International Journal of Artificial Intelegence Research
Vol. 8, No. 1, June 2024, pp. 1-14 ISSN 2579-7298

Assessing Performance Across Various Machine Learning Algorithms with Integrated Feature Selection for Fetal Heart Classification

Laura Rizka Amanda[a,b,1], Mila Desi Anasanti[a,c,d,2*]

[a] Computer Science Master's Study Program, Nusa Mandiri University, Jakarta, Indonesia
[b]Metropolitan Medical Center Hospital, Jakarta, Indonesia
[c] Bart and London Genome Center, Queen Mary University of London, London, United Kingdom
[d]Department of information Studies, University College London, London, United Kingdom
[1]14220017@nusamandiri.ac.id; [2]*mila.mld@nusamandiri.ac.id
* Corresponding author

ARTICLE INFO

Article history
Received
Revised
Accepted

Keywords
Perinatal Mortality,
Cardiotocography (CTG),
Machine Learning,
Fetal Health Prediction,
Feature Selection

ABSTRACT

The global concern over declining perinatal death rates, particularly in low- and middle-income nations, underscores the importance of adopting Cardiotocography (CTG) as a vital fetal monitoring method. Recent strides in machine learning (ML) present promising opportunities to enhance the accuracy of assessing fetal health, providing a viable alternative to traditional approaches. This study aims to evaluate various ML methodologies and feature selection techniques for predicting fetal health using CTG data. The primary objective is to improve ML algorithms' accuracy, precision, recall, and F1 score while selecting the most critical features. The dataset includes 2,126 expectant mothers in the third trimester, with 35 variables related to fetal heart rate (FHR) and uterine contractions (UC). Preprocessing involves feature scaling, data balancing, and outlier elimination. Additionally, a 10-fold stratified cross-validation approach is employed to ensure robust evaluation and generalizability of the model's performance. Six ML algorithms—Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Naive Bayes (NB), Logistic Regression (LR), and K-Nearest Neighbors (KNN)—are employed, optimized through grid search cross-validation. The RF algorithm outperforms with an impressive 99% accuracy, closely followed by DT at 98.7%. Optimizing 15 features from the original 35 using Simultaneous Perturbation Feature Selection and Ranking (spFSR) yields a remarkable accuracy of 99%, mirroring the full feature set. This underscores the vital role of selected features in improving predictive power and overall model performance. The study emphasizes the efficacy of tree-based classification algorithms, especially RF, in predicting fetal health and highlights the impact of preprocessing on model performance. These findings suggest avenues for future research, including exploring alternative feature engineering methods and assessing algorithm performance in diverse scenarios.

1. Introduction

According to the United Nations International Children's Emergency Fund (UNICEF) definition, In the first week of life, the number of stillbirths and infant deaths for every 1,000 live births is known as the perinatal mortality rate [1]. In 2022, the rate of live births per 1,000 population in low- and middle-income countries was 19; in high-income countries, it was 3 per 1,000, and in upper-middle-income countries, it was 7 per 1,000 [2]. South Asia and Sub-Saharan Africa had the highest rates of perinatal mortality, according to UNICEF, at 26% and 28%, respectively [3].

The adoption of thorough guidelines for cesarean sections, appropriate prenatal care, and the integration of perinatal screening technology contributed significantly to the remarkable decline in the perinatal mortality rates of high-income countries at the beginning of the 20th century. Fetal electrocardiography (ECG), amnioscopy, amniocentesis, CTG, and ultrasound were a few of these technologies [4]. Identification of fetuses at risk of mortality and morbidity is the main objective of fetal monitoring during childbirth, assuring prompt intervention. The most often used external monitoring device is the CTG, which continually captures the uterine contractions (UC) and the fetal heart rate (FHR) to provide a visual display that can be printed on thermal paper or electronically. For the past 60 years, CTG has been used routinely in clinical settings, allowing medical personnel to identify early indicators of fetal impairment.

On the other hand, studies show that a needless rise in interventions and cesarean sections has resulted from ongoing CTG monitoring in low-risk pregnancies, suggesting alternative strategies must be taken into account to avoid this [5] international standards like the International Federation of Gynecology and Obstetrics (FIGO) [6] National Institute for Health and Care Excellence (NICHT, NICE) [7], etc., have recommended using CTG exclusively for pregnancies that pose a high risk. Fetal heart rate (FHR) is expressed as beats per minute (BPM). CTG is a tool used in tracking heart rate, fetal activity, and contractions of the uterus during the baby's time within the womb through clinical prenatal health diagnosis. This examination allows medical professionals to evaluate the fetus's health both before and after delivery. CTG results can lower the risk of perinatal mortality and prevent preterm birth by giving obstetricians vital physiological and pathological information. The FIGO divides the findings of CTG tests into three categories: abnormal, suspicious, and routine. These classes are based on accelerations and decelerations, FHR variability, and FHR[6]. Reported findings indicate that, in comparison to traditional CTG, computerized CTG usage resulted in a significant decrease in perinatal mortality, with a relative risk of 0.20 and a 95% confidence interval. Additional research is required to assess the impact of CTG on perinatal outcomes. However, it is essential to note that the evidence from this study is only of intermediate quality [8].

Artificial intelligence has been employed in signal processing technologies in recent years to transform data from the human body into a diagnostic. Although medical experts are trying to develop an automated CTG interpretation, the results have not been able to identify any suspicious fetal abnormalities [9]. As a result, many researchers started attempting to conduct studies by utilizing different machine learning (ML) algorithms to forecast the condition of the fetus inside the mother's stomach (see Table 1). The goal of these studies is to create an MLmodel that can identify high-risk or pathologically suspect pregnancies as accurately as possible in conjunction with qualified medical personnel.

Table 1. Performance metrics and machine learning methods across multiple studies for fetal heart classification.

