I.J. Information Technology and Computer Science, **2021, 5, 30-40**
Published Online October 2021 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2021.05.03

Cardiotocography Data Analysis to Predict Fetal Health Risks with Tree-Based Ensemble Learning

Pankaj Bhowmik
Dept. of Computer Science & Engineering, Hajee Mohammad Danesh Science and Technology University, Bangladesh
E-mail: pankaj.cshstu@gmail.com

Pulak Chandra Bhowmik[1], U. A. Md. Ehsan Ali[2] and Md. Sohrawordi[3]
[1,2,3]Dept. of Computer Science & Engineering
[1]Stamford University Bangladesh, Dhaka, Bangladesh
[2,3]Hajee Mohammad Danesh Science and Technology University, Dinajpur-5200, Bangladesh
E-mail: [1]pulokbhowmik@gmail.com, [2]ehsan_cse@hstu.ac.bd, [3]mdsohrawordicse@gmail.com

Received: 26 May 2021; Accepted: 29 July 2021; Published: 08 October 2021

Abstract: A sizeable number of women face difficulties during pregnancy, which eventually can lead the fetus towards serious health problems. However, early detection of these risks can save both the invaluable life of infants and mothers. Cardiotocography (CTG) data provides sophisticated information by monitoring the heart rate signal of the fetus, is used to predict the potential risks of fetal wellbeing and for making clinical conclusions. This paper proposed to analyze the antepartum CTG data (available on UCI Machine Learning Repository) and develop an efficient tree-based ensemble learning (EL) classifier model to predict fetal health status. In this study, EL considers the Stacking approach, and a concise overview of this approach is discussed and developed accordingly. The study also endeavors to apply distinct machine learning algorithmic techniques on the CTG dataset and determine their performances. The Stacking EL technique, in this paper, involves four tree-based machine learning algorithms, namely, Random Forest classifier, Decision Tree classifier, Extra Trees classifier, and Deep Forest classifier as base learners. The CTG dataset contains 21 features, but only 10 most important features are selected from the dataset with the Chi-square method for this experiment, and then the features are normalized with Min-Max scaling. Following that, Grid Search is applied for tuning the hyperparameters of the base algorithms. Subsequently, 10-folds cross validation is performed to select the meta learner of the EL classifier model. However, a comparative model assessment is made between the individual base learning algorithms and the EL classifier model; and the finding depicts EL classifiers' superiority in fetal health risks prediction with securing the accuracy of about 96.05%. Eventually, this study concludes that the Stacking EL approach can be a substantial paradigm in machine learning studies to improve models' accuracy and reduce the error rate.

Index Terms: Ensemble Learning, Stacking, Cardiotocography, Hyperparameter Tuning, Feature Selection, Cross Validation, Random Forest classifier.

1. Introduction

Cardiotocograms provide essential facts to observe fetal health conditions by recording fetus heart-rate signals, maternal uterine contractions pressure, fetus movements in mothers' wombs, and other types of monitoring simultaneously [16,25]. By analyzing CTG data, the forthcoming possible risks of the fetus can be avoided. The clinical CTG test is a straightforward and affordable way to become aware of fetal current health situation. Besides, an early diagnosis of any health issue will allow the medical team to take proper steps, and perhaps, they would able to prevent the offspring and maternal mortality. In general, the antepartum CTG test is conducted after the succeeding 28th weeks (at the 7th month) of pregnancy to monitor fetal wellbeing state [3]. The outcome of this test suggests if there is any abnormality in fetus growth, which consequently helps the obstetricians to propose clinical decisions. Actually, the CTG test diagnoses the fetal wellbeing by observing if the fetus tissues are getting ample volume of oxygen— in other words, it monitors the potential Hypoxia or Acidosis.

In this 21st century, the advanced healthcare system is now frequently engaging machine learning models for predicting different health issues and notably, these robust models are much reliable in most cases. As the clinical results are so crucial, a tiny glitch in decision-making can cause serious health threats. In the CTG test [4], typically a decision support system (DSS) provides valuable real-time clinical data as visual reports, and interpretation of the report depends on the expertise of obstetricians. Although these conventional DSSs are supporting the field of obstetrics for years now to predict fetal wellbeing, they are less likely to handle uncertain situations. Therefore, the doctors or experts

in obstetrics analyze these sensitive data by hand to make the final decisions. But unintentional slips make this practice vulnerable and possibly lead to true-negative (misdiagnosis) results. However, many pieces of research [5,6,7,8] have been conducted with distinct machine learning approaches to employ some feasible systems that can predict fetal health status efficiently by analyzing CTG data. These systems are less prone to errors and can improve the prediction accuracy significantly.

This article aims to establish a predictive machine learning system that can predict the fetal wellbeing status from CTG data, perhaps can act as a DSS. However, the study intended to use only tree-based algorithms for developing the model. Why tree-based algorithms? Well, the CTG data used in this study is an imbalanced dataset, and most of the previous researches [4,6,13,14] that are experimented with tree-based algorithms showed comparatively higher efficiency on this dataset. Those prior studies suggest that tree-based algorithms can deal with the imbalanced data effectively without even resampling. Therefore, this study selected some particular popular tree-based algorithms for the experiment. Besides, the Neural Network (NN) models showed [7,8] significant performance in fetal wellbeing prediction, but they are so computationally expensive. Considering a small dataset, the NNs are prone to overfitting and without proper care, the result can be misleading.

This study strove to achieve satisfactory accuracy, concurrently reducing the number of features with feature engineering. A tree-based Stacked ensemble learning approach is considered to develop the classifier model which can predict fetal health risks on the CTG dataset. Stacked ensemble learning is a layered-structure machine learning approach that combines the predictive competency of diverse base learning algorithms to attain higher prediction accuracy compared to the individual weak learners [12]. The study investigated four tree-based algorithms to employ the Stacking EL model. Besides, the study also attempted to perform a comparative model evaluation to determine if the ensemble learning model can beat the solo learning models. The study designed a compendious analysis of each step in the Stacked EL approach and followed the corresponding steps accordingly while developing the predictive model.

2. Related Works

A comprehensive study of various contemporary research works related to fetal health classification is performed. Besides, this study found most of the research works applied tree-based machine learning algorithms to accomplish this particular task. The study has also explored the prior approaches such as traditional learning, ensemble learning, deep learning, and data mining to classify fetal health status using different algorithms.

