ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

# Modelling Fetal Morphologic Patterns through Cardiotocography Data: A Random Forest based Approach.

**Kamath RS[1]*, and Kamat RK[2]**

[1]Department of Computer Studies, Chhatrapati Shahu Institute of Business Education and Research, University Road, Kolhapur 416004
[2] Department of Electronics, Shivaji University, Kolhapur – 416 004

## ABSTRACT

We report Random Forest modeling of fetal morphologic patterns by analyzing Cardiotocography. Present study exhibits performance estimation of various random forest configurations and compares the classification accuracy. The reported investigation depicts optimum random forest architecture achieved by tuning the number of trees and choice of variables for partitioning the dataset. A classification model, thus derived entails 600 trees in the forest with 5 partitioning variables. Furthermore the performance of the model is evaluated with reference to Out-of-bag estimate of error rate. The results showcases prediction of the fetal morphologic patterns based on the Cardiotocography data by using the random forest modeling.

**Keywords**: random forest, CTGs, classification, fetal morphologic pattern, decision tree, R and Rattle

**Corresponding author*

## INTRODUCTION

Cardiotocography (CTG) comprising of fetal heart rate (FHR) and tocographic (TOCO) measurements, is useful to evaluate fetal well-being [17]. The technique uses ultrasound waves to measure the aforesaid parameters. The CTG is generally indicated since 27 weeks of pregnancy and it measures heart activity, uterine contraction and fetal movement. FHR patterns are observed by obstetricians during the process of CTG analysis. Results of the CTG allow recognizing of three basic different fetal states such as physiological, suspect and pathological. The acquired information is necessary to visualize unhealthiness of the embryo and gives an opportunity for early intervention prior to happening of a permanent impairment to the embryo [9]. There are several signal processing and computer programming based techniques for interpreting a typical CTG data [15, 16]. The machine learning methods can also be employed on these data to classify as pathological or normal.

Literature review reveals that there are several reported instances of using the machine learning methodologies in the field of CTG data analysis [10-14]. Sundar et al have designed artificial neural network (ANN) model for the classification of cardiogram data [1]. This classifier was capable of identifying Normal, Suspicious and Pathologic condition with less error. Performance metrics such as Precision, Recall, F-Score and Rand Index were used to evaluate the performance aforesaid model. Thomas et al have reported random forest (RF) algorithm for automatic recognition of three basic different fetal states such as physiological, suspect and pathological [2]. This system especially used in prenatal care as a support decision system. Karabulut and Ibrikci have revealed a computer-based approach for analyzing cardiotocogram data by employing decision tree and various other machine learning algorithms [3]. Out of which decision tree contributes to the final decision of the system with accuracy 95.01%. Kamath et al have presented random forest modeling of expression levels of proteins critical to learning in a mouse model of Down syndrome [19]. The reported investigation depicts optimum random forest architecture achieved by tuning the number of trees and choice of variables for partitioning the dataset.

Yet another paper by Arif presented random forest classifier to classify the cardiotocograms into normal, suspicious and pathological classes [4]. Feature importance index is applied here to identify important features of the dataset. The classification accuracy of aforesaid model was 93.6%. Sundar et al have effectively demonstrated research challenges and solutions for classification of cardiotocogram data [5]. The traditional clustering methods such as Fuzzy C-mean and k-mean clustering can identify the Normal CTG patterns but they were incapable of finding Suspicious and Pathologic patterns. Whereas ANN based classifier is able to classify the CTG data with less error. Magenes et al have described neural classifiers to discriminate among fetal behavioral states on the basis of CTG signals [6]. These classifiers are fed by indexes extorted from fetal heart rate signal. Research confirmed promising performance towards the prediction of fetal behavioral states on the set of collected FHR signals. Sahin and Subasi have reported the research that evaluates the performances of various machine-learning methods on the CTG data [7]. The research revealed that classification is necessary to predict newborn health, especially for the critical cases.

Thus, the international scenario of modelling depicts the researchers striving hard to come out with an all-encompassing model for the purpose of analysis and experimentation of CTG data. In the backdrop of the research endeavors portrayed above, the present paper reports the random forest based approach for modelling fetal morphologic pattern through CTG Data. The dataset consists of measurements of fetal heart rate (FHR) and uterine contraction (UC) features on CTG data classified by expert obstetricians. The dataset with 2126 samples of fetal CTGs is selected for modeling [8]. The reported experiment is simulated in R and Rattle. Random forest approach builds multiple decision trees, using a concept called bagging [20]. Bagging is the idea of collecting a random sample of observations into a bag. The results of the modeling are encouraging and show that the derived RF model efficiently classifies CTG samples into the given ten classes with very less error.

