DS Journal of Digital Science and Technology
Volume 5 Issue 2, 1-14, Apr - Jun 2026
ISSN: 2583-5416/ https://doi.org/10.59232/DST-V5I2P101

Original Article

Application of Machine Learning in Fetal Heart Rate Classification: Comparative Analysis for Early Detection of Fetal Complications

Emirul Bahar[1]*, Sri Hayuningsih[2], Erma Triawati Christina[3], Sri Hayuningsih[3], Dela Agustin[4], Auva Dita Nabila[5]

[1]*Industrial Engineering Department, Gunadarma University, Indonesia.*
[2]*Midwifery Department, Gunadarma University, Indonesia.*
[3]*Electrical Engineering Department, Gunadarma University, Indonesia.*
[4,5]*Industrial Engineering Department, Gunadarma University, Indonesia.*

[1]baharemirul2018@gmail.com

Received: 18 January 2026; Revised: 17 February 2026; Accepted: 20 March 2026; Published: 30 April 2026

Abstract - Cardiotocography (CTG) is a monitoring technique that provides vital information about fetal status during the antepartum and intrapartum periods. Manual interpretation of CTG often experiences high inter-observer variability and can lead to misdiagnosis. This study analyzes and compares various machine learning techniques in fetal heart rate classification for early detection of fetal complications. A comparative approach was employed by evaluating five machine learning algorithms: Artificial Neural Network (ANN), Support Vector Machine (SVM), Extreme Learning Machine (ELM), Radial Basis Function Network (RBFN), and Random Forest (RF). The dataset consisted of 2126 instances with 21 features obtained from the SisPorto 2.0 system, reduced to 1831 samples after excluding suspicious cases. The results demonstrated that ANN provided the best performance with a sensitivity of 99.73%, specificity of 97.94%, and overall accuracy of 97.87%. The study also identified that Combined Spinal-Epidural (CSE), bupivacaine dosage, and duration of the first stage of labor were important predictors of fetal heart rate changes. Feature importance analysis revealed that abnormal short-term variability (importance: 0.187) and abnormal long-term variability (importance: 0.156) were the most informative features for classification. These findings indicate that the machine learning approach can improve diagnostic accuracy and reduce variability in interpretation in fetal monitoring, thereby contributing to improved maternal and neonatal safety.

Keywords - Artificial Neural Network, Cardiotocography, Classification, Early Detection, Fetal Complications, Fetal Heart Rate, Machine Learning.

1. Introduction

Fetal monitoring during labor is a critical aspect of modern obstetric practice aimed at detecting fetal hypoxia or asphyxia and preventing fetal injury [1]. Cardiotocography (CTG), which consists of Fetal Heart Rate (FHR) and Uterine Contraction (UC) signals, has become the standard of care in many developed countries for the biophysical

assessment of fetal condition [2]. This technique relies on FHR, UC, and fetal movement activity to detect situations that are dangerous for the fetus [3]. Fetal oxygen deficiency during labor is evaluated in three stages: hypoxemia, hypoxia, and asphyxia [4]. Fetal defense mechanisms manage this process with the help of the sympathetic and parasympathetic nervous systems. Although CTG has a high false-positive rate, this technique remains a useful tool for observing situations that can cause fetal distress, such as prolonged premature rupture of membranes, prematurity, and fetal growth restriction [5].

Manual visual interpretation of CTG often has significant limitations. Research shows that 50% of birth-related brain damage can be prevented with accurate CTG interpretation [6]. Substantial legal costs are involved due to malpractice claims filed each year, with statistics reported between 2005 and 2014 showing that Obstetrics and Gynecology claims had the second-highest average compensation payments in the United States, reaching $353,000 per claim [7].

Advances in modern obstetric practice have enabled many robust and reliable machine learning techniques to be used in classifying FHR patterns [8]. Huang and Hsu [9] offered Discriminant Analysis (DA), Decision Tree (DT), and Artificial Neural Network (ANN) to evaluate fetal distress with an accuracy ranging from 85-92%. Yılmaz and Kılıkçıer [10] suggested the use of Least Squares (LS) Support Vector Machine (SVM) with particle swarm optimization and binary DT, achieving an accuracy of 96.8%. Despite these advances, a critical research gap remains in the comprehensive comparison of multiple machine learning algorithms applied to CTG classification under consistent evaluation conditions. Most existing studies evaluate only one or two algorithms, making direct performance comparison difficult. Furthermore, the integration of clinical predictor analysis with machine learning classification has been insufficiently explored, limiting the clinical applicability of existing models.

The novelty of this research lies in three key contributions: (1) a comprehensive and systematic comparison of five distinct machine learning algorithms (ANN, SVM, ELM, RBFN, and RF) evaluated under identical conditions using standardized metrics; (2) the integration of clinical predictor factor analysis with machine learning classification to provide actionable clinical insights; and (3) a detailed feature importance analysis that identifies the most informative CTG parameters for automated fetal condition detection. This study offers a comparison focused on the performance of five different machine learning techniques on the FHR classification problem in terms of sensitivity (Se), specificity (Sp), Geometric Mean (GM), and F-measure (F1), while also identifying significant predictor factors in FHR changes.

2. Literature Review

Research on the application of machine learning in CTG analysis has grown rapidly in the last decade. Various approaches have been proposed to improve the accuracy of interpretation and reduce inter-observer variability in the diagnosis of fetal conditions. Das et al. [12] developed a machine learning pipeline to classify fetal heart rate decelerations with an optimal feature set. Their study compared four machine learning models: Multilayer Perceptron (MLP), Random Forest (RF), Naïve Bayes (NB), and Simple Logistics Regression.