Reference	Type of model	total participants	dataset used	Best Classifier Evaluation
Rahmayanti et al (2022) [10].	XGB, SVM, KNN, LGBM, RF, ANN, LSTM	399	BV, AC, FM, UC, LD, SD, PD, ASTV, ALTV, MLTV, HW, Hmax, Hmin, NP, NZ, Hmo, Hme, Hmed, HV, HT, NSP	XGB: 0.99 LGBM: 0.99 RF: 0.98 ANN: 0.17 LSTM:0.34
Park et al (2022) [11].	LGBM	1456	UC, DL, DS, DP, DR, AC	Internal Validation Dataset: - AUROC: 0.89 - AUPRC: 0.73 External Validation Dataset: - Average AUROC: 0.73 - Average AUPRC: 0.40
Das et al(2023) [12].	MLP, RF, SVM, ANN, ELM, XGB	2126	BV, AC, FM, UC, LD, SD, PD, ASTV, MSTV, ALTV, MLTV, HW, HMAX, HMIN, NP, NZ, HMO, HME, HMED, HV, HT, NSP	SVM dan RF memiliki akurasi di atas 0.96.

Baseline Value (BV); Accelerations (AC); Fetal Movement (FM); Uterine Contractions (UC); Light Decelerations (LD); Severe Decelerations (SD); Prolonged Decelerations (PD); Abnormal Short-Term Variability (ASTV); Mean Value of Short-Term Variability (MSTV); Percentage of Time with Abnormal Long-Term Variability (ALTV); Mean Value of Long-Term Variability (MLTV); Histogram Width (HW); Histogram Max (HMax); Histogram Min (HMin); Number of Histogram Peaks (NP); Number of Histogram Zeroes (NZ); Histogram Mode (HMo); Histogram Mean (HMe); Histogram Median (HMed); Histogram Variance (HV); Histogram Tendency (HT); Fetal Health (NSP); Uterine Contractions (UC); Duration of Uterine Contractions (seconds) (DL); Duration of Uterine Contractions (pulses) (DS); Delivery Risk (DP); Fetal Breathing Rate (DR); Accelerations (AC); Baseline Value (BV); Accelerations (AC); Fetal Movement (FM); Uterine Contractions (UC); Light Decelerations (LD); Severe Decelerations (SD); Prolonged Decelerations (PD); Abnormal Short-Term Variability (ASTV); Mean Value of Short-Term Variability (MSTV); Percentage of Time with Abnormal Long-Term Variability (ALTV); Mean Value of Long-Term Variability (MLTV); Histogram Width (HW); Histogram Max (HMax); Histogram Min (HMin); Number of Histogram Peaks (NP); Number of Histogram Zeroes (NZ); Histogram Mode (HMo); Histogram Mean (HMe); Histogram Median (HMed); Histogram Variance (HV); Histogram Tendency (HT); Fetal Health (NSP).

The study by Rahmayanti et al. [10] utilized various machine learning and neural network models for a specific purpose. The experimental results indicate that models such as XGB, LGBM, and RF exhibit high Accuracy, with scores of 0.99, 0.99, and 0.98, respectively. Meanwhile, deep learning models such as ANN and LSTM show lower Accuracy, namely 0.17 and 0.34. These findings suggest that ensemble models like XGB, LGBM, and RF tend to provide more accurate results than deep learning models such as ANN and LSTM. This information serves as the basis for this research to explore the performance comparison of various models in the context further described in the study.

The study by Park et al. (2022) [11] focused on the LGBM model, which achieved an Area Under the Receiver Operating Characteristic Curve (AUROC) of 0.89 and an Area Under the Precision-Recall Curve (AUPRC) of 0.73 in the internal validation dataset. Meanwhile, the study by Das et al. (2023) [12] highlights that SVM and RF models demonstrated nearly identical performance, with Accuracy above 0.96 and sensitivity and specificity exceeding 0.96. Although there are limitations in distinguishing classes, indicated by a low Discriminant Power (DP) value, Matthews Correlation Coefficient (MCC) and Cohen's kappa values above 0.8 suggest a near-perfect fit. In this context, SVM and RF exhibit excellent performance; however, it should be noted that there are limitations in distinguishing certain classes.

These collective findings underscore the growing role of machine learning in advancing perinatal care. The application of advanced ML algorithms has the potential to significantly reduce the risk of perinatal mortality and morbidity. By employing these models, healthcare professionals can enhance the Accuracy of fetal monitoring, leading to timely interventions and better outcomes for both mothers and newborns.

In this study, we implemented two main approaches to enhance the quality of data analysis: the interquartile range (IQR) treatment to identify and address outliers and feature selection using the Simultaneous Perturbation Feature Selection and Ranking (spFSR) method. Our focus was on assembling an optimal subset of features with at least reducing more than half of features yet still maintaining the same Accuracy as using the full features. By discovering the most important features for prediction, we aimed to improve the quality of the analysis by reducing the dimensionality of the data while ensuring the relevance and diversity of the represented information. Overall, this approach is expected to make a positive contribution and provide more detailed findings in the context of this research.

2. Methods

2.1. Data description

The dataset utilized in this study was taken from the publicly accessible UCI Machine Learning Repository [13]. The dataset contains data on 2,126 expectant moms who were in the third trimester of their pregnancy. Thirty-seven features in this dataset are used to measure UC and the FHR.

According to the National Institute of Child Health and Human Development's standards [14], several important characteristics are used to assess the fetus's state based on the FHR's description. Among these are the baseline heart rate, baseline variability, accelerations per second, extended decelerations per second, early, late, and variable decelerations per second, and the existence of a sinusoidal pattern. In addition, variables, including the tone of the uterine floor and the frequency, duration, and severity of contractions, are taken into account while evaluating uterine contractions. Three obstetricians evaluated the interpretations of CTG data for expectant mothers, and their conclusions served as the standard for categorization. The automated SisPorto 2.0 program, created by Speculum in Lisbon, Portugal, was used to generate fetal CTG data. It was intended for use in the study of CTG results. All of the features of the data are described in Table 2.

Table 2. Features of Cardiotocography (CTG) data for expectant mothers in the third trimester of pregnancy.

