Fetal Health Risk Classification Using Important Feature Selection and Cart Model on Cardiotocography Data

Ahmad Ilham[1,2], Thahta Ardhika Prabu Nagara[1,2], Mudyawati Kamaruddin[3,4], Laelatul Khikmah[2,5], Teddy Mantoro[6]
[1]Department of Informatics, Universitas Muhammadiyah Semarang, Semarang, Indonesia
[2]Intelligent Data Science Research Group, Universitas Muhammadiyah Semarang, Semarang, Indonesia
[3]Postgraduate Program of Medical Laboratory Science, Universitas Muhammadiyah Semarang, Semarang, Indonesia
[4]Interdisciplinary Research Laboratory for Experimental Plasma Medicine, Universitas Muhammadiyah Semarang, Semarang, Indonesia
[5]Department of Statistics, Institut Teknologi Statistika dan Bisnis Muhammadiyah Semarang, Semarang, Indonesia
[6]School of Computer Science, Nusa Putra University, West Java, Indonesia
E-mail: ahmadilham@unimus.ac.id, mudyawati@unimus.ac.id, c2c023166@student.unimus.ac.id, laelatul.khikmah@itesa.ac.id, teddy@ieee.org

Keywords: Cardiotocography, fetal health risk classification, important feature selection, cart, machine learning in healthcare

Received: January 22, 2024

Fetal mortality and newborn health issues require urgent attention because of high maternal and infant mortality rates during labor, highlighting the critical need for accurate fetal condition monitoring to reduce complications. This study proposes the development of a fetal health risk classification model based on Important Feature Selection (IFS) and a Classification and Regression Tree (CART) using cardiotocography (CTG) data from the UCI Machine Learning Repository. The IFS method was used to select the most relevant features, reduce model complexity, and increase generalization to prevent overfitting. The IFS-CART model was tested with 10-fold cross-validation and showed an accuracy of 94.50%, superior to the conventional CART, which only reached 93.83%. In addition, the average values of the True Positive Rate (TPR) and True Negative Rate (TNR) also increased, indicating that this model is effective in distinguishing normal, suspected, and pathological fetal conditions. Evaluation using the area under the curve receiver operating characteristic (AUC-ROC) showed that the model had high performance in detecting at-risk conditions, with an AUC of 0.981 for the "suspect" class. This finding confirmed that IFS-CART is not only accurate but also has high interpretability, making it easy for medical personnel to use for clinical decision support. The results of this study show that IFS-CART can serve as a reliable decision support system for real-time fetal health monitoring. Further implementation is expected to improve diagnostic accuracy and prevent complications during pregnancy and labor

Povzetek: Opisan je nov model za klasifikacijo tveganja za zdravje ploda z uporabo izbire pomembnih značilnosti in metode CART na podatkih kardiotokografije.

1 Introduction

Fetal hypoxia occurs when oxygen supply to the fetus is insufficient during labor. This condition can cause severe consequences, including intrapartum stillbirth, asphyxia, neonatal encephalopathy, neonatal death, and neurodevelopmental impairment [1]. The incidence of fetal hypoxia in European hospitals ranges from 0.06% to 2.8% [2]. Globally, intrapartum fetal hypoxia results in approximately 1.3 million stillbirths, 0.9 million neonatal deaths, and 0.6 to 1 million cases of long-term disability from neonatal hypoxic-ischemic encephalopathy annually [3]. These statistics underscore the urgency of addressing this issue to prevent further cases. Labor naturally induces a degree of hypoxic stress as uterine contractions (UC) potentially impair maternal placental perfusion, compromising fetal oxygen delivery. Clinicians face the challenge of identifying the small number of cases where physiological protective mechanisms fail to compensate for labor-induced hypoxic stress, leading to significant cerebral injury [4]. Effective fetal monitoring during labor is crucial to prevent the devastating effects of fetal hypoxia. However, it must also be sufficiently discriminatory to minimize unnecessary iatrogenic interventions, such as caesarean sections, which carry their own risks to both mother and baby [5].

Cardiotocography (CTG) is a commonly used screening tool for monitoring fetal conditions, such as fetal heart rate (FHR) and uterine contractions (UC). Unfortunately, the interpretation of CTG is subjective and depends on the clinician's experience, which often leads to erroneous diagnoses and unnecessary medical interventions [6]. This has led to an increased interest in machine learning to provide a more

objective and accurate analysis. A promising approach is the classification and regression tree (CART), which has a structure that is easy to interpret [7] .

However, CART is prone to overfitting, particularly when all features in the dataset are used without selection. To overcome this drawback, this study proposes the integration of an important feature selection (IFS) method with CART. IFS helps select the most relevant features, reduce model complexity, and increase robustness to overfitting, thereby improving classification accuracy. In addition, IFS increases the resistance to overfitting, thereby improving classification accuracy.

The main contribution of this research is the development of an IFS-CART model that is not only accurate but also easy to interpret. The CART model is not only accurate but also easy to interpret, providing practical benefits for medical personnel performing risk assessments without the need for an in-depth understanding of computational models. This research also demonstrates how the combination of IFS and CART can overcome the limitations of previous state-of-the-art models, such as SVM and Naive Bayes, which, although accurate, lack high interpretability and require complex computations.

The remainder of this paper is organized as follows. Section 2 presents a comprehensive review of related studies, highlighting previous research and machine learning techniques. Section 3 outlines the research methodology, including the dataset, preprocessing steps, model development, and validation procedures, to ensure the reliability of the results. Section 4 presents the experimental findings with an in-depth discussion, provides critical insights, and presents the results in the context of existing literature. Finally, Section 5 draws conclusions from the study and proposes recommendations for future research to strengthen the findings presented.

2 Related works

Various machine learning techniques have been applied to analyze cardiotocography (CTG) data to improve the accuracy and reliability of fetal health condition diagnosis. Innovations in these techniques are evolving along with the increasing need for models that are not only accurate but also interpretative and easy to use in a clinical context.

Sahin and Subasi [8] initially investigated the integration of a support vector machine (SVM) with an empirical mode decomposition (EMD) metaheuristic technique to enhance the accuracy of fetal condition classification. The model attained an accuracy of 86% by utilizing a limited dataset of 90 samples. Nevertheless, they encounter challenges regarding computational complexity and constraints in the interpretation of the results. In the same study, the authors evaluated the naive bayes (NB) model, which is recognized for its simplicity and computational efficiency, achieving an accuracy of 96.77%. However, NB performance frequently deteriorates when applied to datasets containing irrelevant or redundant features, indicating the necessity for more effective feature selection in fetal health predictive models.

