Scalogram-Based Multiclass Fetal State Classification Using Expert-Annotated CTG and SE-ResNet-50

Trie Maya Kadarina[1], Basari[1], Dadang Gunawan[1,*], Abraham Auzan[2]
[1]Dept. of Electrical Engineering, Universitas Indonesia, 16424, Indonesia
[2]Elektra Medical Research and Development, PT Elektra Abimantrana Internasional, Indonesia
E-mail: guna@eng.ui.ac.id
*Corresponding author

Keywords: Cardiotocography, expert validation, fetal state classification, residual network, scalograms

Received: April 24, 2025

Accurate classification of cardiotocographic (CTG) signals plays a critical role in the early detection of fetal health conditions, enabling timely and appropriate medical interventions. A more precise understanding of cardiotocography patterns, particularly in suspicious cases, can help minimize unnecessary interventions and reduce healthcare costs. This study proposes a multiclass classification framework using 552 expert-annotated CTG records containing fetal heart rate (FHR) and uterine contraction (UC) signals. Time-domain augmentation, including cyclic temporal shifting, Gaussian noise, and segmented Gaussian noise, was applied to address class imbalance. The augmented FHR and UC signals were then transformed into scalograms using Continuous Wavelet Transform (CWT), producing dual-channel RGB-encoded images. Data were split into 90% for stratified five-fold cross-validation and 10% for independent testing. The proposed ResNet50, enhanced with SE-based channel attention and dropout layers, was compared against several baselines, including CNN, MobileNet, EfficientNet-B0, ResNet18, and ResNet50. It achieved the best performance with an F1-score of 0.7267 and AUC of 0.7489 on the test set, outperforming all baselines. These results highlight its potential for integration into intelligent clinical decision support systems in prenatal care.

Povzetek: Članek obravnava klasifikacijo fetalnega stanja iz kardiotokografije v povezavi s subjektivnimi ocenami in prekrivanjem vzorcev, zlasti pri sumljivih primerih. Predlaga metodo SE-ResNet-50, ki FHR in UC signale pretvori v skalograme s CWT. Model doseže najboljši rezultat med primerjanimi pristopi.

1 Introduction

Ensuring fetal well-being is critical to maintaining the health and safety of babies during pregnancy and childbirth. The World Health Organization (WHO) reports that more than 800 women die every day from complications of pregnancy and childbirth, especially in areas with limited access to quality health services [1]. Furthermore, according to the WHO report, effective monitoring of maternal and fetal health can lower newborn mortality rates by up to 20% in regions where adequate healthcare services are available [2].

Health care professionals can take precautions to reduce the risk if fetal complications arise from antenatal care evaluations [3]. Cardiotocography (CTG) is a widely used tool for assessing fetal conditions during pregnancy and labor by recording fetal heart rate (FHR) and uterine contraction (UC) signals. A key advantage of CTG is its ability to provide continuous and real-time monitoring, allowing early detection of fetal distress and uterine activity changes [4]. By identifying fetal hypoxia, CTG monitoring reduces the risk of newborn asphyxia and lowers perinatal mortality rates [5, 6].

CTG interpretation follows three main guidelines from NICE (National Institute of Health and Care Excellence), FIGO (International Federation of Gynecology and Obstetrics), and ACOG (American College of Obstetricians and Gynecologists) [7]. However, CTG assessment often varies due to subjectivity and differences in expertise among healthcare professionals. Fatigue, stress, complex cases, and time constraints can further hinder accurate and timely decision-making [8]. To address these challenges, artificial intelligence (AI) and machine learning (ML) techniques have been integrated into CTG analysis. ML enhances diagnostic accuracy by detecting disease patterns and risk factors, enabling early fetal distress detection and proactive interventions [9, 10]. AI-driven models allow healthcare providers to analyze large volumes of fetal monitoring data, leading to faster and more informed decisions, while also reducing unnecessary interventions such as cesarean sections [11, 12].

The classification of the fetal state in CTG is essential for early detection, particularly when rapid intervention is required during pregnancy or labor [13]. This classification includes normal, suspicious, and pathological states. Distinguishing between these categories is inherently

challenging, as suspicious cases often present intermediate features between normal and pathological patterns, causing overlaps in FHR and UC characteristics that obscure decision boundaries and hinder classification accuracy [14]. Misclassification can lead to unnecessary interventions or delayed responses, affecting maternal and fetal safety.A deeper understanding of borderline CTG patterns can help optimize clinical decisions, reduce unnecessary procedures, and lower healthcare costs.