The rest of paper is structured as follows; after a brief introduction, second section deals with the materials and methods exploited in the present investigation. The third section outlines our computational details of the RF model with results and discussions. The conclusion at the end discusses aptness of the RF for modelling the fetal morphologic patterns.

## MATERIALS AND METHODS

The dataset for RF modeling contains 2126 samples of fetal CTGs is taken from UCI data repository [8]. It consists of measurements of FHR and UC features on Cardiotocograms. The CTGs were classified by expert obstetricians and classification was both with respect to a morphologic pattern and to a fetal state. Present research reports analysis of CTGs data for classifying it in to ten classes of morphologic patterns. These patterns are described based on fetal heart rate and uterine contraction features [8]. Table 1 lists set of classes and corresponding number of observations in the dataset. Fig. 1 shows density of these classes described in the dataset.

**Table 1. Fetal morphologic patterns class details of CTG data**

| Abbreviation | Class Details | No. of Observations |
|---|---|---|
| A | calm sleep | 384 |
| B | REM sleep | 579 |
| C | calm vigilance | 53 |
| D | active vigilance | 81 |
| SH | shift pattern (A or SUSP with shifts) | 72 |
| AD | accelerative/decelerative pattern (stress situation) | 332 |
| DE | decelerative pattern (vagal stimulation) | 252 |
| LD | largely decelerative pattern | 107 |
| FS | flat-sinusoidal pattern (pathological state) | 69 |
| SUSP | suspect pattern | 197 |

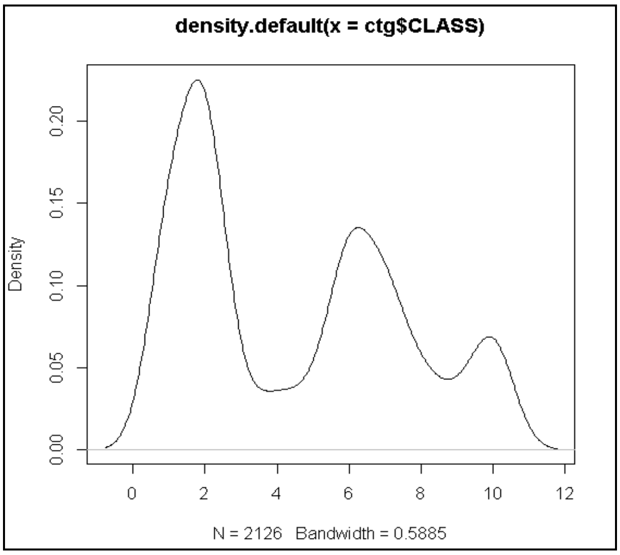


**Figure 1: CTG data projection**

In the present investigation we have employed random forest modeling for classifying CTGs data in to ten classes of morphologic patterns. RF modeling is often used when there is a very large training datasets and a very large number of input variables. It is a collection of unpruned decision trees. Rather than growing a single very deep tree, random forest relies on aggregating the output from many shallow trees that are tuned and pruned without much oversight. Random forest employs randomization in two places:

1. Each tree is trained using a random sample with replacement from the given dataset
2. While training individual trees, subset of features are chosen randomly for searching of splits. This can reduce the correlations among trees in the forests thus achieves improved performance in prediction.

RF model is simulated in R and Rattle environment [20]. The model is conceived as a Multi-Input Single-Output configuration. It works basically with 21 inputs viz. values of FHR and UC features. Morphologic pattern class is considered as an output variable. Bagging concept applied here for the construction of multiple decision trees. Bagging is the idea of collecting a random sample of observations into a bag [19]. Each bag of observations is then used as the training dataset for building a decision tree. The performance of the resulting model is evaluated by out-of-bag (OOB), estimate of the error rate is calculated using the observations that are not included in the bag. Performance of the model can be pictorially represented using Receiver Operating Characteristic (ROC) curve. It plots the true positive rate against the false positive rate.