In 2016, Yılmaz [9] introduced and compared three neural network architectures: multilayer perceptron neural network (MLPNN), probabilistic neural network (PNN), and general regression neural network (GRNN). Each of these architectures was designed to address specific challenges in fetal health data analysis, especially in detecting complex nonlinear patterns in CTG data. The MLPNN model demonstrated the ability to detect nonlinear relationships with an accuracy of 90.35%, recall of 78.71%, specificity of 90.50%, precision of 84.44%, and F1-score of 81.47%. However, MLPNN require a long training time and are prone to overfitting, especially when the data features are large or poorly structured. In contrast, PNNs excel at classification speed and are well suited for processing biomedical data, such as physiological signals on CTG, with an accuracy of 92.15%, recall of 82.82%, specificity of 92.24%, precision of 87.63%, and F1-score of 85.16%. However, PNN require large memory resources, which is an obstacle in large-scale clinical applications. In contrast, the GRNN models offer continuous prediction capabilities with 91.86% accuracy, 83.92% recall, 92.62% specificity, 85.81% precision, and an 84.85% F1-score. However, the proposed GRNN model is less optimal for distinct classifications and requires careful parameterization to achieve the best performance. These limitations suggest that although neural networks are capable of good performance, their computational complexity and susceptibility to overfitting limit their applicability in clinical contexts where quick and intuitive interpretation is required.

More recently, Chuatak et al. [7] applied a classification and regression tree (CART) to a CTG dataset with 2,126 samples from the UCI Repository. The CART model is known for its easy-to-understand tree structure and ability to produce clear classification rules for medical personnel without requiring a deep understanding of the computational models. In this study, CART achieved an accuracy of 93.65%. However, the main limitation of CART is its susceptibility to overfitting when all dataset features are used without selection. Thus, additional techniques, such as boosting or bagging, are required to achieve optimal performance, especially on large datasets with many features.

In 2024, Shalini et al. [10] extended the exploration of machine learning models on CTG data by comparing various models, including Linear Regression, Random Forest, K-Nearest Neighbors (KNN), Gradient Boosting Classifier, Decision Tree, and Support Vector Classifier. Based on the same dataset of 2,126 samples, the Random Forest and Decision Tree models achieved 93% and 85% accuracy, respectively. The Linear Regression, KNN, and gradient boosting classifier models demonstrated varying performances, with accuracies ranging from 89% to 90%, whereas the support vector classifier obtained a lower accuracy of 81%. This study showed that although some of these models provide reasonably good results, they have limita-

tions in both accuracy and interpretability, making them less than ideal in clinical contexts that require quick and accurate decisions.

To facilitate a clear comparison, Table 2 summarizes the performance of the various machine learning models on the CTG data.

Through these studies, it can be seen that neural network-based models, such as MLPNN, PNN, and GRNN, perform quite well in the classification and prediction of fetal conditions. However, their high computational requirements and interpretation complexity are obstacles in clinical applications that require fast and reliable decisions. In addition, traditional models such as SVM and NB demonstrate fast performance; however, they have limitations when handling data with redundant or overlapping features, which often results in decreased accuracy under certain conditions.

This study selected CART because of its ability to generate classification rules that are clear and easy to interpret by medical personnel, without the need for in-depth knowledge of computational models. However, without proper feature selection, CART is prone to overfitting when applied to large datasets with many features. To address this issue, this study proposes the integration of an essential feature selection (IFS) method with CART. IFS allows the model to retain only the most relevant features, thereby reducing complexity, increasing generalizability, and ensuring reliable results in clinical practice.

3 Methodology

3.1 Dataset

In this study we used the cardiotocography (CTG) dataset from the UCI Machine Learning Repository [11]. The dataset comprises 2126 entries with information on third-trimester pregnant women. It includes 21 features and one feature class (NSP) used to determine the fetal heart rate (FHR) and uterine contractions (UC) in the CTG dataset. A comprehensive overview of the CTG dataset used in this study is presented in Table 3.1.

3.2 Decision tree-based learning

Decision tree-based learning represents an effective approach for addressing regression and classification problems. Support Vector Machines (SVM) and Decision Trees (DT) are among the machine learning techniques utilized for these tasks [12]. DT offers distinct advantages, including independence from predictor parameter distribution assumptions and computational efficiency. Furthermore, decision trees can effectively handle missing data [13].

A hypothetical study employs a vector of two independent variables (X_1, X_2) to construct a tree. Figure 1 illustrates a model comprising four internal nodes and five leaves, used to estimate the target parameter. The tree's growth progresses from top to bottom, with initial comparisons made between X_1 and the threshold value T_1. Subsequent steps depend on whether X_1 exceeds T_1, determining the branch selection.

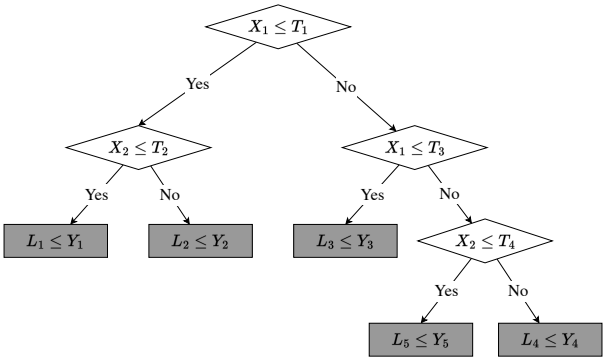

Figure 1: A typical decision trees

Various techniques exist for creating decision tree-based regressor classification models, including fuzzy ID3 [14], ID3 [15], C4.5 [16], and CART [17]. DT offers superior interpretability and visualization compared to black-box models like artificial neural networks, facilitating easier comprehension of results in regression and classification problems.

However, DT presents certain limitations. In regression analysis, it can only estimate continuous values. Additionally, the tree structure may become complex due to numerous branches in classification and regression tasks. The modeling data significantly influences the DT structure, potentially leading to inconsistent results and structural variations across datasets.

The CART model, widely adopted for its efficiency in processing qualitative and quantitative data [17], employs recursive binary splitting. This study utilizes the CART algorithm with the Gini splitting rule. The permutation approach and mini measure identify relevant dataset characteristics.