This study aims to address existing gaps in CTG analysis and presents two key contributions in advancing fetal state classification using deep learning and signal processing:

(1) Multiclass fetal state classification with expert annotations. Unlike previous studies that classify fetuses into only normal and pathological categories, this study introduces a suspicious category, creating a three-class classification (normal, suspicious, pathological). Expert annotations are used instead of relying solely on biochemical indicators like cord blood pH values, ensuring a more clinically relevant and interpretable classification.

(2) Combining FHR and UC signals as input for a SE-ResNet-50 model in scalogram-based CTG classification. This study combines FHR and UC signals, transforming them into scalograms using Continuous Wavelet Transform (CWT) to capture temporal and frequency domain features which allows a detailed representation of signal variations critical for accurate fetal state classification. The scalograms are represented in RGB format, enabling the model to leverage multi-channel color information for richer feature extraction. An SE-ResNet50 model with squeeze and excitation attention blocks is used to enhance feature extraction and improve generalization in the classification of three classes of fetal state.

2 Related works

Several studies have applied machine learning algorithms to predict fetal abnormalities by extracting features from CTG signals [15, 16]. These studies have primarily focused on classifying fetal conditions into normal and abnormal (pathological) categories, as well as predicting specific complications such as fetal hypoxia and preterm birth [17–21]. Previous studies have used different labelling approaches based on data availability. Expert annotations involves obstetricians clinically assessing fetal conditions, while biochemical indicators such as cord blood pH and Base Excess (BE) evaluate acid-base balance and hypoxia risk. The Apgar score is also used to validate fetal distress predictions based on newborn health.

Previous studies have classified fetal status into three classes: normal, suspicious, and pathological as shown in Table 1. However, most of these studies are based on traditional machine learning models trained in the UCI Machine Learning Repository dataset [22]. These studies applied a variety of algorithms, including Artificial Neural Networks (ANN), Rough Neural Networks, Naïve Bayes, Logistic Regression (LR), K-Nearest Neighbors (KNN), Decision Tree (DT), Classification and Regression Tree (CART), Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB), and Light Gradient Boosting Machine (LGBM). Although these models achieve high accuracy (0.89-0.99) [23–28], the dataset has significant limitations, as it contains only 21 static features and lacks the temporal information necessary to capture the dynamic nature of CTG signals. This limitation hinders the ability of the model to fully assess fetal conditions, which depend on time-dependent variations in FHR and uterine contractions for a more accurate and reliable classification.

Deep learning offers a more robust approach by eliminating the need for manual feature selection and enabling automatic extraction of important features directly from raw signals [29]. Several studies have explored deep learning for CTG classification [17, 19, 29–31]. Despite these developments, most studies only use FHR signals, ignoring UC signals, which play a crucial role in assessing fetal response to labour stress. The combination of FHR and UC provides a more comprehensive view of fetal well-being, particularly in detecting late or variable decelerations, which are early indicators of fetal distress and hypoxia [16, 32]. Table 2 shows the summary of related work using CTU-UHB dataset.

A number of research employ both FHR and UC signals with varying outcome predictions utilizing various signal processing approaches. For instance, Zeng, et al. classified fetal state in normal and abnormal classes based on pH values and Base Excess (BE) values used to assess the oxygenation status and acid-base balance of the fetus for indications of hypoxia. They used the Time-Frequency (TF) and Ensemble Cost-Sensitive Support Vector Machine (ECSVM) features which resulted in a sensitivity of 0.852, a specificity of 0.661, and a quality index of 0.75.0 [33]. Liang et al. utilized a CNN-RNN model to detect fetal hypoxia based on pH values, achieving an accuracy of 0.9515 [34]. Similarly, Ogasawara et al. developed CTG-Net, a CNN-based model that classifies fetal conditions using Apgar scores and umbilical artery pH, resulting in an F1 score of 0.67 [35]. Meanwhile, Saini et al. applied a 2D-CNN model to classify fetal conditions into normal, mild hypoxia, and severe hypoxia, obtaining an accuracy of 0.70 [36].

