IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 13, No. 2, February, 2026, pp. 371-380
Journal homepage: http://iieta.org/journals/mmep

# Optimizing Fetal Heartbeat Detection with Feature Selection in Medical Data Using Machine Learning Algorithms

Yuli Wahyuni[1,2*], Hadiyanto[3], Ridwan Sanjaya[4], Nendar Herdianto[5]

[1] Doctoral Program in Information Systems, School of Postgraduate Studies, Universitas Diponegoro, Semarang 50241, Indonesia
[2] Department of Computer Engineering, Vocational School, Pakuan University, Bogor 16143, Indonesia
[3] Chemical Engineering Department, Faculty of Engineering, Universitas Diponegoro, Semarang 50241, Indonesia
[4] Department of Information Systems, Universitas Katolik Soegijapranata, Semarang 50234, Indonesia
[5] Center for Composites and Biomaterials Research, Badan Riset dan Inovasi Nasional (BRIN), Banten 15314, Indonesia

Corresponding Author Email: yuli_wahyuni@unpak.ac.id



https://doi.org/10.18280/mmep.130213

**Received:** 10 October 2025
**Revised:** 20 December 2025
**Accepted:** 4 January 2026
**Available online:** 15 March 2026

***Keywords:***
*optimizing, machine learning, fetal heartbeat, detection, feature selection, classification models, predictive modeling*

**ABSTRACT**

The interpretation of cardiotocography (CTG) for fetal health monitoring is often hampered by subjective assessments and inter-observer variability, which highlights the need for more objective computational techniques. This study aims to determine whether feature selection can improve the classification of prenatal health conditions using machine learning techniques. Using a publicly available CTG dataset with 2,126 recordings, Analysis of Variance (ANOVA) F-test and Recursive Feature Elimination (RFE) were performed to reduce the number of variables from 21 to 10. The models trained using the full feature set and the reduced feature set include Random Forest, Logistic Regression, and Support Vector Machine (SVM). The evaluation metrics used include F1-score, recall, accuracy, and precision. After feature selection, the Random Forest model showed the most significant improvement, with accuracy increasing from 92.72% to 93.19%, the macro F1-score rising from 86% to 87%, and the micro F1-score also demonstrating improved performance. Although the magnitude of this improvement is relatively small, these findings suggest that feature reduction can effectively reduce computational complexity while providing incremental improvements in model performance. In addition, balanced classification between Normal and Pathological classes is maintained with the smaller feature set. Overall, the results indicate that a focused feature selection approach can improve the effectiveness of fetal health categorization models without compromising prediction quality; however, further research is needed to validate the clinical use of these features.

## 1. INTRODUCTION

Fetal health monitoring during pregnancy and childbirth is very important to prevent complications such as fetal hypoxia, intrapartum asphyxia, and fetal death, which are often preceded by abnormal fetal heart rate (FHR) patterns, including bradycardia, tachycardia, and variability disorders [1]. Cardiotocography (CTG) remains the primary clinical tool for assessing FHR and uterine contractions; however, because different professionals have different experiences, their interpretations tend to be subjective, influenced by environmental conditions and clinical standards. The high inter-observer variability and uneven diagnostic accuracy resulting from this subjectivity highlight the need for more objective and automated methods of CTG interpretation [2].

Machine learning is increasingly being used to support CTG analysis. Previous studies have shown that machine learning can analyze complex physiological data and identify abnormal fetal patterns [3]. Baseline heart rate, acceleration, and deceleration, and short- and long-term variability indices are some of the numerical characteristics found in CTG datasets, and metrics derived from histograms, many of which may be redundant or contain minimal information. Feature redundancy can increase computational load, reduce generalization, and hinder interpretability. As a result, feature selection has emerged as a critical step for identifying the most relevant CTG attributes, reducing overfitting, and improving the efficiency and transparency of models in clinical decision-making [4]. When combined with machine learning, feature selection has been shown to improve the reliability and consistency of classification in fetal health assessment [5, 6].

However, the quality and representativeness of input significantly affect the effectiveness of machine learning models. Various feature selection methods, including ensemble-based techniques, mutual information, SelectKBest, and Recursive Feature Elimination (RFE), have been evaluated to address potential challenges in CTG datasets, such as noisy data, redundant variables, and imbalanced class distributions [7]. Although performance improvements have been reported for algorithms such as Support Vector Machine

(SVM) and Random Forest after feature selection, variations in datasets and analytical methods continue to produce inconsistent results, highlighting the need for broader validation and more standardized methodologies [8].

Various machine learning techniques have been applied to CTG analysis, including FHR pattern recognition, metabolic acidosis prediction, and perinatal risk assessment [3, 4]. Ensemble methods, Artificial Neural Networks (ANNs), and complex CTG patterns that are difficult for humans to see visually can be comprehensively modeled by deep learning architectures such as Random Forest, SVM, K-Nearest Neighbors (KNNs), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), hybrid CNN LSTM, and transformer-based models [9, 10]. Yet, CTG-based machine learning systems still face challenges related to signal noise, device variability, class imbalance, interpretability requirements, and the technical demands related to real-time fetal monitoring systems and emerging IoT [11-13].

Despite significant progress, several key research gaps remain. These include: (1) lack of diverse and representative CTG datasets across different populations and types of equipment, (2) limited external validation of machine learning models in real-world clinical settings, (3) the absence of a systematic feature selection framework that balances performance and interpretability, (4) inadequate integration between artifact-handling and signal-quality techniques, and (5) limited efforts to ensure that machine learning models are explainable and clinically acceptable [14-17]. Previous studies have highlighted that improvements in data quality, feature selection, model validation, readability, and clinical integration are essential to optimize the potential of AI-based CTG monitoring [18-21].

Given these challenges, this study investigates the role of feature selection in fetal health classification using CTG data. To develop a more robust and clinically viable fetal health detection model, this study aims to: (1) identify the most relevant CTG features using the Analysis of Variance (ANOVA) F-test and RFE; (2) compare the performance of Logistic Regression, SVM, and Random Forest classification; and (3) evaluate the impact of feature selection on accuracy, precision, recall, macro/micro F1-scores, and computational efficiency [3, 22-25].

## 2. METHODS

Figure 1 presents a flowchart illustrating the research process, in which machine learning algorithms are applied both with and without feature selection to evaluate and enhance FHR detection. This approach aims to improve model efficiency while maintaining accuracy, with feature selection serving to identify and optimize the most relevant variables for effective FHR categorization.