The percentage of samples in category k (0 or 1) at a node is expressed in Equation (1):

$$p_k = \frac{n_k}{n} \tag{1}$$

where p represents the data class probability at node τ. The mini impurity $i(\tau)$ is calculated using Equation (2).

$$i(\tau) = 1 - \sum_k p_k^2 \tag{2}$$

For a two-class problem in (3):

$$i(\tau) = 1 - p_1^2 - p_0^2 \tag{3}$$

The Gini impurity change when nodes split into subnodes τ_1 and τ_2 is formulated in Equation (4):

Table 1: Performances of several machine learning models on CTG data based on previous studies

Studies	Models	Accuracy (%)	Precision (%)	Recall (%)	Dataset Size
[8]	SVM + EMD	86.0	-	-	90
[8]	NB	96.77	-	-	90
[9]	MLPNN	90.35	84.44	78.71	2,126
[9]	PNN	92.15	87.63	82.82	2,126
[9]	GRNN	91.86	85.81	83.92	2,126
[7]	CART	93.65	-	-	2,126
[10]	Linear Regression	89	82.67	77	2,126
[10]	Random Forest	93	87.33	85.67	2,126
[10]	KNN	90	82.67	76.67	2,126
[10]	Gradient Boosting Classifier	90	80.67	77.67	2,126
[10]	Decision Tree	85	85	85	2,126
[10]	Support Vector Classifier	81	80.67	79.67	2,126

$$\Delta i(\tau) = i(\tau) - p_l i(\tau_1) - p_r i(\tau_2) \tag{4}$$

The algorithm tracks and aggregates each node's $i(\tau)$ drop. Using a single CART Gini Importance is formulated in (5):

$$I_G(\theta) = \sum_{\tau} \sum_{T} \Delta i_\theta(\tau, T) \tag{5}$$

The Gini importance identifies relevant features for the classification objective function, indicating feature frequency as a separator and its discriminative value. In equation (5), T represents the number of trees in the model, and $I_G(\theta)$ denotes the corresponding Gini importance.

3.3 Proposed model

Figure 2 illustrates the block diagram of our proposed model. We initiated the process by selecting the CTG dataset, comprising 2216 original samples. Subsequently, we divided and categorized the data to enhance system transparency. We implemented decision trees and rule-based classifiers using 10-fold cross-validation for modeling categorization. The folds served as training data to develop the model. We utilized the remaining data as a test set to validate the resulting model and determine performance metrics, such as accuracy.

3.3.1 Data preparation

Each feature in the cardiotocography (CTG) dataset exhibits distinct value ranges and units, such as fetal heart rate (FHR) and uterine contractions (UC). This scale disparity among features can lead to imbalances. Classification algorithms like CART may overemphasize features with larger values, compromising model performance. Consequently, normalization becomes essential to ensure equal feature contribution in machine learning processes and prevent model bias.

This study employs zero-mean normalization to transform numerical features. This technique ensures a mean of zero and uniform variance for each feature, enhancing data distribution consistency. Zero-mean normalization proves particularly valuable when handling high-variance data, enabling optimal model performance unaffected by feature scale differences.

The zero-mean normalization equation is expressed as in Equation (6):

$$x' = \frac{x - \mu}{\sigma} \tag{6}$$

where: x' represents the normalized feature value, x denotes the original feature value, μ signifies the average feature value, and σ indicates the feature's standard deviation.

This normalization step is crucial for maintaining data integrity and improving overall model accuracy in the analysis of cardiotocography dataset.

3.3.2 Feature preprocessing and selection

Feature selection using Important Feature Selection (IFS) follows the normalization process. This step is crucial in analyzing medical data like CTGs, which often contain numerous characteristics, not all relevant for fetal health status classification. IFS selects features that significantly contribute to classifying fetal status (normal, suspect, or pathological).

The proposed method employs the Information Gain criterion to assess feature relevance based on its impact in reducing data entropy. Features with minimal contributions are eliminated, retaining only important ones for model training. This approach reduces dataset dimensionality, enhances computational efficiency, and accelerates training. It also minimizes overfitting risk, resulting in a more accurate and interpretable model.

Equation (7) calculates feature significance in IFS using the Information Gain criterion:

Table 2: CTG dataset description

Features	Information
BV	FHR baseline (beats per minute)
AC	FHR of acceleration per second
FM	FHR of fetal movements per second
UC	FHR of uterine contractions per second
LD	FHR of light deceleration per second
SD	FHR of severe deceleration per second
PD	FHR during prolonged deceleration per second
ASTV	Percentage of time with abnormal short-term variability
MSTV	Mean short-term variability
ALTV	Percentage of time with abnormal long-term variability
MLTV	Mean long-term variability
HW	Width of the FHR histogram
HMin	Minimum FHR histogram
HMax	Maximum FHR histogram
HNMax	FHR of histogram peaks
NZ	The FHR of the histogram zeros is given by
HMo	Histogram mode
HMean	Histogram mean
HMed	Histogram median
HV	Histogram variance
HT	Histogram tendency
NSP	Fetal state class code (N=normal; S=suspect; P=pathologic)

$$I_G(S,A) = E(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} \times E(S_v) \quad (7)$$

Where $I_G(S,A)$ represents feature A's information gain relative to dataset S, S_v is the data subset based on a particular value of feature A, and $E(S)$ measures uncertainty or entropy in dataset S.

Information gain quantitatively evaluates each feature's contribution to reducing classification target uncertainty. It provides a direct indication of a feature's ability to improve model class separation by calculating entropy change before and after feature use.

This method was chosen for its ability to explicitly measure individual feature relevance to the classification target, offering advantages over techniques like Principal Component Analysis (PCA). While PCA reduces dataset dimensionality, it transforms features into a new dimensional space without considering specific contributions to the classification target, often losing original feature interpretability.

The Information Gain criterion ensures retained original features directly relate to the classification target, maintaining model result interpretability and clarity. Features with low I_G values are eliminated, while those with high values are retained. This process reduces dataset complexity, improves computational efficiency, and minimizes overfitting risk.

By retaining the most relevant features, the model produces more accurate predictions while maintaining a clear relationship between important features and classification results. This approach is vital in clinical applications and data-driven analyses where interpretability and efficiency are key factors for model effectiveness.

3.3.3 Overfitting control and hyperparameter optimization

IFS-CART implements critical steps to maintain high accuracy while resisting overfitting. These steps include hyperparameter optimization and cross-validation. Overfitting occurs when a model performs exceptionally well on training data but fails to predict new data accurately, leading to decreased performance in real-world conditions. To mitigate this issue, the study optimizes key hyperparameters such as maximum tree depth (`max_depth`), pruning, and minimum samples per leaf. These optimizations reduce model complexity without sacrificing accuracy.

