Autonomous Fetal Distress Detection based on performance of Machine learning algorithms

A. Venkata Sriram[1], P. Rajesh Kumar[2], L. Alekhya[3]

[1]Lecturer, Electronics and Communication Department, Government Polytechnic, Parvathi Puram, Andhra Pradesh, India, Email: avsriram406@gmail.com
[2]Professor, Electronics and Communication Department, Andhra University College of Engineering, Visakhapatnam, India, Email: rajeshauce@gmail.com
[3]Assistant Professor, Electronics and Communication Department, Lendi Institute of Engineering and Technology, Vizianagaram, Andhra Pradesh, India, Email: chinni2788@gmail.com

Received: 13.04.2024 Revised: 15.05.2024 Accepted: 25.05.2024

ABSTRACT
Cardiotocography (CTG) is a medical tracking procedure used to assess the well-being of the foetus by monitoring the patterns of its heart ratein response to the mother's signal. Although CTG is predominantly used tool to monitor and detect the health of the foetus, the increase in the results of false alarm rates due to visual deciphering highly constitute to unnecessary operative delivery or delayed intrusion. A novel automatic process is proposed here for early diagnose and detect of foetus abnormality using machine learning approach. The dataset is taken from Cardiotocography UCI repository which holds 2126 instances with normal, suspect, and pathological (N, S, P) classes obtained from measurements of uterine contraction (UC) and fetal heart rate (FHR)features. Following feature scaling and normalization, the feature data is fed into machine learning models like Naive Bayes, Support Vector Machine, k-Nearest Neighbor, and Random Forest techniques to classify the imbalanced data into multiclass categories N, S, P.The various performance metrics were calculated for four algorithms and the results show that Random Forest within computational time of 6.32shas obtained overall Accuracy of 90.82%, weighted F_1 score of 91.24%, mean MCC (Mathews Correlation Coefficient) of about 74%Mean Kappa Score of 72.82% and Averaged Area under the ROC of 0.8766 which is better when compared to other algorithms. Hence Random Forest method can be used to autonomously detect the fetaldistress during pregnancy.

Keywords: Cardiotocography, Support Vector Machine, Random Forest, Mathews Correlation Coefficient, Naive Bayes, k-Nearest Neighbour.

INTRODUCTION

Fetal heart activity provides crucial information about the health of the fetus both before and after birth. Cardiotocography (CTG) is a widely utilized method globally to assess maternal uterine pressure (UP) and fetal heart rate (FHR) simultaneously. Information about maternal uterine contractions was captured through a qualitative method known as Tocography [1], which furnishes essential insights into the intensity and duration of these contractions. Tocography employs an external tocodynamometer [2], functioning as a pressure transducer [3]. Unlike other techniques such as fetal stethoscope, Doppler ultrasound, and electronic fetal monitoring (EFM), which may not reliably record fetal heart rate (FHR) and necessitate skilled interpretation [4], cardiotocography (CTG) has gained prominence. Consequently, Cardiotocograms, depicting uterine contractions (UC) and FHR, are predominantly utilized for identifying fetal distress and categorizing fetal health status as normal, pathological, or suspect through feature extraction. By analyzing features based on both linear and nonstationary interactions of FHR and UC signals in CTGs, health risks for both mother and infant are minimized.

Cardiotocography was introduced in the late 1960s and remains the primary method for detecting intrapartum hypoxia and fetal distress. However, despite its widespread use, the improvements in delivery outcomes have not met expectations when compared to the previously utilized intermittent auscultation method. In 1986, the International Federation of Gynecology and Obstetrics (FIGO) introduced general guidelines for evaluating macroscopic morphological features of fetal heart rate (FHR) and their correlation with tocographic measurements. Despite these guidelines being available for nearly thirty-five years and the introduction of the first unified FIGO guidelines aimed at facilitating computer evaluation of CTG signals, unsatisfactory interpretations of CTG persist. Nevertheless, advancements in automatic systems for CTG analysis, particularly those relying on automatically extracted morphological features, have significantly enhanced the ability to detect fetal distress early. These systems contribute to the overall improvement of maternal and fetal health outcomes.

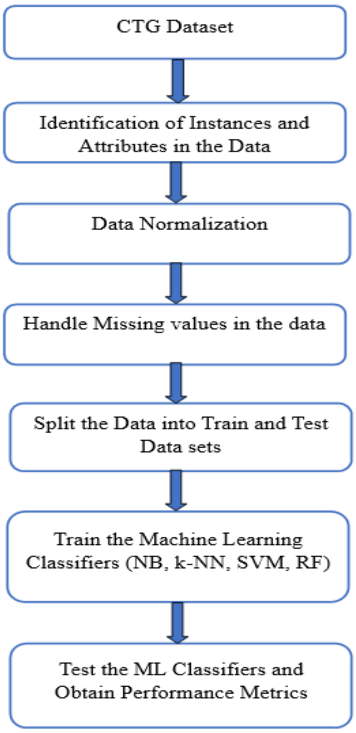

CTG Dataset
Identification of Instances and Attributes in the Data
Data Normalization
Handle Missing values in the data
Split the Data into Train and Test Data sets
Train the Machine Learning Classifiers (NB, k-NN, SVM, RF)
Test the ML Classifiers and Obtain Performance Metrics

Figure 1. Block Diagram of Proposed Method.