### 2.1 Cardiotocography dataset

Figure 1 begins with the CTG dataset from the University of California, Irvine (UCI) Machine Learning Repository. This dataset consists of 2,126 recordings and includes 21 variables, including FHR, variability, acceleration, and other factors relevant to assessing fetal health.

The three target classes in the CTG dataset—Normal, Suspect, and Pathological—represent different levels of fetal risk. There is a significant imbalance between these classes, with 1,655 samples in the Normal category compared to only 295 in the Pathological category. This imbalance directly affects classification performance, especially for minority classes, making it an essential methodological consideration. To address this issue, stratified splitting was applied during data splitting to ensure consistent class proportions between the training and testing sets. In addition, evaluation metrics such as macro F1-scores were prioritized to accurately capture performance across classes rather than being dominated by the majority class.

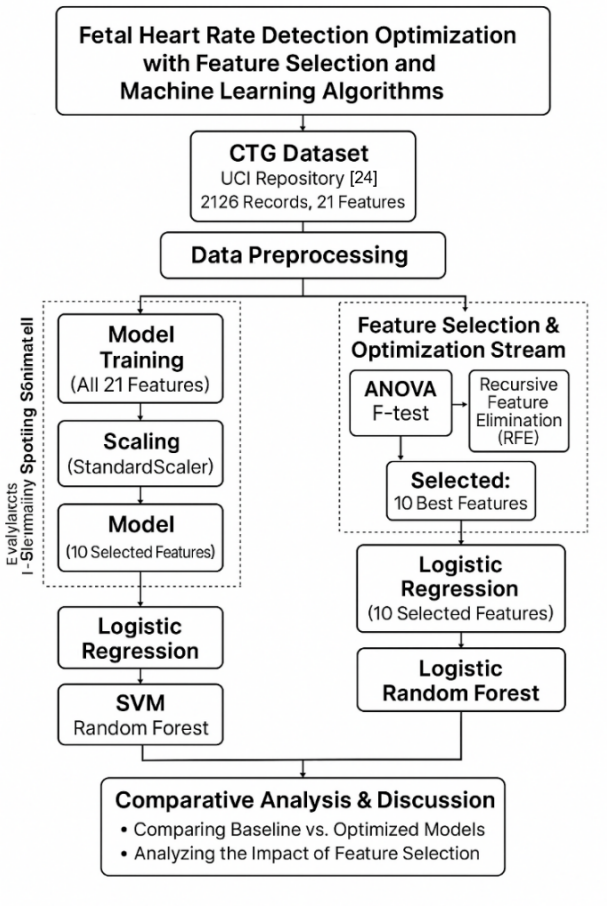


**Figure 1.** The optimization of fetal heart rate detection through feature selection and machine learning algorithms [24]

### 2.2 Data preprocessing

The data underwent a preparation stage before being used for model training. This stage included:

a) Stratified data division (80/20): To maintain the same class distribution (1,655/295/176) in both the training and testing subsets, the dataset was divided into two parts.
b) Scaling (StandardScaler): All features were scaled using StandardScaler to normalize variance among predictors, allowing algorithms such as SVM and Logistic Regression to work optimally.

### 2.3 Separation of analysis paths

Once the data were ready, the process was divided into two parallel paths:
(1) Baseline evaluation path

On this path, all 21 features were used without feature selection. Before optimization, the main goal was to achieve the baseline performance of the model. The procedure included:

a. Data model, consisting of all 21 characteristics, was trained using three different methodologies: Random Forest, SVM, and Logistic Regression.
b. Baseline performance assessment, consisting of metrics such as recall, accuracy, precision, and F1-score to evaluate model performance before feature reduction.

(2) Feature selection and workflow optimization

Feature selection was employed to ensure the most informative subset of predictors to improve classification efficiency and accuracy. The ANOVA F-test evaluated the statistical significance of individual features, whereas RFE determined traits with negligible contributions. After identifying the ten most significant characteristics, the Random Forest, SVM, and Logistic Regression models were reconfigured using the reduced feature set. Subsequently, the effectiveness of the models was reevaluated to determine the impact of feature selection on classification results.

### 2.4 Final analysis

The final part of the flowchart shows the comparative analysis stage:

a. Comparing how well the optimized model (using only 10 characteristics) performed compared to the baseline model (using 21 features).
b. Analyzing the impact of feature selection on model accuracy, computational efficiency, and generalization comprehensively.

Figure 1 illustrates the systematic process of developing and optimizing an FHR classification model. This study focuses not only on model training but also evaluates the impact of feature selection on final performance. The two-path approach (baseline vs. optimization) helps ensure that performance improvements were not coincidental, but rather the result of appropriate feature selection.

### 2.5 Feature selection method

Feature selection was employed to ensure the most informative subset of predictors to improve classification efficiency and accuracy. The ANOVA F-test evaluated the statistical significance of individual features, while RFE determined traits with negligible contributions. After identifying the ten most significant features, the Random Forest, SVM, and Logistic Regression models were retrained using the reduced feature set. The ANOVA F-test was chosen as an effective univariate filter method to measure statistical differences in feature distributions among fetal health categories. This method provides quick and objective rankings for numerical variables, making it suitable for an initial screening stage. RFE was chosen as a wrapper method to complement the screening approach by considering multivariate interactions that cannot be captured by the ANOVA F-test. Compared to other alternatives, such as Mutual Information or SHapley Additive exPlanations (SHAP), the ANOVA F-test, and RFE, were preferred due to (1) lower computational cost, (2) suitability for limited sample sizes such as CTG datasets, and (3) better alignment with the study objective to balance performance and readability. SHAP, while highly interpretable, requires substantial computational resources and is generally used for post-analysis explanation rather than feature selection. Mutual Information was not selected because it is less effective with correlated features, which are commonly found in CTG measurements.

2.5.1 Analysis of Variance F-test (filter method)

For statistical purposes, the relationship between each numerical characteristic and the categorical fetal health outcomes was analyzed using the ANOVA F-test. A reduced feature subset was initially formed by selecting features with the largest F-statistic values using SelectKBest from Scikit-learn.

2.5.2 Recursive Feature Elimination (wrapper method)

RFE was implemented using Logistic Regression as an estimator. This algorithm repeatedly removed features with the lowest importance until a predetermined number of selected features was reached. As a wrapper method, RFE evaluated feature combinations in conjunction with the algorithm, enabling the selection of a more representative multivariate feature subset.