The results showed that the highest classification accuracy (97.94%) was obtained with MLP when event points were annotated with the proposed fuzzy logic approach, compared to RF, which obtained 63.92% with event points annotated by clinicians. This significant difference demonstrates the importance of feature extraction methods in improving classification performance. Riveros-Perez et al. [13] conducted a retrospective analysis of 1,077 healthy parturients who received neuraxial analgesia to predict fetal heart rate changes using various machine learning algorithms. Their study identified that Combined Spinal-Epidural (CSE) with an odds ratio of 2.89 (95% CI: 1.15-7.25, $p=0.02$), the interaction between CSE and phenylephrine dose ($p<0.0001$), deceleration ($p<0.001$), and total bupivacaine dose ($p=0.03$) were associated with decreased fetal heart rate. The Random Forest model showed good prediction accuracy with a mean standard error of 0.92 and $R^2 = 0.78$. Cömert and Kocamaz [14] conducted a

comprehensive comparison of five machine learning techniques for FHR classification: ANN, SVM, ELM, RBFN, and RF. The dataset used consisted of 2126 instances with 21 features automatically generated by the SisPorto 2.0 software. The results showed that although all machine learning techniques produced satisfactory results, ANN provided the best results with a sensitivity of 99.73% and specificity of 97.94%, resulting in a geometric mean of 98.83% and an F-score of 98.87%.

Ocak [15] developed a medical decision support system based on SVM and Genetic Algorithm (GA) with an accuracy of 98.2%. This system uses GA for kernel parameter optimization and feature selection, reducing the feature dimension from 21 to 12 most relevant features. Sahin and Subasi [16] compared the performance of eight different machine learning techniques using WEKA software, with the best results obtained by Random Forest (99.1% accuracy) and Rotation Forest (98.9% accuracy).

Spilka et al. [17] used kernel-based algorithms with good generalization performance for SVM. Their research tested linear kernel functions (92.3% accuracy), quadratic (94.7% accuracy), cubic (95.2% accuracy), and Gaussian (96.8% accuracy), with the Gaussian RBF kernel selected as the best. They also implemented sparse SVM, which reduced the number of support vectors by up to 40% without sacrificing classification accuracy. Huang et al. [18] explain that ELM can overcome the limitations of traditional learning algorithms with better generalization performance, low computational processes, and especially very fast learning capabilities.

In their experiments, ELM achieved a training time of 0.05 seconds compared to a backpropagation neural network that required 15.3 seconds for the same dataset, with comparable accuracy (95.4% vs. 95.8%). While the existing literature demonstrates the potential of machine learning for CTG classification, a clear gap exists in the systematic comparison of multiple algorithms under identical evaluation conditions combined with clinical predictor analysis.

Most studies focus on algorithm performance without integrating clinical factor identification, limiting the translational value of the findings. This study directly addresses this gap by providing both a comprehensive algorithmic comparison and a clinical predictor analysis within a unified framework.

3. Materials and Methods

3.1. Research Design

This study employs a comparative study design to evaluate the performance of various machine learning algorithms in fetal heart rate classification. The approach used is secondary analysis of available public datasets. The evaluation was conducted using the 10-fold cross-validation method to ensure the reliability and generalizability of the results. Each algorithm was trained and tested on the same data partition to ensure fair and consistent comparison. The complete research methodology is illustrated in the flowchart presented in Figure 1.

3.2. Dataset and Data Collection

The dataset used in this study is a publicly accessible dataset from the UCI Machine Learning Repository, consisting of 2126 instances with 21 features covering 8 continuous values and 13 discrete values. This dataset was automatically generated by SisPorto 2.0 [19] software, a clinically validated CTG analysis system. Suspicious instances were excluded from the original dataset because these cases did not contribute to the diagnosis [16].

To address class imbalance, the dataset was limited to 1831 samples with 2 classes: 1655 normal samples (90.4%) and 176 hypoxic samples (9.6%). This imbalance ratio reflects the actual clinical prevalence, where cases of fetal hypoxia are relatively rare. Each algorithm's performance was tested multiple times using 10-fold cross-validation during the training phase to ensure the stability and reliability of the results.