Cost-complexity pruning serves as a primary technique. This method eliminates tree branches that minimally contribute to model performance, maintaining model simplicity and preventing overfitting by reducing tree size. The pruning process is optimized using Equation (8).

$$R_\alpha(T) = R(T) + \alpha \cdot |T| \quad (8)$$

Here, $R_\alpha(T)$ represents the tree's cost after considering its complexity, $R(T)$ denotes the total classification er-

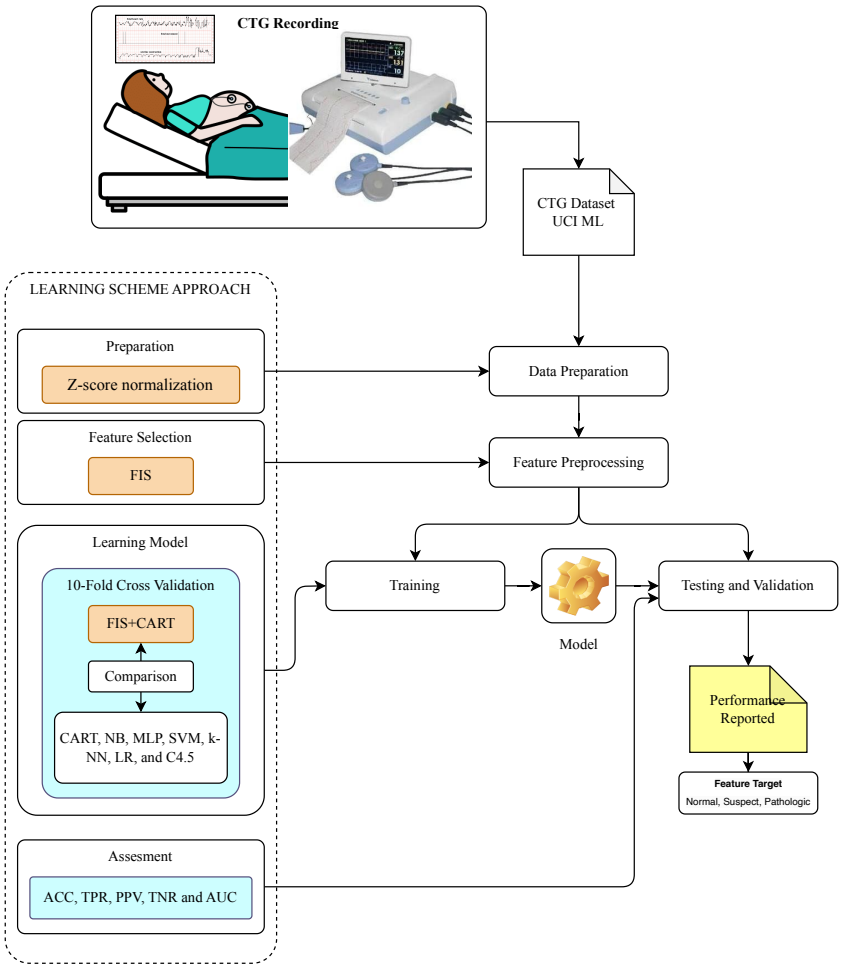

Figure 2: Schematic model of fetal health risk status classification using the CTG dataset

ror, and $|T|$ signifies the number of tree nodes. The parameter α controls complexity; a higher α value results in a smaller, simpler tree. This optimization retains only nodes that significantly contribute to classification results, enhancing model efficiency and interpretability.

3.3.4 Model stability and generalizability via cross-validation

Cross-validation ensures model stability and generalizability. This study employed 10-fold cross-validation to optimize tree structure and maintain performance consistency. The dataset was divided into 10 subsets. Each subset served as test data, while the remaining nine were used for training. This process repeated until all folds were tested.

Cross-validation tests the model across various data configurations, providing a comprehensive performance assessment. It reduces overfitting risk and enhances result reliability. By averaging performance metrics from each fold, the model demonstrates consistent performance independent of specific data.

This validation approach is crucial in medical applications. It increases confidence in the model's ability to function optimally beyond training data, including real clinical scenarios [18, 19, 20, 21, 22].

3.3.5 Evaluation of the implications for clinical applications

To ensure the comprehensive performance of the IFS-CART model, this study used several confusion matrix-based metrics. In addition to accuracy (ACC), metrics such as recall or true positive rate (TPR), precision or positive predictive value (PPV), and specificity or true negative rate (TNR) were evaluated to provide a deeper understanding of the model's ability to classify different classes (normal, suspect, and pathological). In Equations (9), (10), (11), and (12), provideive the appropriate formulas for the ACC, TPR, PPV, and TNR.

$$ACC = \frac{TP + TN}{TP + FP + FN + TN} \times 100 \quad (9)$$

$$TPR = \frac{TP}{TP + FN} \times 100 \quad (10)$$

$$PPV = \frac{TP}{TP + FP} \times 100 \quad (11)$$

$$TNR = \frac{TN}{TN + FP} \times 100 \quad (12)$$

Where TP (True Positive) denotes the number of correct predictions for each target class (normal, suspect, pathological), and FP (False Positive) and FN (False Negative) describe the prediction errors that may affect clinical decisions. Using these metrics, the model is not only measured based on overall accuracy but also assessed in terms of its ability to distinguish high-risk cases (suspect and pathological) from normal cases.

In addition, this study also applied the area under the curve receiver operating characteristic (AUC-ROC) to provide an additional evaluation of the model's ability to distinguish positive and negative classes at various classification thresholds. The AUC-ROC is particularly relevant in the medical context as it can assess how well the model predicts suspect and pathological conditions compared to normal classes. The ROC curve plots the TPR against the false positive rate at various thresholds, and AUC values close to 1.0 indicate that the model has excellent and consistent classification ability.

4 Results and discussion

A computing platform with an Intel Core i5 2.5GHz dual-core CPU, 16 GB of RAM, and the 64-bit operating system macOS Catalina was used for the experiments. KNIME version 4.7.0 produced model performance as the calculation output, including accuracy, recall, and Precision.

4.1 Results

First, we applied the CART model without importance feature selection (IFS) on the CTG dataset. All features in this dataset are used for the optimal classification analysis of the data. The experimental results are presented in Table 4.1. The model produced good ACC (93.83%), and the average values of the TPR (93.83%), PPV (93.83%), and TNR (96.91%) measures also obtained good average values. The results are quite good, but this model still indicates overfitting, which can be seen in some incorrect predictions that should be predicted correctly.

In the second experiment, the importance feature selection (IFS) method was implemented to select important and influential features to tackle overfitting in the CART model. Figure 3 presents the results of analyzing relevant features using IFS obtained from the CTG dataset. We found that the MSTV feature had the greatest effect with an importance value of more than 20.50% when using all features in the CTG dataset. Some features, including NZ and SD, had no impact on the CART model's development, where they were of 0% importance in tree building. The HT feature had the lowest effect (0.95%) compared to the other features. The results show that MSTV, ASTV, ALTV, and HMe have the highest significance. Based on this result, all features with less than 2% importance were removed from the new dataset; thus, 18 features remained in this process. The new CTG data is then used for reclassification using the CART-based important feature selection model (IFS-CART). The aim of this stage is to obtain logic classification results based on important features. The results of the second set of experiments are presented in Table 4.