The primary aim of this paper is to develop a precise and resilient artificial intelligence system capable of identifying and assessing fetal distress during labor, thereby categorizing the health status as normal, suspect, or pathological. Figure 1 illustrates the block diagram depicting the proposed methodology. Utilizing the CTG dataset sourced from the UCI repository, the evaluation of fetal distress is conducted. The dataset undergoes scrutiny for any missing values, with appropriate measures taken to address them if found. Subsequently, normalization and feature scaling are applied to prepare the data for deployment across four different Machine Learning classification algorithms. The data is split into training and testing sets in4:1 ratio, with the training set utilized for model training. The test set is then employed to classify the fetal distress state as Normal, Suspect, or Pathological. Given the unbalanced nature of the data, the Mean of Matthews Correlation Coefficient (MCC), Kappa score and ROC-AUC are calculated along with remaining metrics to evaluate the performance of the classification model. The primaryfocus of this study encompass:

1) Data preprocessing of the CTG dataset using MATLAB version 2023b program which include data cleaning, handling the missing values and data normalization for the implementation of Autonomous Fetal distress detection algorithms.
2) Train and Testdata set of the preprocessed CTG data is applied to the four Machine learning based algorithms namely Naive Bayes, SVM, k-NN and Random Forest and verify and compare the performance of these models with various metrics evaluated from confusion matrix of multiclass classification problem.
3) Random Forest algorithm has achieved an overall accuracy of 90.82%, weighted F_1 score of 91.24%, Mean MCC of 74%, Mean Kappa Score of 72.82% and averaged area under the ROC is 0.8766were obtained in this study which is best when compared to remaining ML based algorithms.

The subsequent section will introduce relevant literature and theories associated with the proposed study. Section III will detail the research methodology and the proposed approach. Section IV will present the experimental procedures and outcomes. Discussions and conclusions will be provided in Sections V and VI, respectively.

2. Related Works

Cardiotocography serves as a crucial tool in identifying fetal distress during both the antepartum and intrapartum stages. Recent studies have witnessed enhanced automatic analysis of digitized CTG data, thanks to the emergence of advanced algorithms within the realm of artificial intelligence. The digitization of CTG data has significantly contributed to the automation of fetal distress detection.

In a study by S. Öztürk et al. [9], the utilization of the empirical mode decomposition technique facilitated the digitization of CTG paper, enabling the extraction of various spectral entropy features. By employing a Support Vector Machine classifier with ReliefF as the feature selection algorithm, an accuracy of approximately 90% was attained in classifying the computerized CTG data into normal and abnormal classes. Additionally, Verburg et al. [10] conducted a comparative analysis between computerized CTG and analog CTG in detecting fetal distress. Their findings indicate a notable enhancement in accuracy and sensitivity, with digitalized CTG demonstrating a remarkable 94.9% sensitivity compared to traditional CTG methods.

H Sahin et al. [11] have used eight various Machine learning algorithms to classify the CTG dataset from UCI repository as normal and pathological on 1831 instances, a binary classification approach and obtained a highest accuracy of 99.2% random forest algorithm. Y. Zhang et al. [12] has used Adaboost classifier algorithm with PCA as feature selection on the same CTG UCI repository dataset and obtained an accuracy of 98.6% and compared the same with SVM classifier with an accuracy of 97.7%.

S. Dash et al. [13]used Bayesian with Generative models for the improvement of automatic FHR classification approach and obtained a best weighted relative accuracy (WRA) of 0.425 for GM-MM or NB-C.The prediction of fetal acidemia by analyzing digitalized CTG traces usingsignal-processing algorithm is developed by Ayres-de-Campos et al. [14]. This study reviewthat the technique had a sensitivity and specificity of 88.2%,85.9% respectively in predicting fetal acidemia.

Rana et al. [15] employed amachine learning classifiers with ensemble average to categorize CTG signals in three prescribed classes achieving an accuracy of 98.4%. This underscores the ability of utilizing machine learning algorithms for automated CTG interpretation. Chudacek et al. [16] developed an algorithm based on FHR and UC patterns to classify CTG signals, yielding a classification accuracy of 87.3%, indicative of its medical utility. Similarly, Chudacek et al. [17] investigated FHR variability and acceleration, devising a technique that obtained an accuracy of 88.8%, suggesting the probability for enhanced accuracy with more specific CTG signal analysis.

In the proposed method, four machine learning algorithms—Naive Bayes, k-Nearest Neighbor, Support Vector Machine, and Random Forest—are employed to classify the CTG dataset into Normal, Suspect, and Pathological categories. Given the dataset's multiclass and unbalanced nature, performance metrics such as accuracy, sensitivity, specificity, Matthews Correlation Coefficient (MCC), and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) were calculated to evaluate the algorithms. Notably, the Random Forest algorithm demonstrated superior performance in this study.The development of accurate methods for CTG-based irregularity detection is crucial for providing early warnings to both patients and clinicians. These findings highlight the ability of machine learning algorithms as valuable tools for assessing unfavourablefetal outcomes.

3. RESEARCH METHODOLOGY

In this study, the analysis focuses on the CTG dataset sourced from the UCI Machine Learning Repository, originally obtained from the SisPorto 2.0 software adhering to FIGO guidelines. The objective is to evaluate the effectiveness of machine learning models. The dataset undergoes preprocessing, encompassing the handling of missing values and data normalization. Subsequently, the data is split into training and testing datasets, as depicted in Figure 1. The training data is utilized to train four distinct machine learning algorithms. Various performance metrics, including ROC-AUC curves obtained through Confusion Matrix analysis, are then examined to address the multiclass classification problem of fetal distress detection.

A. Preprocessing

The CTG data set considered from UCI machine learning repository contains 2126 instances with 21 related features. The features values of this feature set are in different ranges and some values may be missed or may contain NaN values. The steps involved in the data preprocessing stage is shown in the Figure 2.To apply these values to next stage i.e., classification stage the missing values must be handled by either removing themor replacing them with mean of the feature column where that value is present.

Once the missing values are handled, the feature scaling is applied using z-score normalisation to improve convergence speed of gradient descent algorithms and enhances the potential of the machine learning models. After proper preprocessing, the data set can now be applied to classification stage.