The CTG dataset is imbalanced, and it's tough to achieve higher prediction accuracy, especially on the minor categories (Suspicious, Pathological). However, the findings of several research work suggest that the tree-based algorithms handled this situation more effectively and outperformed the other traditional algorithms by reducing the error rate in classification [7,9,10,11]. In many research, the F1-score and sensitivity rate of the Suspicious case varies mostly between 45%–82% and 39%–90% respectively, on the other hand, in Pathological cases, it varies between 64%–97% and 59–97% respectively [13,18,19,20]. Considering clinical fetal health monitoring, it may cause potential risks if those models (low F1-score and sensitivity) are applied in real-life applications.

Several ensemble machine learning approaches such as Bagging (Bootstrap aggregating) and Boosting have been applied to classify fetal wellbeing risks. In most cases, the base learners are considered as tree-based algorithms including, Decision Tree, Random Forest, J48, CART, C4.5, REP Tree, and so on. However, the ensemble learning models performed significantly well compared to the conventional algorithms [4,16,21]. Rafael M.O. Cruz et al. [27] proposed an ensemble classifier 'META-DES' that can predict fetal wellbeing status with an accuracy of 84.6%.

A comparative study by Septian Eko Prasetyo, Pulung Hendro Prastyo and Shindy Arti, shows the outcome of different classifier models with and without using the feature selection method. The CTG dataset contains 21 features, in their study, they selected 7, 10, and 15 features from the dataset using CFS Subset Evaluation, Info Gain, and Chi-Square methods respectively. Afterward, they experimented with the selected features on seven machine learning algorithms and made a comparative evaluation between the models. However, the average precision, F1-score, and sensitivity rate in most of the models remained under the 90% mark [9].

In another study, by Sahana Das, Himadri Mukherjee, Sk. Md. Obaidullah, Kaushik Roy and Chanchal Kumar Saha showed how feature selection can improve classification accuracy. The study reduced the number of features to 9 with the 'Minimum Redundancy Maximum Relevancy (MRMR)' method and secured an accuracy of 99.91% applying Random Forest classifier [5].

Yandi Chen et al. proposed a Deep Forest algorithm based intelligent model that can predict fetal abnormality with securing high accuracy. They performed the experiment with two datasets, on the public CTG dataset the model achieved a prediction accuracy of 95.07%, average F1 score of 0.920, and AUC score of 0.989. Moreover, they have integrated four basic classifiers namely, Random Forest, Weighted Random Forest, Completely Random Forest, and Gradient Boosting Decision Tree during the cascade forest phase [15]. In general, hyperparameter optimization of the learning algorithms is mostly overlooked in the existing models conducted to predict fetal health states.

Although several quality research works have been implemented to predict fetal health conditions using CTG data, from top of the authors' knowledge, this study did not find any Stacked ensemble learning approach for solving the task. Consequently, to overcome the research gap, this article proposed to develop a tree-based Stacked ensemble learning

classifier model to predict the fetal health condition. However, most of the existing studies did not consider dimensionality reduction techniques in their experiments. Hence, this study planned to apply statistical feature selection method for reducing the number of input variables i.e., to select only the topmost important features. In addition, the study intended to perform a comparative performance evaluation between the base learning models and the EL model.

3. Research Methodologies

Fetal health abnormality in the course of pregnancy can induce deficiencies in the infants' early development, or even worse. Therefore, to deal with such unwanted situations, precise monitoring of fetal health status is a must. In this study, a robust Stacking ensemble learning classifier model is proposed to predict the prospective fetus health risks. The intended models' flow chart is shown in Fig. 1, and the system architecture is depicted in Fig. 2. The study followed the precise guidelines illustrated in the flow chart and system architecture to establish the proposed Stacking EL model. This experiment is commenced by collecting the CTG dataset from the UCI Machine Learning Repository.