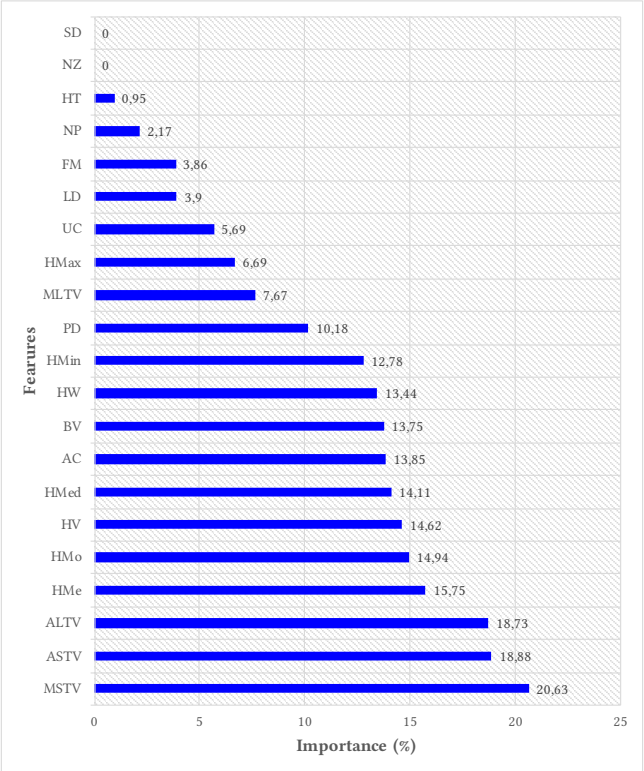

Figure 3: Feature importance of the proposed CART model on the CTG dataset using all features

As can be seen in Table 4, this model resulted in an ACC (94.50%) which is quite impressive as it is up 0.67% from the first experiment results. Meanwhile, the average values of TPR (94.49%), PPV (94.47%), and TNR (97.25%) across all classes also improved compared to the first experimental results. In this case, the proposed model mostly produced relatively good classification averages and obtained impressive class prediction results. Based on these results, the proposed model is quite promising because it can improve the classification performance of the CART model on the CTG dataset with a superb average value.

A more detailed comparison of the first and second ex-

Table 3: Confusion matrix of CART without IFS on the CTG dataset

		Predicted Class			ACC (%)	TPR (%)	PPV (%)	TNR (%)
		(1)	(2)	(3)				
Targeted Class	(1)	446	35	6	93.83	91.58	90.84	95.51
	(2)	37	467	4		91.93	92.66	96.23
	(3)	8	2	485		97.98	97.98	98.99
Average						93.83	93.83	96.91

Note: (1) Normal; (2) Suspect; (3) Pathological

Table 4: Confusion matrix of CART with IFS on the CTG dataset

		Predicted Class			ACC (%)	TPR (%)	PPV (%)	TNR (%)
		(1)	(2)	(3)				
Targeted Class	(1)	443	33	11	94.50	90.97	92.87	96.61
	(2)	29	476	3		93.70	93.33	96.54
	(3)	5	1	489		98.79	97.22	98.59
Average						94.49	94.47	97.25

Note: (1) Normal; (2) Suspect; (3) Pathological

periments is presented in Table 5. The best model performance results are indicated by bold numbers. As shown in Table 5, the second experiment with ACC value (94.50%) outperformed the first experiment. TPR outperformed the first experiment in classes 2 and 3 with values of (2 = 93.70%) and (3 = 98.79%), respectively. In addition, PPV and TNR also outperformed in class (1) and (2), with the respective values of PPV in class (1 = 92.87%) and class (2 = 93.33%), while the respective values of TNR in class (1 = 96.61%) and class (2 = 96.54%). In contrast to the first experiment, the classification performance was superior only in class TPR (1= 91.58%) and only for PPV and TNR in class 3 with values of 97.98% and 98.99%, respectively. In the second experiment, the average TPR, PPV, and TNR values were superior to those in the first experiment, with average values of 94.49%, 94.47%, and 97.25%, respectively. Based on the experimental results, overall, the proposed model in the second experiment outperforms the first experiment in which the ACC and average values of TPR, PPV, and TNR were quite impressive.

In the third experiment, we compared the proposed model with other models such as NB, MLP, SVM, k-NN, LR, and C4.5, as shown in Table 6. The results show that the accuracy (ACC), error classification (E-CL), correct classification (C-CL), and incorrect classification (I-CL) of our proposed model gives better results than the other models with respective values of ACC (94.73%), E-CL (5.27%), C-CL (2014), and I-CL (112). The second, third, and fourth best models based on accuracy were C4.4 with ACC (92.24%), k-NN with ACC (90.31%), and Logistic Regression with ACC (89.24%). The best result was MLP with ACC (81.70%).

To ensure that the IFS-CART model is not only accurate but also resistant to overfitting and has good generalizability, an evaluation was conducted on the training data and separate test datasets. This evaluation involves the calculation of accuracy metrics as the main performance indicator, which is reinforced by AUC-ROC analysis to provide a more comprehensive understanding of the model's ability to distinguish each target class: normal (1.0), suspect (2.0), and pathological (3.0). The results are shown in Figures 4 and 5.

As shown in Figure 4, the performance of the IFS-CART model remained consistent between the training and test data, with an accuracy of 94.73% on the training data and 93.80% on the test data. This minimal difference in performance indicates that the model does not experience overfitting and can generalize well to new data. In addition to accuracy, similar consistency was observed in the precision, recall, and specificity metrics, indicating that the model maintained high performance on both datasets. These results demonstrate that hyperparameter optimization, including pruning and limiting the maximum tree depth, successfully prevented the model from learning noise or irrelevant patterns from the training data. As shown in Figure 4, the performance of the IFS-CART model remained consistent between the training and test data, with an accuracy of 94.73% on the training data and 93.80% on the test data. This minimal difference in performance indicates that the model does not experience overfitting and can generalize well to new data. In addition to accuracy, similar consistency was observed in the precision, recall, and specificity metrics, indicating that the model maintained high performance on both datasets. These results demonstrate that hyperparameter optimization, including pruning and limiting the maximum tree depth, successfully prevented the model from learning noise or irrelevant patterns from the training data.

Table 5: Comparison results between the first and second experiment of CTG dataset

		Predicted Class				ACC (%)	TPR (%)	PPV (%)	TNR (%)
			(1)	(2)	(3)				
CART	Targeted Class	(1)	446	35	6	93.83	91.58	90.84	95.51
		(2)	37	467	4		91.93	92.66	96.23
		(3)	8	2	485		97.98	97.98	98.99
	Average						93.83	93.83	96.91
CART+IFS	Targeted Class	(1)	443	33	11	94.73	90.97	92.87	96.61
		(2)	29	476	3		93.70	93.33	96.54
		(3)	5	1	489		98.79	97.22	98.59
	Average						94.49	94.47	97.25

Note: (1) Normal; (2) Suspect; (3) Pathological

Table 6: Comparative effectiveness of proposed and traditional machine learning models on the CTG dataset

Approaches	Models	ACC (%)	E-CL* (%)	C-CL**	I-CL***
Machine Learning Traditional	NB	83.90	16.04	1785	341
	MLP	81.70	18.30	1737	389
	SVM	87.58	12.42	1862	264
	k-NN	90.31	9.69	1920	206
	LR	89.37	10.63	1900	226
Decision Tree-Based	C4.5	92.24	7.76	1961	165
	CART	93.83	7.79	1962	164
	Proposed Model	94.73	5.27	2014	112

Note: *Error Classification (E-CL), **Correctly Classified (C-CL), ***Incorrectly Classified (I-CL)

The evaluation of the AUC-ROC also confirms that the model is excellent in distinguishing the positive and negative classes. Figure 5 shows the ROC curves for each target class.