Deep learning shows promise for improving fetal state classification, but challenges remain in accuracy and clinical reliability. Most studies use binary classification (normal vs. abnormal) and rely on biochemical markers for labeling, often overlooking the suspicious class, which can lead to unnecessary interventions or delayed responses. Furthermore, a large proportion of existing work processes CTG as raw time-series signals, which may not capture important frequency-domain features, and often relies solely on FHR without incorporating UC. The use of scalograms (TF representations) offers richer feature extraction, enhancing deep learning performance. In particular, CWT-based scalograms provide high-resolution localization of

Table 1: Summary of related studies using UCI CTG dataset

Work	Input	Label Type	Classes	**Method**	**Performance**
[23]	21 statistical features	Expert Annotations	Normal, Suspicious, Pathological	Rough Neural Network	Accuracy = 0.905
[24]	21 statistical features	Expert Annotations	Normal, Suspicious, Pathological	AdaBoost with RF	Accuracy = 0.976
[25]	21 statistical features	Expert Annotations	Normal, Suspicious, Pathological	LR, KNN, RF, and GBM	Highest accuracy = 0.99 (RF)
[26]	21 statistical features	Expert Annotations	Normal, Suspicious, Pathological	RF, Naïve Bayes, SVM	Highest accuracy = 0.9993 (RF)
[27]	21 statistical features	Expert Annotations	Normal, Suspicious, Pathological	RF, LR, DT, SVM, Voting Classifier, KNN	Highest accuracy = 0.9751 (RF)
[28]	21 statistical features	Expert Annotations	Normal, Suspicious, Pathological	Classification and Regression Tree (CART)	Accuracy=0.945

Table 2: Summary of related studies using CTU-UHB dataset

Work	Input	Label Type	Classes	Method	Performance
[17]	FHR Signals	Delivery mode & pH	Normal (vaginal delivery), Pathological	1DCNN-MLP	Sensitivity=0.80, Specificity=0.791
[19]	FHR Signals	pH	Normal, Hypoxia	Ensemble Learning, CNN, DenseNet	CNN: F1-score=0.9 (Normal), 0.23 (Hypoxia); Bagging Tree + NB: F1-score=0.76 & 0.45
[29]	FHR Signals	pH	Normal, Distressed	TF Morse Wavelet, ResNet	Accuracy=0.987, Sens=0.97, Spec=0.100
[30]	FHR Signals	pH	Normal, Hypoxia	DenseNet	Precision=0.50, Sensitivity=0.43, F1-score=0.46
[31]	FHR Signals	pH	Normal, Abnormal	SE-ResNet50, XGBooster	Accuracy=0.9634, Precision=0.967, Sens=0.973, Spec=0.96
[33]	FHR and UC	pH & BE	Normal, Abnormal	TF, ECSVM	Sens=0.852, Spec=0.661
[34]	FHR and UC	pH	Normal, Abnormal	CNN, RNN	Accuracy=0.9515, Sens=0.962, Spec=0.9409
[35]	FHR and UC	Apgar Score & pH	Normal, Abnormal	CTG-Net	F1-score=0.67
[36]	FHR and UC	pH	Normal, Mild Hypoxia, Severe Hypoxia	2D-CNN	Accuracy=0.70

signal patterns in both time and frequency domains [29], allowing the detection of subtle and transient variations in FHR and UC that are clinically relevant for distinguishing fetal states. Our study addresses these gaps by integrating both FHR and UC signals, encoding them as RGB scalogram images, adopting a three-class scheme according to FIGO guidelines, and applying time-frequency representations based on scalograms through CWT for fully automated feature learning with a deep learning model.

3 Methods

In this study, expert labeling was chosen over physiological metrics such as the pH value of cord blood, as it considers multiple clinical factors for a more comprehensive

fetal evaluation. By using both FHR and UC for a comprehensive assessment and employing multiclass classification, this study enhances the detection of the suspicious category for early risk identification. This study employs a systematic methodology, as shown in Figure 1 to classify fetal conditions using CTG signals and deep learning.

The research methodology consists of three main stages: (1) data selection and annotation analysis, (2) data preprocessing and feature extraction, and (3) model training and evaluation. In the first stage, CTG data from the publicly available CTU-UHB dataset [37] is annotated by experts, with consistency assessed using Pearson correlation and Fleiss' kappa. The second stage involves cleaning and augmenting FHR and UC signals, transforming them into scalograms via Continuous Wavelet Transform (CWT), and encoding them into dual-channel RGB format. In the final stage, scalogram images and annotations are split into a 10% independent test set and a 90% training–validation set, with stratified five-fold cross-validation to preserve class balance. An SE-ResNet50 model with ImageNet-pretrained weights and Squeeze-and-Excitation blocks is trained using focal loss, class weighting, and dropout regularization. Evaluation includes comparison with baseline models, performance assessment on cross-validation and test sets, statistical significance testing, ablation studies, and Grad-CAM interpretability analysis. The following describes the algorithm used for the classification of CTG signals.