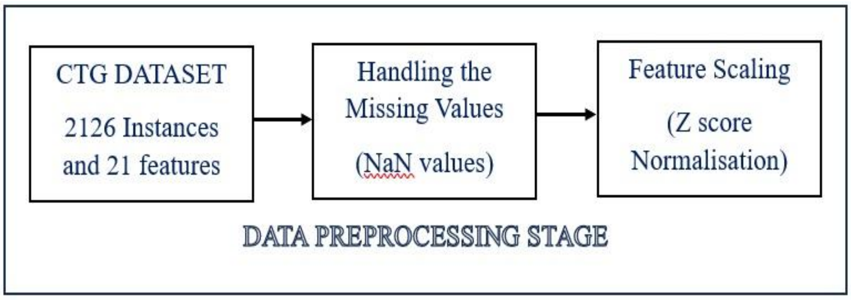


Figure 2. Steps involved in Data Preprocessing Stage

B. Dataset Description

Cardiotocography serves as a method for monitoring both fetal heart activity and uterine contractions. The dataset utilized in this study is sourced from the UCI Machine Learning Repository, recorded via the Sisporto version 2.0 program [18], which can be installed on any personal computer for signal acquisition. This automated program adheres closely to FIGO guidelines for the analysis of fetal distress detection. The dataset comprises 2126 instances and 23 attributes [19], primarily derived from fetal heart rate (FHR) baseline, uterine contractions per second (UC), and fetal movements per second (FM). Additionally, several other attributes contribute to the recognition of fetal health status. Among these attributes, four are deemed fundamental and critical in CTG data analysis: Fetal Heart Rate Baseline (BL), Accelerations (ACC), Decelerations (DCL), and Variability.

C. Naive Bayes

Naive Bayes classifier is probability based classier which uses the Bayes Theorem which is simple and most efficient. It is relevantto Bayesian network in which all attributes are independent of given class variable. This conditional independency of the attributes of Bayes theorem can be called as Naive bayes [20]. The probability function for Naive Bayes classifier for multiclass classification is given as;

$$p(A = k|B_1, B_2, \dots.. B_p) = \frac{\pi(A=k)\prod_{j=1}^{p} p(B_j|A=k)}{\sum_{k=1}^{k}\pi(A=k)\prod_{j=1}^{p} p(B_j|A=k)} \quad (1)$$

Equation (1) computes the probability value [21], where k represents the number of classes for classification. Here, A denotes the random variable associated with the class index of an observation, while B_1, B_2, ..., B_P represent the random predictors of an observation. Additionally, π(A=k) signifies the prior probability that a class index is k.

D. k-Nearest Neighbour

The k-Nearest Neighbor algorithm is a straightforward supervised machine learning technique that falls under the category of example-based learning. It classifies data by assessing the similarity of each data point to others [22]. Initially, the dataset under consideration is divided into training and testing subsets, with the algorithm learning from the training data and grouping it into predefined categories. To classify the test data into categories, one must specify the number of k neighbors to consider. Compute the k neighbors of the test data point according to some distance measure such as Euclidean distance given by Equation (2)

$$d_{xy} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} (2)$$

Where (x_2,y_2) is training object coordinates and (x_1,y_1) is testing object coordinates.

The algorithm counts number of data points from each category among the k neighbours computed previously. The test data point is assigned to the category with most neighbours and the process endureson all test data points for chosen k neighbours.

E. Support Vector Machine

Support Vector Machine (SVM) is widely recognized for its effectiveness in solving classification problems, particularly for datasets with high dimensions (i.e., numerous features) [23]. The fundamental principle of SVM revolves around identifying the optimal maximum margin hyperplane (MMH). In constructing this hyperplane, SVM selects the most extreme points or vectors, termed support vectors, hence lending its name to the algorithm. Error-Correcting Output Codes (ECOC) is a technique employed to address multi-class classification problems by transforming them into multiple binary classification tasks [24]. Therefore, SVM can be adapted to handle multiclass classification problems through the utilization of the ECOC method.

F. Random Forest

A random forest is a classifier comprising multiple decision trees. A training forest which is composed of multiple classification trees is used to create a random forest. It is more accurate and stable prediction. It trains the model with the bagging method. The classification result of the test data is obtained by the score formed by the classification tree voting [25].

G. Multiclass classification Metrics

Multiclass classification is such scenario where the number of output units/classes are greater than two. In this classification problem we need evaluate the parameter based on One vs Rest algorithm. One vs Rest is based on considering one class and evaluate the parameters of that one class with respect to the rest of all class parameters from the confusion matrix only. In assessing the performance of machine learning models, the Confusion Matrix serves as a pivotal tool, comprising an N x N matrix where N represents the number of target classes. In this study, N equals 3. Unlike binary class classification, parameters are evaluated individually for each class in a multiclass scenario. Consequently, their measurement varies and is contingent upon the specific class. These parameters are calculated separately for each class, and subsequently, the relevant performance metrics are determined using the formulas outlined in Table 1 below.

Table 1. Performance Metrics Formulae for the Evaluation of the ML models

Performance Metrics	Formulae
Accuracy of the Class	$\frac{(\mathrm{TP_k} + \mathrm{TN_k})}{(\mathrm{TP_k} + \mathrm{TN_k} + \mathrm{FP_k} + \mathrm{FN_k})}$
Sensitivity	$\frac{\mathrm{TP_k}}{(\mathrm{TP_k} + \mathrm{FN_k})}$
Specificity	$\frac{\mathrm{TN_k}}{(\mathrm{TN_k} + \mathrm{FP_k})}$
Precision	$\frac{\mathrm{TP_k}}{(\mathrm{TP_k} + \mathrm{FP_k})}$
$\mathrm{F_1}$ Score	$\frac{2 * \mathrm{TP_k}}{(2 * \mathrm{TP_k} + \mathrm{FP_k} + \mathrm{FN_k})}$
Mathews Correlation Coefficient	$\frac{(\mathrm{TP_k} * \mathrm{TN_k} - \mathrm{FP_k} * \mathrm{FN_k})}{\sqrt{(\mathrm{TP_k} + \mathrm{FP_k})(\mathrm{TP_k} + \mathrm{FN_k})(\mathrm{TN_k} + \mathrm{FP_k})(\mathrm{TN_k} + \mathrm{FN_k})}}$
Cohen Kappa Score	$\frac{2 * (\mathrm{TP_k} * \mathrm{TN_k} - \mathrm{FP_k} * \mathrm{FN_k})}{(\mathrm{TP_k} + \mathrm{FP_k}) * (\mathrm{TN_k} + \mathrm{FP_k}) + (\mathrm{TP_k} + \mathrm{FN_k}) * (\mathrm{TN_k} + \mathrm{FN_k})}$
Classification Error	1-Accuracy

Since the study focuses on multiclass classification problem with three classes, the value of k is 3 in the above metrics and are calculated for each class individually. The overall accuracy of the model is obtained from Equation (3) given below.