2.5.3 Classification algorithm

This study utilized three machine learning algorithms, Random Forest, Logistic Regression, and SVM, to identify fetal health problems. Grid search with 5-fold stratified cross-validation was used to adjust the hyperparameters of SVM and Random Forest. With this optimization, we ensured that the selected parameters were robust, reduced the possibility of overfitting, and helped achieve the goal of improving model performance.

**(1) Logistic Regression**

Logistic Regression was used as a baseline classifier due to its ease of interpretation and efficiency. The model employed L2 regularization penalty with max_iter = 1000 to ensure convergence.

**(2) SVM**

Instead of using default parameters, SVM with the RBF kernel was optimized using grid search.

Explored grid parameters:

- C: {0.1, 1, 10, 100}
- gamma: {1e-3, 1e-2, 1e-1, ‘scale’}
- kernel: {‘rbf’}

Optimal parameters:

- C = 10
- gamma = 0.01
- kernel = 'rbf'

These adjusted parameters improved class separation and enhanced performance for minority fetal-health classes.

**(3) Random Forest**

Random Forest hyperparameters were adjusted to increase tree diversity and reduce overfitting.

Explored grid parameter:

- n_estimators: {100, 200, 300, 500}
- max_depth: {None, 10, 20, 30}
- min_samples_split: {2, 5, 10}
- min_samples_leaf: {1, 2, 4}
- max_features: {‘sqrt’, ‘log2’}

Optimal parameters:

- n_estimators = 300
- max_depth = 20
- min_samples_split = 5

- min_samples_leaf = 2
- max_features = 'sqrt'

This optimization produced a more stable ensemble with better generalization.

2.5.4 Evaluation metrics

The metrics included a confusion matrix, accuracy, precision, recall, macro/micro F1-score, and algorithm performance evaluation. Due to the class imbalance, greater emphasis was placed on the macro F1-score and per-class recall, especially for the minority classes (Suspect and Pathological), which are clinically critical.

2.5.5 Feature selection implementation

This software identified the ten most significant features from the normalized data with the ANOVA F-test and RFE. The selection approach sought to improve generalization, minimize redundancy, and augment computing efficiency. Eleven attributes were deliberately chosen. An initial assessment was conducted by analyzing several k values, including 5, 8, 10, 12, and 15 characteristics, to determine the ideal number of features. The ANOVA F-test and RFE were utilized for this purpose. The results showed that k = 10 consistently produced the best balance between computational cost, accuracy, and macro F1-score (especially for minority classes). Retaining fewer than 10 features resulted in the loss of important clinical information, while selecting more than 10 features increased redundancy and computation time without meaningful performance improvement. Additionally, the decision aligns with prior CTG feature-selection studies, in which 8–12 features are commonly retained to achieve interpretability, avoid overfitting, and maintain relevant clinical variables. Therefore, the selection of 10 features is methodologically justified and based on in empirical testing and consistency with the existing literature.

2.5.6 Model training without feature selection

To evaluate the initial performance of each classifier using all 21 characteristics, model training without feature selection was performed as a baseline. Unlike the previous version, this revised baseline integrated optimized SVM and Random Forest hyperparameters to ensure a fair comparison between models before and after feature selection (Table 1).

In this first phase, all classifiers were trained on all scaled features using optimized hyperparameters for SVM and Random Forest. The test set was used to generate predictions, and performance was evaluated by computing the F1-score, recall, accuracy, and precision. The optimized configuration ensured that the baseline reflected the performance capability of each algorithm without relying on default parameters.

**Table 1.** Model training without feature selection

| Algorithm | Library/Function | Configuration (Updated) | Evaluation Metrics |
|---|---|---|---|
| Logistic Regression | Logistic Regression (max_iter=1000) | Penalty=L2 | Accuracy, Precision, Recall, F1-score |
| Support Vector Machine | Support Vector Classifier () | C=10, γ=0.01, kernel=Radial Basis Function (RBF) (Grid search optimized) | Accuracy, Precision, Recall, F1-score |
| Random Forest | Random Forest Classifier () | n_estimators=300, max_depth=20, min_samples_split=5, min_samples_leaf=2, max_features=sqrt (Grid search optimized) | Accuracy, Precision, Recall, F1-score |

**Table 2.** Model training with feature selection

| Algorithm | Configuration (Updated) | Evaluation Metrics |
|---|---|---|
| Logistic Regression | max_iter=1000; L2 penalty | Accuracy, Precision, Recall, F1-score |
| Support Vector Machine | C=10; γ=0.01; kernel=Radial Basis Function (RBF) | Accuracy, Precision, Recall, F1-score |
| Random Forest | n_estimators=300; max_depth=20; min_samples_split=5; min_samples_leaf=2; max_features=sqrt | Accuracy, Precision, Recall, F1-score |

2.5.7 Model training with feature selection

The top ten features chosen using an ANOVA F-test and RFE were used to train the model with feature selection. To ensure consistency and determine how feature reduction affected model performance, the dataset was reprocessed using improved SVM and Random Forest configurations (Table 2).

All models were trained using the identically optimized hyperparameters used during the baseline phase, incorporating a condensed 10-feature dataset. This ensured that performance variances were only attributable to feature selection rather than alterations in parameters. Performance metrics were evaluated to determine whether feature reduction improved accuracy, computational efficiency, and class-balance sensitivity.

**2.6 Feature optimization in machine learning**

Feature optimization is a critical pre-modeling step in machine learning that improves algorithm performance through the selection of the most relevant and informative features. By eliminating redundant or low-contributing variables, this process reduces model complexity, mitigates overfitting, and improves computational efficiency. Common feature optimization approaches include filter, wrapper, and embedded methods. When applied correctly, these techniques enable machine learning models to produce more accurate and interpretable predictive results.

2.6.1 Importing algorithm library

Importing several important libraries to support the entire process, this application was the initial stage in data processing and analysis using machine learning techniques. Data manipulation was accomplished using libraries such as pandas and numpy, while data visualization was executed using matplotlib.pyplot, and seaborn. Features were standardized using the sklearn.model_selection and sklearn.preprocessing modules, and the data were partitioned into training and test sets. Two common techniques for feature selection are RFE

and SelectKBest, which use the f_classif statistical approach. Various leading methods, including XGBoost, Random Forest, Logistic Regression, and SVM, were employed in the classification phase. Finally, to assess model performance, the sklearn.metrics module employed evaluation measures such as accuracy, confusion matrix, and classification report (Table 3).