We hypothesize that combining fetal heart rate (FHR) and uterine contraction (UC) signals in scalogram form, encoded as dual-channel RGB images, and training a modified SE-ResNet50 model with focal loss and class weighting will improve multiclass CTG classification performance compared to baseline models.

Algorithm 1 CTG_Classification

Input: Selected Raw CTG signals (FHR, UC) with expert-validated labels
Output: Classified fetal state (Normal, Suspicious, Pathological)

Step 1: Preprocessing Data
Denoise and clean FHR and UC signals
Remove outliers, fill missing value
Apply augmentation: cyclic shift, Gaussian noise, segmented Gaussian noise

Step 2: Continuous Wavelet Transform (CWT)
Convert each FHR and UC signal into separate scalograms

Step 3: Dual-Channel RGB Encoding
Map FHR→R, UC→G, B=0 to form RGB scalogram

Step 4: Model Training (SE-ResNet-50)
Initialize SE-ResNet50 with Squeeze-and-Excitation blocks + dropout
Compile with Adam, Focal Loss, and class weights
Perform stratified five-fold cross-validation on 90% data
Select fold with highest validation F1-score

Step 5: Testing and Evaluation
Test best fold on 10% hold-out data
Report Accuracy, Precision, Sensitivity, F1-Score, and AUC

End Algorithm

3.1 Dataset

We utilized the public CTG dataset from the Czech Technical University (CTU), containing 552 CTG recordings [37], available at: https://physionet.org/content/ctu-uhb-ctgdb/1.0.0/. This CTG data was annotated by nine obstetricians following FIGO guidelines at each evaluation step as shown in Figure 2. The CTG dataset annotation is available at https://people.ciirc.cvut.cz/ spilkjir/data.html [38]. In the CTU-CHB Intrapartum Cardiotocography dataset, the collected signals focused mainly on the first- and second-stage labor phases of the labour process. The first stage labour includes the opening of the cervix from the beginning of contraction to the full opening of the cervix, reflecting variations in the fetal heart rate as the intensity of contractions increases. The dataset also includes the second phase of labor, when the cervix has fully opened and the birth process begins, showing an FHR response to stronger contractions as the mother strains.

In this study, data annotations were determined based on the majority voting technique of nine experts on each CTG recording. A technique in data labeling in which the final decision is determined based on the label that is most chosen among several experts [39]. This technique is often used when there are multiple evaluations from different sources or individuals, for example, in a situation where a number of experts label the data, and we need to determine a single consensus label to use. To evaluate annotation consistency, each step was assessed using two complementary agreement metrics: (1) average pairwise Pearson correlation between expert ratings, which reflects the similarity in scoring tendencies; and (2) Fleiss' kappa, which measures exact categorical agreement while accounting for chance agreement. For the correlation analysis, a 9×9 matrix was generated for each step, where each element represents the correlation between a pair of experts, and the average value was calculated across all pairs. Fleiss' kappa was calculated based on the categorical labels assigned by the nine experts for each case. By combining these two metrics, we identified the step with the highest and most consistent agreement across both scoring trends and categorical label matches, ensuring that the final dataset is derived from the most reliable annotation stage.