## Computational details, results and discussions

This section explores details of experiment conducted for the classification of fetal morphologic patterns with different RF architectures. RANDOMFOREST package in R environment is used to analyze model structure, number of trees in the forest and choice of variables for partitioning the dataset [19]. We used the training data set for the parameter adjustment of model whereas validation set to control learning process. We carried out performance evaluation for various RF configurations and compare the classification accuracy. RF builds many decision trees using random subset of data and variables [20]. This method is proven for assessing proximities among data points in unsupervised mode.

In the present investigation, RF model is tuned with two parameters $n_{tree}$ and $n_{try}$ to get optimized forest architecture. The parameter $n_{tree}$ specifies number of trees is to be built to populate the random forest where as $n_{try}$ specifies the how many variables that will be considered in deciding partitioning of the dataset. We have demonstrated RF modeling per variation in $n_{tree}$ and $n_{try}$. The whole experiment is summarized in table 2. We have varied value for $n_{tree}$ from 100 to 1000 and value for $n_{try}$ from 4 to 8. Table 2 shows performance of RF model with reference to OOB estimate of error rate. Random forest has selected 1488 observations randomly to build the model. We have explored error plot and ROC curve as useful analytic tool for our random forest modeling. Figure 2(a-d) presents error plots for the execution of RF models per variant in $n_{tree}$ and $n_{try}$. Error plot is useful for deciding optimal number of trees to build since the plot error rate gradually for the number of trees built. The plot reports the accuracy of the forest of trees in terms of error rate on the y-axis against the number of trees that have been included in the forest. Figure 3(a-d) presents the ROC curves for different architectures based on the out-of-bag predictions for each observation in the training dataset.

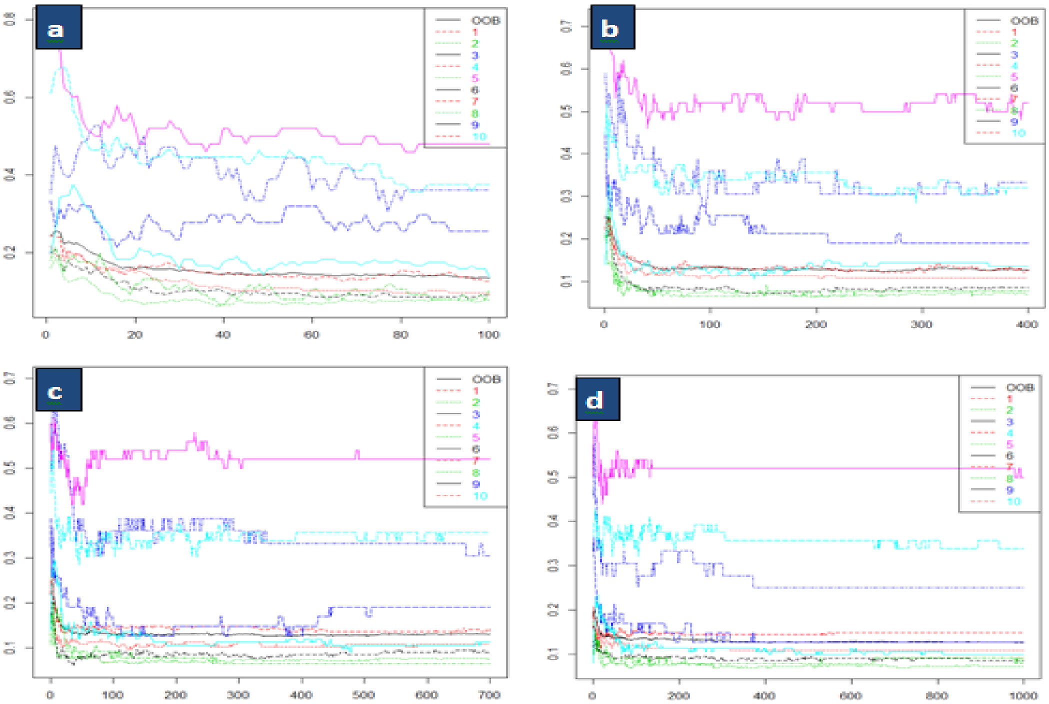


**Figure 2: Error plots for RF models per variation in $n_{tree}$ and $n_{try}$. Figure (a) represents mean square error of RF model with $n_{tree}$ is 100 and $n_{try}$ is 4; Figure (b) represents mean square error of RF model with $n_{tree}$ is 400 and $n_{try}$ is 5; Figure (c) represents mean square error of RF model with $n_{tree}$ is 700 and $n_{try}$ is 6; Figure (d) represents mean square error of RF model with $n_{tree}$ is 1000 and $n_{try}$ is 8.**

