PREDICTING PROSTATE CANCER IN EARLY STAGE

Nirmala Nambikai Mary.S
M.E Computer Science and Engineering
KLN College of Engineering
Pottapalayam
Sivaganga dist, India.
nirmalajustin@gmail.com

Jeyadurga.P
Asst.Prof. Computer Science and Engineering
KLN College of Engineering
Pottapalayam
Sivaganga dist, India
jeyaocteves@gmail.com

Abstract— Precise Outcome prediction is crucial for the cancer patients to get optimal care. Tools available to predict cancer in advance remains deficient and also there is considerable false positive rate. In order to avoid this, the proposed work focuses on incorporating Clinical Data and Image processing done on Ultra sound images to predict cancer at initial stages. In the proposed work, the classification, clustering and carcinoma prediction are done by using decision tree (DT) classifier, Random Forest (RF) classifier. The error rate, efficiency and classification accuracy are compared for both the methodologies.

Keywords-*Prostate Cancer, Decision Tree, Random Forest, Risk levels, Prediction.*

*****

I. INTRODUCTION

Prostate cancer is the second most frequent cancer in men. More than 98% are classified as adenocarcinomas. Other rare malignancies are neuroendocrine tumors, urothelial cancer, squamous-cell carcinomas, lymphomas. Cancer of the prostatic gland accounts for 26% of all cancerous diseases in males. The median age is 69 years. The incidence rate has increased continually since 1980. Age-standardized mortality has decreased by 20% in the same time interval. Only relatively few, widely accepted risk factors have been identified.

The digital rectal examination has a high specificity in detecting prostate cancer, but a low sensitivity. Regular DRE does not decrease prostate cancer specific mortality. The quantitative determination of the prostate–specific antigen (PSA) is a suitable parameter for the follow-up of patients with prostate cancer. For more than 20 years PSA has also been used to screen asymptomatic males. The sensitivity and specificity of this parameter depends on the definition of the threshold value. The sensitivity in detecting prostate cancer is high at a limit of 4ng/ml. Specificity decreases with increasing age.

II. DIAGNOSIS

The first step consists of diagnosis in the confirmation of the suspected. Diagnostic procedures are recommended if the result is relevant to therapy based on the patient's decision or severe impact. Ultrasound-guided biopsy is considered standard of care. As a rule, 10 to 12 core biopsies should be taken.. Quality-assured patho histological processing and reporting are the basis for the ensuing treatment recommendations. There are several screening tests available such as DRE, quantitative PSA determination and biopsies. Once the diagnosis confirms, then Therapy recommendations are based on stage of the disease. At present, different therapeutic options are considered equivalent at all stages of prostate cancer. It is the task of multidisciplinary tumor boards to recommend the most appropriate treatment. Patients need access to full and comprehensive information as the basis for his autonomous decision. Patients should be treated in the scope of clinical trials based on the risk levels in order to make the treatment work. The risk levels could be low, intermediate and high risk based on the stratification and therapy will vary accordingly and dependently. Among all three risk levels, patients with low risk have high chance of cure.

III. LITERATURE SURVEY

A. Survey 1

"Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral Data" by Antonia Vlahou, John O. Schorge proposes biomarker patterns software (BPS) classification algorithm which is based on Classification and Regression Tree (CART). This algorithm shows better performance in discriminating ovarian cancer from benign diseases. Decision Tree is formed with five protein peaks and resulted 81.55 in the validation group and 80% in the study group.

B. Survey 2

"Suite of decision tree-based classification algorithms on cancer gene expression data" by Mohmad Badr Al Snousy, Hesham Mohamed El-Deeb, proposes two major classification categories such as single decision trees and ensample decision trees includes nine decision trees. There are five single decision trees and four ensample decision trees are considered. The analysis and comparison between two broad categories are performed. The classification accuracy is measured among nine decision trees. The ensample methods enhances the classification accuracy of single decision trees.

C. Survey 3

"Prostate cancer prediction using the random forest algorithm that takes into account transrectal ultrasound findings, age, and serum levels of prostate-specific antigen" by Li-Hong Xiao, Pei-Ran Chen proposes Random Forest algorithm that combines ultra sound images, age and PSA. The proposed method shows accuracy of 81.10%, sensitivity of 65.64% and specificity of 93.83%.The proposed method shows better diagnostic performance.

