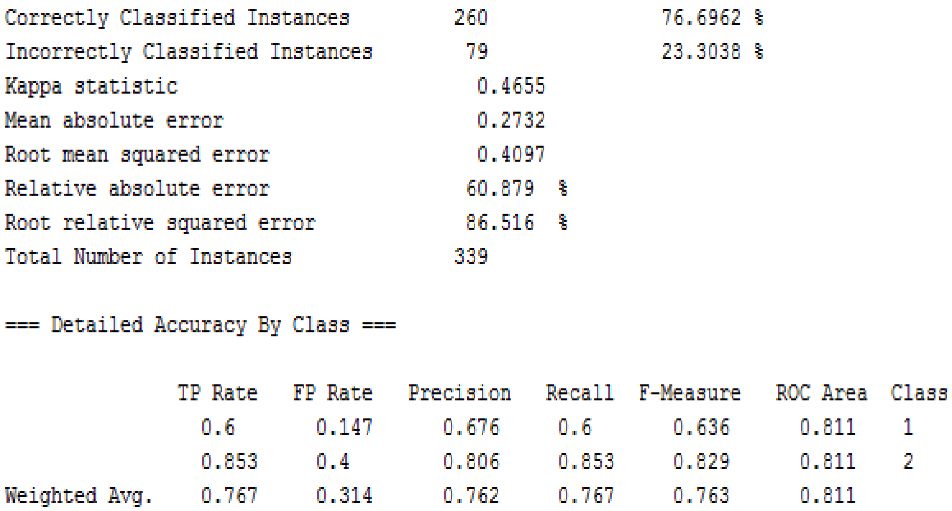
```
Correctly Classified Instances         251               74.0413 %
Incorrectly Classified Instances        88               25.9587 %
Kappa statistic                          0.3732
Mean absolute error                      0.3352
Root mean squared error                  0.42
Relative absolute error                 74.696  %
Root relative squared error             88.691  %
Total Number of Instances              339

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.461     0.116      0.671     0.461     0.546      0.782    1
                 0.884     0.539      0.762     0.884     0.818      0.782    2
Weighted Avg.    0.74      0.396      0.731     0.74      0.726      0.782

=== Confusion Matrix ===

   a   b   <-- classified as
  53  62 |   a = 1
  26 198 |   b = 2
```

Figure 3:Results obtained by Bagging of J48



```
Correctly Classified Instances         256               75.5162 %
Incorrectly Classified Instances        83               24.4838 %
Kappa statistic                          0.4433
Mean absolute error                      0.2448
Root mean squared error                  0.4948
Relative absolute error                 54.5676 %
Root relative squared error            104.501  %
Total Number of Instances              339

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.6       0.165      0.651     0.6       0.624      0.717    1
                 0.835     0.4        0.803     0.835     0.818      0.717    2
Weighted Avg.    0.755     0.32       0.751     0.755     0.753      0.717

=== Confusion Matrix ===

   a   b   <-- classified as
  69  46 |   a = 1
  37 187 |   b = 2
```

Figure 4: Results obtained by SMO



```
Correctly Classified Instances         260               76.6962 %
Incorrectly Classified Instances        79               23.3038 %
Kappa statistic                          0.4655
Mean absolute error                      0.2732
Root mean squared error                  0.4097
Relative absolute error                 60.879  %
Root relative squared error             86.516  %
Total Number of Instances              339

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.6       0.147      0.676     0.6       0.636      0.811    1
                 0.853     0.4        0.806     0.853     0.829      0.811    2
Weighted Avg.    0.767     0.314      0.762     0.767     0.763      0.811
```

Figure 5: Results obtained by Bagging of SMO

	Precision	Recall	F measure	ROC	RMS Error
J48	.692	.711	.696	.67	.4488
Bagging of J48	.731	.74	.726	.782	.4200
SVM	.751	.755	.753	.751	.4948
Bagging of SVM	.762	.767	.763	.811	.4097

Table 2: Comparison of Decision Tree, Support Vector Machine (SVM), Bagging of Decision Tree and Support Vector Machine (SVM) on basis of Precision,Recall,F measure,ROC and RMS Error