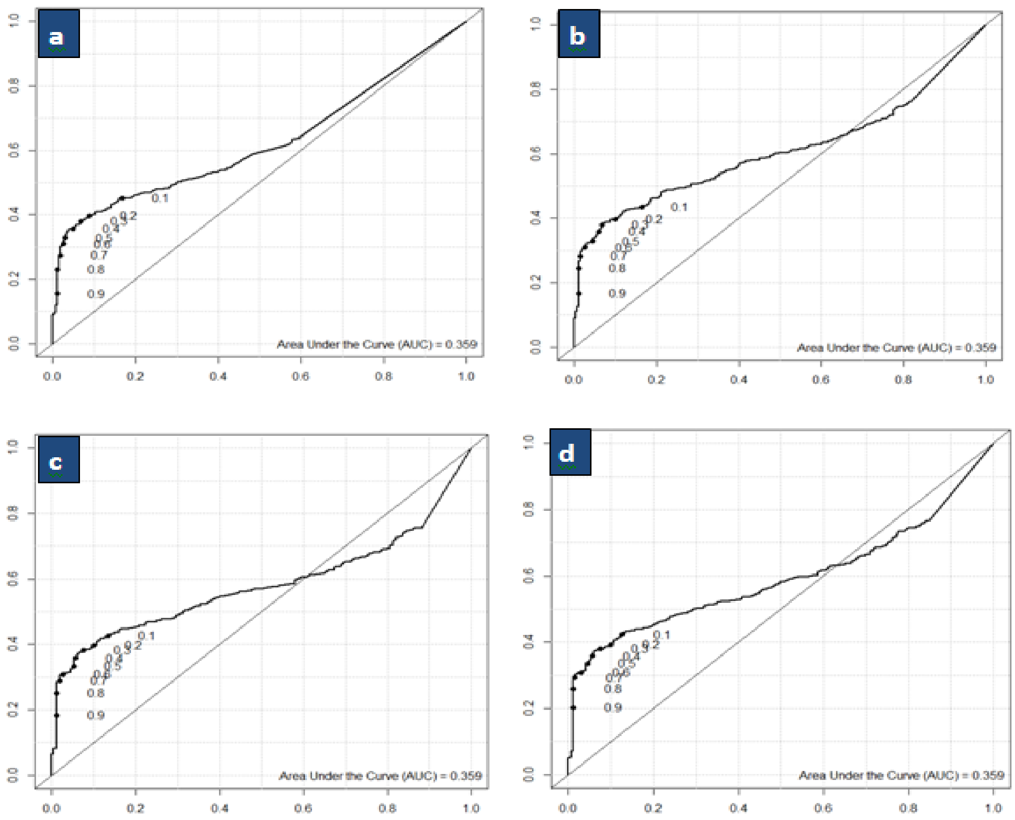
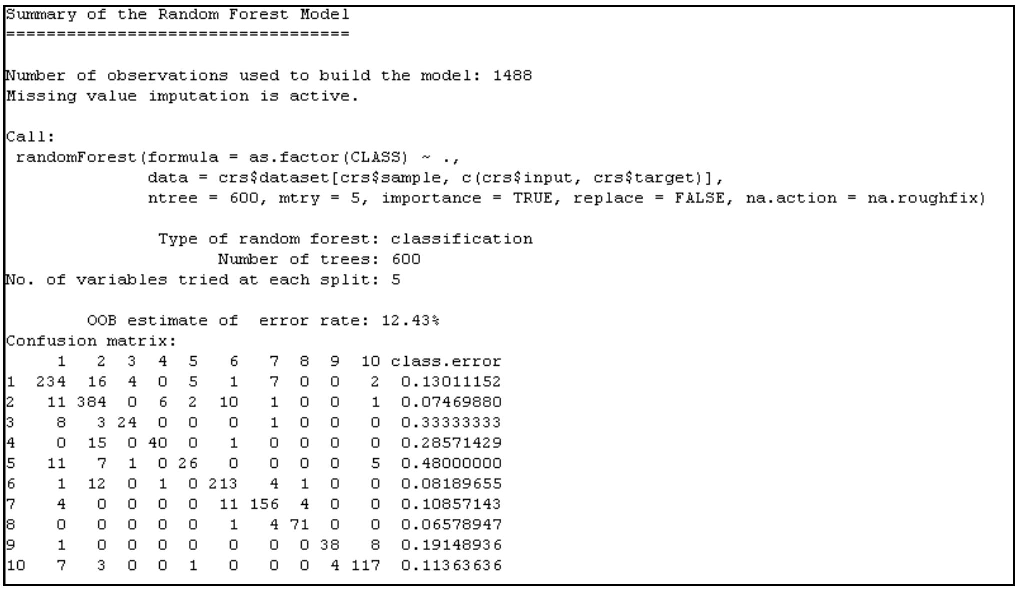