As shown in Figure 5, for the suspect class, the proposed model performed best with an AUC of 0.981. The ROC curve for this class is close to the upper left corner of the graph, indicating that the model achieved a high detection rate with minimal error. This performance is particularly relevant in a clinical context because the suspect class requires more intensive monitoring and early intervention to avoid the development of more serious conditions. In the pathological class, the model also performed reasonably well, with an AUC of 0.778. Although these results were sufficient to detect most high-risk conditions, some prediction errors indicated a feature overlap between the suspect and pathological classes. This reduces the accuracy of classification in certain cases, but it still provides a strong basis for medical personnel to detect critical conditions early. Further improvements to feature selection and threshold optimization can improve accuracy and reduce errors in this class. In contrast, the model performed very poorly in the normal class, exhibiting an AUC of 0.097. The ROC curve for this class almost follows a random line, indicating that the model has difficulty distinguishing between the normal and suspect classes. This low performance is likely due to an imbalance in the number of samples or feature overlap between the two classes, which makes prediction under normal conditions inaccurate. This emphasizes the importance of improving feature selection and dataset balancing to improve prediction accuracy in the normal class and reduce the possibility of unnecessary false positives. Overall, the AUC-ROC evaluation results showed that the IFS-CART model has great potential for detecting high-risk conditions, such as suspicious and pathological conditions, with good accuracy. Despite the weakness in normal class detection, this model remains relevant as a reliable diagnostic tool for real-time monitoring of fetal health. With additional optimization, the model can further strengthen support for rapid and accurate clinical decision-making, ensuring that at-risk conditions are detected in time to prevent more serious complications.

Finally, we compared the proposed model with previous studies. The results are presented in Table 4.1.

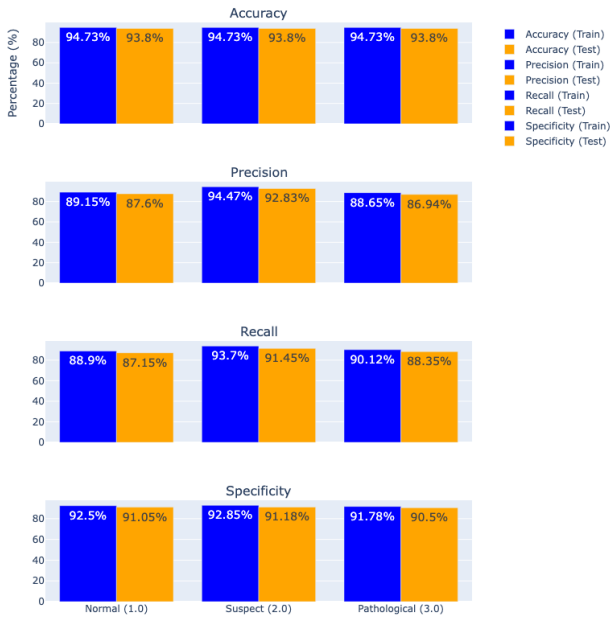

Figure 4: Comparison of the performance of the IFS-CART model with training and test data for each class (normal, suspect, and pathological)

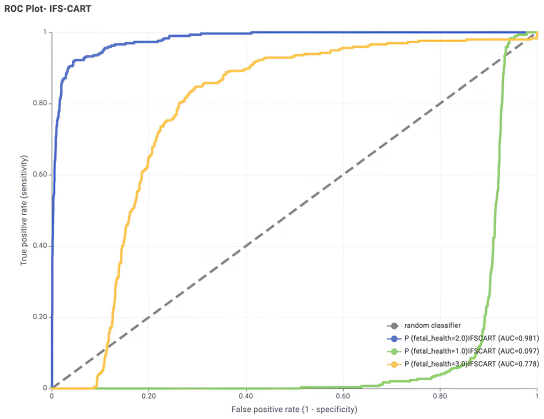

Figure 5: Receiver operating characteristic curves and area under the curve of the IFS-CART model for each fetal health class: normal (1.0), suspect (2.0), and pathological (3.0)

The comparison results presented in Table 7 show that the proposed IFS-CART model achieved the highest accuracy of 94.73%, outperforming various machine learning models reported in previous studies. The conventional CART model reported in reference [7] achieved an accuracy of 93.65%, whereas the random forest and logistic regression models reported in reference [10] only achieved an accuracy of 85%. The Decision Tree and Gradient Boosting Classifier models, with 93% and 90% accuracy, respectively, and the K-Nearest Neighbor (k-NN) model, with 90% accuracy [10], also fall below the accuracy level of the proposed model. In addition, some neural network-based models, such as the Multi-Layer Perceptron Neural Network (MLPNN) with an accuracy of 90.35%, the Probabilistic Neural Network (PNN) with an accuracy of 92.15%, and the General Regression Neural Network (GRNN) with an accuracy of 91.86% [9], although performing quite well, still show lower accuracy than the IFS-CART model. Based on these data, the proposed model showed a significant improvement in accuracy compared with other existing models, making it the highest performing model for fetal health risk classification in the context of this study.

Table 7: Comparison between the proposed model and prior studies

Studies	ACC (%)
CART [7, 11]	93.65
Random Forest [10]	85
Decision Tree [10]	93
K-Nearest Neighbor [10]	90
Logistic Regression [10]	85
Gradient Boosting Classifier [10]	90
Support Vector Machine [10]	81
MLPNN [9]	90.35
PNN [9]	92.15
GRNN [9]	91.86
Proposed Model	**94.73**

4.2 Discussions

The experimental results demonstrate that integrating the Important Feature Selection (IFS) method with the Classification and Regression Tree (CART) algorithm significantly improves fetal health classification using cardiotocography (CTG) data. The application of IFS successfully reduces dataset dimensionality by identifying the most relevant features for classification, resulting in a simpler model with reduced risk of overfitting. The IFS-CART model not only achieves high accuracy but also maintains interpretability, which is crucial in clinical applications, allowing medical professionals to understand and effectively use this model in decision-making.