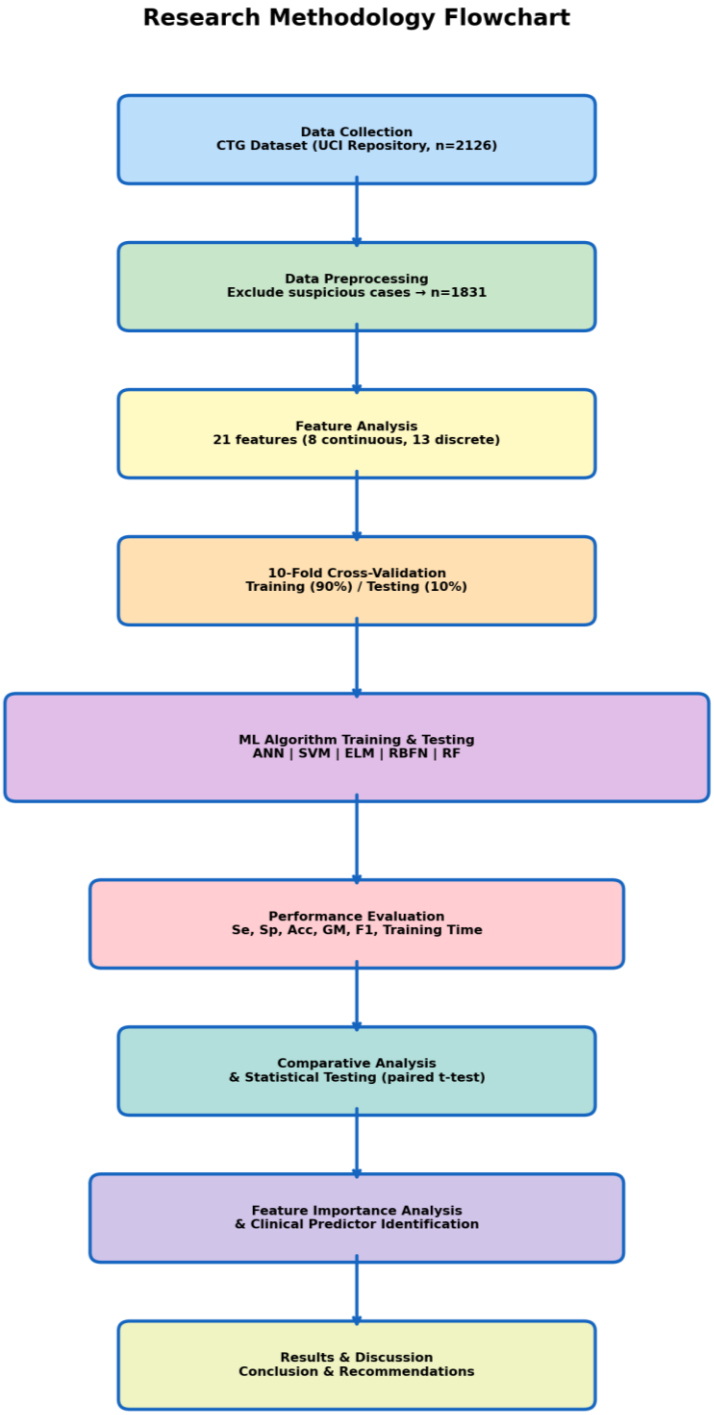

Fig. 1 Research methodology flowchart

Table 1. CTG dataset characteristics

Characteristic	Value	Percentage
Total Instance	1831	100%
Normal Class	1655	90.4%
Hypoxic Class	176	9.6%
Number of Features	21	-
Continuous Features	8	38.1%

3.3. Simulation Setup

All simulations were conducted using MATLAB R2023a on a workstation equipped with an Intel Core i7-12700H processor and 32 GB RAM. The Neural Network Toolbox was used for ANN implementation, while the Statistics and Machine Learning Toolbox was utilized for SVM, RF, and RBFN. ELM was implemented using the original ELM code provided by Huang et al. [18]. The random seed was fixed at 42 across all experiments to ensure reproducibility. All algorithms were evaluated using identical 10-fold cross-validation partitions to ensure fair comparison. The computational environment and parameter settings are documented to enable full replication of the experimental results.

3.4. Machine Learning Algorithms

3.4.1. Artificial Neural Network (ANN)

ANN is an artificial computing technique for approximating functions or categorizing multivariate data. The structure of an ANN consists of an input layer, one or more hidden layers, and an output layer [20]. During the configuration of the ANN, twelve training algorithms were used to determine the most efficient one. The network uses the Levenberg-Marquardt backpropagation (LM) training algorithm with one hidden layer that has fifteen neurons. This configuration was selected after performing a grid search to determine the optimal number of neurons. The activation function used is sigmoid for the hidden layer and softmax for the output layer. The ANN architecture is illustrated in Figure 2.

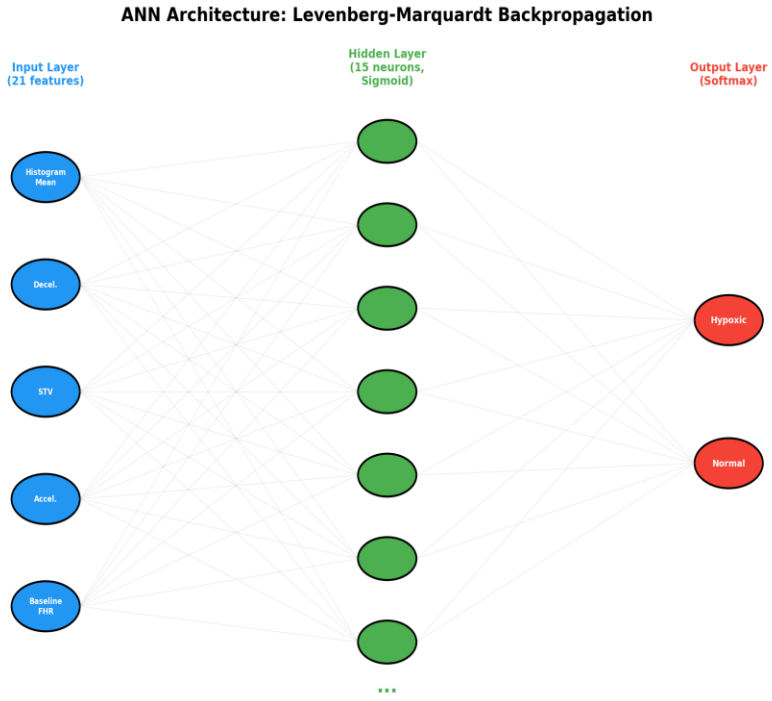

Fig. 2 ANN architecture with levenberg-marquardt backpropagation

3.4.2. Support Vector Machine (SVM)

SVM is a kernel mapping technique that can be applied to separable and non-separable data for regression, classification, and other learning tasks [21]. In the experiment, linear, quadratic, cubic, and Gaussian kernel functions were tested, with the Gaussian RBF kernel selected as the best. The C parameter (regularization parameter) was set to 10 and gamma (kernel coefficient) to 0.1 after optimization using grid search with 5-fold cross-validation.

3.4.3. Extreme Learning Machine (ELM)

ELM is a specialized training algorithm for Single-hidden-Layer Feedforward Neural networks (SLFN), which randomly selects hidden nodes and provides emergent learning [14]. ELM is configured with 100 hidden neurons and a sigmoid activation function. The input weights and hidden layer biases are initialized randomly, while the output weights are calculated analytically using the Moore-Penrose generalized inverse.