IV. CHALLENGES IN PREDICTING CANCER

There are several challenges identified in classifying and predicting cancer. Noise present in the data set poses a challenge as it includes clinical images such as ultrasound findings. Data set as a whole includes study as well as validation group poses another challenge as it includes only minimal portion of relevance to the suspected disease. An

application criterion such as accuracy, sensitivity and specificity with respect to results poses another challenge. Huge amount of dataset with multiple variety of validation group with minimal amount of study group may lead to less accuracy.

V. METHODOLOGIES

Inaccurate classification leads to considerable no of false positive tests. In the proposed work, decision tree and random forest tree are used to classify and clustering. Then the prediction is performed.

A. Decision Tree Algorithm

Decision Tree is to create a training model which can use to predict class or value of target variables by learning decision rules inferred from prior data which is also referred to as training data. The best attribute of the dataset is placed at the root of the tree. The training set is split into multiple subsets. Subsets are made in such a way that each subset contains data with the same value for an attribute. The process is repeated until finding the leaf nodes or the terminal nodes. The nodes which hold the decision rules are called decision nodes. The decision tree is constructed from a series of samples of attributes, each has a specified outcome of either true or false irrespective of each sample either has each of the attributes or not. The resulting decision tree is a binary tree where each leaf node represents the presence or absence of each attribute named along the path to the root node and the resulting outcome for the set of decisions.

After the tree is built, all the data are run down the tree, and proximities are computed for each pair of cases. If two cases occupy the same terminal node, their proximity is increased by one. At the end of the run, the proximities are normalized by dividing by the number of trees. Proximities are used in replacing missing data, locating outliers, and producing low-dimensional views of the data. The sample attributes are PSA values, Gleason score and age.

Fig. 1 builds the decision tree with PSA levels and Gleason score. If the attribute PSA value is greater than 16, then the cancer is suspected. If the PSA value is lesser than the peak value, then the Gleason score is considered to make the decision. If the Gleason score is greater than 8 then PCa is suspected. If the score is lesser than 8 then it is confirmed as Non PCa.

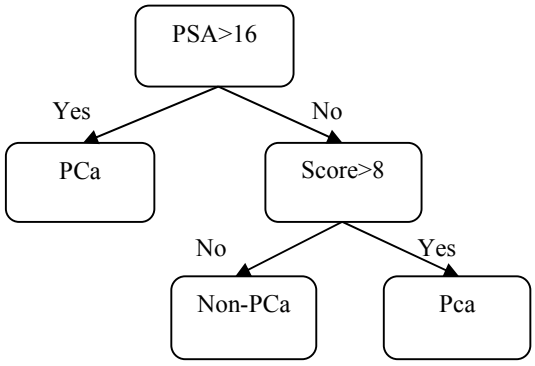

Figure1.Decision Tree with PSA and Gleason

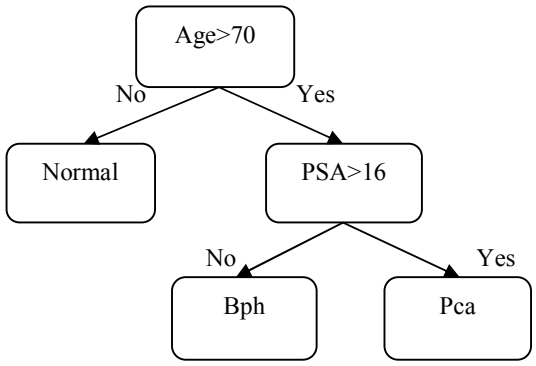

Figure2. Decision Tree with Age and PSA

Fig. 2 forms the decision tree with Age and PSA. If the attribute age value is lesser than 70, then the cancer is not suspected. If the age value is greater than 70, then the PSA is considered to make the decision. If the PSA is greater than 16 then PCa is suspected. If the PSA is lesser than the peak value then it is confirmed as Non PCa or Bph.