To prevent data leakage during model training and eval-

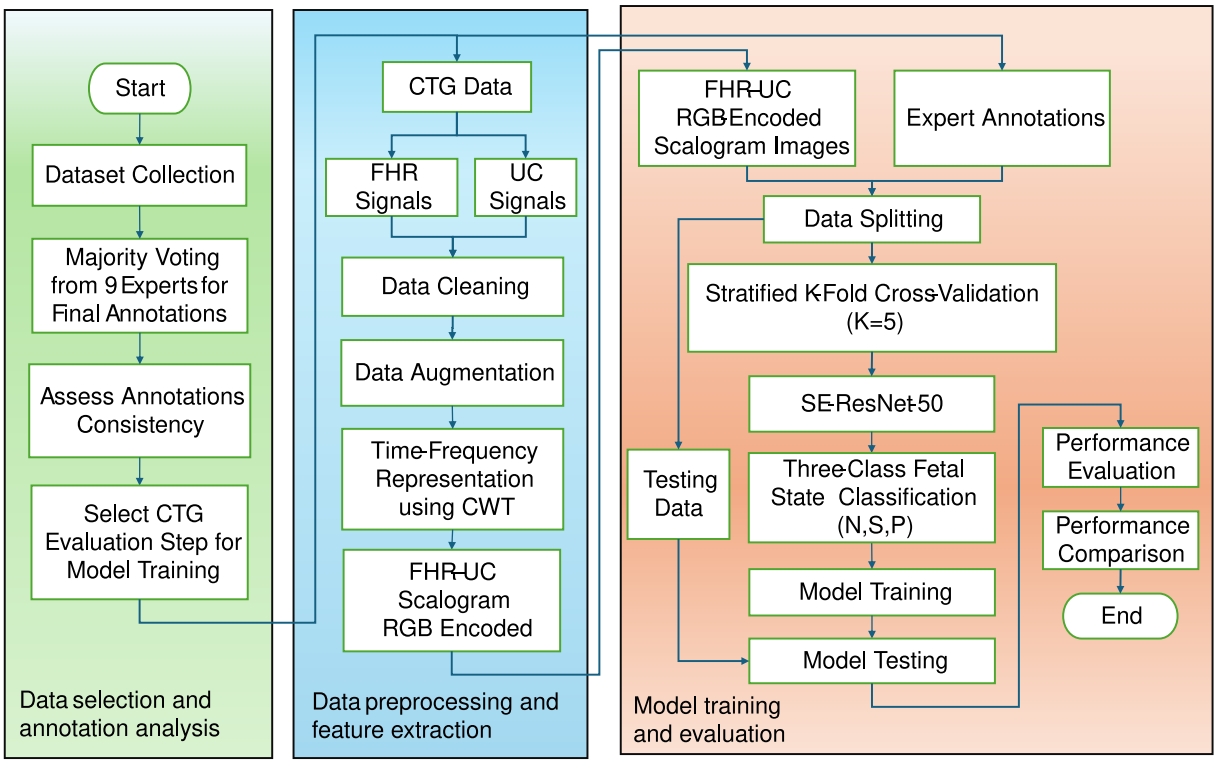

Figure 1: Proposed methodology for CTG classification

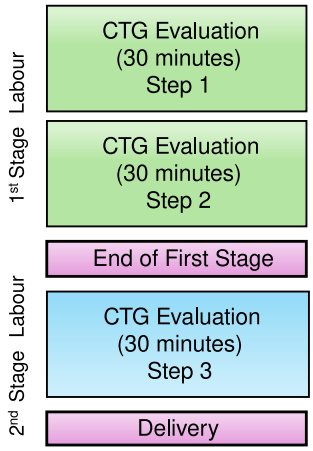

Figure 2: CTG evaluation step in CTG CTU-UHB dataset

uation, the dataset was split using a unique identifier (base_id) for each original CTG signal, ensuring that all augmented versions of the same signal remained within a single subset. Splits were stratified by base_id to preserve class balance, and the same strategy was applied in cross-validation to maintain independence between subsets.

After determining the most reliable annotation stage (detailed in Section 4.1), the dataset was prepared for model development. The selected step was used as the basis for data augmentation and splitting. Table 3 summarizes the dataset distribution before and after augmentation, including the number of samples in the training, validation, and test sets, as well as the cases excluded from the study.

3.2 Preprocessing steps

Preprocessing techniques in this study include data cleaning and augmentation. The data cleaning process consists of several stages. First, all zero values in the dataset are replaced with NaN (Not a Number) to indicate to indicate potentially invalid or unrecorded data. If NaN values occur for more than 15 consecutive seconds, they are retained to prevent excessive interpolation across long missing segments. Next, outlier detection is applied to remove physiologically implausible values. For FHR, values outside the range of 50–200 bpm are discarded [29], while for UC, values outside 0–100 mmHg are excluded [40]. Spike detection is also performed to identify sudden unrealistic changes, defined as variations greater than 25 bpm for FHR [29] or 40 mmHg for UC [40] between consecutive samples. Such spikes are assumed to be artifacts and are replaced with NaN. Then the proportion of missing data is calculated. If the NaN ratio exceeds 20% of the total signal length, the recording is discarded to maintain data quality. For signals that pass this threshold, missing values are interpolated using a two-step approach: linear interpolation to preserve local trends, followed by cubic spline interpolation for smooth transitions. Any remaining NaN values are filled using backward and forward filling methods. Finally, all NaN and infinite values are replaced with zero to ensure

Table 3: Dataset distribution before and after augmentation (Step 1 only), including uninterpretable cases, and split into train/validation/test sets.