**Figure 3: the ROC curves for RF models per variation in $n_{tree}$ and $n_{try}$ based on the out-of-bag predictions for each observation in the training dataset. Figure (a) represents ROC curve of RF model with $n_{tree}$ is 100 and $n_{try}$ is 4; Figure (b) represents ROC curve of RF model with $n_{tree}$ is 400 and $n_{try}$ is 5; Figure (c) represents ROC curve of RF model with $n_{tree}$ is 700 and $n_{try}$ is 6; Figure (d) represents ROC curve of RF model with $n_{tree}$ is 1000 and $n_{try}$ is 8.**

**Table 2: Performance evaluation for accuracy of Random forest Configurations**

| | OOB estimate of error rate (%) | | | |
|---|---|---|---|---|
| No. of Variables ($n_{try}$) →<br>No. of Tree ($n_{tree}$) ↓ | 4 | 5 | 6 | 8 |
| 100 | 13.37 | 13.44 | 12.9 | 13.58 |
| 200 | 13.24 | 12.97 | 13.24 | 13.51 |
| 300 | 12.9 | 12.84 | 13.1 | 13.1 |
| 400 | 12.9 | 12.84 | 12.9 | 12.7 |
| 500 | 12.97 | 12.7 | 12.84 | 12.77 |
| 600 | 12.9 | **12.43** | 13.1 | 12.97 |
| 700 | 12.7 | 12.63 | 13.17 | 12.77 |
| 800 | 12.7 | 12.63 | 13.17 | 12.9 |
| 900 | 12.7 | 12.9 | 12.97 | 12.77 |
| 1000 | 12.84 | 12.57 | 12.97 | 12.63 |

The optimized RF architecture chosen for the modeling of CTGs entails 600 trees in the forest with 5 partitioning variable. RF Model has used 1488 observations randomly to build the forest. A detail of aforesaid optimized RF model is given in fig. 4. The performance of RF modeling pertaining to this is shown in figure 5(a-b). In this case, OOB estimate error rate found to be 12.43%. This overall measure of accuracy is then followed by a confusion matrix that records the disagreement between the final model's predictions and the actual outcomes of the training observations. Thus derived RF architecture efficiently classifies new CTG samples with very less error. We have tested model with known CTG samples. Fig. 6 shows the result obtained in terms of confusion matrix by applying the test dataset on the derived RF model. Result concludes that RF modeling is a suitable approach since the resulting analysis is much more accurate and precise.



```
Summary of the Random Forest Model
==================================

Number of observations used to build the model: 1488
Missing value imputation is active.

Call:
 randomForest(formula = as.factor(CLASS) ~ .,
              data = crs$dataset[crs$sample, c(crs$input, crs$target)],
              ntree = 600, mtry = 5, importance = TRUE, replace = FALSE, na.action = na.roughfix)

               Type of random forest: classification
                     Number of trees: 600
No. of variables tried at each split: 5

        OOB estimate of  error rate: 12.43%
Confusion matrix:
     1   2  3  4  5   6   7  8  9  10 class.error
1  234  16  4  0  5   1   7  0  0   2  0.13011152
2   11 384  0  6  2  10   1  0  0   1  0.07469880
3    8   3 24  0  0   0   1  0  0   0  0.33333333
4    0  15  0 40  0   1   0  0  0   0  0.28571429
5   11   7  1  0 26   0   0  0  0   5  0.48000000
6    1  12  0  1  0 213   4  1  0   0  0.08189655
7    4   0  0  0  0  11 156  4  0   0  0.10857143
8    0   0  0  0  0   1   4 71  0   0  0.06578947
9    1   0  0  0  0   0   0  0 38   8  0.19148936
10   7   3  0  0  1   0   0  0  4 117  0.11363636
```