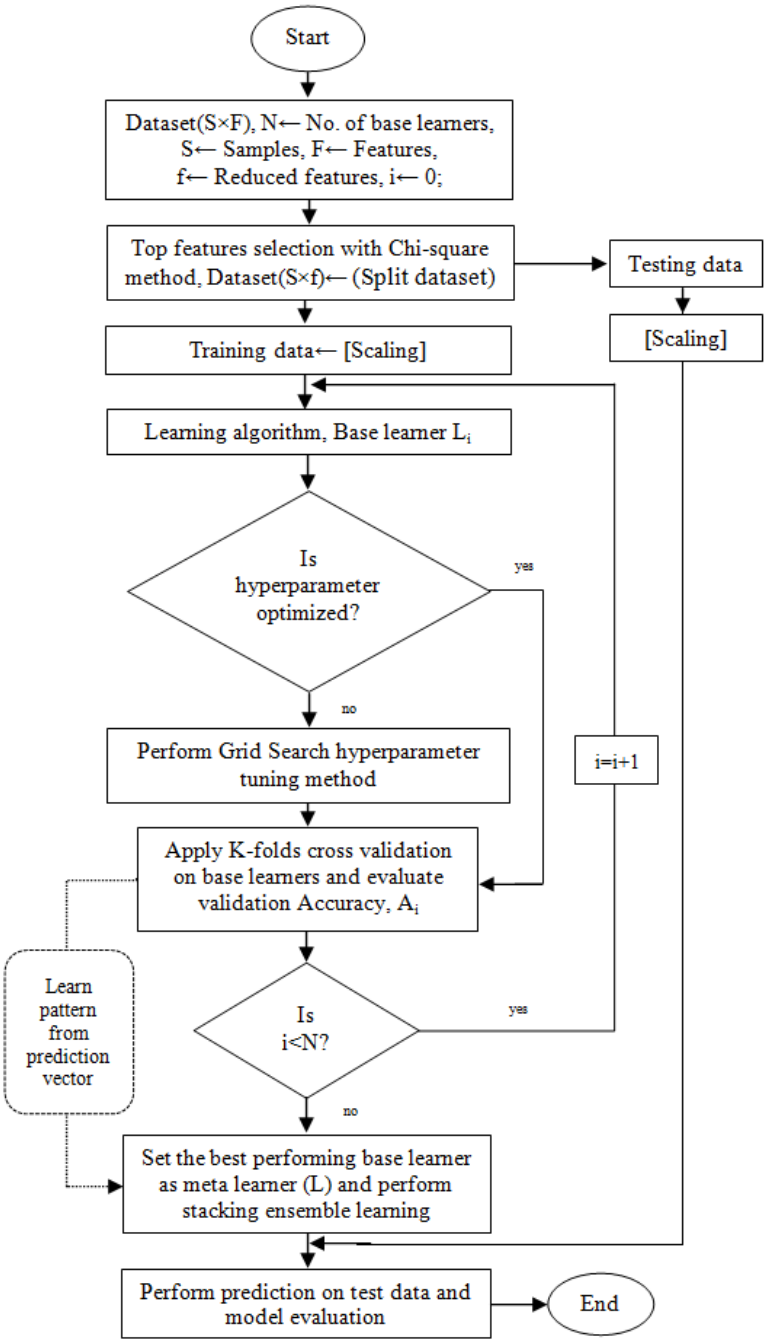


Fig.1. Flow chart of proposed Ensemble Learning model

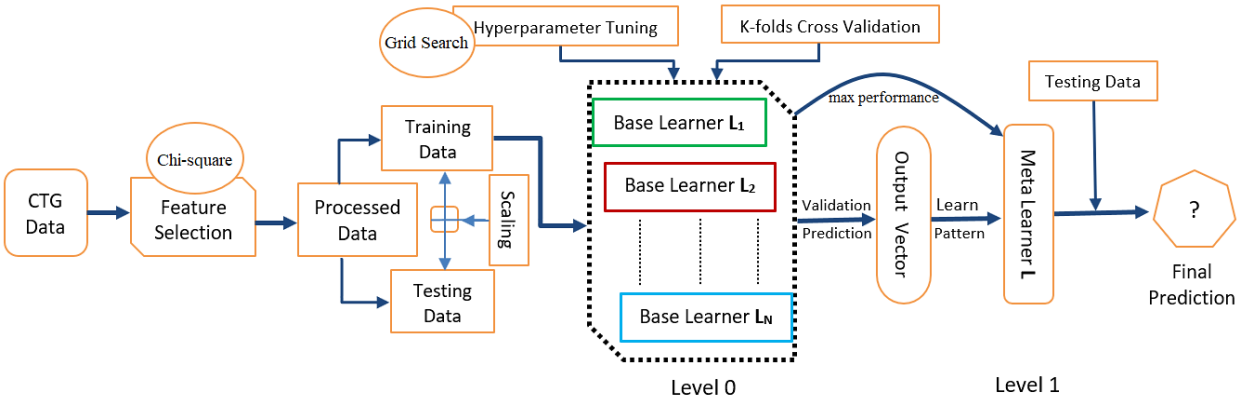


Fig.2. Proposed system structure of Stacked Ensemble Learning classifier

The K-best features are selected from the data by calculating the feature importance with the Chi-square method. Afterward, the reduced dataset is split into training and testing data with maintaining the ratio of 0.75 and 0.25. The train and test data are normalized by applying the Min-Max scaling approach. Four tree-based machine learning algorithms, including Random Forest classifier, Decision Tree classifier, Extra Trees classifier, and Deep Forest classifier are considered as base learners in the proposed Stacking ensemble learning mechanism. Besides, hyperparameter tuning of these algorithms is carried out using the Grid Search method. K-folds cross validation is performed on the base learners using the training data, and the prediction accuracy score on validation data is recorded; the process is marked as 'Level 0' in the system architecture. The base learner model, which comes with the maximum accuracy is employed as the meta learner in the EL model. Consequently, the output vector (prediction← validation) of the individual learners is stacked and fed to the meta learner for learning and getting trained as an EL classifier model; this module is labeled as 'Level 1' in the system architecture. Finally, the classifier model is ready to perform prediction on testing data. The Stacked ensemble learning approach applied in this study is discussed in Algorithm 1.

Algorithm 1 Stacked Ensemble Learning Classifier

Input: CTG dataset D (training data D_{tr}, validation data D_{val}, testing data D_{ts}), Base learning algorithms $\{L_1, L_2, \ldots, L_n\}$, N← No. of base learner, Stacking approach S_N, i← 0;

Output: Ensemble Learning Model, E_M

Process:

Step 1: $\{L_1, L_2, \ldots\ldots, L_n\} \leftarrow S_N(D_{tr})$

Step 2: **while** i< N

do

Perform K-folds cross validation on L_i, $\{L_1, L_2, \ldots, L_n\}$

Evaluate validation accuracy A_i, $\{A_1, A_2, \ldots, A_n\}$ on D_{val}

end while

Select outperformed base learner L_{max},

$$L_{max} \leftarrow \underset{Ai \in Li}{argmax}\{L_1(A_1), L_2(A_2), \ldots., L_n(A_n)\}$$

Step 3: Set meta learner, L← L_{max} and perform stacking ensemble with the base learner $\{L_1, L_2, \ldots, L_n\}$

Step 4: Return E_M and perform final prediction on D_{ts}

The proposed methodology supports the study to highly focus on the research objectives, and thus the study stuck to the guiding principles to achieve the desired outcomes. However, to combat experimental bias, the study double-checked the proposed methods, and the data flow of the system. The study paid close attention to statistical modeling and numerical simulation of the model. Different machine learning libraries available in Python programming language such as Scikit-learn, Pandas, Numpy, Mlxtend, Matplotlib, Seaborn are applied to implement the proposed system. The Jupyter Notebook, an interactive web tool, is used to simulate and deploy the machine learning models.

4. Result Analysis and Discussion

4.1. Dataset description

The CTG dataset, used to investigate the outcome of the proposed EL model, is gathered from the Machine Learning Repository of the University of California Irvine (UCI). It is an open-source standard dataset and was found by the SisPorto project of the University of Porto, Portugal. The dataset was labeled by three professional obstetricians [1,2].