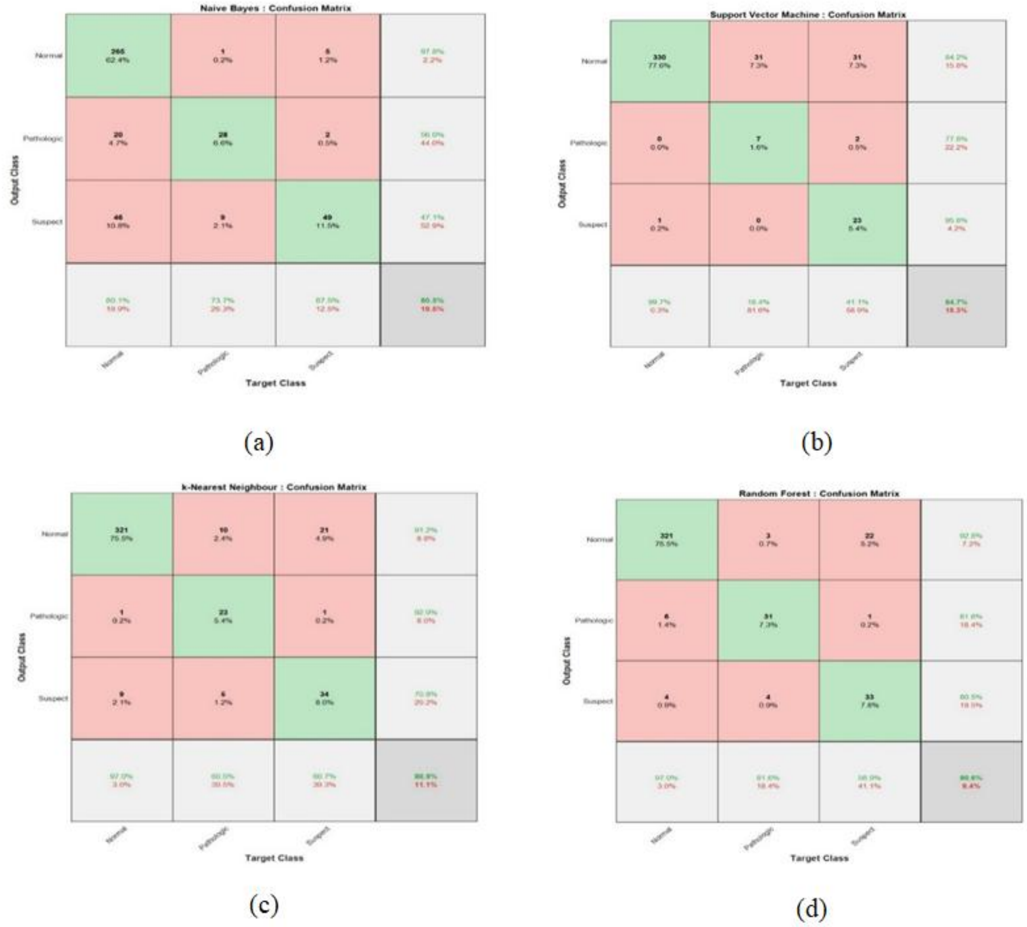
Overall Accuracy of the model= $\frac{\text{Correctly predicted samples}}{\text{Total samples}}$ (3)

Another essential metric frequently utilized in analyzing false positive rates is the Receiver Operating Characteristic (ROC) curve, which plots the false positive rate (FPR) against the true positive rate (TPR), where TPR corresponds to sensitivity and FPR is given as 1-specificity. This curve aids in selecting an optimal cut-off value for determining the classes utilized in the study. Additionally, the Area Under the Curve (AUC) is calculated from the ROC curve, with its value ranging between 0 and 1. A higher AUC value, closer to 1, signifies superior performance of the classifier model.

In addition to overall accuracy, mean Matthews Correlation Coefficient (MCC), weighted F1 score, mean Cohen Kappa score, and misclassification error were computed for analyzing the machine learning models. For unbalanced data, Mathew's correlation coefficient (MCC) and Cohen Kappa score are preferred over accuracy and F1 score, as they provide more appropriate measurements of the model's ability to classify the problem. MCC values range between -1 and 1, with 1 indicating the best prediction model and -1 indicating the worst case scenario. Cohen Kappa score values help assess the level of agreement between actual and predicted classes in the classification problem, with ranges indicating various levels of agreement: 0.21-0.4 (Fair agreement), 0.41-0.6 (Moderate agreement), 0.61-0.8 (Substantial agreement), and 0.81-0.99 (Almost perfect agreement). These metrics are calculated and tabulated in the Results and Discussions section.

4. Experimental Analysis

Experiments were conducted on a computer having an Intel i5 microprocessor and 16 GB Ram and Nvidia GPU of 2 GB, while coding were carried out on MATLAB 2023b version software. The study was conducted on the CTG dataset of size 2126 x 23 which include 21 and 1 for 3 class and the other 1 for 10 class classification labels together constitute 23 columns in the dataset. Out of 2126 instances 1655 belongs to Normal, 295 to Suspect and 176 to pathologic classes respectively. Hence the data size becomes 2126 x 21 with 3 class classification columns of the data taken as labels in this study. This data along with labels are divided into train and test dataset in the ratio of 80 and 20 percent respectively. The train dataset is applied as input to four Machine learning algorithms namely Naïve bayes, SVM, k-NN and Random Forest to train the models and the test dataset is used evaluate these models and calculate various performance metrics using Confusion plotas discussed in Section II.