**Figure 4: Textual representation of selected RF model**

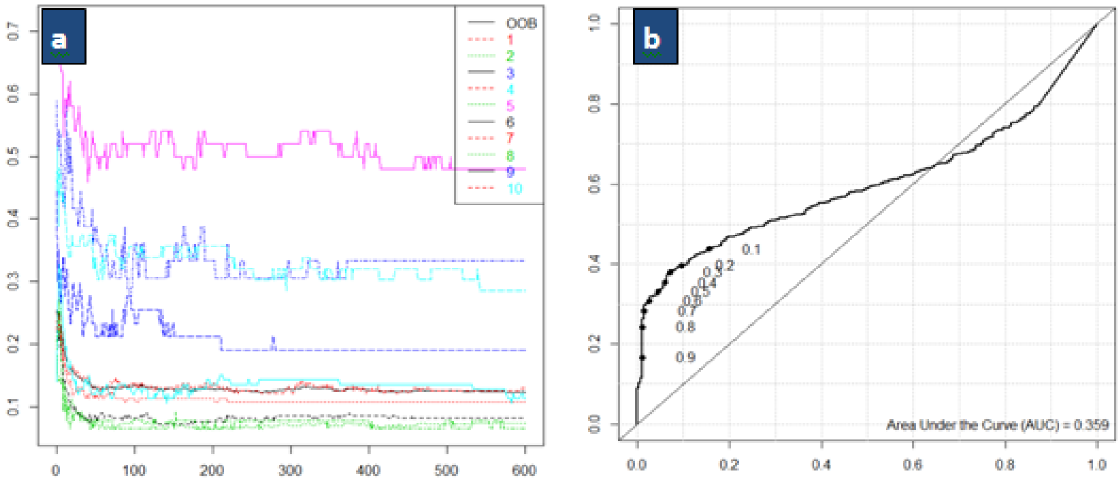
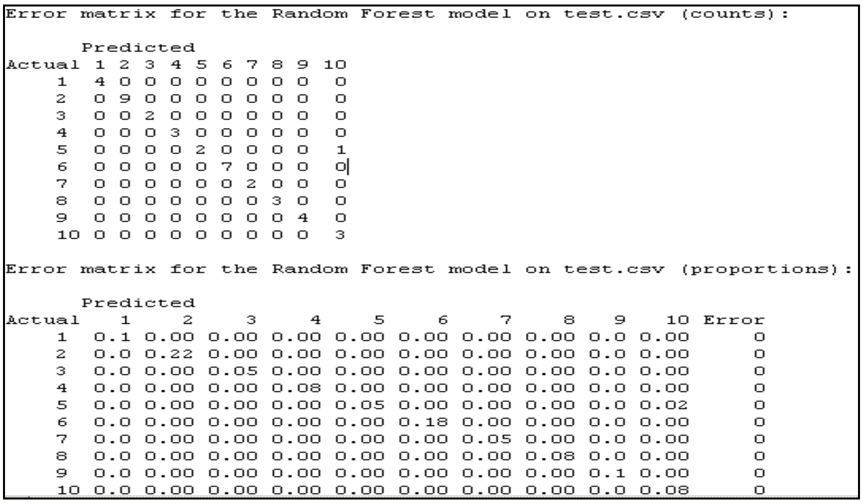


**Figure 5: Performance of selected RF model with $n_{tree}$ is 50 and $n_{try}$ is 8; Figure (a) represents mean square error plot; Figure (b) represents ROC curve based OOB**



```
Error matrix for the Random Forest model on test.csv (counts):

      Predicted
Actual 1 2 3 4 5 6 7 8 9 10
     1 4 0 0 0 0 0 0 0 0  0
     2 0 9 0 0 0 0 0 0 0  0
     3 0 0 2 0 0 0 0 0 0  0
     4 0 0 0 3 0 0 0 0 0  0
     5 0 0 0 0 2 0 0 0 0  1
     6 0 0 0 0 0 7 0 0 0  0
     7 0 0 0 0 0 0 2 0 0  0
     8 0 0 0 0 0 0 0 3 0  0
     9 0 0 0 0 0 0 0 0 4  0
    10 0 0 0 0 0 0 0 0 0  3

Error matrix for the Random Forest model on test.csv (proportions):

      Predicted
Actual   1    2    3    4    5    6    7    8   9   10 Error
     1 0.1 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.0 0.00     0
     2 0.0 0.22 0.00 0.00 0.00 0.00 0.00 0.00 0.0 0.00     0
     3 0.0 0.00 0.05 0.00 0.00 0.00 0.00 0.00 0.0 0.00     0
     4 0.0 0.00 0.00 0.08 0.00 0.00 0.00 0.00 0.0 0.00     0
     5 0.0 0.00 0.00 0.00 0.05 0.00 0.00 0.00 0.0 0.02     0
     6 0.0 0.00 0.00 0.00 0.00 0.18 0.00 0.00 0.0 0.00     0
     7 0.0 0.00 0.00 0.00 0.00 0.00 0.05 0.00 0.0 0.00     0
     8 0.0 0.00 0.00 0.00 0.00 0.00 0.00 0.08 0.0 0.00     0
     9 0.0 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.1 0.00     0
    10 0.0 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.0 0.08     0
```