Feature	Explanation
b	Fetal biometric index of fetal weight
e	Fetal biometric index of head circumference
LBE	Fetal biometric index of mother's pelvic bone
LB	Fetal weight in grams
AC	Fetal abdominal circumference in mm
FM	Number of fetal movements per minute
UC	Uterine contractions per minute
ASTV	Short-term fetal heart rate variability (in ms)
MSTV	Mean short-term fetal heart rate variability (in ms)
ALTV	Long-term fetal heart rate variability (in ms)
MLTV	Maximum long-term fetal heart rate variability (in ms)
DL	Fetal breathing rate in breaths per minute
DS	Duration of uterine contractions in seconds
DP	Duration of uterine contractions in pulses
Width	Width of uterine contractions in ms
Min	Minimum value of uterine contractions in ms
Max	Maximum value of uterine contractions in ms
Nmax	Number of uterine contractions in one hour (max)
Nzeros	Number of zeros in the measurement signal
Mode	Mode of the measurement signal
Mean	The mean value of the measurement signal
Median	Median value of the measurement signal
Variance	Variance of the measurement signal
Tendency	Tendency of the measurement signal
A	Score A (accelerations) in cardiotocography assessment
B	Score B (bradycardia) in cardiotocography assessment
C	Score C (contractions) in cardiotocography assessment
D	Score D (decelerations) in cardiotocography assessment
E	Score E (excessive accelerations) in cardiotocography assessment
AD	Accumulated scores of A and D in cardiotocography assessment
DE	Accumulated scores of D and E in cardiotocography assessment
LD	Accumulated scores of A, D, and E in cardiotocography assessment
FS	Cumulative score that includes all features in cardiotocography assessment
SUSP	Suspicious (SUSP) score in cardiotocography assessment
CLASS	Class or label that may be used in machine learning or statistical analysis
NSP	Apgar score at 1 minute after birth (in the range 1-3)

2.2. Preprocessing techniques

Preprocessing involves following a set of procedures shown in a flowchart in Figure 1.

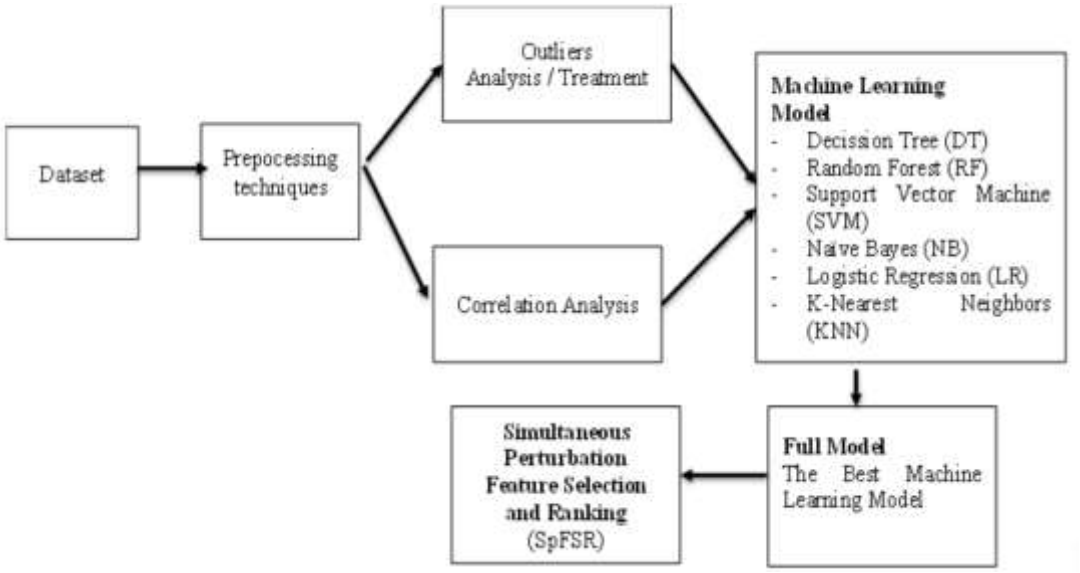

Fig. 1. Model diagram for data preprocessing, feature Selection, prediction, and accuracy check.

The fetal health dataset, which consists of 35 variables and 2,126 data points overall and was taken from the CTG interpretation, is entered to start the process. Next, the first step is to remove outliers that could jeopardize the model's accuracy that is being used. As shown in Table 2, outliers were found throughout this analysis.

3.1 Outliers analysis and treatment

Table 3. presents a comprehensive outlier analysis, detailing the number of removed outliers for each feature. This analysis provides valuable insights into the extent of outliers within the dataset, highlighting specific features that contain such anomalies. The number of removed outliers per feature is meticulously documented, shedding light on the data cleansing process and aiding in a more robust understanding of the dataset's overall quality and reliability. When outliers were replaced with mean values, it retained the original size of the data, 2,126 rows and 35 columns in total. The IQR approach is used in data analysis to handle outliers for each feature in the data frame. Values deemed to be outliers were identified by analyzing each feature, including "b," "e," "LBE," and others. Calculating the first quartile (Q1), third quartile (Q3), and interquartile range (IQR) was the first step in identifying outliers. Next, a threshold was applied using the rule of 1.5 times the IQR. Values above this cutoff were regarded as anomalies and were substituted with the column's mean value. By ensuring the dataset's consistency and stability, this procedure helps to reduce the possible influence of extreme results on later analysis [15].