**Table 3.** Libraries, modules, and algorithms used in the optimization workflow

| Category | Library/Module | Function/Algorithm | Description |
|---|---|---|---|
| Data and Visualization | pandas, numpy, matplotlib, seaborn | DataFrame, numerical computation, data visualization | Used for data management, analysis, and visualization of experimental results |
| Preprocessing | sklearn.model_selection | train_test_split | Splits the dataset into training and testing sets |
| | sklearn.preprocessing | StandardScaler | Normalizes/standardizes features to ensure balanced scaling |
| Feature Selection | sklearn.feature_selection | SelectKBest, f_classif, RFE | Selects the best features using statistical methods and recursive elimination |
| Classification | sklearn.linear_model | LogisticRegression | Logistic regression model for binary and multi-class classification |
| | sklearn.svm | SVC | Support Vector Machine classifier for supervised classification |
| | sklearn.ensemble | RandomForestClassifier | Ensemble algorithm based on decision trees |
| | xgboost | XGBClassifier | High-performance tree-based boosting algorithm |
| Evaluation | sklearn.metrics | accuracy_score, classification_report, confusion_matrix | Model evaluation metrics (accuracy, classification report, and confusion matrix) |

Note: SVC: Support Vector Classifier; XGB: Extreme Gradient Boosting; RFE: Recursive Feature Elimination.

**Table 4.** Data loading and preparation workflow

| Step | Code/Function | Description |
|---|---|---|
| Dataset Input | pd.read_csv('fetal_health2.csv') | Loads the dataset containing fetal health records |
| Dataset Information | df.info(), df['fetal_health'].value_counts() | Displays general dataset information and class distribution of the target |
| Feature–Target Separation | X=df.drop('fetal_health', axis=1); y=df['fetal_health'] | Separates independent variables (features) from the dependent variable (target) |
| Train–Test Split | train_test_split(X, y, test_size=0.2, random_state=42, stratify=y) | Separates data into training and testing sets (80% training, 20% testing) with stratification |
| Feature Scaling | StandardScaler() | Standardizes feature values (essential for algorithms like Support Vector Machine and Logistic Regression) |
| Scaled Datasets | scaler.fit_transform(X_train); scaler.transform(X_test) | Applies scaling to training and testing datasets |

**Table 5.** Baseline data algorithm

| No. | Feature Name | Non-Null Count | Data Type |
|---|---|---|---|
| 1 | baseline value | 2126 | float64 |
| 2 | accelerations | 2126 | float64 |
| 3 | fetal_movement | 2126 | float64 |
| 4 | uterine_contractions | 2126 | float64 |
| 5 | light_decelerations | 2126 | float64 |
| 6 | severe_decelerations | 2126 | float64 |
| 7 | prolonged_decelerations | 2126 | float64 |
| 8 | abnormal_short_term_variability | 2126 | float64 |
| 9 | mean_value_of_short_term_variability | 2126 | float64 |
| 10 | percentage_of_time_with_abnormal_long_term_variability | 2126 | float64 |
| 11 | mean_value_of_long_term_variability | 2126 | float64 |
| 12 | histogram_width | 2126 | float64 |
| 13 | histogram_min | 2126 | float64 |
| 14 | histogram_max | 2126 | float64 |
| 15 | histogram_number_of_peaks | 2126 | float64 |
| 16 | histogram_number_of_zeroes | 2126 | float64 |
| 17 | histogram_mode | 2126 | float64 |
| 18 | histogram_mean | 2126 | float64 |
| 19 | histogram_median | 2126 | float64 |
| 20 | histogram_variance | 2126 | float64 |
| 21 | histogram_tendency | 2126 | float64 |
| 22 | fetal_health (Target) | 2126 | int64 |

2.6.2 Data loading and preparation

Data loading and preparation is the initial stage in machine learning that aims to load and prepare data so that it can be used by the model. This step involved retrieving data from sources, removing errors or gaps, and transforming the data through normalization and dividing it into training and testing

datasets. Data quality is extremely important since it substantially affects the accuracy of model predictions. Various algorithms are available (Table 4).

This program aimed to load and prepare data from the fetal_health.csv dataset for fetal health condition classification analysis. After ensuring the file was available, the program displayed the data structure and distribution to understand the initial condition of the dataset. The data were then divided into features (X) and targets (y), and further divided into training and test datasets using stratification to maintain a balanced class proportion. The features in the dataset were normalized using StandardScaler, which is crucial for improving the performance of models such as SVM and Logistic Regression.

2.6.3 Baseline data

Initial data collected before treatment or intervention function as a reference point for comparing changes or effects of a process, program, or model. In this study, baseline data are important for determining the initial condition of the research object so that the results of the analysis can be measured objectively.

Table 5 presents a complete list of the dataset's numerical attributes, along with the target variable fetal_health, all of which have a uniform data type and complete values with no nulls. These features include FHR characteristics, variability patterns, and other histogram-based markers, which are clinical and statistical metrics derived from CTG signals. The completeness and consistency of the data indicate that the dataset is in optimal condition for machine learning analysis and modeling without requiring additional data-cleaning procedures.

The dataset comprised 2,126 observations and 21 numerical variables, all complete with no missing values. These features capture FHR characteristics and histogram-based descriptors generated from CTG recordings. Table 6 shows that the distribution of fetal health is uneven. Most instances are classified as Normal (1.0), while there are far fewer samples in the Suspect (2.0) and Pathological (3.0) classifications. This pronounced class imbalance has methodological implications, requiring careful interpretation of classification performance, particularly for minority classes, and underscores the importance of model evaluation strategies that account for unequal class representation.

2.6.4 Model training without feature selection as baseline

Feature selection-free model training was performed to establish a baseline as a reference for objectively evaluating the impact of the optimization process. At this stage, all available features were included in the training phase, allowing each algorithm to learn from the complete feature set without any prior dimensionality reduction. The performance of this baseline serves as a reference point to determine whether subsequent feature selection enhances model accuracy, computational efficiency, or generalization ability in classifying fetal health conditions (Table 7).