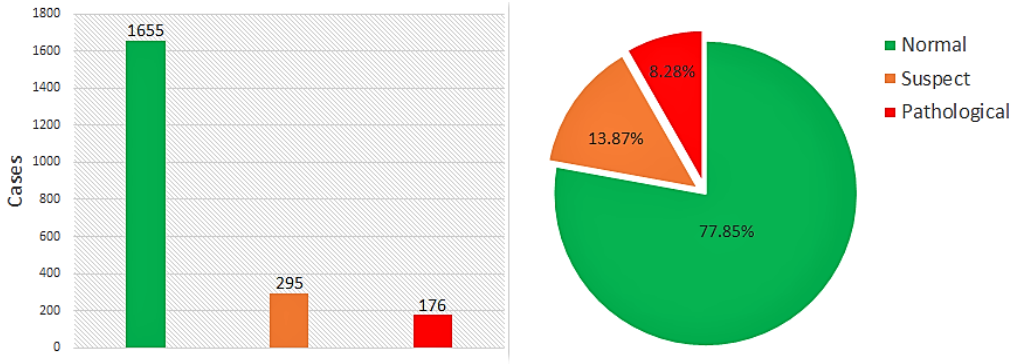


Fig.3. Graphical representation of CTG dataset

There are 21 features recorded in the CTG dataset [1], which were generated by combining the results of fetal heart-rate signal, uterine contraction pressure, and other measurements. There are 2 target variables in this data— the first one is classifying the pattern (1-10), and the other is classifying fetal health status such as Normal (N), Suspect (S), and Pathological (P); this study considered the later one. The dataset contains 2126 observations, among which 1655 samples belong to the N-class, 295 and 176 samples belong to the S and P-class respectively. It is an imbalanced dataset since a significant number of instances possess in a single class, to exemplify, the Normal-class contains 77.85% of the entire samples. Besides, for clear understanding, a visual representation of the CTG data distribution is depicted in Fig. 3.

4.2. Feature Selection and Scaling

Feature selection in machine learning studies is a method of selecting the most relevant features from a dataset. The selected features show a high correlation with the target variable. Besides, the feature selection technique helps to prevent the curse of dimensionality and simplifies the training models [24,26].

The shape (sample, feature) of the CTG dataset is (2126, 21). The dataset contains 21 features but not all of them are equally important to perform prediction. Therefore, the study applied the Chi-square feature selection method to capture the K(=10)-best features from the CTG data and the shape of the data reduced to (2126, 10). Since the feature size is reduced, consequently the degree of computations as well as the complexity of the final classifier model will be minimized. The topmost important features are shown in Fig. 4 with the feature importance score. The Chi-square score specifies the correlation of features with the target class. In this method, frequency distributions are stored in the contingency tables. Let's consider the observed frequency of a feature in i^{th} position of the table is O_i and the expected frequency is E_i. Now, the Chi-square value of that feature calculated as,

$$X^2 = \sum_{i}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{1}$$

In real-world data, the value of observed features frequently varies in an unsteady range. Besides, the features with high weight can have a higher impact on the objective function, can add bias to the experimental results. In machine learning studies, feature scaling is an essential data processing strategy in which the independent features (variables) of a dataset are normalized on a fixed scale. It allows each feature to play an equal contribution in optimizing the objective function.

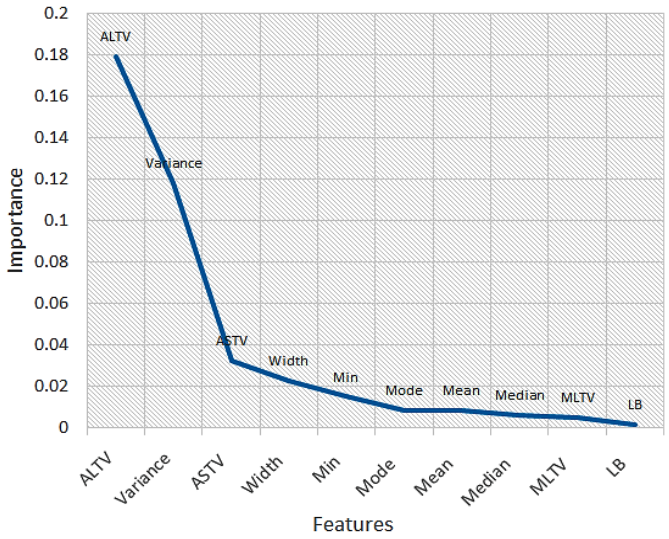


Fig.4. Top 10 features selected with Chi-square

After splitting the dataset, both the training and testing data are normalized with the Min-Max feature scaling method— in a range of [0,1]. The CTG data contains both discrete and continuous values, therefore the features are scaled to perform bias-free and accurate predictions. Let, X is a random feature in the feature space, its actual and normalized value is X_{val}, and X_{scaled} respectively. Here, X_{min} is the minimum and X_{max} is the maximum value of the feature in the list. The equation of Min-Max feature scaling to calculate the normalized value of feature X is defined in Eq^{n}. (2).