Table 3. Outlier analysis with number of removed outliers per feature

Attributes	Upper Bound	Lower Bound	Outliers removed
AC	10.0	-6.0	184
FM	5.0	-3.0	192
UC	11.0	-5.0	184
MSTV	3.2	-0.8	197
ALTV	27.5	-16.5	157
MLTV	20.1	-4.7	176
DL	7.5	-4.5	189
WIDTH	194.5	-57.5	43
MIN	199.5	-12.5	88
MAX	207.0	119.0	201
NMAX	12.0	-4.0	184
MODE	176.5	100.5	201
MEAN	175.0	95.0	195
MEDIAN	176.5	100.5	201
VARIANCE	57.0	-31.0	112
TENDENCY	2.5	-1.5	197
CLASS	14.5	-5.5	181

NSP	1.0	1.0	200

3.2 Correlation analysis

It is clear from using the correlation heat map (Figure 2) that some predictor variables have a substantial correlation. The term multicollinearity refers to this strong correlation between predictor variables. The Variable Inflation Factor (VIF) was used to identify and remove predictor variables that have a substantial correlation to reduce overfitting [16].

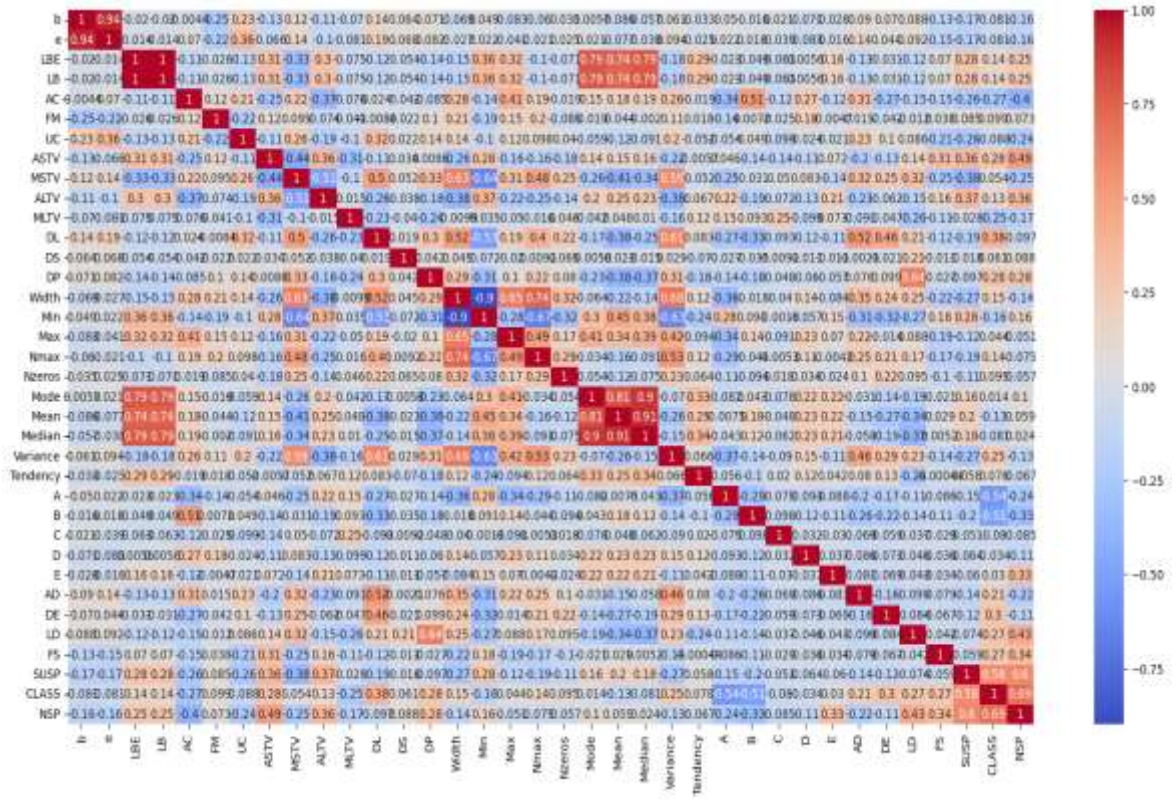

Fig. 2. Correlation Heatmap.

A balance of the data assessment was carried out during this research project. The results show an imbalance, meaning that every class has a very irregular count. As a result, a balancing method known as up-sampling was utilized to guarantee that the data was distributed equally among the normal (N), suspect (S), and pathological (P) classifications. After data balancing, feature scaling was carried out to standardize and convert the data into an appropriate range and format for modeling. When compared to models trained on unscaled data, models trained on scaled data routinely show much better performance. Data scaling is therefore acknowledged as an essential phase in data preprocessing [17].

2.3. Machine Learning Algorithms

Machine learning is a potent and extensively utilized tool in the medical domain, especially in prenatal research studies. A significant application of machine learning in this sphere involves the estimation of fetal weight [18], estimating the likelihood of fetal hypoxia [19], Forecasting fetal development, and estimating gestational age [20]. This work focused on using deep learning and machine learning techniques to analyze CTG data to categorize fetal health. One kind of machine learning called deep learning is notable for its ability to use a flexible learning mechanism to automatically extract complex features from input. Although it performs better on larger datasets than traditional machine learning methods, it has disadvantages such as decreased interpretability, longer training durations because of more parameters, and best results on powerful computers [21].

Various ML algorithms have been used for fetal health classification, including ANN, LSTM, XGB, NB, LR, KNN, SVM, LGBM, and RF. XGB, RF, and LGBM are ensemble algorithms that

incorporate principles from Decision Trees. The effectiveness of these algorithms varies based on the dataset and specific study. However, in certain investigations, RF has shown superior performance compared to other classifiers [22]. KNN classifies data by considering the nearest point's class, whereas SVM bases its classification on a support vector network and a hyperplane [23]. While LSTM is an evolution of neural network design that incorporates advanced deep learning techniques, ANN operates on the core principles of a neural network [24].