Comparison of the first and second experiment results shows significant performance improvement after IFS implementation. Accuracy increased from 93.83% to 94.50%, average True Positive Rate (TPR) improved from 93.83% to 94.49%, and positive predictive value (PPV) rose from 93.83% to 94.47%. These results confirm that proper feature selection is critical for enhancing prediction accuracy by reducing the influence of less relevant features. Additionally, the increase in True Negative Rate (TNR) from 96.91% to 97.25% indicates the model's improved ability

to identify negative cases, which can help reduce unnecessary medical interventions.

In the third experiment, the proposed IFS-CART model was compared with other machine learning models from previous studies to assess its superiority. With the highest accuracy of 94.73%, IFS-CART outperformed various other models. For instance, conventional CART from previous references achieved 93.65% accuracy [7], while random forest and logistic regression had 85% accuracy [10], demonstrating that IFS-CART offers higher accuracy and better efficiency. The Decision Tree model with 93% accuracy [10] approached IFS-CART's performance but still fell short. These results highlight that although traditional decision tree models offer good interpretability, the IFS-CART model provides superior accuracy crucial for reliability in medical applications.

Furthermore, K-Nearest Neighbor and Gradient Boosting Classifier algorithms, each with 90% accuracy [10], as well as Support Vector Machine (SVM) with 81% accuracy [10], showed limitations in handling the complexity of fetal health classification compared to IFS-CART. Neural network-based models, such as multi-layer perceptron neural network (MLPNN) with 90.35% accuracy, Probabilistic Neural Network (PNN) with 92.15%, and General Regression Neural Network (GRNN) with 91.86% [9], also approached IFS-CART's performance but required significantly more computational resources and were prone to overfitting. In contrast, IFS-CART is not only more computationally efficient but also easier to understand, allowing medical professionals to interpret results quickly.

The Area Under Curve-Receiver Operating Characteristic (AUC-ROC) evaluation shows that IFS-CART performs well in distinguishing the "suspect" class with an AUC value of 0.981 and the "pathological" class with an AUC of 0.778. High performance in the "suspect" class is particularly important in clinical settings, where at-risk cases require intensive monitoring and early intervention to prevent serious complications. However, the low AUC value for the "normal" class (0.097) indicates the model's difficulty in distinguishing between normal and at-risk conditions. This difficulty may be due to sample imbalance and feature overlap between normal and at-risk classes, which can affect prediction accuracy for these classes. These results highlight the importance of additional optimization, such as class data balancing and more effective feature selection, to improve model performance in detecting normal cases.

Additionally, the application of cost-complexity pruning and tree depth limitation successfully reduced the risk of overfitting, reflected in the consistency of model performance between training and testing data. From a practical perspective, the IFS-CART model provides significant value to medical professionals due to its high interpretability. This model allows clinicians to utilize prediction results quickly and accurately without requiring a deep understanding of complex computational algorithms. This easy interpretability is highly relevant in clinical contexts, where quick and accurate decisions are crucial for reducing the risk of complications during childbirth.

Furthermore, the model's flexibility in handling data variations and its ability to be integrated into real-time fetal health monitoring systems demonstrate its initial potential as a reliable and effective solution in medical applications. However, these findings require further validation using broader and more diverse datasets to ensure model generalization and reliability across various clinical conditions. Sub optimal performance in detecting normal cases indicates the need for further strategies, such as enhanced feature selection algorithms or class data balancing, to reduce false-positive predictions and improve accuracy.

Overall, this study shows that the combination of IFS and CART is an effective strategy for classifying fetal health risks using CTG data. These findings are in line with previous studies that have demonstrated the benefits of integrating feature selection techniques with machine learning algorithms to improve medical classification accuracy [2]. For example, [23] found that the use of feature selection improved model performance in predicting pregnancy outcomes based on CTG data. Finally, our findings not only improve accuracy, but also offer important interpretability and efficiency in medical applications. With continuous optimization and validation, this model can support faster and more accurate clinical decision-making.

5 Conclusions

In conclusion, this study highlights the theoretical and practical implications of developing a fetal health risk classification model based on Essential Feature Selection (IFS) and Classification and Regression Trees (CART) tested on cardiotocography (CTG) data. The results of this study confirm the importance of relevant feature selection for improving the accuracy of prediction models, which supports more accurate clinical decision-making and potentially reduces the risk of maternal and fetal health complications.

The primary contribution of this study was the finding that the integration of IFS and CART can improve the consistency and interpretability of the model, making it more practical for use by medical personnel without in-depth knowledge of computational algorithms. This finding could encourage the use of predictive models to detect fetal risk conditions, which is clinically crucial in managing cases of labor-related complications. Although the model performed well in the "at-risk" category, this study identified the need for improved accuracy in distinguishing normal fetal conditions, especially to reduce the likelihood of unnecessary medical interventions.

In future research, subsequent investigations should focus on refining feature selection techniques and data balancing methods to enhance prediction accuracy across all fetal condition categories. Furthermore, additional testing with more extensive and diverse datasets will strengthen the generalizability of the model in various clinical contexts.

The exploration of more advanced machine learning models and automation in CTG data assessment are also crucial steps to improve the accuracy and efficiency of predicting real-time fetal health.

Author contributions

Ahmad Ilham: Conceptualized and led the research, designed the methodology, developed models, analyzed performance, supervised the project, and significantly contributed to manuscript writing. Thahta Ardhika Prabu Nagara: Assisted in data collection and preprocessing as part of student mentorship. Mudyawati Kamaruddin: Provided expertise in interpreting medical data, particularly cardiotocography (CTG), and reviewed clinical aspects and implications of research findings. Laelatul Khikmah: Performed statistical analysis, ensured data normalization, and contributed to experimental design and validation. Teddy Mantoro: Provided guidance on machine learning approach, ensured compliance with computational standards, and assisted with manuscript revision by offering technical insights for model refinement.

Acknowledgement

The authors would like to thank to Director of Research Muhammadiyah Higher Education Research and Development Council Muhammadiyah Central Leadership (DRMHER–DCMCL). This research work was funded by DRMHER–DCMCL under Grant No. 1687.249/PD/I.3/D/2022.