(a)

(b)

(c)

(d)

Figure 3. Confusion Matrix plot obtained for four Machine Learning Algorithms (a) Naive Bayes(b)Support vector machine (c) k-Nearest Neighbour (d) Random Forest models

While training Naive Bayes algorithm the train dataset is applied to the model with kernel distribution as specification. For training the SVM algorithm the kernel function of radial basis function (rbf) is used. For training the k-NN algorithm number of neighbors as 9 is taken as specification. Finally for Random Forest model training Bagging method with Maximum number of splits as 3 are taken as specifications. Once the train dataset is trained to all these models with respective specifications, then the test dataset is applied to evaluate the models and obtain confusion matrix and calculate other performance metrics as well. The Confusion matrix plot for each model is shown in the Figure 3. Once the confusion matrix is obtained the performance metrics are evaluated for each model of machine learning algorithms from the parameters obtained in the confusion matrix as discussed in the section II.

Table 2 gives the performance metrics evaluated for each class namely Normal, Suspect and pathologic using Naïve bayes algorithm. This algorithm is good at detecting the pathologic samples aptly with an accuracy of 92.47% and specificity of 94.32% which is better in comparison to other class samples.

Table 2. Performance Metrics evaluated for Naïve Bayes Algorithm

Performance Metrics	**Normal**	**Pathologic**	**Suspect**
Class Accuracy	0.8306	0.9247	0.8541
Sensitivity	0.8006	0.7368	0.8750
Specificity	0.9362	0.9432	0.8509
Precision	0.9779	0.5600	0.4712
F_1 Score	0.8804	0.6364	0.6125
MCC	0.6362	0.6022	0.5711
Cohen Kappa score	0.5997	0.5952	0.5324
AUC	0.9340	0.8311	0.8927

Table 3 gives the performance metrics evaluated using SVM algorithm for each class individually. The class accuracy and specificity are good for both pathologic and suspect class with 92% and 99% each respectively but very poor sensitivity and kappa score due to the impact of unbalanced dataset. The Table 4 gives the performance metrics evaluated for fetal distress detection using k-NN algorithm. This algorithm is good at detecting pathologic samples with a better class accuracy of 96%, specificity of 99.48% and AUC of 0.9784 which good when compared to Naïve Bayes and SVM algorithms. Also, the other class samples are also detected with better accuracy and AUC but poor in MCC and Kappa score evaluation.

Table 3. Performance Metrics evaluated for SVM Algorithm

Performance Metrics	Normal	Pathologic	Suspect
Class Accuracy	0.8518	0.9224	0.9200
Sensitivity	0.9970	0.1842	0.4107
Specificity	0.3404	0.9948	0.9973
Precision	0.8418	0.7778	0.9583
F_1 Score	0.9129	0.2979	0.5750
MCC	0.5233	0.3548	0.5979
Cohen Kappa score	0.4395	0.2730	0.5385
AUC	0.9596	0.9491	0.9579

Table 4. Performance Metrics evaluated for k-NN algorithm

Performance Metrics	**Normal**	**Pathologic**	**Suspect**
Class Accuracy	0.9035	0.9600	0.9153
Sensitivity	0.9698	0.6053	0.6071
Specificity	0.6702	0.9948	0.9621
Precision	0.9119	0.9200	0.7083
F_1 Score	0.9400	0.7302	0.6538
MCC	0.7042	0.7277	0.6083
Cohen Kappa score	0.6956	0.7095	0.6059
AUC	0.9671	0.9784	0.9487

Table 5 gives the performance metrics evaluated for fetal distress detection using Random Forest with Bagging method. This algorithm is good at detection all the three classes Normal, Suspect, Pathologic with a better class accuracy of 92%, 96.71% and 92.94% respectively. Also other performance metrics like AUC, MCC and Kappa score are having better values when compared to Naïve bayes, SVM and k-NN algorithms for all the three classes.

Table 5. Performance Metrics evaluated for Random Forest algorithm

Performance Metrics	Normal	Pathologic	Suspect
Class Accuracy	0.9200	0.9671	0.9294
Sensitivity	0.9698	0.8158	0.6071
Specificity	0.7447	0.9819	0.9783
Precision	0.9304	0.8158	0.8095
F_1 Score	0.9497	0.8158	0.6939
MCC	0.7586	0.7977	0.6636

Kappa score	0.7465	0.7977	0.6404
AUC	0.9343	0.9503	0.9051

Table 6 gives the Overall Performance metrics of all the four Machine learning classifier algorithms for the detection of fetal distress and classify into Normal, Suspect and Pathologic state. Of all the four algorithms the proposed Random Forest algorithm gives better overall accuracy of 90.82%, weighted F_1 score of 91.24%, Mean MCC of 74%, Mean kappa score of 72.82% and averaged AUC of 87.66% and low classification error of 9.18% when compared to other three considered machine learning algorithms namely Naïve Bayes, SVM, k-NN. But in terms of elapsed time taken by the algorithm Random Forest has 6.32 seconds which is high when compared to other models i.e., 0.4s higher to SVM and k-NN and 0.3s to Naïve Bayes models. It can be negligible due to its high performance in terms of other metrics.

Table 6. Overall Performance Metrics of the Machine Learning Models

Performance Metrics	Naïve Bayes	SVM	k-NN	RF
Overall Accuracy	80.47%	84.71%	88.94%	90.82%
Classification Error	19.53%	15.29%	11.06%	9.18%
Weighted F_1 score	78.61%	88.08%	89.53%	91.24%
Mean MCC	60.32%	49.20%	68.01%	74%
Mean Kappa Score	57.58%	41.70%	67.04%	72.82%
Averaged AUC	88.94%	75.80%	92.78%	87.66%
Elapsed Time	6.06s	5.59s	5.53s	6.32s

Also averaged AUC metric is less for Random Forest model when compared to Naïve Bayes and k-NN models. But when 0.5<AUC<1, there is a high chance that the classifier will be able to distinguish the positive class values from the negative ones. So, in this study all the four classifiers are good at distinguishing the positive class from negative class in terms of averaged AUC. The Figure 4 gives the comparison chart of all the four algorithms in terms of Overall Performance metrics. As shown in the Table 8 the Random Forest gives better results when compared to other algorithms in terms of all the performance metrics.