3.3 Logistic Regression (LR)

In LR analysis, the formula P(Y=1| X) = 1/(1 + e-(b0 + b1X)) is used to model the probability of a positive event (Y=1) based on the predictor variable X. Here, b0 and b1 represent the adjusted regression parameters. Here, b0 and b1 represent the regression parameters adjusted to optimize the prediction accuracy. The expression e^-(b0 + b1X) reflects the log-odds function converted into probability through logistic transformation. The use of this formula allows researchers to accurately understand and predict the relationship between predictor variables and the probability of the event of interest [25].

$$P(Y=1|X)=1+e-(b0+b1X)1$$

3.4 Decision Tree (DT)

In the development of the Decision Tree model, we measured the impurity of the set (□) using the Entropy criterion (□(□)) which is calculated as the sum of all classes (□) minus the probability of each class (□$_i$) multiplied by the base 2 logarithm. The formula, □(□) = ∑□ - □$_i$ * log$_2$(□$_i$), reflects the degree of uncertainty or confusion in the dataset □. The use of Entropy criteria helps in selecting the best attributes to split the dataset, with the aim of producing a Decision Tree that provides optimal and efficient classification results [26].

$$E(S) = \sum c-P1log2Pi$$

3.5 Random Forest (RF)

In a RF model for classification, the prediction YRF is generated by taking the average of the class predictions given by each individual decision tree Yi. The number of trees in the forest, Ntrees, plays a key role in formulating the final prediction, which incorporates information from the entire ensemble to improve the stability and accuracy of the model.[27].

$$Y\text{^}RF=Ntrees1\sum i=1NtreesYi$$

3.6 Support Vector Machine (SVM)

It possesses the capability to manage datasets with a high number of dimensions, allowing for both linear and non-linear kernel classification as well as regression tasks. SVM is a dependable choice for classification and regression algorithms. The primary goal is to pinpoint the optimal classification function for effectively segregating the training data into two distinct classes. This segregation is achieved through the use of a separator line referred to as a hyperplane equation, calculated to efficiently discriminate between the two classes [28]. The hyperplane is calculated using this formula:

$$H: wT(x) + b = 0$$

b = the bias term and intercept of the hyperplane equation.

3.7 Naïve Bayes (NB)

In the NB classification method, the formula (P(Ck | X) gives an estimate of how likely the data (X) belongs to the class (Ck). This formula takes into account a number of factors, such as how often we see similar data in that class and how common the class itself is [29].

$$P(Ck|X)=P(X)P(X|Ck)P(Ck)$$

3.8 K-Nearest Neighbor (KNN) Classifier

It is a method for classifying unknown cases by identifying nearby patterns within the pattern space. KNN relies on Euclidean distance to predict the class of a given case [30].

To find the closest instance in the pattern space, the Euclidean distance $d(x, y)$ is employed for distance computation from each feature i to k. The class of the unknown examples is then determined through a majority vote from their neighboring instances.

$$d(x, y) = \sqrt{\sum_{i=1}^{k} (xi - yi)2}$$

3.9 Classifier Evaluation

The study employed multiple criteria to evaluate the efficacy of the produced categorization models. The performance evaluation comprised the metrics (Accuracy, precision, recall, and F1 score) listed in Equations (1) through (4). A confusion matrix was created to examine the classification results of every model for multiclass classification. These confusion matrices were then used to compute Accuracy, precision, recall, and F1 score. Accuracy is the percentage of right predictions made over the whole test dataset, whereas precision and recall measure the model's capacity to locate major data points and identify all pertinent cases within a dataset. High precision points to a low false-positive rate, whereas great recall points to a low false-negative rate. When recall and precision are combined, a high F1 score indicates a strong classification model [32].

The applied algorithms, developed in Python 3 and leveraging libraries like Scikit-learn, Pandas, NumPy, and Matplotlib, assesses its performance through crucial metrics, such as recall, specificity, accuracy, precision, and F1 score. These metrics function as benchmarks for evaluating the efficacy of the proposed algorithm in comparison to other established categorization algorithms detailed in their respective sections.

2.4. Simultaneous Perturbation Feature Selection and Ranking (spFSR)

SpFSR represents a novel approach for FS and ranking, extending the capabilities of a versatile black box stochastic optimization algorithm. The SpFSR process commences with an initial solution, denoted as w0, and employs a recursive procedure to identify the local minimum [31].

$$\omega k + 1 := \omega k - ak\, G(\omega k)$$

Where ak is the order of iteration gain; $ak \geq 0$ and $G(\omega k)$ are estimates of the gradient at k.

Ultimately, 15 of the best features—or roughly 42.86% of the total 35 features—were chosen from the dataset using the SpFSR feature selection technique combined with the best classification method on the entire model. For several reasons, including dimensionality reduction, enhanced model performance, computational efficiency, preventing overfitting, and handling collinearity, feature selection is an essential stage in machine learning. Large numbers of features are included in many datasets used in machine learning. Overfitting may result from certain features being redundant, unnecessary, or noisy. By choosing the most pertinent features, feature selection lowers the dimensionality of the dataset while enhancing model performance and efficiency. Repetitive or irrelevant features in a model can cause it to perform poorly. The model can concentrate on the important relationships and patterns in the data by just choosing the most instructive features, which improves generalization on data that hasn't been seen before.

In general, less feature-rich models are less computationally expensive. This holds significance for practical applications where efficiency is paramount, particularly in situations requiring the real-time processing of substantial volumes of data. Overfitting occurs when a model learns to perform well on training data but is unable to generalize to new, unseen data. This can happen when a model has too many characteristics. By concentrating on the most pertinent features, feature selection helps minimize overfitting and lowers the possibility that the model will learn noise from the data. Models are frequently made more comprehensible by employing a selection of their most crucial elements to simplify them. This is especially critical in sectors like banking and healthcare, where it's just as necessary to comprehend the model's decision-making process as it is to forecast results with Accuracy. Additionally, feature selection can assist in addressing multicollinearity difficulties when predictor variables in a model show significant correlation. As a result, the model's interpretability is improved, and unpredictable coefficient estimations are avoided.

3. Result and discussion

3.10 Performances of ML Classification Model result

This research specifically focuses on the classification of fetal health in pregnant women within the Performance of the ML Classification Model results. The algorithms compared are RF, DT, SVM, NB, LR, and k-NN classifiers.