References

[1] C. E. Wood and M. Keller-Wood, ”Current paradigms and new perspectives on fetal hypoxia: implications for fetal brain development in late gestation,” *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, vol. 317, no. 1, pp. R1–R13, 2019. https://doi.org/10.1152/ajpregu.00008.2019

[2] F. Francis, S. Luz, H. Wu, S. J. Stock, and R. Townsend, ”Machine learning on cardiotocography data to classify fetal outcomes: A scoping review,” *Computers in Biology and Medicine*, vol. 172, p. 108220, 2024, https://doi.org/10.1016/j.compbiomed.2024.108220

[3] D. Ayres-de-Campos, *Obstetric emergencies*. Springer, 2016. Springer International Publishing, 2017. https://doi.org/10.1007/978-3-319-41656-4

[4] J. E. Lawn, H. Blencowe, P. Waiswa, A. Amouzou, C. Mathers, D. Hogan, V. Flenady, J. F. Frøen, Z. U. Qureshi, C. Calderwood, S. Shiekh, F. B. Jassir, D. You, E. M. McClure, M. Mathai, S. Cousens, and others, ”Stillbirths: rates, risk factors, and acceleration towards 2030,” *The Lancet*, vol. 387, no. 10018, pp. 587–603, 2016. https://doi.org/10.1016/S0140-6736(15)00837-5

[5] S. Rajagopal, K. Ruetzler, K. Ghadimi, E. M. Horn, M. Kelava, K. T. Kudelko, I. Moreno-Duarte, I. Preston, L. L. R. Bovino, N. R. Smilowitz, A. Vaidya, and the American Heart Association Council on Cardiopulmonary, Critical Care, Perioperative and Resuscitation, and the Council on Cardiovascular and Stroke Nursing, ”Evaluation and Management of Pulmonary Hypertension in Noncardiac Surgery: A Scientific Statement From the American Heart Association,” *Circulation*, vol. 147, no. 17, pp. 1317–1343, 2023. https://doi.org/10.1161/CIR.0000000000001136.

[6] K. Barnova, R. Martinek, R. Vilimkova Kahankova, R. Jaros, V. Snasel, and S. Mirjalili, “Artificial Intelligence and Machine Learning in Electronic Fetal Monitoring,” *Archives of Computational Methods in Engineering*, vol. 31, no. 5, pp. 2557–2588, Jul. 2024. https://doi.org/10.1007/s11831-023-10055-6

[7] J. V. Y. Chuatak, E. R. C. Comentan, R. L. H. G. Moreno, R. K. C. Billones, R. G. Baldovino, and J. C. V. Puno, “A decision tree-based classification of fetal health using cardiotocograms,” 2023, p. 020003. https://doi.org/10.1063/5.0111194

[8] H. Sahin and A. Subasi, “Classification of the cardiotocogram data for anticipation of fetal risks using machine learning techniques,” *Appl Soft Comput*, vol. 33, pp. 231–238, Aug. 2015. doi:10.1016/j.asoc.2015.04.038

[9] E. Yılmaz, “Fetal State Assessment from Cardiotocogram Data Using Artificial Neural Networks,” *J Med Biol Eng*, vol. 36, no. 6, pp. 820–832, Dec. 2016. doi:10.1007/s40846-016-0191-3

[10] Y. Salini, S. N. Mohanty, J. V. N. Ramesh, M. Yang, and M. M. V. Chalapathi, “Cardiotocography Data Analysis for Fetal Health Classification Using Machine Learning Models,” *IEEE Access*, vol. 12, pp. 26005–26022, 2024. https://doi.org/10.1109/ACCESS.2024.3364755

[11] D. Campos and J. Bernardes. ”Cardiotocography,” *UCI Machine Learning Repository*, 2000. [Online]. Available: https://doi.org/10.24432/C51S4N

[12] F. Mesa, R. Ospina-Ospina, and D. Devia-Narvaez, “Comparison of Support Vector Machines and Classification and Regression Tree Classifiers on the Iris Data Set,” *Journal of Southwest Jiaotong University*, vol. 58, no. 2, 2023. https://doi.org/10.35741/issn.0258-2724.58.2.59

[13] J. Josse, J. M. Chen, N. Prost, G. Varoquaux, and E. Scornet, "On the consistency of supervised learning with missing values," *Statistical Papers*, Sep. 2024. https://doi.org/10.1007/s00362-024-01550-4

[14] S. Chanmee and K. Kesorn, "Semantic decision Trees: A new learning system for the ID3-Based algorithm using a knowledge base," *Advanced Engineering Informatics*, vol. 58, p. 102156, Oct. 2023. https://doi.org/10.1016/j.aei.2023.102156

[15] A. Agarwal, K. Jain, and R. K. Yadav, "A mathematical model based on modified ID3 algorithm for healthcare diagnostics model," *International Journal of System Assurance Engineering and Management*, vol. 14, no. 6, pp. 2376–2386, Dec. 2023. https://doi.org/10.1007/s13198-023-02086-w.

[16] Y.-C. Chiang, Y.-C. Hsieh, L.-C. Lu, and S.-Y. Ou, "Prediction of Diagnosis-Related Groups for Appendectomy Patients Using C4.5 and Neural Network," *Healthcare*, vol. 11, no. 11, p. 1598, May 2023. https://doi.org/10.3390/healthcare11111598

[17] M. Ozcan and S. Peker, "A classification and regression tree algorithm for heart disease modeling and prediction," *Healthcare Analytics*, vol. 3, p. 100130, Nov. 2023. https://doi.org/10.1016/j.health.2022.100130

[18] A. R. Kadhim, R. S. Khudeyer, and M. Alabbas, "Facial Sentiment Analysis Using Convolutional Neural Network and Fuzzy Systems," *Informatica*, vol. 48, no. 12, Sep. 2024. https://doi.org/10.31449/inf.v48i12.6151

[19] C. Pal, S. Das, A. Akuli, S. K. Adhikari, and A. Dey, "Cocoa-Net: Performance Analysis on Classification of Cocoa Beans Using Structural Image Feature," *Informatica*, vol. 48, no. 12, Sep. 2024. https://doi.org/10.31449/inf.v48i12.5762

[20] X. Ying, "An Overview of Overfitting and its Solutions," *J Phys Conf Ser*, vol. 1168, no. 2, p. 022022, Feb. 2019. https://doi.org/10.1088/1742-6596/1168/2/022022

[21] A. Ilham, A. Kindarto, A. Kareem Oleiwi, and L. Khikmah, "CFCM-SMOTE: A Robust Fetal Health Classification to Improve Precision Modelling in Multi-Class Scenarios," *International Journal of Computing and Digital Systems*, vol. 15, no. 1, pp. 1–9, 2024. https://doi.org/10.12785/ijcds/160137

[22] A. Ilham, R. Satria Wahono, C. Supriyanto, and A. Wijaya, "U-control Chart Based Differential Evolution Clustering for Determining the Number of Cluster in k -Means," *International Journal of Intelligent Engineering and Systems*, vol. 12, no. 4, pp. 306–316, Aug. 2019. https://doi.org/10.22266/ijies2019.0831.28

[23] Cömert, Zafer and Şengür, Abdulkadir and Budak, Ümit and Kocamaz, Adnan Fatih, "Prediction of intrapartum fetal hypoxia considering feature selection algorithms and machine learning models," *Health Information Science and Systems*, vol. 12, no. 4, 2019. https://doi.org/10.1007/s13755-019-0079-z