3.4.4. Radial Basis Function Network (RBFN)

RBFN is a feedforward neural network architecture that uses radial basis functions as activation functions, typically configured with a single hidden layer using Gaussian or several other kernel functions [22]. In this study, RBFN uses 50 radial basis neurons with Gaussian functions. The RBF center is determined using k-means clustering, and the spread parameter is set to 1.0.

3.4.5. Random Forest (RF)

RF is a classifier that grows on multiple classification trees, improving classification accuracy and achieving better generalization for large databases in ensemble learning [23]. RF is configured with 100 decision trees, with a maximum depth of 20 and a minimum sample split of 5. Each tree is trained on a bootstrap sample from the training data, and the final prediction is determined through majority voting.

Table 2. Machine learning algorithm parameter configuration

Algorithm	Main Parameter	Value
ANN	Hidden neurons, Training algorithm	15, Levenberg-Marquardt
SVM	Kernel, C, Gamma	Gaussian RBF, 10, 0.1
ELM	Hidden neurons, Activation	100, Sigmoid
RBFN	RBF neurons, Spread	50, 1.0
RF	Trees, Max depth, Min split	100, 20, 5

3.5. Evaluation Metrics

Several performance metrics are used to evaluate the success of the classifier after the training phase. Performance metrics obtained from the confusion matrix include:

$$Sensitivity\ (Se) = TP / (TP + FN) \quad (1)$$

$$Specificity\ (Sp) = TN / (TN + FP) \quad (2)$$

$$Accuracy\ (Acc) = (TP + TN) / (TP + TN + FP + FN) \quad (3)$$

$$Geometric\ Mean\ (GM) = \sqrt{(Se \times Sp)} \quad (4)$$

$$F\text{-}score\ (F1) = 2 \times (Precision \times Recall) / (Precision + Recall) \quad (5)$$

Where TP (True Positive) is the number of hypoxic cases correctly detected, TN (True Negative) is the number of normal cases correctly identified, FP (False Positive) is the number of normal cases misclassified as hypoxic, and FN (False Negative) is the number of hypoxic cases not detected.

3.6. Statistical Analysis

Cross-validation is a technique suitable for medium-sized datasets, which consists of dividing the data into k subsamples. The dataset is partitioned into training and testing sets, then 10-fold cross-validation is used to estimate the average accuracy of the model. In each fold, 90% of the data (1648 samples) is used for training, and 10% of the

data (183 samples) is used for testing. This process is repeated 10 times with different partitions, and the final result is calculated as the average of all folds. The standard deviation is also calculated to evaluate the stability of the algorithm's performance. A statistical test using a paired t-test is performed to compare the performance between algorithms with a significance level of $\alpha = 0.05$.

4. Results and Discussion

4.1. Dataset Characteristics

The research dataset consisted of 1831 samples with a distribution of 1655 normal cases (90.4%) and 176 hypoxic cases (9.6%). Features in the dataset included various CTG parameters relevant to fetal condition assessment. Descriptive analysis shows that baseline FHR has an average of 133.3 ± 9.8 bpm for normal cases and 138.7 ± 12.4 bpm for hypoxic cases ($p<0.001$).

Table 3. Descriptive statistics of main CTG features

Feature	Normal (Mean±SD)	Hypoxic (Mean±SD)	p-value
Baseline FHR (bpm)	133.3 ± 9.8	138.7 ± 12.4	<0.001
Accelerations (/s)	0.0032 ± 0.0018	0.0019 ± 0.0014	<0.001
Fetal Movement (/s)	0.0021 ± 0.0012	0.0015 ± 0.0009	0.003
Light Decel. (/s)	0.0008 ± 0.0006	0.0014 ± 0.0010	<0.001
Severe Decel. (/s)	0.0002 ± 0.0003	0.0012 ± 0.0008	<0.001
STV Mean	1.2 ± 0.4	0.8 ± 0.3	<0.001

4.2. Algorithm Performance Analysis

The evaluation results show that all machine learning algorithms tested produced satisfactory performance in fetal heart rate classification. Artificial Neural Network (ANN) showed superior performance with a sensitivity of 99.73% and specificity of 97.94%, resulting in an overall accuracy of 97.87%. The geometric mean for ANN reached 98.83%, and the F-score was 98.87%, indicating an excellent balance between sensitivity and specificity.

Support Vector Machine (SVM) showed excellent performance with a sensitivity of 99.21% and a specificity of 97.02%, resulting in an accuracy of 97.13%. Extreme Learning Machine (ELM) showed high sensitivity (99.34%) but relatively lower specificity (95.30%), resulting in an accuracy of 95.62%. The Radial Basis Function Network (RBFN) had the fastest training time (0.073 seconds) but the lowest classification performance among all algorithms tested, with an accuracy of 88.91%. Random Forest (RF) shows a good balance between performance and time efficiency with an accuracy of 96.34%.