**Table 6.** Target class distribution

| Class (fetal_health) | Number of Samples |
|---|---|
| 1.0 (Normal) | 1655 |
| 2.0 (Suspect) | 295 |
| 3.0 (Pathological) | 176 |

**Table 7.** Model training without feature selection as baseline

| Algorithm | Library/Function | Configuration | Evaluation Metrics |
|---|---|---|---|
| Logistic Regression | Logistic Regression (max_iter=1000) (scikit-learn) | Maximum iterations=1000 | Accuracy, Precision, Recall, F1-score |
| Support Vector Machine | Support Vector Classifier () (scikit-learn) | Default parameters | Accuracy, Precision, Recall, F1-score |
| Random Forest | Random Forest Classifier (random_state=42) (scikit-learn) | Random state=42 | Accuracy, Precision, Recall, F1-score |

**Table 8.** Features of the Random Forest model algorithm

| Method | Technique | Number of Selected Features |
|---|---|---|
| Filter Method | Analysis of Variance F-test (SelectKBest) | 10 |
| Wrapper Method | Recursive Feature Elimination (RFE) with Logistic Regression | 10 |

**Table 9.** Model training with feature selection algorithms

| Model | Accuracy | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|---|
| Logistic Regression | 0.8850 | 0.81 | 0.76 | 0.78 | 426 |
| Support Vector Machine | 0.8944 | 0.82 | 0.75 | 0.78 | 426 |
| Random Forest | 0.9272 | 0.88 | 0.84 | 0.86 | 426 |
| Macro Average | 0.9022 | 0.84 | 0.78 | 0.81 | 425 |
| Weighted Average | 0.9000 | 0.89 | 0.90 | 0.90 | 426 |

The basic procedure consists of three stages: (1) Training the algorithm with the scaled training dataset (X_train_scaled, y_train); (2) Utilizing the scaled test dataset (X_test_scaled) for predictions; and (3) Evaluating performance through various metrics, including accuracy and a comprehensive classification report (precision, recall, and F1-score for each class). Analysis of Random Forest, SVMs, and Logistic Regression was performed under the same conditions to ensure objectivity. These initial results form the basis for evaluating feature selection in the next optimization phase.

2.6.5 Feature selection using Analysis of Variance F-test and Recursive Feature Elimination

During feature selection implementation, the ten most relevant features from the 21 characteristics in the CTG data were identified using the ANOVA F-test and RFE methods. The objective was to simplify the model, lower the risk of overfitting, and improve accuracy and computational

efficiency in FHR classification (Table 8).

This software employed the ANOVA F-test and RFE to choose the top 10 features from the scaled data. RFE uses the Logistic Regression model to systematically exclude less significant characteristics, whereas the ANOVA F-test selects features based on the robustness of their statistical association with the objective. The selected features aimed at improving the efficiency and accuracy of the classification model, as detailed in the findings.

2.6.6 Model training with feature selection

Feature selection training using RFE results was performed by training classification models such as Logistic Regression, SVM, and Random Forest only on the top 10 features. Classification of FHR problems should be carried out with consideration for improving accuracy while reducing model complexity and the risk of overfitting (Table 9).

Significance of the 0.5% accuracy gain: The observed improvement in Random Forest classification from 92.72% to 93.19% is relatively small, and because the experiment was conducted using a single stratified train-test split, this improvement cannot be interpreted as statistically significant. To make such a claim, repeated k-fold cross-validation combined with a statistical test such as the McNemar test or a paired t-test on repeated runs is necessary.

## 3. RESULTS AND DISCUSSION

### 3.1 Results

3.1.1 Model evaluation without feature selection

To provide an initial assessment of model effectiveness, all attributes were used without any feature selection or filtering at this stage. After training on the scaled data, the three models—Random Forest, SVM, and Logistic Regression—were evaluated using accuracy and other metrics, including recall, F1-score, and precision. These assessment outcomes serve as a baseline to determine whether feature selection can substantially improve model performance (Table 10).

The application evaluated three machine learning models: Random Forest, SVMs, and Logistic Regression, which were trained on a dataset related to fetal health without feature selection. After evaluating each model on the test data, performance measures including accuracy, precision, recall, and F1-score were generated. SVM ranked second with an accuracy of 89.44%, followed by Logistic Regression with 88.50%. Random Forest emerged as the best model, achieving an accuracy of 92.72% and the highest F1-score on both the macro and weighted scales. All models excelled in classifying the dominant class (class 1.0), but performance in minority classes (classes 2.0 and 3.0) tended to be lower, indicating the challenges in classifying unbalanced data. This evaluation serves as an initial reference before feature selection to optimize model performance.

3.1.2 Model evaluation with 10 selected features (Recursive Feature Elimination)

The performance of the classification model was evaluated using a model evaluation with 10 selected features (RFE) after feature reduction, by training and evaluating the model using only the most essential characteristics. This evaluation included calculating accuracy, precision, recall, and F1-score metrics to describe the effectiveness of the model in identifying each category of FHR abnormalities (Table 11).

3.1.3 Visualization of feature importance in Random Forest

Visualization of the Importance of Random Forest Features illustrates the contribution of each feature to the accuracy of the model in classifying FHR conditions by providing an important score obtained from its weight in the ensemble decision tree. The use of horizontal bar charts makes it easier for researchers to identify the most influential features, thereby supporting more efficient feature selection, model simplification, and improved prediction performance.

**Table 10.** Model evaluation without feature selection algorithms

| Model | Accuracy | Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|---|---|
| Logistic Regression | 0.8850 | 1.0 | 0.94 | 0.95 | 0.94 | 332 |
| | | 2.0 | 0.61 | 0.68 | 0.64 | 59 |
| | | 3.0 | 0.88 | 0.66 | 0.75 | 35 |
| | | Macro Average | 0.81 | 0.76 | 0.78 | 426 |
| | | Weighted Average | 0.89 | 0.88 | 0.89 | 426 |
| Support Vector Machine | 0.8944 | 1.0 | 0.93 | 0.97 | 0.95 | 332 |
| | | 2.0 | 0.67 | 0.58 | 0.62 | 59 |
| | | 3.0 | 0.86 | 0.71 | 0.78 | 35 |
| | | Macro Average | 0.82 | 0.75 | 0.78 | 426 |
| | | Weighted Average | 0.89 | 0.89 | 0.89 | 426 |
| Random Forest | 0.9272 | 1.0 | 0.94 | 0.98 | 0.96 | 332 |
| | | 2.0 | 0.85 | 0.68 | 0.75 | 59 |
| | | 3.0 | 0.86 | 0.86 | 0.86 | 35 |
| | | Macro Average | 0.88 | 0.84 | 0.86 | 426 |
| | | Weighted Average | 0.92 | 0.93 | 0.92 | 426 |

**Table 11.** Model evaluation with 10 selected features (Recursive Feature Elimination) algorithm

| Model | Accuracy | Precision (Macro Average) | Recall (Macro Average) | F1-Score (Macro Average) | Weighted Average F1-Score |
|---|---|---|---|---|---|
| Logistic Regression | 0.8967 | 0.82 | 0.78 | 0.80 | 0.90 |
| Support Vector Machine | 0.8944 | 0.81 | 0.77 | 0.79 | 0.89 |
| Random Forest | 0.9319 | 0.90 | 0.84 | 0.87 | 0.93 |

Notes: Results are based on the classification report after applying Recursive Feature Elimination (RFE) with 10 selected features. The macro average represents the unweighted mean across all classes. The weighted average adjusts the metrics according to the class distribution.