$$X_{scaled} = \frac{X_{val} - X_{min}}{X_{max} - X_{min}} \tag{2}$$

4.3. Base Learning Algorithms: Stacked Ensemble Learning

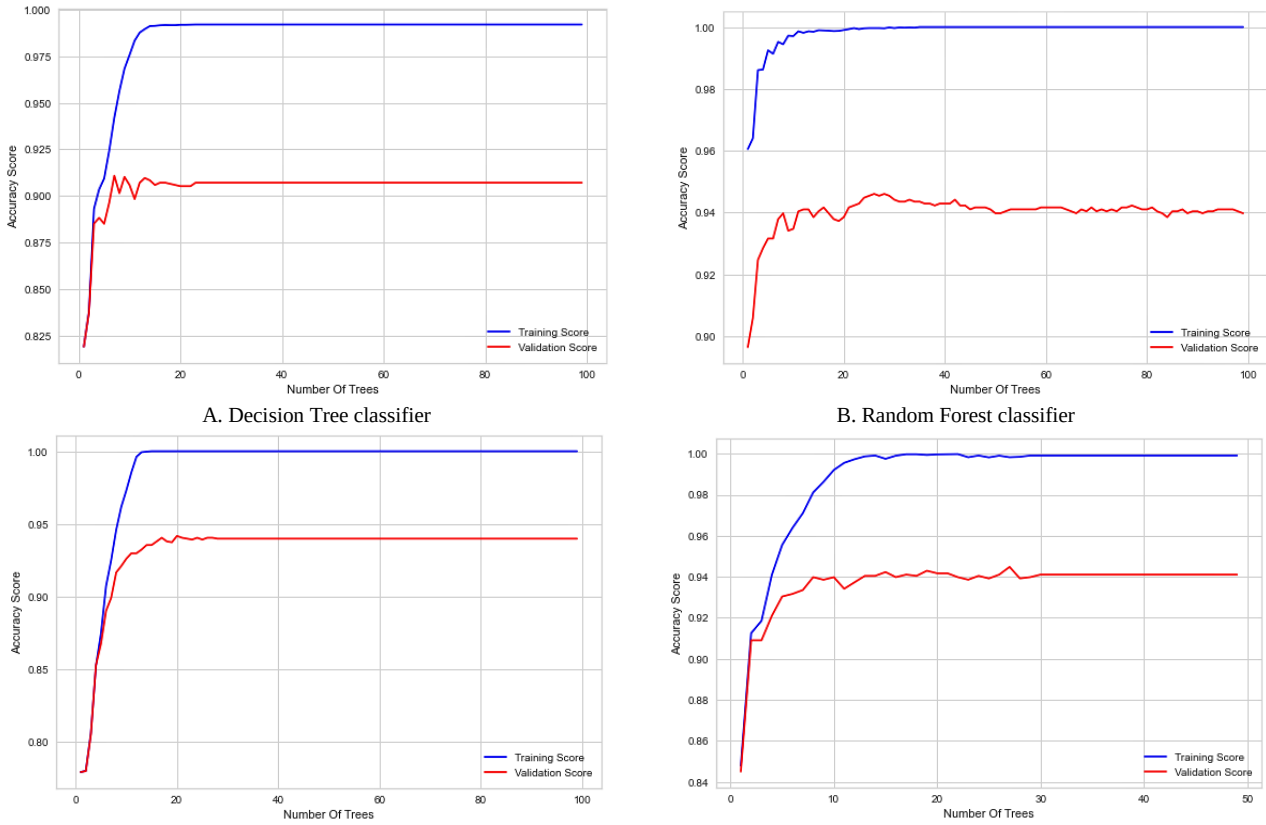
The study approached to experiment with four tree-based machine learning algorithms, namely, Random Forest classifier, Decision Tree classifier, Extra Trees classifier, and Deep Forest classifier as base learning algorithms for developing an ensemble learning model. However, hyperparameter optimization of these base learner algorithms is carried out with the Grid Search hyperparameter tuning technique. In machine learning studies, hyperparameter optimization is the process of tuning the initialized (default) parameters of an algorithm, and it helps the algorithm to choose the best match hyperparameters for a given dataset [22]. Previous studies [22,23] showed that an algorithm 'A_{op}' with optimized hyperparameters can perform better than the same algorithm 'A_{df}' with having the default hyperparameter. The tuned hyperparameters of the base algorithms tabulated in Table 1.

Table 1. Tuned Hyperparameter of Base Learners

Base Learner	Optimized Hyperparameter	Method
Decision Tree	criterion='entropy', max_depth=10, max_features='log2', min_samples_split=3	Grid Search
Random Forest	criterion='entropy', n_estimators=100, min_samples_leaf=1, min_samples_split=2	Grid Search
Extra Trees	criterion='entropy', n_estimators=100, max_depth=15, min_samples_split=2	Grid Search
Deep Forest [17]	n_estimators=50, max_layers=4, min_samples_leaf=1	Manual

A. Level 0: K-Folds Cross Validation

At 'Level 0' of the system, K (=10)-folds cross validation is performed on each base learner algorithm using the training dataset. Here, the training data is divided into 10-folds, and during cross validation (K-1)-folds i.e., 9 out of 10-folds are claimed as train data, and the rest of the fold is claimed as validation data. Cross validation is an iterative process, and the number of iterations is equal to the number of folds (K), still in each iteration the validation fold changes. The meta classifier of 'Level 1' is selected based on the performance of base learners on validation data.



A. Decision Tree classifier B. Random Forest classifier

C. Extra Trees classifier D. Deep Forest classifier

Fig.5. Train VS Validation curve of base learning algorithms (A, B, C, D) in CV.

Therefore, the study made a comprehensive assessment of these base learners' performance throughout the cross validation. The training and validation curve of these base learners is shown in the following Fig. 5.

However, the performance evaluation of each base learner in cross validation (CV) is shown in Table 2. The Random Forest classifier model outweighs the other three models and the model showed about 93.92% accuracy on validation data, where Decision Tree classifier, Extra Trees classifier, and Deep Forest classifier obtains 91.03%, 93.35%, and 93.73% respectively. The study also estimated the Standard Deviation (Std.) value of the base learners. Since the base learning algorithm Random Forest came up with the maximum outcome in validation, it is set as the meta classifier for the ensemble learning model.

Table 2. Performance Evaluation of Base Learners in CV

Base Learner	Precision	Recall	F1-score	Accuracy	Std.
Decision Tree	0.8559	0.8231	0.8351	91.03%	0.023
Random Forest	**0.9181**	**0.8647**	**0.8864**	**93.92%**	**0.020**
Extra Trees	0.9018	0.8529	0.8732	93.35%	0.025
Deep Forest	0.9065	0.8675	0.8830	93.73%	0.022

B. Level 1: Stacking and Building EL Classifier

After selecting the meta learner, all the output vectors propagated from the base learning algorithms during validation are stacked and approached to pass them into the ensemble learning classifier model. Consequently, the meta learner learns the patterns from the predictions (←validation) of individual base learners. The EL approach ensembles all the base learners and stacks their decisions, as a result, the meta learner gets well trained and reduces the error rate. For instance, let a base learner *A* makes an error on a test sample *T* in prediction space and another base learner *B* provides a correct output on *T*. Now, the meta learner in ensemble learning will make a predictive analysis thus it can resolve the possible errors (minimize loss function) that can encounter in individual the base learners (A or B). However, after being ensembled and trained, the Stacked EL classifier model is ready to perform prediction on the testing dataset.

C. Model Evaluation and Comparison

The experimental result illustrates, the EL model provides a classification accuracy of about 96.05% on the test dataset. The model evaluation metric of the EL classifier model shows the average F1-score of 0.9320, Precision score of 0.9627, Recall score of 0.9063. Besides, the Kappa score of the model is 0.8875, the zero-one loss score is 0.0395

and the area under curve (AUC) score is 0.9595. The confusion matrix depicts, 511 samples out of 532 CTG testing samples were truly classified (True Positive) and the rest of the samples were misclassified (False Positive). However, the model secured a sensitivity rate of 99.52% on the Normal-class, 80.49% and 91.89% on the Suspicious and Pathological-class respectively. Besides, the F1-score of Normal, Suspicious and Pathological class is 97.74%, 87.42% and 94.44% respectively. Fig. 6 illustrates the receiver operating characteristic (ROC) curve and AUC score of the three classes (Normal = 0.95, Suspicious = 0.94, and Pathological = 0.98) predicted by the EL model.