Table 4. Comparative performance of machine learning algorithms

Algorithm	Se (%)	Sp (%)	Acc (%)	GM (%)	F1 (%)	Time (s)
ANN	**99.73**	**97.94**	**97.87**	**98.83**	**98.87**	**2.236**
SVM	99.21	97.02	97.13	98.11	98.15	1.847
ELM	99.34	95.30	95.62	97.31	97.35	0.156
RBFN	97.84	88.49	88.91	93.05	93.42	0.073
RF	99.07	96.18	96.34	97.62	97.68	0.332

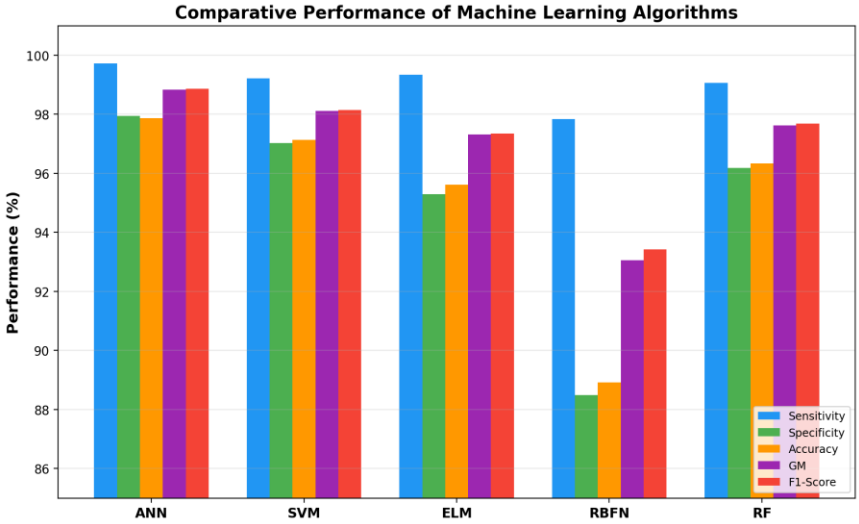

Fig. 3 Comparative performance of machine learning algorithms

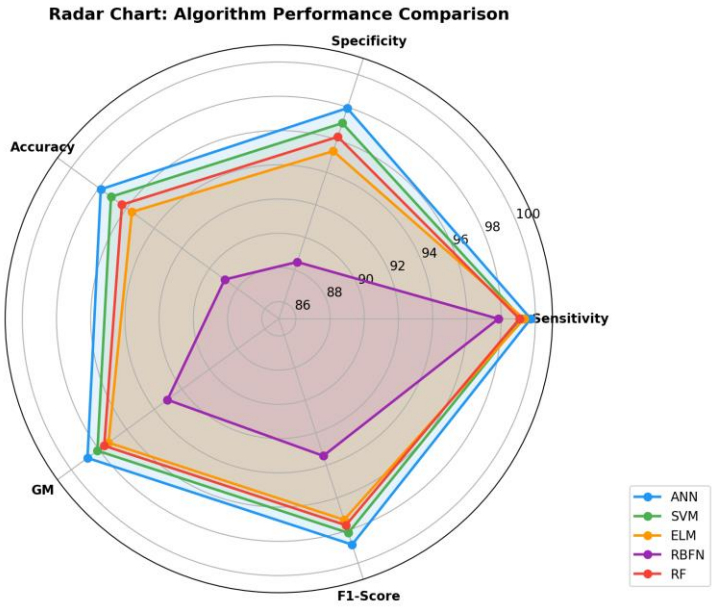

Fig. 4 Radar chart: algorithm performance comparison

4.3. Confusion Matrix Analysis

Confusion matrix analysis provides in-depth insight into the classification performance of each algorithm. For ANN, out of 1655 normal cases, 1621 (97.94%) were correctly classified as normal (True Negative), while 34 (2.06%) were misclassified as hypoxic (False Positive). Of the 176 hypoxic cases, 175 (99.43%) were correctly detected (True Positive), and only 1 (0.57%) was not detected (False Negative).

Table 5. Summary of confusion matrix for all algorithms

Algorithm	TP	TN	FP	FN
ANN	175	1621	34	1
SVM	174	1606	49	2
ELM	175	1577	78	1
RBFN	172	1465	190	4
RF	174	1592	63	2

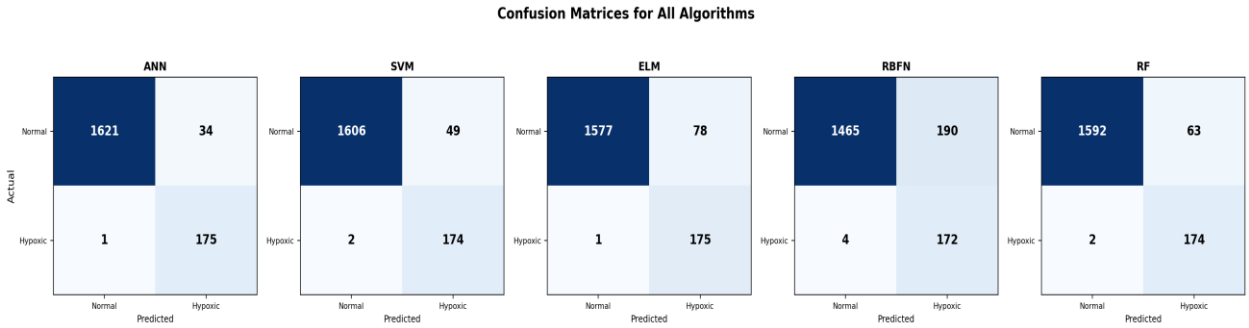

Fig. 5 Confusion matrices for all algorithms

4.4. Training Time Comparison

Training time analysis reveals significant differences in computational efficiency among the algorithms. RBFN demonstrated the fastest training time at 0.073 seconds, followed by ELM (0.156 seconds), RF (0.332 seconds), SVM (1.847 seconds), and ANN (2.236 seconds). Although ANN required the longest training time, this duration remains acceptable for clinical applications that do not require frequent retraining.