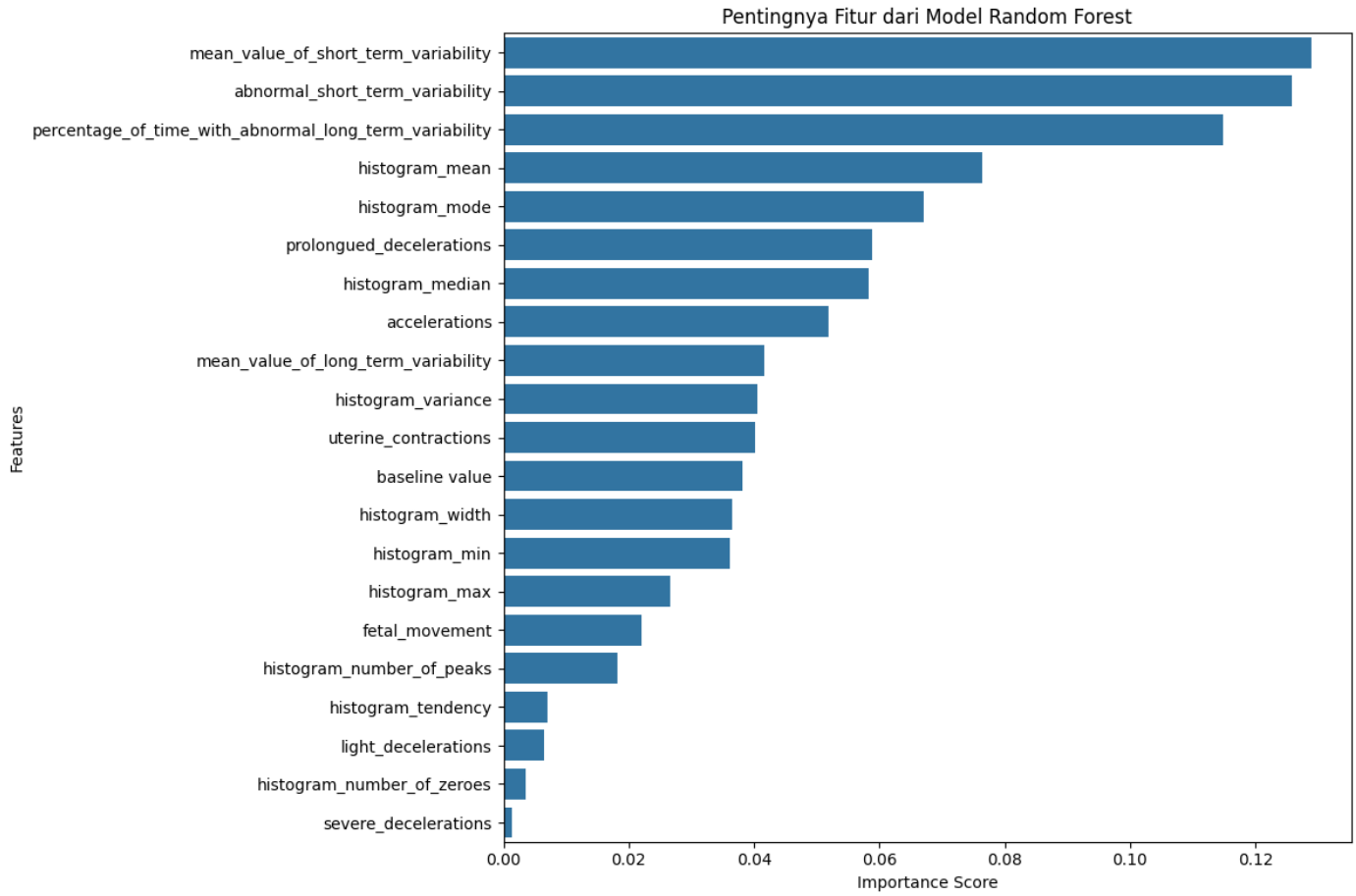


**Figure 2.** Random Forest model feature
Note: Generated by the authors using Google Colab.

Figure 2 illustrates the feature importance ranking derived from the Random Forest model for fetal health classification based on FHR medical data.

Characteristics such as abnormal short-term variability, average short-term variability values, and time fractions exhibiting anomalous long-term variability, due to their significant relevance ratings, are the most critical variables for generating predictions. Considerable decreases and many zeros, however, have little effect on the model's results. During the feature selection phase, researchers may utilize this information to focus on the most relevant characteristics and improve model accuracy and efficiency.

3.1.4 Confusion matrix visualization for the best model

To evaluate the accuracy and inaccuracy of models in identifying FHR conditions, confusion matrix visualization was used to demonstrate the number of accurate and inaccurate predictions made by the best model.

When the model's output contains actual confusion matrix values, those values can be entered into Table 12. The confusion matrix of the best Random Forest model trained with the given features can be viewed in this software to evaluate the prediction accuracy of three fetal condition classes through a heat map.

Figure 3 illustrates the confusion matrix of the Random Forest model, equipped with feature selection, which aims to evaluate the accuracy of grouping fetal conditions into three categories: normal, suspicious, and unknown/pathological. This model accurately classified 326 instances as normal, but there were some classification errors, such as 16 suspicious cases predicted as normal and five normal cases predicted as suspicious. Meanwhile, the pathological class was classified relatively well with 30 correct predictions from the total data, although there were a few errors. This matrix shows that the model is highly accurate for the normal class, relatively effective for the pathological class, and still needs improvement in the classification of the suspicious class.