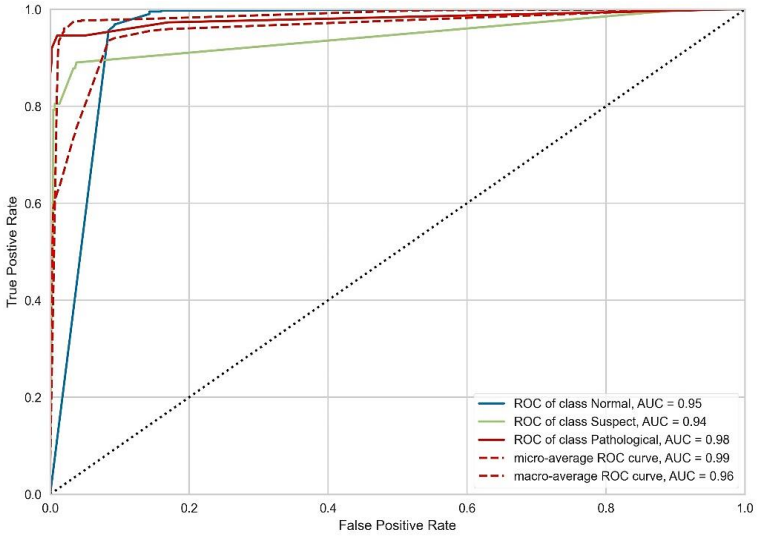


Fig.6. The ROC curve of Stacked classifier model

However, to perform a comparative assessment, the study also evaluates the performance metrics of the individual base learner algorithms on the testing dataset. The outcome of the models and comparison with the proposed EL classifier is tabulated in Table 3. The study compared the evaluation matrices, mean square error (MSE), coefficient of determination (R^2), Kappa score, AUC score, and error rate of the models.

Table 3. Performance Comparison of Classifier Models

Classifier Model	MSE ↓	R^2 ↑	Kappa ↑	Precision ↑	Recall ↑	F1-score ↑	Accuracy ↑	Error Rate ↓	AUC ↑
Decision Tree	0.092	0.925	0.7987	0.8734	0.8877	0.8804	92.48%	0.0752	0.9127
Random Forest	0.077	0.951	0.8635	0.9321	0.9055	0.9178	95.11%	0.0489	0.9257
Extra Trees	0.062	0.949	0.8567	0.9395	0.8900	0.9132	94.92%	0.0508	0.9165
Deep Forest	0.085	0.944	0.8430	0.9095	0.8826	0.8954	94.36%	0.0564	0.9128
Ensemble Learning	**0.051**	**0.961**	**0.8875**	**0.9627**	**0.9063**	**0.9320**	**96.05%**	**0.0395**	**0.9595**

The performance evaluation metrics in Table 3, demonstrate the individual models— Decision Tree classifier, Random Forest classifier, Extra trees classifier, and Deep Forest classifier obtain prediction accuracy of about 92.48%, 95.11%, 94.92%, and 94.36% respectively. Besides, the models achieved promising 'coefficient of determination' and AUC score. The Random Forest classifier model outruns the other typical base learning models in fetal health risks prediction. Although the individual classifier models showed reasonable outcomes, the Stacking ensemble learning approach has boosted the performance to the next level. Comparing with the Random Forest model, the Kappa, Precision, F1-measure, and Accuracy score in the ensemble learning approach is increased by 2.40%, 3.06%, 1.42%, and 0.94% respectively. The AUC score, on the other hand, is improved nearly by 3.50%. Evidently, it can be argued that the Staking ensemble learning approach has notably enhanced the performance of the base learning model. Performance comparison of the previous research models with the proposed model is cataloged in Table 4.

The premise of this study is to develop a Stacking ensemble learning system that can effectively predict fetal health issues, and to investigate the performance metrics along with traditional tree-based models. Overall, the proposed methodology facilitated the learning models to achieve satisfactory performance. Conforming the research flow, the study reduced the dimensionality of data significantly with the Chi-square feature selection method. Hyper-parameters of base learning algorithms were optimized with Grid Search that made the algorithms more compatible. Besides, K-fold cross validation was performed to select the meta learner of the final model, and it ensures no existence of selection bias or data overfitting. The comparative assessment concludes that the proposed Stacked ensemble model outweighed the individual models and performed significantly well on testing the CTG samples. In fact, it surpassed the solo

learning models in all evaluation parameters considered in this assessment.

Table 4. Performance comparison between the proposed method and the existing works

Research Reference	Best Method	Feature, Target class	Performance
Yandi Chen et al. [15], 2021	Deep Forest	21,3	Accuracy 95.07 %, F1-score 0.9201
Septian Eko Prasetyo et al. [9], 2021	Random Forest	10,3	Accuracy 93.74%, Kappa score 0.8262, F1-score 0.937
K. Agrawal and H. Mohan [19], 2019	Support Vector Machine	21,3	Accuracy 92.39%, F1-score 0.8424
M. M. Imran Mollaet al. [13], 2019	Random Forest	21,3	Accuracy 94.80%, F1-score 0.948
Razman Afridi et al. [28], 2019	Naïve Bayes	16,3	Accuracy 85.88%, F1-score 0.8950
Syifa Fauziyah Nurul Islam and Intan Nurma Yulit [14], 2019	Random Forest	21,3	Accuracy 95.11%
M. Ramlaet al. [20], 2018	CART with Gini Index	21,3	Accuracy 90.12%, F1-score 0.90
Sumedh Anand Sontakke et al. [10], 2018	Random Forest	21,3	Accuracy 93.40%, Kappa score 0.817
Rafael M.O. Cruz et al. [27], 2015	META-DES, ensemble classifier	21,3	Accuracy 84.62%
S. A. A. Shah et al. [21], 2015	Bagging based Random Forest	21,3	Accuracy 94.73%, F1-score 0.9047
Pankaj Bhowmik et al. [This Paper]	**Stacking Ensemble Learning**	**10,3**	**Accuracy 96.05%, Kappa score 0.8875, F1-score 0.9320, AUC score 0.9595**

5. Conclusion

Pregnancy is a natural process, but sometimes it does not run smoothly. Nowadays, fetal abnormalities, including a congenital heart defect, fetal distress, hypoxia, or acidosis are becoming a growing worldwide issue, but early detection can be helpful to combat the forthcoming risks. The CTG data provides crucial information to make a precise diagnosis of the fetus's health state. However, this critical data is inspected manually by the professionals in obstetrics, sometimes which can be misleading and perhaps can lead to life-threatening conditions. Therefore, in this article, a tree-based ensemble learning model is proposed that can be deployed as an automated DSS to predict fetal abnormality on the CTG dataset. The study considered a Stacking ensemble learning approach to build the classifier model. The meta learner of the model is selected by applying 10-folds cross validation on the base learning algorithms. Afterward, building the ensemble learning model, it is employed to predict the CTG samples and the model achieved a favorable accuracy of about 96.05%. The model also proved its superiority in a comparative assessment with base learners.