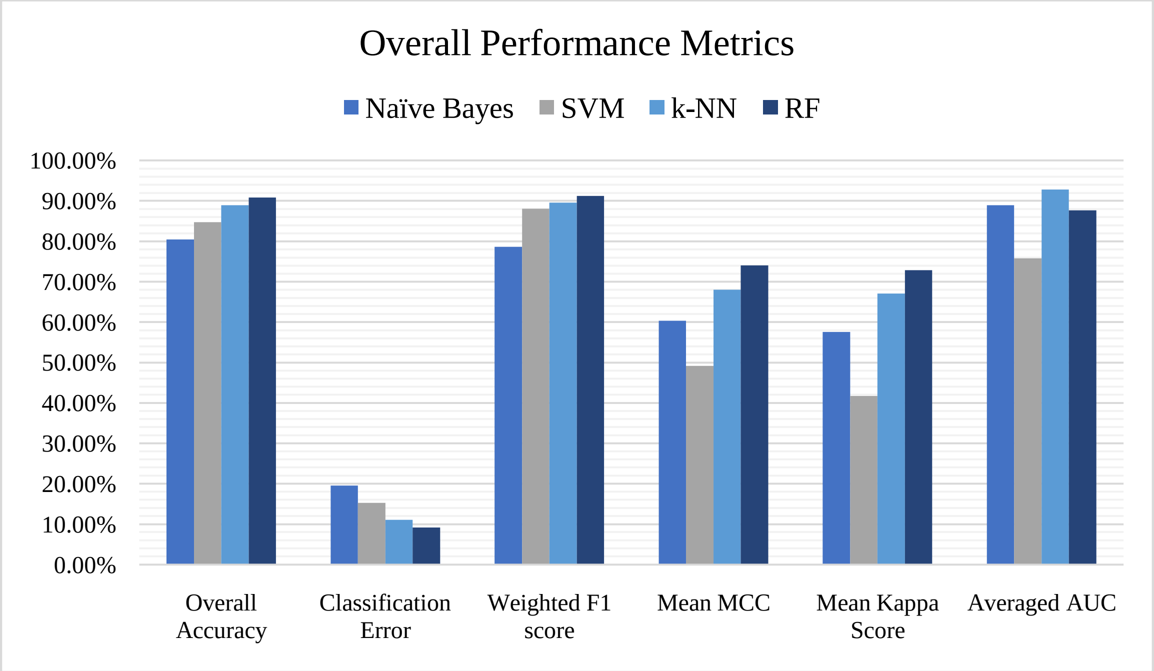
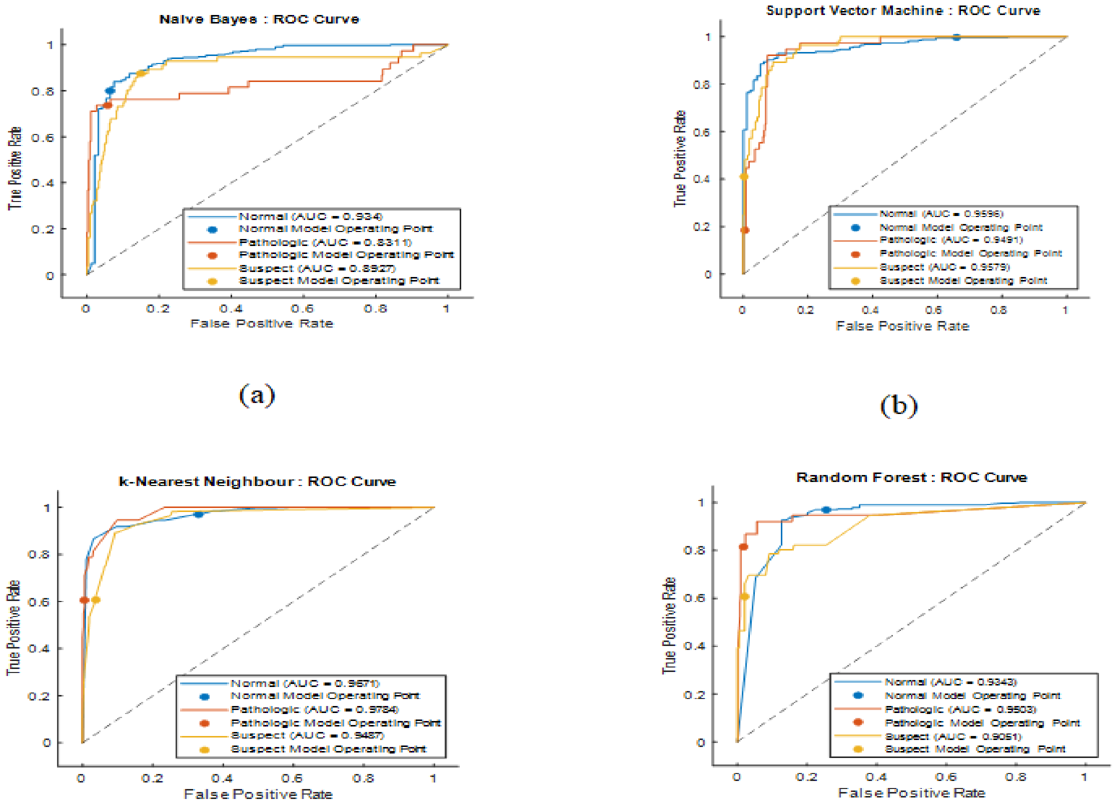


Figure 4. Comparison Chart of Overall Performance Metrics of Machine learning models

Figure 5 given below shows the various plot of ROC curves for four Machine learning models. The area calculated under these curves gives the averaged AUC metric tabulated in the Table 8 for all the four machine learning models. This ROC-AUC metric plays an important role in distinguishing between positive and negative classes. The ROC graph is between True positive rate and False positive rate for each class as study is on the multiclass classification problem. So, there are three individual graphs for three classes and Area under the curve is obtained individually for each class for all the four algorithms as given in the Figure 5 (a), (b), (c),(d) and are tabulated respectively above.



(a)

(b)

(c)

(d)

Figure 5. Receiver Operating Characteristic Curves for four Machine learning models (a) Naïve Bayes (b) SVM (c) k-Nearest Neighbour (d) Random Forest.