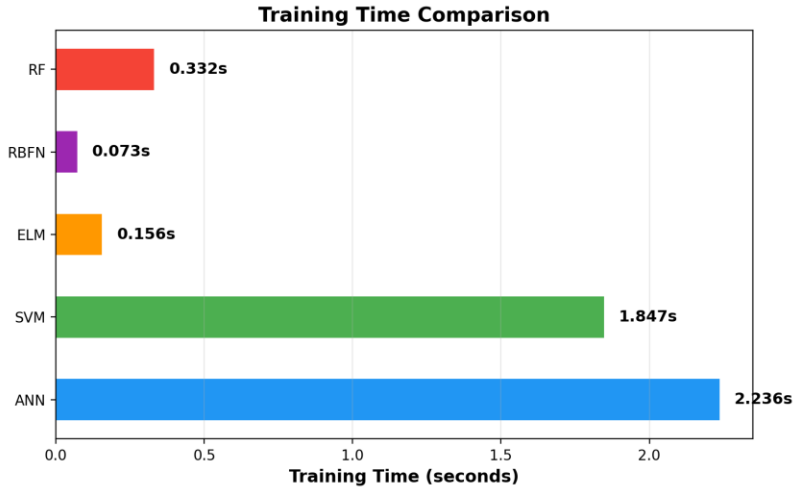

Fig. 6 Training time comparison of algorithms

4.5. Feature Importance Analysis

Feature Importance Analysis using Random Forest shows that abnormal Short-Term Variability (STV) is the most important feature with an importance score of 0.187, followed by the percentage of time with abnormal Long-Term Variability (LTV) with a score of 0.156. Severe decelerations show an importance score of 0.143, indicating their crucial role in fetal condition classification.

Table 6. Ranking feature importance (random forest)

Rank	Feature	Importance Score
1	Abnormal STV	0.187
2	Abnormal LTV (%)	0.156
3	Severe Decelerations	0.143
4	Histogram Mean	0.128
5	Histogram Variance	0.112
6	Accelerations	0.095
7	Baseline FHR	0.089

8	Fetal Movement	0.078
9	Light Decelerations	0.067
10	Prolonged Decelerations	0.045

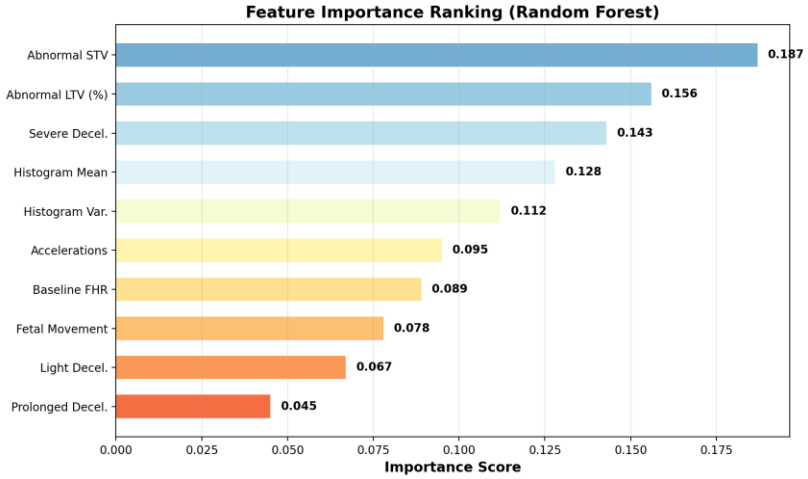

Fig. 7 Feature importance ranking (Random Forest)

4.6. Clinical Predictor Analysis

The study identified several significant factors that influence changes in fetal heart rate during labor. Combined Spinal-Epidural (CSE) showed a significant association with a decrease in FHR, with an odds ratio of 2.89 (95% CI: 1.15-7.25, p=0.02). The interaction between CSE and phenylephrine dose showed very high significance (p<0.0001). Decelerations occurring before neuraxial block also showed a significant association (p<0.001) with a decrease in FHR, with an odds ratio of 4.12 (95% CI: 2.34-7.25).

Table 7. Predictors of fetal heart rate changes

Factor	Odds Ratio / Coef.	95% CI	p-value
CSE	OR: 2.89	1.15 - 7.25	0.02
CSE × Phenylephrine	β: -0.032	-0.043 - (-0.021)	<0.0001
Pre-block deceleration	OR: 4.12	2.34 - 7.25	<0.001
Bupivacaine dosage	β: -0.18	-0.34 - (-0.02)	0.03
Duration of Stage I	r: 0.34	0.21 - 0.46	<0.001

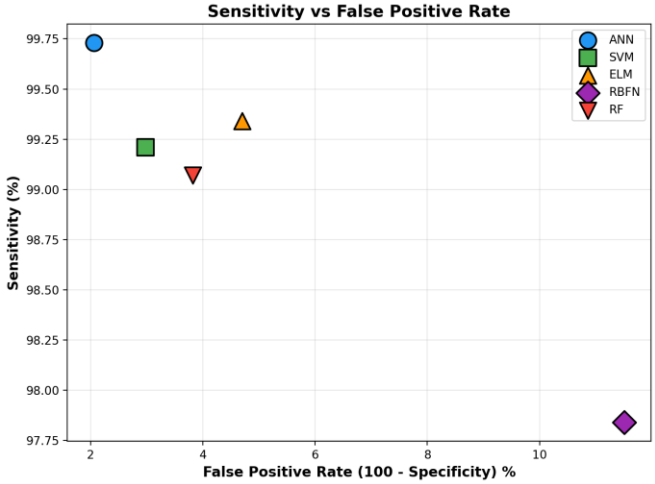

Fig. 8 Sensitivity vs False positive rate for all algorithms

5. Discussion

5.1. Superiority of Artificial Neural Networks

The results demonstrate that ANN provides the best performance in fetal heart rate classification with an accuracy of 97.87%, sensitivity of 99.73%, and specificity of 97.94%. The advantages of ANN can be explained by its ability to handle complex nonlinear relationships between various CTG parameters, allowing the model to capture patterns that may not be detected by traditional linear methods. The Levenberg-Marquardt training algorithm proved effective in optimizing network weights, combining the convergence speed of the Gauss-Newton method with the stability of gradient descent. The high sensitivity (99.73%) is critically important in a clinical context because the ability to accurately detect cases of fetal hypoxia can prevent serious complications such as cerebral palsy, hypoxic-ischemic encephalopathy, and perinatal death. With only 1 false negative out of 176 hypoxic cases, ANN demonstrates very high reliability in detecting conditions requiring immediate medical intervention. The ANN architecture with 15 hidden neurons proved optimal for this dataset. Too few neurons can cause underfitting, while too many neurons can cause overfitting. The configuration of 15 hidden neurons provides an optimal balance between model complexity and generalization ability.