Table 4. Performance comparison of different machine learning algorithms using the full model

Algorithms	Accuracy	Precision	Recall	F1 Score
DT	0.987	0.977	0.964	0.97
RF	**0.99**	**0.991**	**0.966**	**0.978**
SVM	0.989	0.99	0.962	0.975
NB	0.985	0.981	0.953	0.966
LR	0.989	0.994	0.955	0.973
KNN	**0.99**	**0.982**	**0.973**	**0.977**

The metrics for performance measurements covered in the preceding section were used to evaluate the effectiveness of the recommended algorithms. The outcomes are RF, DT, SVM, NB, LR, and k-NN when compared to standard classification methods. We divided the current data into two categories: 30% was utilized for training and the remaining 70% was used for testing, as this was the first training on the data. The first step of the process yields the following findings after all of our data's attributes are applied to the collection of algorithms shown in the table below. Because it has been constructed on the Python platform, this helpful component is regarded as a comprehensive and integrated platform. It displays each algorithm's accuracy given that it accepts algorithms. DT accuracy is 0.987, RF accuracy is 0.99, SVM accuracy is 0.989, NB accuracy is 0.985, LR accuracy is 0.989, and KNN Classifier accuracy is 0.99. These results lead to the conclusion that the RF algorithm's accuracy is quite good, comparable to that of KNN. The majority of the categorization algorithms that we employ for fetal health disorders have shown to be beneficial in helping to diagnose these conditions, which has had a significant positive effect on the healthcare system as a whole.

A classifier's efficiency can only be increased by carefully choosing its features. Modern devices send millions of data points, which results in datasets containing hundreds of undesirable properties. Consequently, these attributes heighten the possibility of overfitting, choking the model, and experiencing exponential growth in training time. Feature selection approaches allow one to decrease the average training and prediction time while retaining all the information. These well considered characteristics were then used for training and testing to reduce costs and time. The classification results are significantly impacted by these techniques [32].

3.11 Performances of ML feature selection result

We carefully selected the number of features using spFSR and RF algorithm as a wrapper, the best ML method on the full model, then compared their performance as can bee seen in Table 5.

Table 5. Performance comparison of different subset of feature selection models

Methods	Feature	Result
SpFSR	30	0.99
	15	0.99
	10	0.92

An accuracy score of 0.99 was obtained by applying spFSR and making use of all features. Remarkably, the accuracy score stayed at 0.99 even when a mere 15 features were taken into account. Consequently, we chose to employ a strategy of utilizing only 15 features, which produced results that were on par with 35 features in terms of accuracy. Without sacrificing the caliber of the outcomes, this choice was made with efficiency and simplicity of model interpretation in mind.

3.12 Feature importance identification

The regularization procedure, which aims to find and keep the most significant variables while reducing redundancy, determines the relevance of features in the context of spFSR. Selecting a subset of features that make a significant contribution to model predictions is made possible by spFSR's use of regularization approaches that promote sparsity, in contrast to linear regression models. Features in spFSR are assessed based on how relevant they are to the dataset as a whole. Our model found and highlighted 15 variables that are essential to improving the efficiency of model performance as can be seen in Table 6. SpFSR strives to maximize predicted accuracy while reducing model complexity by concentrating on the most useful variables.

Table 6. Fifteen features with the highest importance factor

Feature	Description	Importance Factor
LD	Long-term variability of FHR	0.3085
FS	Short-term variability of FHR	0.0813
AD	Accelerations per minute	0.0791
DE	Decelerations per minute	0.0760
Tendency	Tendency of FHR to be predominantly reactive or nonreactive	0.0679
D	Abnormal short-term variability	0.0486
DL	Prolonged decelerations	0.0338
Variance	Variability in FHR signal	0.0319
E	Presence of episodic changes	0.0217
B	Baseline FHR	0.0228
Median	Baseline FHR median	0.0199
MLTV	Percentage of time with abnormal long-term variability	0.0190
ALTV	Percentage of time with abnormal short-term variability	0.0185
UC	Uterine contractions per minute	0.0180
Min	Minimum FHR	0.0145

The classification outcomes are markedly influenced by the choice of feature selection approach [32]. One primary rationale behind this is its ability to trim more than half of the features while maintaining an accuracy comparable to that of the full model, achieving an impressive 99% accuracy.

A classifier's efficiency can only be increased by carefully choosing its features. Modern devices send millions of data points, which results in datasets containing hundreds of undesirable properties. Consequently, these attributes heighten the possibility of overfitting, choking the model, and experiencing exponential growth in training time. Feature selection approaches allow one to decrease the average training and prediction time while retaining all the information. These well considered characteristics were then used for training and testing to reduce costs and time.

We carefully chose and improved our predictive models in our study using the spFSR approach. The top 15 features—which were found via an extensive spFSR analysis—were essential in figuring out how well our models performed overall. Notably, the feature "LD," which stands for the cardiotocography evaluation's accumulated scores of A, D, and E, stood out as a major contribution and received the highest important score. In the spFSR framework, these variables are arranged hierarchically according to their weights. This hierarchical structure is illustrated graphically in Figure 3.

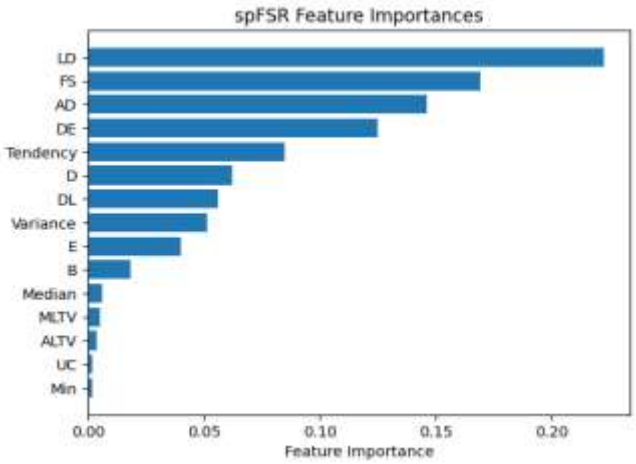

Fig. 3. Ranking the ten most important features.