The existing studies [10,13,14,15,19,20,21] generally used all the available features (21) of the CTG dataset to build their models. Mohammad Saber Iraji [8] implemented several neural network models, among them the DSSAEs (deep stacked sparse auto-encoders) achieved the maximum accuracy of 96.77%, yet the model applied all the 21 features. In contrast, this study utilized less than 50% (10 features) of the defined features from the CTG dataset and achieved an outstanding performance.

Septian Eko Prasetyo et al. [9] applied the Chi-square feature selection technique to select the 10 most important features from CTG data and secured 93.46% accuracy with Random Forest Classifier. But with the same feature selection method and number of features, this study has improved the prediction accuracy by 2.59%. Razman Afridi et al. [28] chose 16 high correlated attributes with a tool called 'Rapid Miner Studio' and achieved the highest accuracy of 85.88% with the Naïve Bayes classifier model.

The outcome of this study reveals that by combining the predictive power of base learning algorithms, the Stacked ensemble learning approach has boosted the overall accuracy. Besides, the overall methodology proposed to develop the ensemble learning model can be applied in several other clinical decision-making applications such as the prediction of diabetes, cancer, heart attack, and so forth. However, the proposed study used only the tree-based algorithms, and the model achieved 'an average' sensitivity and specificity rate (80%–95%) in the minor classes. In the future, the study will attempt to cover a variety of heterogeneous algorithms for ensemble learning and will highly focus on improving the accuracy rate of minor classes in the imbalanced dataset.

References

[1] Dua, D. and Graff, C., "UCI machine learning repository," [Online]. Available: archive.ics.uci.edu/ml/datasets/Cardiotocography

[2] Ayres de Campos, Diogo, et al. "SisPorto 2.0: a program for automated analysis of cardiotocograms", *Journal of Maternal-Fetal Medicine*, vol. 9, no.5, pp. 311-318, Sep-Oct 2000.

[3] Md Zannatul Arif, Rahate Ahmed, Umma Habiba Sadia, Mst Shanta Islam Tultul and Rocky Chakma, "Decision Tree Method Using for Fetal State Classification from Cardiotography Data," *Journal of Advanced Engineering and Computation*, vol. 4, no. 1, pp. 64-73, March 2020.

[4] Abdulhamit Subasi, Bayader Kadasa and Emir Kremic, "Classification of the Cardiotocogram Data for Anticipation of Fetal Risks using Bagging Ensemble Classifier," *Procedia Computer Science*, vol. 168, pp. 34-39, 2020.

[5] Sahana Das, Himadri Mukherjee, Sk. Md. Obaidullah, Kaushik Roy and Chanchal Kumar Saha, "Ensemble based technique for

the assessment of fetal health using cardiotocograph – a case study with standard feature reduction techniques," *Multimedia Tools and Applications*, vol. 79, issue 47-48, pp. 35147 - 35168, April 2020.

[6] Jia-ying Chen, Xiao-cong Liu, Hang Wei, Qin-qun Chen, Jia-ming Hong, Qiong-na Li and Zhi-feng Hao, "Imbalanced Cardiotocography Multi-classification for Antenatal Fetal Monitoring Using Weighted Random Forest," *International Conference, ICSH 2019*, Shenzhen, China, pp. 75-85, July 2019, Springer, Cham.

[7] Hakan Sahin and Abdulhamit Subasi, "Classification of the cardiotocogram data for anticipation of fetal risks using machine learning techniques," *Applied Soft Computing*, vol. 33, pp. 231-238, August 2015.

[8] Mohammad Saber Iraji, "Prediction of fetal state from the cardiotocogram recordings using neural network models," *Artificial Intelligence in Medicine*, vol. 96, pp. 33-44, May 2019.

[9] Septian Eko Prasetyo, Pulung Hendro Prastyo and Shindy Arti, "A Cardiotocographic Classification using Feature Selection: A Comparative Study," *Journal of Information Technology and Computer Engineering (JITCE)*, vol. 5, no. 01, pp. 25-32, March 2021.

[10] Sumedh Anand Sontakke, Jay Lohokare, Reshul Dani and Pranav Shivagaje, "Classification of Cardiotocography Signals using Machine Learning," *2018 Intelligent Systems Conference (IntelliSys)*, pp. 439-450, Springer, Cham.

[11] E. Kannan, S. Ravikumar, A. Anitha, Sathish A. P. Kumar and M. Vijayasarathy, "Analyzing uncertainty in cardiotocogram data for the prediction of fetal risks based on machine learning techniques using rough set," *Journal of Ambient Intelligence and Humanized Computing*, January 2021.

[12] Susan Yuhou Xia, "Using a Stacking Model Ensemble Approach to Predict Rare Events," *Conference Talks, SciPy 2019, 18th annual Scientific Computing with Python Conference*, in Austin, Texas, USA. [Online]. Available: youtube.com/watch?v=6oD5K0x1k7c&t=551s

[13] M. M. Imran Molla, Julakha Jahan Jui, Bifta Sama Bari, Mamunur Rashid and Md Jahid Hasan, "Cardiotocogram Data Classification Using Random Forest Based Machine Learning Algorithm," *Proceedings of the 11th National Technical Seminar on Unmanned System Technology 2019*, *Lecture Notes in Electrical Engineering*, vol 666. Springer, Singapore.

[14] Syifa Fauziyah Nurul Islam and Intan Nurma Yulita, "Predicting Fetal Condition from Cardiotocography Results Using the Random Forest Method," *7th Mathematics, Science, and Computer Science Education International Seminar, MSCEIS 2019*, Bandung, West Java, Indonesia.

[15] Yandi Chen, Ao Guo, Qinqun Chen, Bin Quan, Guiqing Liu, Li Li, Jiaming Hong, Hang Wei and Zhifeng Hao, "Intelligent classification of antepartum cardiotocography model based on deep forest," *Biomedical Signal Processing and Control*, vol. 67, Article 102555, May 2021.