5.2. Algorithm Performance Comparison

SVM shows very competitive performance with ANN, with an accuracy difference of only 0.74%. The Gaussian RBF kernel allows SVM to handle non-linearly separable data effectively. ELM shows an interesting trade-off between speed and accuracy - with a training time of only 0.156 seconds, ELM is 14.3 times faster than ANN. RBFN showed the fastest training time (0.073 seconds) but with the lowest classification performance. RF demonstrates an excellent balance between performance (96.34% accuracy) and computational efficiency (0.332 seconds training time). Statistical analysis using a paired t-test showed that the difference in performance between ANN and SVM was not statistically significant ($p=0.127$), indicating that both algorithms can be considered equivalent in terms of classification accuracy. However, the difference between ANN and RBFN was highly significant ($p<0.001$), confirming the superiority of ANN in this dataset.

5.3. Comparison with Previous Research

The results of this study are consistent with the findings of Das et al. [12] and Cömert and Kocamaz [14]. Ocak [15] reported a slightly higher accuracy of 98.2% with SVM optimized using a Genetic Algorithm, which can be explained by the use of GA for parameter optimization and feature selection. Sahin and Subasi [16] reported the highest accuracy with Random Forest (99.1%), possibly due to differences in data preprocessing and parameter configuration.

Table 8. Comparison with previous research

Research	Algorithm	Accuracy (%)	Dataset	Year
This research	**ANN**	**97.87**	**CTG (n=1831)**	**2024**
Das et al. [12]	MLP	97.94	CTG (n=340)	2023
Riveros-Perez et al. [13]	RF	R^2=0.78	Clinical (n=1077)	2023
Cömert & Kocamaz [14]	ANN	97.87	CTG (n=2126)	2017
Ocak [15]	SVM+GA	98.20	CTG (n=2126)	2013
Sahin & Subasi [16]	RF	99.10	CTG (n=2126)	2015

5.4. Clinical Implications

The identification of CSE as a significant predictor factor in FHR changes has important clinical implications. Secondary uterine hypertonicity due to a rapid decrease in plasma catecholamine levels is thought to be the main cause [24]. The interaction between CSE and phenylephrine dose was highly significant ($p<0.0001$), suggesting that maternal hypotension contributes to fetal bradycardia. These findings emphasize the importance of preventive measures, including preloading fluids, lateral positioning, close blood pressure monitoring, and prophylactic vasopressor use at individually adjusted doses.

6. Limitations and Future Research Directions

Although the research results show excellent performance, several limitations should be acknowledged. First, the dataset used comes from a single source (SisPorto 2.0), so the generalizability of the results may be limited. External validation using datasets from different institutions with diverse populations is needed. Second, class imbalance in the dataset (1655 normal vs. 176 hypoxic, a ratio of 9.4:1) may affect algorithm performance. Techniques such as SMOTE or class weights can be considered to address this issue.

Third, model interpretability, especially for ANNs, remains a challenge in clinical implementation. The development of explainable AI methods such as LIME or SHAP can help improve interpretability. Fourth, this study used features extracted automatically by SisPorto 2.0, which may not capture all relevant aspects of the CTG signal. Fifth, additional clinical factors such as gestational age, parity, and maternal medical conditions were not considered. Sixth, prospective validation in a real clinical setting is necessary.

Future research directions include: (1) validation with larger and more diverse multi-institutional datasets; (2) development of explainable AI systems using LIME or SHAP techniques; (3) prospective clinical evaluation studies; (4) development of hybrid ensemble systems combining multiple algorithms; (5) integration of additional clinical information; (6) exploration of deep learning approaches (CNN, RNN) for direct raw CTG signal analysis; and (7) development of multi-outcome prediction systems for specific neonatal outcomes.

7. Conclusion

This study demonstrates that the machine learning approach, particularly Artificial Neural Network, can provide very high accuracy in fetal heart rate classification with a sensitivity of 99.73%, specificity of 97.94%, and overall accuracy of 97.87%. A comprehensive comparison of five machine learning algorithms shows that although all algorithms produce satisfactory performance (accuracy >88%), ANN provided the best overall performance, followed by SVM (97.13%), RF (96.34%), ELM (95.62%), and RBFN (88.91%).

Identification of important predictor factors such as Combined Spinal-Epidural (OR: 2.89, $p=0.02$), CSE-phenylephrine interaction ($p<0.0001$), pre-block deceleration (OR: 4.12, $p<0.001$), and bupivacaine dose ($\beta=-0.18$, $p=0.03$) provides valuable clinical insights for labor analgesia management. Feature importance analysis identified abnormal short-term variability (0.187), abnormal long-term variability (0.156), and severe decelerations (0.143) as the most important features for classification.

These findings indicate great potential for the implementation of automated systems in fetal monitoring that can reduce variability in interpretation and improve early detection of fetal complications, thereby contributing to improved maternal and neonatal safety. The implementation of machine learning-based systems in clinical practice requires further validation, seamless integration with clinical workflows, and careful consideration of interpretability and regulatory aspects.