Recognizing the importance of the "LD" feature and its prominent role constitutes a departure from prior studies. A direct alignment with earlier research, specifically outlined in [33], underscores our emphasis on "LD" and its established predictive capabilities. Incorporating "LD" compared to the other pertinent features not only optimized the predictive model's accuracy but also its contextual relevance. This thoughtful combination of ML and spFSR demonstrates our approach to expanding the capabilities of predictive modeling.

Furthermore, the spFSR methodology helps identify relevant features and offers a thorough understanding of their respective influences on prediction outcomes. Our ML-spFSR framework is illustrated in Figure 1 and Figure 3, which also provides a graphical depiction of the unique contribution of each feature, enhancing the interpretive depth of our research findings. Our work is at the forefront of predictive modeling techniques thanks to our meticulous and thorough methodology, which adds to the current discussion in this area. However, it's important to note that our ML-spFSR framework's generalizability to other datasets remains a limitation and requires thorough testing for broader applicability.

By assessing the predictive model for diagnosing fetal health, this study creates opportunities for additional research that may have a major impact on the advancement of clinical procedures. In the future, a number of research avenues could be investigated to improve our comprehension and utilization of this predictive model. First, additional validation utilizing a wider variety of datasets may be a part of future study. The model's generalizability can be enhanced and the findings more readily applicable in many clinical scenarios by evaluating the model across age groups and geographic conditions. Subsequently, the analysis of algorithm performance comparison across various clinical settings can be investigated. Evaluating the merits and demerits of every algorithm in various medical contexts can offer a more profound understanding of the usefulness of this model. Future studies may concentrate on other feature engineering methods. The management of high-dimensional data can be aided by the further development of techniques like Principal Component Analysis (PCA), which will increase the efficiency of dimensionality reduction.

Lastly, more thorough research on particular fetal health issues can offer insightful information. Extensive studies focusing on particular facets of fetal health will enhance our comprehension and utilization of this prognostic model. Future studies in this area can continue to improve clinical procedures and significantly enhance the diagnosis and care of fetal health by investigating these avenues. It is anticipated that these actions will provide a solid basis for future advancements in the field of fetal health.

4. Conclusion

In conclusion, this study addresses the global concern of declining perinatal death rates by exploring CTG as a vital fetal monitoring method with the aid of ML. Our research enhances predictive accuracy, with the RF algorithm leading at 99%. Notably, feature selection using spFSR mirrors full feature accuracy, emphasizing its pivotal role. This study underscores the effectiveness of tree-based classification algorithms, particularly RF, in predicting fetal health, while also

highlighting the impact of preprocessing on overall model performance. Future research should explore alternative feature engineering methods and assess algorithm performance in diverse scenarios. Additionally, our ML-spFSR framework's generalizability to other datasets remains a crucial area for further testing.

Declarations

Author contribution: MDA directed the research objectives and methodology. LRA drafted the manuscript. The final draft was collectively reviewed by MDA and approved by both MDA and LRA.
Funding statement: there was no external funding involved in this project.
Conflict of interest: no conflict of interest was declared by the authors.
Additional information: no additional information is available for this paper.

Data and Software Availability Statements

The data utilized in this study was acquired from the publicly accessible UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/193/cardiotocography). All analyses were conducted using Python 3.12.1 on Google Colab, utilizing various modules, including those from scikit-learn, numpy, and matplotlib.

References

[1] nina, "Levels and trends in child mortality," UNICEF DATA. Accessed: Oct. 20, 2023. [Online]. Available: https://data.unicef.org/resources/levels-and-trends-in-child-mortality/
[2] U. Syed et al., "Advancing maternal and perinatal health in low- and middle-income countries: A multi-country review of policies and programmes," Front. Glob. Womens Health, vol. 3, 2022, Accessed: Dec. 23, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fgwh.2022.909991
[3] Y. G. Robi and T. M. Sitote, "Neonatal Disease Prediction Using Machine Learning Techniques," J. Healthc. Eng., vol. 2023, pp. 1–16, Feb. 2023, doi: 10.1155/2023/3567194.
[4] "Acta Obstet Gynecol Scand - 2015 - Goldenberg - Reducing stillbirths in low-income countries.pdf."
[5] Z. Alfirevic, G. M. Gyte, A. Cuthbert, and D. Devane, "Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour," Cochrane Database Syst. Rev., vol. 2019, no. 5, Feb. 2017, doi: 10.1002/14651858.CD006066.pub3.
[6] D. Lewis, S. Downe, and FIGO Intrapartum Fetal Monitoring Expert Consensus Panel, "FIGO consensus guidelines on intrapartum fetal monitoring: Intermittent auscultation," Int. J. Gynecol. Obstet., vol. 131, no. 1, pp. 9–12, Oct. 2015, doi: 10.1016/j.ijgo.2015.06.019.
[7] G. A. Macones, G. D. V. Hankins, C. Y. Spong, J. Hauth, and T. Moore, "The 2008 National Institute of Child Health and Human Development workshop report on electronic fetal monitoring: update on definitions, interpretation, and research guidelines," Obstet. Gynecol., vol. 112, no. 3, pp. 661–666, Sep. 2008, doi: 10.1097/AOG.0b013e3181841395.
[8] R. M. Grivell, Z. Alfirevic, G. M. Gyte, and D. Devane, "Antenatal cardiotocography for fetal assessment," Cochrane Database Syst. Rev., vol. 2019, no. 5, Sep. 2015, doi: 10.1002/14651858.CD007863.pub4.
[9] A. Sheikhtaheri, M. R. Zarkesh, R. Moradi, and F. Kermani, "Prediction of neonatal deaths in NICUs: development and validation of machine learning models," BMC Med. Inform. Decis. Mak., vol. 21, no. 1, p. 131, Apr. 2021, doi: 10.1186/s12911-021-01497-8.
[10] N. Rahmayanti, H. Pradani, M. Pahlawan, and R. Vinarti, "Comparison of machine learning algorithms to classify fetal health using cardiotocogram data," Procedia Comput. Sci., vol. 197, pp. 162–171, 2022, doi: 10.1016/j.procs.2021.12.130.
[11] T. J. Park et al., "Machine Learning Model for Classifying the Results of Fetal Cardiotocography Conducted in High-Risk Pregnancies," Yonsei Med. J., vol. 63, no. 7, p. 692, 2022, doi: 10.3349/ymj.2022.63.7.692.
[12] S. Das, H. Mukherjee, K. Roy, and C. K. Saha, "Fetal Health Classification from Cardiotocograph for Both Stages of Labor—A Soft-Computing-Based Approach," Diagnostics, vol. 13, no. 5, p. 858, Feb. 2023, doi: 10.3390/diagnostics13050858.
[13] J. B. D. Campos, "Cardiotocography." UCI Machine Learning Repository, 2000. doi: 10.24432/C51S4N.
[14] "Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD)," National Institutes of Health (NIH). Accessed: Dec. 23, 2023. [Online]. Available:

https://www.nih.gov/about-nih/what-we-do/nih-almanac/eunice-kennedy-shriver-national-institute-child-health-human-development-nichd
[15] "3.2 - Identifying Outliers: IQR Method | STAT 200." Accessed: Dec. 24, 2023. [Online]. Available: https://online.stat.psu.edu/stat200/lesson/3/3.2
[16] N. Shrestha, "Detecting Multicollinearity in Regression Analysis," Am. J. Appl. Math. Stat., vol. 8, pp. 39–42, Jun. 2020, doi: 10.12691/ajams-8-2-1.
[17] "Technologies | Free Full-Text | Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance." Accessed: Dec. 24, 2023. [Online]. Available: https://www.mdpi.com/2227-7080/9/3/52
[18] A. I. Naimi, R. W. Platt, and J. C. Larkin, "Machine Learning for Fetal Growth Prediction," Epidemiol. Camb. Mass, vol. 29, no. 2, pp. 290–298, Mar. 2018, doi: 10.1097/EDE.0000000000000788.
[19] Z. Arain, S. Iliodromiti, G. Slabaugh, A. L. David, and T. T. Chowdhury, "Machine learning and disease prediction in obstetrics," Curr. Res. Physiol., vol. 6, p. 100099, May 2023, doi: 10.1016/j.crphys.2023.100099.
[20] Y. Yin and Y. Bingi, "Using Machine Learning to Classify Human Fetal Health and Analyze Feature Importance," BioMedInformatics, vol. 3, no. 2, Art. no. 2, Jun. 2023, doi: 10.3390/biomedinformatics3020019.
[21] T. Tiwari, T. Tiwari, and S. Tiwari, "How Artificial Intelligence, Machine Learning and Deep Learning are Radically Different?," Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 8, p. 1, Mar. 2018, doi: 10.23956/ijarcsse.v8i2.569.
[22] A. Kuzu and Y. Santur, "Early Diagnosis and Classification of Fetal Health Status from a Fetal Cardiotocography Dataset Using Ensemble Learning," Diagnostics, vol. 13, no. 15, p. 2471, Jul. 2023, doi: 10.3390/diagnostics13152471.
[23] "Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review." Accessed: Dec. 07, 2023. [Online]. Available: https://www.scirp.org/journal/paperinformation?paperid=104256
[24] "Fig. 8. Artificial neural network (ANN) and long short-term memory...," ResearchGate. Accessed: Dec. 07, 2023. [Online]. Available: https://www.researchgate.net/figure/Artificial-neural-network-ANN-and-long-short-term-memory-LSTM-network-training-and_fig6_351535690
[25] G. S. Prayoga, "Logistic Regression: A Classifier With a Sense of Regression," Medium. Accessed: Dec. 16, 2023. [Online]. Available: https://python.plainenglish.io/logistic-regression-a-classifier-with-a-sense-of-regression-83ce49ba3f5b
[26] "Decision Tree," GeeksforGeeks. Accessed: Dec. 24, 2023. [Online]. Available: https://www.geeksforgeeks.org/decision-tree/
[27] D. Lazar, "Understanding Logistic Regression," Medium. Accessed: Dec. 16, 2023. [Online]. Available: https://towardsdatascience.com/understanding-logistic-regression-81779525d5c6
[28] M. Jain, "Support Vector Machine vs K Nearest Neighbours," Stack Overflow. Accessed: Dec. 07, 2023. [Online]. Available: https://stackoverflow.com/q/19421954
[29] "Bayes' Theorem: The Idea Behind Naive Bayes Algorithm | by Soner Yıldırım | Towards Data Science." Accessed: Dec. 16, 2023. [Online]. Available: https://towardsdatascience.com/bayes-theorem-the-idea-behind-naive-bayes-algoritm-f7068834a4d7
[30] "What Is a K-Nearest Neighbor Algorithm? | Built In." Accessed: Dec. 24, 2023. [Online]. Available: https://builtin.com/machine-learning/nearest-neighbor-algorithm
[31] Mila Desi Anasanti, Khairunisa Hilyati, and Annisa Novtariany, "The Exploring feature selection techniques on Classification Algorithms for Predicting Type 2 Diabetes at Early Stage," J. RESTI Rekayasa Sist. Dan Teknol. Inf., vol. 6, no. 5, pp. 832–839, Nov. 2022, doi: 10.29207/resti.v6i5.4419.
[32] M. Hossin and S. M.N, "A Review on Evaluation Metrics for Data Classification Evaluations," Int. J. Data Min. Knowl. Manag. Process, vol. 5, pp. 01–11, Mar. 2015, doi: 10.5121/ijdkp.2015.5201.
[33] "A Generic Machine Learning Approach for IoT Device Identification | IEEE Conference Publication | IEEE Xplore." Accessed: Jan. 13, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/9702983