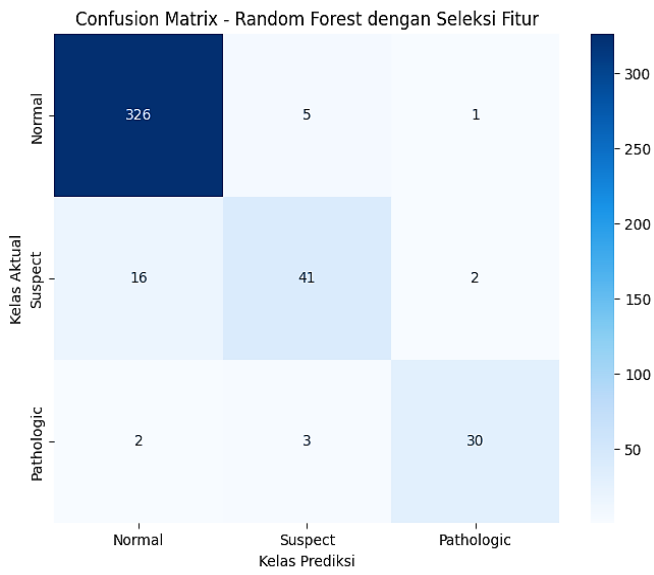


**Figure 3.** Random Forest model feature
Note: Generated by the authors using Google Colab.

**Table 12.** Confusion matrix visualization for the best model algorithm

| Actual/Predicted | Normal | Suspect | Pathological |
|---|---|---|---|
| Normal | TN | $FP_1$ | $FP_2$ |
| Suspect | $FN_1$ | TS | $FP_3$ |
| Pathological | $FN_2$ | $FN_3$ | TP |

Notes: TN = True Negative (Normal correctly predicted as Normal); TS = True Suspect (Suspect correctly predicted as Suspect); TP = True Pathological (Pathological correctly predicted as Pathological); and FN = False Negative, FP = False Positive.

### 3.2 Discussion

The results demonstrate that feature selection strongly affects machine learning performance in CTG-based fetal health categorization. When all 21 characteristics were used in the baseline findings, Random Forest exhibited the best accuracy (92.72%) and the highest F1-score (0.86), while SVM and Logistic Regression performed worse, especially for the minority classes (Suspect and Pathological), which is consistent with previous CTG-related studies on class imbalance challenges. This reinforces that class imbalance and overlapping feature distributions remain major challenges in CTG classification. Following the application of feature selection using ANOVA F-test and RFE, model performance improved. Random Forest showed the most notable gains, with accuracy increasing to 93.19% and macro F1-score rising to 0.87. These improvements indicate that the selected 10 features offer a more informative representation of fetal health, lowering noise and redundancy while improving class-wise balance, in accordance with findings reported in previous feature-selection studies in biomedical classification [26-28]. The most influential predictors identified, such as short-term and long-term variability indicators, correspond with established clinical determinants of fetal well-being, supporting the interpretability of the selected feature subset.

The models are also more efficient and fit for real-time or near real-time clinical applications due to reduced computing complexity resulting from feature selection, agreeing with previous findings on dimensionality reduction in biomedical machine learning [29]. Nevertheless, the Suspect class remains difficult to classify accurately, reflecting its inherent clinical ambiguity and overlap with the Normal and Pathological categories, as reported in previous CTG studies [30]. Overall, the findings indicate that combining feature selection with Random Forest yields better performance, improved interpretability, and greater computational efficiency, supporting its potential use in clinical decision-support systems for CTG interpretation, which aligns with established evidence regarding the clinical applicability of CTG-based machine learning models [3, 5].

## 4. CONCLUSION

This study demonstrates that feature selection can enhance model efficiency and interpretability in CTG-based fetal health classification, although its impact varies across algorithms. Random Forest shows the most substantial improvement, with accuracy improving from 92.72% to 93.19% and the macro F1-score rising from 86% to 87%, indicating that reducing redundant features enhanced its predictive capability. In contrast, Logistic Regression and SVM exhibited a slight decrease in performance after feature reduction, suggesting that feature selection is model-dependent rather than universally advantageous. Overall, feature selection remains a useful approach for simplifying models and improving clinical applicability, particularly when combined with robust algorithms such as Random Forest.


## ACKNOWLEDGMENT

For the encouragement, financial support, and assistance in conducting the research necessary to complete my doctoral dissertation, I would like to express my deepest gratitude to the National Research and Innovation Agency (BRIN), Universitas Diponegoro's Doctoral Program in Information Systems, and my Promoter as well as Co-Promotors. Their dedication to moving the research forward and completing the output efforts was crucial to the success of this research.


## REFERENCES


[1] Turner, J.M., Mitchell, M.D., Kumar, S.S. (2020). The physiology of intrapartum fetal compromise at term. American Journal of Obstetrics and Gynecology, 222(1): 17-26. https://doi.org/10.1016/j.ajog.2019.07.032

[2] Ben M'Barek, I., Ben M'Barek, B., Jauvion, G., Holmström, E., Agman, A., Merrer, J., Ceccaldi, P.F. (2024). Large-scale analysis of interobserver agreement and reliability in cardiotocography interpretation during labor using an online tool. BMC Pregnancy and Childbirth, 24(1): 136. https://doi.org/10.1186/s12884-024-06322-4

[3] Salini, Y., Mohanty, S.N., Ramesh, J.V.N., Yang, M., Chalapathi, M.M.V. (2024). Cardiotocography data analysis for fetal health classification using machine learning models. IEEE Access, 12: 26005-26022. https://doi.org/10.1109/ACCESS.2024.3364755

[4] Bai, J., Wang, W., Kang, X., Zhou, B., et al. (2024). Machine learning-based prediction of fetal health using cardiotocography. ESS Open Arch Eprints, 560: 56010945. https://doi.org/10.22541/au.172146395.56010945/v1

[5] Alkurdi, A., Abdulazeez, A.M. (2024). Comprehensive classification of fetal health using cardiotocogram data based on machine learning. The Indonesian Journal of Computer Science, 13(1): 277-300. https://doi.org/10.33022/ijcs.v13i1.3718

[6] Olayemi, O.C., Olasehinde, O.O. (2024). Machine learning prediction of fetal health status from cardiotocography examination in developing healthcare contexts. Journal of Computer Science Research, 6(1): 43-53. https://doi.org/10.30564/jcsr.v6i1.6242

[7] Ali, M.Z., Abdullah, A., Zaki, A.M., Rizk, F.H., Eid, M.M., El-Kenway, E.M. (2024). Advances and challenges in feature selection methods: A comprehensive review. Journal of Artificial Intelligence and Metaheuristics, 7(1): 67-77. https://doi.org/10.54216/JAIM.070105

[8] Theng, D., Bhoyar, K.K. (2024). Feature selection techniques for machine learning: A survey of more than two decades of research. Knowledge and Information Systems, 66(3): 1575-1637. https://doi.org/10.1007/s10115-023-02010-5