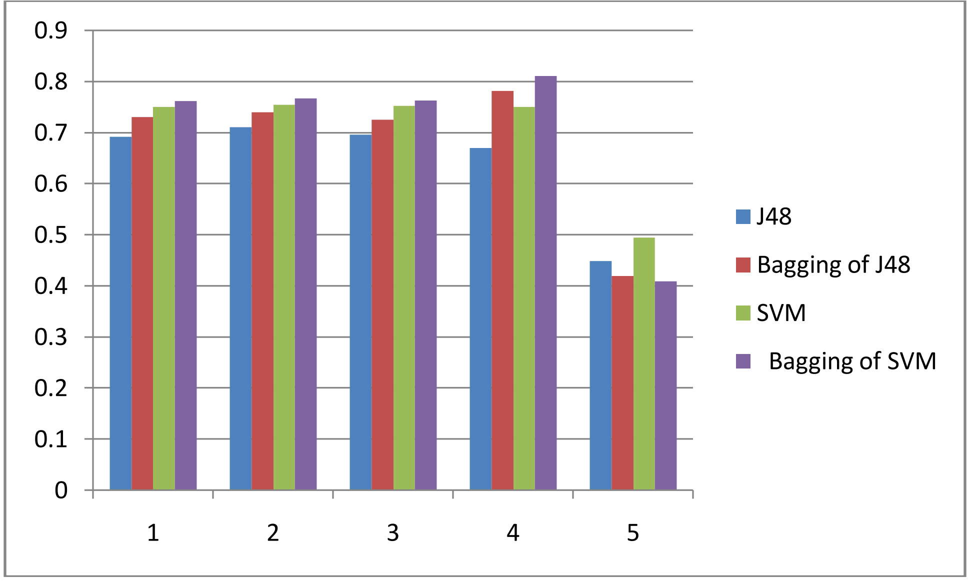


Figure5: Graphical representation ofComparison of Decision Tree, Support Vector Machine (SVM) ,Bagging of Decision Tree and Bagging of Support Vector Machine (SVM) on basis of Precision,Recall,F measure,ROC and RMS Error

	Accuracy
J48	71.09
Bagging of J48	74.04
SVM	75.51
Bagging of SVM	76.69

Table 4: Comparison of Decision Tree, Support Vector Machine (SVM) ,Bagging of Decision Tree and Support Vector Machine (SVM) on basis of Accuracy

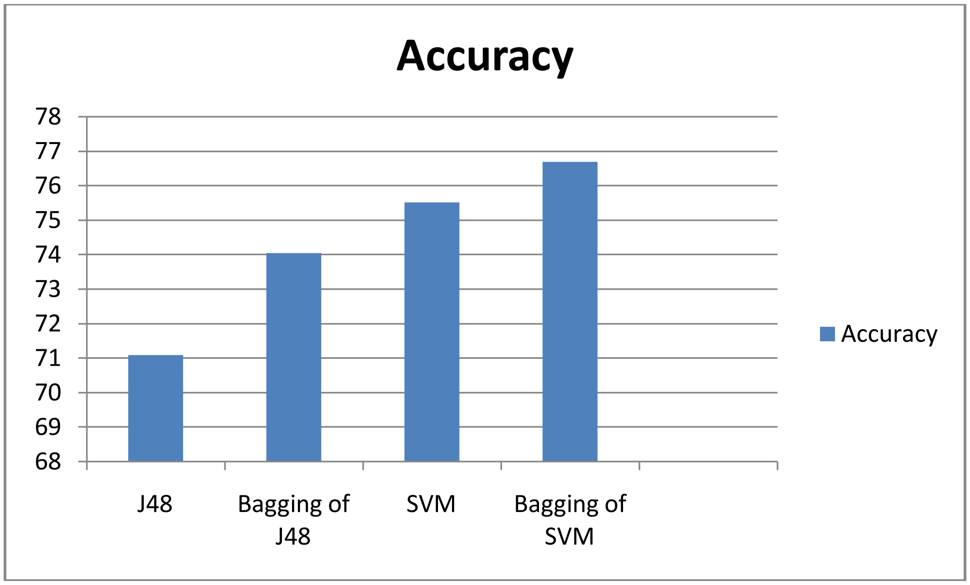


Figure 6: Graphical representation of Comparison of Decision Tree, Support Vector Machine (SVM) ,Bagging of Decision Tree and Bagging Support Vector Machine (SVM) on basis of accuracy