5. RESULTS AND DISCUSSIONS

In this study four different supervised machine learning algorithms were applied to CTG dataset taken from UCI repository. These algorithms were trained with the 80% samples of the preprocessed CTG data and remaining 20% of the data were used to test the trained machine learning models namely Naïve Bayes, SVM, k-NN and Random Forest models. Confusion matrix for each model is obtained which is multiclass classification problem. From this confusion matrix various performance metrics were evaluated and a satisfied performance was obtained for the proposed Random Forest algorithm with bagging method.

Table 7shows the comparison of the proposed method with various other studies conducted so far. Ricciardiet al. [26] used SVM model on binary class classification as Normal and Suspicious samples of the CTG data and achieved and Overall Accuracy of 92% and F_1 score of 0.919. Krupa et al.[27] used Empirical mode decomposition for CTG data preprocessing and SVM for the classification of the samples into Normal and At Risk, a binary class problem and achieved an Overall accuracy of 81.5% and Kappa score of 0.684.M. Ajiraket al. [28] used AdaBoost and Ensemble learning methods for both preprocessing and classification of binary class in to normal and hypoxic class and obtainedOverall accuracy of 79.8%, AUC of 0.796 and F_1 score of 78.5% respectively.

Georgoulas et al. [29] used SVM model with RBF as Kernel function on binary class data and achieved an Overall accuracy of 81.25% and AUC of 0.78 in classification of fetal health state into Normal and Suspicious classes respectively. N. Chamidahet al.[30] used Hybrid k-means for preprocessing and SVM for classification on multiclass classification of CTG data to classify into Normal, Suspect and Pathological classes and achieved an Overall accuracy of 90.64%. The proposed method used statistical preprocessing techniques and standard normalization on data and Random Forest for classification and achieved an

Overall accuracy of 90.82%, mean MCC of 74%, kappa score of 0.728, AUC of 0.876 and mean F_1 score of 91.24% for the autonomous classification of fetal health distress into Normal, Suspect and Pathological classes.

Table 7. Summary of Comparison with State-of-the-art models.

References	Method	Performance	Class
Ricciardiet al. [26]	SVM	A = 92% F_1 score=0.919	Normal, Suspicious (Binary class)

Krupa et al. [27]	EMD + SVM	A = 81.5%
		Kappa = 0.684
			Normal,
			At Risk
			(Binary class)

M. Ajiraket al. [28]	AdaBoost + Ensemble learning	A = 79.8%
		AUC = 0.796
		F_1 score=78.5%
			Normal, Hypoxic (Binary Class)

Georgoulas et al. [29]	SVM (RBF kernel)	A = 81.25%
		AUC = 0.78
			Normal, Suspicious (Binary class)

N. Chamidahet al.[30]	Hybrid k-Means + SVM	A = 90.64%	Normal, Suspect, Pathological (Multi class)

Proposed Method*	Random Forest
	(DT with Bagging)
		A = 90.82%
		MCC = 74%
		Kappa = 0.728
		AUC = 0.876
		F_1 score=91.24%
			Normal, Suspect, Pathological (Multi class)

6. Conclusions and Future Work

The Cardiotocography dataset, sourced from the UCI repository, comprises 2126 instances categorized into normal, suspect, and pathological (N, S, P) classes, derived from measurements of fetal heart rate (FHR) and uterine contraction (UC) features. Following feature scaling and normalization, the feature data is inputted into machine learning models such as Naive Bayes, Support Vector Machine, k-Nearest Neighbor, and Random Forest to classify the unbalanced data into multiclass categories N, S, P.Upon evaluating various performance metrics for the four algorithms, the results indicate that Random Forest, with a computational time of 6.32 seconds, achieves an overall accuracy of 90.82%, a weighted F1 score of 91.24%, a mean Matthews Correlation Coefficient (MCC) of approximately 74%, a mean Kappa Score of 72.82%, and an averaged Area under the Receiver Operating Characteristic (ROC) curve of 0.8766, outperforming other algorithms. Thus, the Random Forest method demonstrates potential for autonomously detecting fetal distress during pregnancy.Furthermore, to enhance autonomous fetal distress detection, feature selection procedures can be adopted to reduce the number of features, thereby improving model performance. Additionally, the implementation of Deep Learning and Recurrent Neural Networks presents promising avenues for enhancing the autonomous detection of fetal health status.

REFERENCES

[1] P. A. Warrick, E. F. Hamilton, D. Precup, and R. E. J. I. T. o. B. E. Kearney, "Classification of normal and hypoxic fetuses from systems modeling of intrapartum cardiotocography," vol. 57, no. 4, pp. 771-779, 2010.

[2] D. M. Sherer, C. B. Caverly, J. S. J. A. J. o. O. Abramowicz, and Gynecology, "Severe obstructive sleep apnea and associated snoring documented during external tocography," vol. 165, no. 5, pp. 1300-1301, 1991.

[3] Z. Alfirevic, G. M. Gyte, A. Cuthbert, and D. J. C. d. o. s. r. Devane, "Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour," no. 2, 2017.

[4] S. Saleem, S. S. Naqvi, T. Manzoor, A. Saeed, and J. J. F. i. p. Mirza, "A Strategy for Classification of “Vaginal vs. Cesarean Section” Delivery: Bivariate Empirical Mode Decomposition of Cardiotocographic Recordings," vol. 10, p. 246, 2019.

[5] Alfirevic Z, Devane D, Gyte GML: Continuous cardiotocography (CTG)as a form of electronic fetal monitoring (EFM) for fetal assessmentduring labour. Cochrane Database Syst Rev 2006, 3(3):CD006066.

[6] FIGO: Guidelines for the use of fetal monitoring. Int J GynecolObstet1986, 25:159–167.

[7] Blackwell SC, Grobman WA, Antoniewicz L, Hutchinson M, GyamfiBannerman C: Interobserver and intraobserver reliability of theNICHD 3-Tier Fetal Heart Rate Interpretation System. Am J ObstetGynecol 2011, 205(4):378 e1–378.e5.