**Figure 6: Execution result of RF model on test dataset**

## CONCLUSION

In the present paper we have reported modeling of fetal morphologic patterns by analyzing Cardiotocography, a diagnostic method which is widely used in prenatal care. The dataset with 2126 samples of CTGs were selected for aforesaid modeling. The present investigation demonstrated optimum RF architecture by varying its various attributes such as number of trees and choice of variables for partitioning the dataset. The resulted RF architecture entails 600 trees in the forest with 5 partitioning variable. RF Model has selected 1488 observations randomly to build the forest. Thus derived RF model efficiently classifies CTG samples into the given ten morphologic pattern classes with very less error. The result suggests that the RF has the potential to exhibit as the best tool for modeling of CTG samples.

## REFERENCES

[1] Sundar C, Chitradevi M, Geetharamani G. International Journal of Computer Applications 2012; 47(14): 19-25.

[2] Tomas P, Krohova J, Dohnalek P, Gajdos P. 36th International Conference Telecommunications and Signal Processing 2013; 620 – 923.

[3] Karabulut EM, Ibrikci T. Journal of Computer and Communications 2014; 2: 32-37.

[4] Arif M. Biomaterials and Biomechanics in Bioengineering 2015; 2(3): 173-183.

[5] Sundar C, Chitradevi M, Geetharamani G. Journal of Computer Science 2013; 9(2): 198-206.

[6] Magenes G, Signorini MG, Arduini D. Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks 2000; 3: 637 – 641.

[7] Sahin H, Subasi A. Applied Soft Computing 2015; 33: 231-238.

[8] Lichman M. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California 2013; School of Information and Computer Science.

[9] Alfirevic Z, Devane D, Gyte, Gillian ML. In Alfirevic, Zarko. Cochrane Database of Systematic Reviews 2006; doi:10.1002/14651858.CD006066.

[10] Huang M, Hsu Y. Journal of Biomedical Science and Engineering 2012; 5: 526-533.

[11] Ocak H. J. Med. Syst. 2013; 37(2): 1-9.

[12] Menai MEB, Mohder FJ, Al-mutairi F. J. Med. Bioeng. 2013; 2(1).

[13] Karabulut EM, Ibrikci T. J. Comput. Commun. 2014; 2(09): 32.

[14] Czabanski R, Jezewski M, Wrobel J, Horoba K, Jezewski J. In 14th Nordic-Baltic Conference on Biomedical Engineering and Medical Physics 2008; Springer Berlin Heidelberg.

[15] Martinez AM, Kak AC. IEEE Transactions on Pattern Analysis and Machine Intelligence 2001; 23(2):228-233.

[16] Ayres-de-Campos D, Bernardes J, Garrido A, Marques-de-Sa J, Pereira-Leite L. J. Maternal-Fetal Neonatal Med. 2000; 9(5): 311-318.

[17] Grivell RM, Alfirevic Z, Gyte GM, Devane D. Cochrane Database Syst. Rev. 2010; 1.

[18] Breiman L. Machine Learning 2001; 45(1): 5-32.

[19] Kamath RS, Dongale TD, Pawar P, Kamat RK. Research journal of Pharmaceutical, Biological and Chemical Sciences 2016; 7(4): 830-836.

[20] Kamath R, Kamat R. Educational Data Mining with R and Rattle, River Publishers, Netherland, 2016, pp. 65-67.

[21] Graham W. Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery. Springer, UK, 2011, pp. 245-268.