This result shows that Support Vector Machine performs better than Decision Tree in aspect of all parameters like of Precision,Recall,F measure,ROC, RMS Error and accuracy. Because SVM is more stable technique than Decision Tree. Ensembles models further provides better prediction and more stable models. This aggregate will be less noisy than other models. Ensemble models are better than individual.Further more performance is enhanced in Ensemble of Decision Tree as compared to Ensemble of SVM. Therefore,Ensemble of stable learners like SVM is less advantageous than ensemble of Decision Treesince the ensemble will not help to improve generalization performance

V. Conclusion:

ML techniques and their ensembles that can be used for the detection ofPrimary Tumor.In this paper twoML techniques and their ensembles are used to predict primary tumor These techniques are compared on behalf of Precision,Recall,F measure,ROC , RMS Error and accuracy. Our studies showed that Support Vector Machine model turned out to be best classifier for primary tumor detection. Ensemble of SVM outperforms all other. In future we intend to improve performance of ensemble by using some hybrid approach ..

References

1. S. Godara and R. Singh, "Evaluation of Predictive Machine Learning Techniques as Expert Systems in Medical Diagnosis", Indian Journal of Science and Technology, Vol. 910, March 2016.
2. Chaplot, S.; Patnaik, L.M.;Jagannathan, N.R. (2006). Classification of magnetic resonance brain images using wavelets as input to support vector machine and neural network, Biomed. Signal Process Control, 1, 86–92.
3. Sunila, Rishipal Singh and Sanjeev Kumar. "A Novel Weighted Class based Clustering for Medical Diagnostic Interface." Indian Journal of Science and Technology Vol 9, Issue 44, Nov.2016.
4. Yao, Z.; Lei, L.; Yin, J. (2005). “R-C4.5 Decision tree model and its applications to health care dataset”. Proceedings of International Conference on Services Systems and Services Management 2005, p 1099- 1103.
5. Zhang, Y.; Wu, L. (2012). An MR brain images classifier via principal component analysis and Kernel support vector machine. Progress Electromagnetic Res., 130, 369–388.
6. Das, R.; Abdulkadir, S. (2008). “Effective diagnosis of heart disease through neural networks ensembles”. Elsevier.
7. Kumari, M.; Godara, S.(2011). Comparitive Study of Data Mining Classification Methods in Cardiovascular Disease Prediction. IJCST, Vol. 2, ISSN: 2229-4333.
8. Zhang, Y.; Lu,S.; Zhou, X. et al. (2016). Comparison of machine learning methods for stationary wavelet entropy-based multiple sclerosis detection: decision tree, k-nearest neighbors and support vector machine. Simulation, Vol. 92(9),861–871.
9. Ian Witten H, Eibe Frank, Mark Hall A., “Data Mining Practical Machine Learning Tools and Techniques”.
10. YosvanyLierena Rodriguez, Antonio Teixeira “On the Detection and Classification of Frames from European Portuguese Oral and Nasal Vowels”FALA2010, VI JornadasenTecnología del Habla and II Iberian SL Tech Workshop, Pg.no: 205-208
11. Chau, M.; Shin, D.(2009). A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms. Proceedings of IEEE International Conference on Dependable, Autonomic and Secure Computing, pp. 183-187.
12. Patil, S.; Kumaraswamy, Y. (2009). Intelligent and effective Heart Attack prediction system using data mining and artificial neural networks. European Journal of Scientific Research, Vol. 31, pp. 642- 656.
13. Han, J.; Kamber, M. Data Mining Concepts and Techniques. 2nd Edition, Morgan Kaufmann, San Francisco.
14. Chau, M.; Shin, D. (2009). “A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms”. Proceedings of IEEE.
15. Godara, Sunila and Singh, Rishipal (2016). Evaluation of Predictive Machine Learning Techniques as Expert Systems in Medical Diagnosis. Indian Journal of Science and Technology, vol. 910.
16. Godara, Sunila and Verma, Amita (June 2013). Analysis of Various Clustering Algorithms. International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-3, Issue-1.
17. Geoffrey I.WEBB,” MultiBoosting A Technique for Combining Boosting and Wagging”, Machine Learning, 40, 2000, Pg.no:159-196.
18. DivyaSravani et.al ,” Protein Fold Classification Using Sequence Features”, Volume 3, Issue 6, June,2013,www.ijarcsse.com,Pg.no:1209-1218.