[8] de Campos DA, Sousa P, Costa A, Bernardes J: Omniview-SisPorto®3.5 - A central fetal monitoring station with online alerts based oncomputerized cardiotocogram+ST event analysis. J Perinat Med 2008,36(3):260–264.

[9] S. Öztürk, S. A. Şahin, A. N. Aksoy, B. Ari and A. Akinbi, "A Novel Approach for Cardiotocography Paper Digitization and Classification for Abnormality Detection," in IEEE Access, vol. 11, pp. 42521-42533, 2023, doi: 10.1109/ACCESS.2023.3271137.

[10] B. O. Verburg, C. J. M. de Groot, and R. H. Stigter, ''Computer analysis of intrapartum fetal heart rate patterns compared to visual cardiotocography: A retrospective analysis,'' Acta Obstetricia Gynecologica Scandinavica, vol. 97, no. 4, pp. 457–465, 2018, doi: 10.1111/aogs.13281.

[11] Hakan Sahin, and Abdulhamit Subasi, Classification of the cardiotocogram data for anticipation of fetal risks using machine learning techniques, Applied Soft Computing, Elsevier,33 (2015) 231-238

[12] Y. Zhang and Z. Zhao, "Fetal state assessment based on cardiotocography parameters using PCA and AdaBoost," 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Shanghai, China, 2017, pp. 1-6, doi: 10.1109/CISP-BMEI.2017.8302314.

[13] S. Dash, J. G. Quirk and P. M. Djurić, "Fetal Heart Rate Classification Using Generative Models," in IEEE Transactions on Biomedical Engineering, vol. 61, no. 11, pp. 2796-2805, Nov. 2014, doi: 10.1109/TBME.2014.2330556.

[14] D. Ayres-de-Campos et al., ''SisPorto 2.0: A program for automated analysis of cardiotocograms,'' J. Maternal-Fetal Neonatal. Med., vol. 29, no. 2,

[15] R. Zeng, Y. Lu, S. Long, C. Wang, and J. Bai, ''Cardiotocography signal abnormality classification using time-frequency features and ensemble cost-sensitive SVM classifier,'' Comput. Biol. Med., 2021, doi:10.1016/j.compbiomed.2021.104218.

[16] V. Chudáèek, J. Andén, S. Mallat, P. Abry, and M. Doret, ''Scattering transform for intrapartum fetal heart rate characterization and acidosis detection,'' in Proc. 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Osaka, Japan, 2013, pp. 2898–2901, doi: 10.1109/EMBC.2013.6610146.

[17] V. Chudáèek et al., ''Low dimensional manifold embedding for scattering coefficients of intrapartum fetale heart rate variability,'' in Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., Chicago, IL, USA, 2014, pp. 6373–6376, doi: 10.1109/EMBC.2014.6945086.

[18] Ayres-de Campos D, Bernardes J, Garrido A, Marques-de-Sá J, Pereira-Leite L. SisPorto 2.0: a program for automated analysis of cardiotocograms. J Matern Fetal Med. 2000 Sep-Oct;9(5):311-8. doi: 10.1002/1520-6661(200009/10)9:5<311::AID-MFM12>3.0.CO;2-9. PMID: 11132590.

[19] Campos,D. and Bernardes,J.. (2010). Cardiotocography. UCI Machine Learning Repository. https://doi.org/10.24432/C51S4N.

[20] S.-B. Kim, K.-S. Han, H.-C. Rim and S. H. Myaeng, "Some effective techniques for Naive Bayes text classification", IEEE Trans. Knowl. Data Eng., vol. 18, no. 11, pp. 1457-1466, Nov. 2006.

[21] Hastie, T., R. Tibshirani, and J. Friedman. The Elements of Statistical Learning, Second Edition. NY: Springer, 2008.

[22] T. Denoeux, "A k-nearest neighbor classification rule based on Dempster–Shafer theory", IEEE Trans. Syst. Man. Cybern., vol. 25, no. 5, pp. 804-813, May 1995.

[23] L. B. Marinho, N. D. M. M. Nascimento, J. W. M. Souza, M. V. Gurgel, P. P. R. Filho and V. H. C. de Albuquerque, "A novel electrocardiogram feature extraction approach for cardiac arrhythmia classification", Future Gener. Comput. Syst., vol. 97, pp. 564-577, Aug. 2019.

[24] Mikel Galar, Alberto Fernández, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. 2011. An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. Pattern Recogn.44, 8 (August, 2011), 1761–1776. DOI:https:// doi.org/10.1016 /j. patcog.2011.01.017.

[25] A. Liaw and M. Wiener, "Classification and regression by random forest", R Newslett., vol. 2, no. 3, pp. 18-22, 2002.

[26] Ricciardi C, Amato F, Tedesco A, Dragone D, Cosentino C, Ponsiglione AM, Romano M.Detection of Suspicious Cardiotocographic Recordings by Means of a Machine LearningClassifier. Bioengineering (Basel). 2023 Feb 15;10(2):252.doi:10.3390/bioengineering10020252. PMID: 36829746; PMCID: PMC9952623.

[27] Krupa, N.; MA, M.A.; Zahedi, E.; Ahmed, S.; Hassan, F.M. Antepartum Fetal Heart Rate Feature Extraction and ClassificationUsing Empirical Mode Decomposition and Support Vector Machine. Biomed. Eng. OnLine 2011, 10, 6.

[28] M. Ajirak, C. Heiselman, J. G. Quirk and P. M. Djurić, "Boost Ensemble Learning for Classification of CTG SIGNALS," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 1316-1320, doi: 10.1109/ICASSP43922.2022.9746503.

[29] Georgoulas, G.; Stylios, D.; Groumpos, P. Predicting the Risk of Metabolic Acidosis for Newborns Based on Fetal Heart RateSignal Classification Using Support Vector Machines. IEEE Trans. Biomed. Eng. 2006, 53, 875–884.

[30] N. Chamidah and I. Wasito, "Fetal state classification from cardiotocography based on feature extraction using hybrid K-Means and support vector machine," 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Depok, Indonesia, 2015, pp. 37-41, doi: 10.1109/ICACSIS.2015.7415166.