[9] Kumari, S., Prabha, C., Karim, A., Hassan, M.M., Azam,

S. (2024). A comprehensive investigation of anomaly detection methods in deep learning and machine learning: 2019–2023. IET Information Security, 2024(1): 8821891. https://doi.org/10.1049/2024/8821891

[10] Ozcanli, A.K., Yaprakdal, F., Baysal, M. (2020). Deep learning methods and applications for electrical power systems: A comprehensive review. International Journal of Energy Research, 44(9): 7136-7157. https://doi.org/10.1002/er.5331

[11] Mesbah, M., Khlif, M.S., Layeghy, S., East, C.E., et al. (2021). Automatic fetal movement recognition from multi-channel accelerometry data. Computer Methods and Programs in Biomedicine, 210: 106377. https://doi.org/10.1016/j.cmpb.2021.106377

[12] Ullah, A., Ul Haq, Q.M., Ullah, Z., Frnda, J., Shahid Anwar, M. (2025). AI-driven fetal distress monitoring SDN-IoMT networks. PloS One, 20(7): e0328099. https://doi.org/10.1371/journal.pone.0328099

[13] Chen, C., Xie, W., Cai, Z., Lu, Y. (2023). Deep learning for cardiotocography analysis: Challenges and promising advances. In International Conference on Intelligent Computing, pp. 354-366. https://doi.org/10.1007/978-981-99-4742-3_29

[14] Ahmed, S.S., Mahmoud, N.M. (2025). Early detection of fetal health status based on cardiotocography using artificial intelligence. Neural Computing and Applications, 37(21): 16753-16779. https://doi.org/10.1007/s00521-025-11343-x

[15] Chiou, N., Young-Lin, N., Kelly, C., Cattiau, J., et al. (2025). Development and evaluation of deep learning models for cardiotocography interpretation. npj Women's Health, 3(1): 21. https://doi.org/10.1038/s44294-025-00068-w

[16] Jahan, S., Nowsheen, F., Antik, M. M., Rahman, M.S., Kaiser, M.S., Hosen, A.S., Ra, I.H. (2023). AI-based epileptic seizure detection and prediction in internet of healthcare things: A systematic review. IEEE Access, 11: 30690-30725. https://doi.org/10.1109/ACCESS.2023.3251105

[17] Tronstad, C., Amini, M., Bach, D.R., Martinsen, Ø.G. (2022). Current trends and opportunities in the methodology of electrodermal activity measurement. Physiological Measurement, 43(2): 02TR01. https://doi.org/10.1088/1361-6579/ac5007

[18] Aeberhard, J.L., Radan, A.P., Delgado-Gonzalo, R., Strahm, K.M., Sigurthorsdottir, H.B., Schneider, S., Surbek, D. (2023). Artificial intelligence and machine learning in cardiotocography: A scoping review. European Journal of Obstetrics & Gynecology and Reproductive Biology, 281: 54-62. https://doi.org/10.1016/j.ejogrb.2022.12.008

[19] O'Sullivan, M.E., Considine, E.C., O'Riordan, M., Marnane, W.P., Rennie, J.M., Boylan, G.B. (2021). Challenges of developing robust AI for intrapartum fetal heart rate monitoring. Frontiers in Artificial Intelligence, 4: 765210. https://doi.org/10.3389/frai.2021.765210

[20] Klumpp, M., Hintze, M., Immonen, M., Ródenas-Rigla, F., et al. (2021). Artificial intelligence for hospital health care: Application cases and answers to challenges in European hospitals. Healthcare, 9(8): 961. https://doi.org/10.3390/healthcare9080961

[21] Liu, L., Pu, Y., Fan, J., Yan, Y., et al. (2024). Wearable sensors, data processing, and artificial intelligence in pregnancy monitoring: A review. Sensors, 24(19): 6426. https://doi.org/10.3390/s24196426

[22] Quin, C. (2025). A path to improved fetal cardiovascular health outcomes using machine learning. In 2025 IEEE International Conference on AI and Data Analytics (ICAD), Medford, MA, USA, pp. 1-8. https://doi.org/10.1109/ICAD65464.2025.11114073

[23] Togunwa, T.O., Babatunde, A.O., Abdullah, K.U.R. (2023). Deep hybrid model for maternal health risk classification in pregnancy: Synergy of ANN and random forest. Frontiers in Artificial Intelligence, 6: 1213436. https://doi.org/10.3389/frai.2023.1213436

[24] Mehbodniya, A., Lazar, A.J.P., Webber, J., Sharma, D.K., et al. (2022). Fetal health classification from cardiotocographic data using machine learning. Expert Systems, 39(6): e12899. https://doi.org/10.1111/exsy.12899

[25] Sufriyana, H., Husnayain, A., Chen, Y.L., Kuo, C.Y., et al. (2020). Comparison of multivariable logistic regression and other machine learning algorithms for prognostic prediction studies in pregnancy care: Systematic review and meta-analysis. JMIR Medical Informatics, 8(11): e16503. https://doi.org/10.2196/16503

[26] Sultan, N., Hasan, M., Wahid, M.F., Saha, H., Habib, A. (2023). Cesarean section classification using machine learning with feature selection, data balancing, and explainability. IEEE Access, 11: 84487-84499. https://doi.org/10.1109/ACCESS.2023.3303342

[27] Priyadharshni, S., Ravi, V. (2025). Dynamic graph-based quantum feature selection for accurate fetal plane classification in ultrasound imaging. Scientific Reports, 15(1): 41743. https://doi.org/10.1038/s41598-025-26835-y

[28] Särestöniemi, M., Taparugssanagorn, A., Iinatti, J., Myllylä, T. (2024). Detection of intestinal tumors outside the visibility of capsule endoscopy camera utilizing radio signal recognition. Nordic Conference on Digital Health and Wireless Solutions, 2084: 426-440. https://doi.org/10.1007/978-3-031-59091-7_28

[29] Alhassan, A.M., Zainon, W.M.N.W. (2021). Review of feature selection, dimensionality reduction and classification for chronic disease diagnosis. IEEE Access, 9: 87310-87317. https://doi.org/10.1109/ACCESS.2021.3088613

[30] Islam, M.M., Rokunojjaman, M., Amin, A., Akhtar, M.N., Sarker, I.H. (2022). Diagnosis and classification of fetal health based on CTG data using machine learning techniques. In Machine Intelligence and Emerging Technologies, pp. 3-16. https://doi.org/10.1007/978-3-031-34622-4_1