B. Random Forest Algorithm

Random Forests grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and the tree votes for that class. The forest chooses the classification having the most votes considering over all the trees in the forest. If the number of cases in the training set is N, sample N cases at random - but with replacement, from the original data. This sample will be the training set for growing the tree

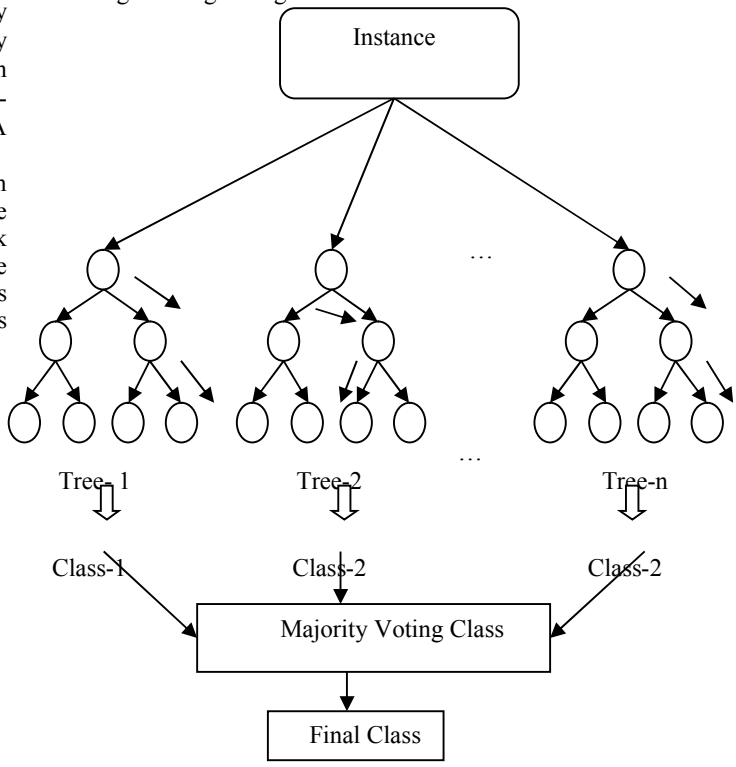

Figure 3.Random Forest Tree

. Fig.3 explains that how the random forest tree is formed and the final decision is made. If there are M input variables, a

number m<<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing. Each tree is grown to the largest extent possible. The forest error rate depends on two things. The correlation between any two trees in the forest. The strength of each individual tree in the forest. Increasing the correlation increases the forest error rate. A tree with a low error rate is a strong classifier. The approach in random forests is to consider the original data as class 1 and to create a synthetic second class of the same size that will be labeled as class 2.

The synthetic second class is created by sampling at random from the univariate distributions of the original data. Here is how a single member of class two is created - the first coordinate is sampled from the N values {x(1,n)}. The second coordinate is sampled independently from the N values {x(2,n)}, and so forth.

VI. COMPARITIVE ANALYSIS

The Decision tree and Random Forest tree are analyzed and compared for the error rate, classification accuracy and efficiency. The classification error rate of the training dataset should be approximately equal to the test dataset; if not, the model may be too particular for the training dataset and not sufficient for generic dataset which includes test or validation. For a classifier, classification accuracy or the capability for separating classes is a central evaluation metric of its performance.

A. Error rate

To prevent over-fitting and to reduce the error rate of the DT, the full grown tree is cut back in the pruning phase. Pruning phase removes subtrees that do not reduce the error rate. The pruning methods are based on minimizing a classification error rate. A tree with a low error rate is a strong classifier. In Random Forest, increasing the strength of the individual tree decreases the forest error rate. Fig.4 shows that the error rates for the Random forest tree is comparitively lesser than the Decision Tree.

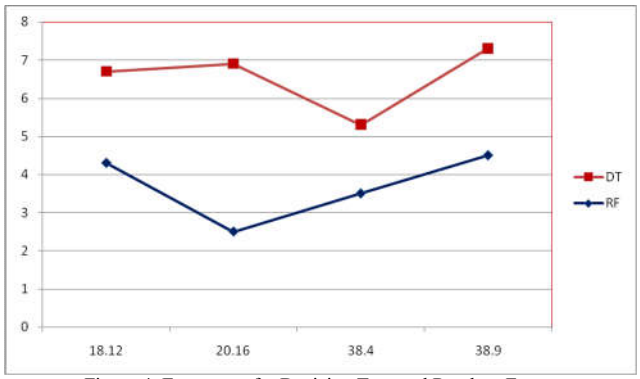

Figure 4. Error rates for Decision Tree and Random Forest

B. Classification accuracy

In the model build or training process, a classification algorithm finds relationships between the values of the predictors and the values of the target. Classification models are tested by comparing the predicted values to known target values in a set of test data. The historical data for a classification project is typically divided into two data sets: one for building the model; the other for testing the model.