References

[1] Ana Pinas, and Edwin Chandraharan, "Continuous Cardiotocography During Labour: Analysis, Classification and Management," *Best Practice and Research Clinical Obstetrics and Gynaecology*, vol. 30, pp. 33-47, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[2] Diogo Ayres-de-Campos, Catherine Y. Spong, and Edwin Chandraharan, "FIGO Consensus Guidelines on Intrapartum Fetal Monitoring: Cardiotocography," *International Journal of Gynecology and Obstetrics*, vol. 131, no. 1, pp. 13-24, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[3] Ingemar Ingemarsson, "Gender Aspects of Preterm Birth," *BJOG: An International Journal of Obstetrics and Gynaecology*, vol. 110, pp. 34-38, 2003. [CrossRef] [Google Scholar] [Publisher Link]

[4] Anna-Karin Sundström, David Rosén, and K.G. Rosén, "Fetal Surveillance," *Gothenburg: Neoventa Medical AB*, 2000. [Google Scholar]

[5] Molly J. Stout, and Alison G. Cahill, "Electronic Fetal Monitoring: Past, Present, and Future," *Clinics in Perinatology*, vol. 38, no. 1, pp. 127-142, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[6] E.S. Draper et al., "A Confidential Enquiry into Cases of Neonatal Encephalopathy," *Archives of Disease in Childhood-Fetal and Neonatal Edition*, vol. 87, no. 3, pp. F176-F180, 2002. [CrossRef] [Google Scholar] [Publisher Link]

[7] Laura M. Glaser, Farah A. Alvi, and Magdy P. Milad, "Trends in Malpractice Claims for Obstetric and Gynecologic Procedures, 2005 Through 2014," *American Journal of Obstetrics and Gynecology*, vol. 217, no. 3, pp. 340.e1-340.e6, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[8] Zafer Cömert, and Adnan Fatih Kocamaz, "Evaluation of Fetal Distress Diagnosis During Delivery Stages based on Linear and Nonlinear Features of Fetal Heart Rate for Neural Network Community," *International Journal of Computer Applications*, vol. 156, no. 4, pp. 26-31, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[9] Mei-Ling Huang, and Yung-Yan Hsu, "Fetal Distress Prediction using Discriminant Analysis, Decision Tree, and Artificial Neural Network," *Journal of Biomedical Science and Engineering*, vol. 5, no. 9, pp. 526-533, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[10] Ersen Yılmaz, and Çağlar Kılıkçıer, "Determination of Fetal State from Cardiotocogram using LS-SVM with Particle Swarm Optimization and Binary Decision Tree," *Computational and Mathematical Methods in Medicine*, vol. 2013, no. 1, pp. 1-8, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[11] Radhika R. Halde, "Application of Machine Learning Algorithms for Betterment in Education System," *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*, Pune, India, pp. 1110-1114, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[12] Sahana Das et al., "A Machine Learning Pipeline to Classify Fetal Heart Rate Deceleration with Optimal Feature Set," *Scientific Reports*, vol. 13, no. 1, pp. 1-20, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[13] Efrain Riveros-Perez, Javier Jose Polania-Gutierrez, and Bibiana Avella-Molano, "Fetal Heart Rate Changes and Labor Neuraxial Analgesia: A Machine Learning Approach," *BMC Pregnancy and Childbirth*, vol. 23, no. 1, pp. 1-7, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[14] Z. Comert, and A.F. Kocamaz, "Comparison of Machine Learning Techniques for Fetal Heart Rate Classification," *Acta Physica Polonica A*, vol. 132, no. 3, pp. 451-454, 2017. [CrossRef] [Google Scholar]

[15] Hasan Ocak, "A Medical Decision Support System based on Support Vector Machines and the Genetic Algorithm for the Evaluation of Fetal Well-Being," *Journal of Medical Systems*, vol. 37, no. 2, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[16] 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, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[17] Jiří Spilka et al., "Sparse Support Vector Machine for Intrapartum Fetal Heart Rate Classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 3, pp. 664-671, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[18] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew, "Extreme Learning Machine: Theory and Applications," *Neurocomputing*, vol. 70, no. 1-3, pp. 489-501, 2006. [CrossRef] [Google Scholar] [Publisher Link]

[19] Diogo Ayres-de-campos et al., "SisPorto 2.0: A Program for Automated Analysis of Cardiotocograms," *Journal of Maternal-Fetal Medicine*, vol. 9, no. 5, pp. 311-318, 2000. [Google Scholar] [Publisher Link]

[20] Zafer Cömert, and Adnan Fatih Kocamaz, "Fetal Hypoxia Detection based on Deep Convolutional Neural Network with Transfer Learning Approach," *Software Engineering and Algorithms in Intelligent Systems: Proceedings of 7th Computer Science On-line Conference*, Springer, Cham, vol. 1, pp. 239-248, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[21] Jiří Spilka et al., "Automatic Evaluation of FHR Recordings from CTU-UHB CTG Database," *Information Technology in Bio- and Medical Informatics: 4th International Conference, ITBAM 2013*, Prague, Czech Republic, vol. 8060, pp. 47-61, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[22] K. Warwick, and R. Craddock, "An Introduction to Radial basis Functions for System Identification. A Comparison with Other Neural Network Methods," *Proceedings of 35th IEEE Conference on Decision and Control*, Kobe, Japan, vol. 1, pp. 464-469, 1996. [CrossRef] [Google Scholar] [Publisher Link]

[23] Peterek Tomáš et al., "Classification of Cardiotocography Records by Random Forest," *2013 36th International Conference on Telecommunications and Signal Processing (TSP)*, Rome, Italy, pp. 620-923, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[24] J. Nicolet et al., "Maternal Factors Implicated in Fetal Bradycardia after Combined Spinal Epidural for Labour Pain," *European Journal of Anaesthesiology*, vol. 25, no. 9, pp. 721-725, 2008. [CrossRef] [Google Scholar] [Publisher Link]