[16] Satish Chandra Reddy Nandipati, "Classification and Feature Selection Approaches for Cardiotocography by Machine Learning Techniques", *Journal of Telecommunication, Electronic and Computer Engineering*, vol. 12, no. 1, pp. 7-14, January - March 2020.

[17] Zhi-Hua Zhou and Ji Feng, "Deep Forest," *National Science Review*, vol. 6, no. 1, pp. 74-86, Jan. 2019.

[18] Sundar.C, M. Chitradevi and G. Geetharamani, "Classification of Cardiotocogram Data using Neural Network based Machine Learning Technique", *International Journal of Computer Applications*, vol. 47, no. 14, pp.19-25, June 2012.

[19] K. Agrawal and H. Mohan, "Cardiotocography Analysis for Fetal State Classification Using Machine Learning Algorithms," *2019 International Conference on Computer Communication and Informatics (ICCCI)*, 2019, pp. 1-6, doi: 10.1109/ICCCI.2019.8822218

[20] M. Ramla, S. Sangeetha and S. Nickolas, "Fetal Health State Monitoring Using Decision Tree Classifier from Cardiotocography Measurements," *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2018, pp. 1799-1803, doi: 10.1109/ICCONS.2018.8663047

[21] S. A. A. Shah, W. Aziz, M. Arif and M. S. A. Nadeem, "Decision Trees Based Classification of Cardiotocograms Using Bagging Approach," *2015 13th International Conference on Frontiers of Information Technology (FIT)*, 2015, pp. 12-17, doi: 10.1109/FIT.2015.14

[22] P. Bhowmik, M. Sohrawordi, U. A. M. Ehsan Ali, M. N. Hasan and P. K. Roy, "Analysis of Social Media Data to Classify and Detect Frequent Issues Using Machine Learning Approach," *2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT)*, 2020, pp. 394-399, doi: 10.1109/ICAICT51780.2020.9333452

[23] Xianping Du, Hongyi Xu and Feng Zhu, "Understanding the Effect of Hyperparameter Optimization on Machine Learning Models for Structure Design Problems," *Computer-Aided Design*, vol. 135, Article 103013, 2021, doi: 10.1016/j.cad.2021.103013

[24] Abdullah Al Imran, Ananya Rahman, Humayoun Kabir, Shamsur Rahim, "The Impact of Feature Selection Techniques on the Performance of Predicting Parkinson's Disease", *International Journal of Information Technology and Computer Science (IJITCS)*, Vol.10, No.11, pp.14-29, 2018. DOI: 10.5815/ijitcs.2018.11.02

[25] Sahana Das, Kaushik Roy, Chanchal K. Saha, "Establishment of Automated Technique of FHR Baseline and Variability Detection Using CTG: Statistical Comparison with Expert's Analysis", *International Journal of Information Engineering and Electronic Business (IJIEEB)*, Vol.11, No.1, pp. 27-35, 2019. DOI: 10.5815/ijieeb.2019.01.04

[26] Kemal Akyol, Baha Şen, "Diabetes Mellitus Data Classification by Cascading of Feature Selection Methods and Ensemble Learning Algorithms", *International Journal of Modern Education and Computer Science (IJMECS)*, Vol.10, No.6, pp. 10-16, 2018. DOI: 10.5815/ijmecs.2018.06.02

[27] Rafael M.O. Cruz, Robert Sabourin, George D.C. Cavalcanti, Tsang Ing Ren, "META-DES: A dynamic ensemble selection framework using meta-learning", *Pattern Recognition*, vol. 48, no. 5, pp. 1925-1935, May 2015, DOI: 10.1016/j.patcog.2014.12.003

[28] R. Afridi, Z. Iqbal, M. Khan, A. Ahmad, and R. Naseem, "Fetal Heart Rate Classification and Comparative Analysis Using Cardiotocography Data and Known Classifiers", *International Journal of Grid and Distributed Computing (IJGDC)*, ISSN: 2005-4262 (Print); 2207-6379 (Online), NADIA, vol. 12, no. 1, pp. 31-42, Jun 2019.

Authors' Profiles

Pankaj Bhowmik received B. Sc. (Engg.) degree in Computer Science and Engineering from Hajee Mohammad Danesh Science and Technology University (HSTU), Dinajpur, Bangladesh in 2018. He has prodigious enthusiasm to explore the contemporary horizon of Computer Science fields. His leading research areas are Machine Learning and its real-life applications, Data Science, Natural Language Processing, Artificial Intelligence, Data Mining and Study of sustainable development. He achieved two consecutive Dean's Awards for excellent academic performance while pursuing his Bachelor's degree.

Pulak Chandra Bhowmik received his Bachelor's degree in Computer Science and Engineering from Stamford University Bangladesh, Dhaka, Bangladesh in 2019. He is currently working as an IT support Engineer at Flight Expert, Dhaka, Bangladesh. His major working interest is based on Software Engineering, Computer Hardware, Computer Networking, Machine Learning and Cloud Computing.

U. A. Md. Ehsan Ali received his B. Sc. degree in Computer Science and Engineering from Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh in 2013. Now, he is pursuing M. Sc. degree in Computer Science and Engineering from Rajshahi University of Engineering & Technology (RUET), Rajshahi, Bangladesh. His main working interest is based on Image Processing, Expanding the Applications of Artificial Intelligence, Machine Learning, Data Mining, Data Security etc. Currently, he is working as an Assistant Professor in Dept. of Computer Science and Engineering in Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh. He has several scientific research publications in various aspects of Computer Science and Engineering.

Md. Shohrawordi is working as an Assistant Professor in Dept. of Computer Science and Engineering in Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh. He received his B. Sc. degree in Computer Science and Engineering from Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh in 2013. Now, he is pursuing M. Sc. degree in Computer Science and Engineering from Rajshahi University of Engineering & Technology (RUET), Rajshahi, Bangladesh. His main working interest is based on Image Processing, Artificial Intelligence, Data Mining, Mobile Networks, cryptography etc. He has several scientific research publications in various aspects of Computer Science and Engineering.

How to cite this paper: Pankaj Bhowmik, Pulak Chandra Bhowmik, U. A. Md. Ehsan Ali, Md. Sohrawordi, "Cardiotocography Data Analysis to Predict Fetal Health Risks with Tree-Based Ensemble Learning", International Journal of Information Technology and Computer Science(IJITCS), Vol.13, No.5, pp.30-40, 2021. DOI: 10.5815/ijitcs.2021.05.03