A confusion matrix displays the number of correct and incorrect predictions made by the model compared with the actual classifications in the test data. The matrix is n-by-n, where n is the number of classes. Table 1 shows the confusion matrix for the decision tree. The correct predictions made are 1221. The incorrect predictions made are 65. The overall predictions are 1286. The overall accuracy rate for the decision tree is 7.383.

Table 2 shows the confusion matrix for random Forest tree. The correct predictions made are 1241.The incorrect predictions made are 35. The overall predictions made are 1276. The overall accuracy rate for the random forest tree is 0.9725. The classification accuracy for the Random Forest tree is better than decision tree.

Table I. Confusion Matrix for Decision Tree

Actual Vs Predicted class	**PCa**	**Non -PCa**
Training set	496	35
Testing set	30	725

Table II. Confusion Matrix for Random Forest

Actual Vs Predicted class	**PCa**	**Non -PCa**
Training set	516	25
Testing set	10	725

C. Prilimainary data

The input data for the classification and prediction are collection of records. Each record is known an instance and characterized by a tuple(x,y), where x is attribute set and y is designated as class label also known as category. The attribute set includes Age, PSA levels and Gleason score. The attribute set contain continuous features and the class label must be discrete value. The data set is categorized as training set or study group and testing set or validation group.

VII. CONCLUSION

The decision tree and random forest tree are analyzed and Random Forest tree is comparatively better than decision tree with respect to classification accuracy and prediction error rate. The Random Forest helps in clear assessment and reduces the no of false positive rates.

ACKNOWLEDGMENT

The author would like to thank Prof. Miruna Joe Amali , Asst. Prof .S.Brindha and Asst. Prof.P.Jeyadurga for the fruitful discussion and multiple reviews for their important remarks that helped to improve the shape of the paper. The author has many thanks to Prof.Dr. P.R.Vijayalakshmi for the wonderful support.

REFERENCES

[1] Yaozong Gao, Yeqin Shao, Jun Lian, "Accurate Segmentation of CT Male Pelvic Organs via Regression-Based Deformable Models and Multi-Task Random Forests", IEEE Transactions on Medical Imaging., Vol. 35, Issue. 6, pp.1531–1542, 2016.

[2] Kemal Hakan , İsmail Turker,"Prediction of prostate cancer using decision tree algorithm", Turkish Journal of Medical Sciences, Vol.40, Issue. 5, pp.681-686,2010.

[3] Li-Hong Xiao,Pei-Ran Chen, "Prostate cancer prediction using the random forest algorithm that takes into account transrectal ultrasound findings, age, and serum levels of prostate-specific antigen", Asian Journal of Andrology, Vol. 19, Issue.5, pp.586-590, 2017.

[4] Stephan, C., Meyer, H.A., Kwiatkowski, M., Recker, F., Cammann, H., Loening, S.A., et al.,"A (-5, -7) propsa based artificial neural network to detect prostate cancer", European Urology , Vol.50, Issue.5, pp.1014–1020 ,2006.

[5] Vickers, A.J., Cronin, A.M., Aus, G., Pihl, C., Becker, C., Pettersson, K., Scardino, P.T., Hugosson, J., Lilja, H.,"Impact of recent screening on predicting the outcome of prostate cancer biopsy in men with elevated PSA: data from the european randomized study of prostate cancer screening in gothenburg, Sweden",Journal of Cancer, Vol. 116, Issue.11,pp.2612–2620 ,2010.

[6] Lisboa, P.J., "A review of evidence of health benefit from artificial neural networks in medical intervention", Journal of Neural Networks, Vol. 15, Issue.1,pp. 11–39 , 2002.

[7] Lawrentschuk, N., Lockwood, G., Davies, P., Evans, A., Sweet, J., Toi, A., Fleshner, N.E.,"Predicting prostate biopsy outcome: artificial neural networks and polychotomous regression are equivalente models",Journal of International Urology and Nephrology, Vol. 43,Issue.1,pp. 23–30, 2010.

[8] Kattan, M.W., Scardino, P.T.: Prediction of progression: nomograms of clinical utility",Journal of Clinical Prostate Cancer , Vol.1, Issue.2, pp.90–96, 2002.

[9] National Cancer Institute" Treatment choices for men with early-stage prostate cancer" ,2011.

[10] Çinar, M., Engin, M., Engin, E.Z., Ateşçi, Y.Z.,"Early prostate cancer diagnosis by using artificial neural networks and support vector machines", Journal of Expert Systems with Applications , Vol.36, Issue.3, pp. 6357–6361, 2009.