Class	Before Aug.	After Aug.	Test Set	Train+Val Set	Note
Normal	296	636	66	570	Used
Suspicious	220	606	66	540	Used
Pathological	30	210	15	195	Used
Uninterpretable	6	–	–	–	Excluded
Total	552	1452	147	1305	

numerical stability before time–frequency transformation.

Data augmentation in the time domain employed three techniques: cyclic temporal shifting, Gaussian noise injection, and segmented Gaussian noise injection. For normal and suspicious classes, cyclic shifts equivalent to 5 and 10 minutes (corresponding to 1,200 and 2,400 sample points at 4 Hz) and Gaussian noise (noise factor = 0.1) were applied to the original and shifted signals. For the pathological class, more extensive augmentation was applied, consisting of cyclic shifts of 5, 10, 15, and 20 minutes (1,200, 2,400, 3,600, and 4,800 sample points), the addition of Gaussian noise, and segmented Gaussian noise with varying noise levels (0.05, 0.1, 0.5) applied to both the original and shifted signals. This strategy enhanced data diversity and mitigated class imbalance.

3.3 Continuous wavelet transform

The next stage is the signal transformation into the time-frequency domain using continuous wavelet transformation (CWT). Continuous Wavelet Transform (CWT) is a method for analyzing signals by breaking them into different parts based on scale or frequency [41]. CWT is very useful for signals whose frequency changes over time. CWT works by shifting and changing the size of the mother wavelet along the signal and seeing how similar the signal is to the wavelet at each point. The result is numbers (coefficients) that show how similar the signal is to a wavelet at a given scale and position. In this way, changes in the frequency of the signal overtime can be seen, providing a more complete picture compared to traditional Fourier analysis, which assumes the signal does not change. The mathematical equation for the wavelet function (1) is as follows [42]:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \tag{1}$$

With a representing the scaling parameter for dilation and b representing the moving parameter for translation across the signal location. CWT follows two properties which are represent in equation (2) and (3) [43]:

$$\int_{-\infty}^{\infty} \psi(t)\, dt = 0 \tag{2}$$

$$\int_{-\infty}^{\infty} [\psi(t)]^2\, dt = 0 \tag{3}$$

For CWT, the mathematical equation $C(a,b)$ is obtained by integrating the input function with the wavelet which is stated in equation (4) [42]:

$$C(a,b) = \int_{-\infty}^{\infty} f(t)\frac{1}{\sqrt{|a|}}\psi\left(\frac{t-b}{a}\right) dt \tag{4}$$

In this study, a Morse wavelet in the frequency domain is used with the following formulation, as shown in equation (5) [43]:

$$\psi_{\beta,\gamma}(\omega) = U(\omega)a_{\beta,\gamma}\omega^{\beta}e^{\omega^{\gamma}} \tag{5}$$

β and γ are parameters that control the shape of the wavelet. β controls the shape and width of the wavelet while γ controls the asymmetry or shift of the wavelet shape. ω is a variable frequency in the frequency domain, which indicates the frequency analyzed at each point in time. The central frequency (ω_o) of the wavelet that regulates the number of oscillations. If we put Morse wavelets into the CWT formula, then the formula becomes as shown in equation (6) [43]:

$$C(a,b) = \int_{-\infty}^{\infty} f(t)\frac{1}{\sqrt{|a|}}U\left(\frac{t-b}{a}\right)a_{\beta,\gamma}\left(\frac{t-b}{a}\right)^{\beta}e^{-\left(\frac{t-b}{a}\right)^{\gamma}}e^{i\omega_o\frac{t-b}{a}}dt \tag{6}$$

Equation (7) shows how CWT uses Morse wavelets for scale and frequency analysis of signals. Appropriate wavelet parameters are selected to obtain optimal scalogram representation. In this study, γ was set to 3 for both signals, following recommendations that this value yields a near Gaussian spectral shape and provides balanced time–frequency localization suitable for biomedical signals. The β parameter was set to 50 for FHR to capture rapid heart rate variations and 100 for UC to emphasize slower sustained contraction patterns, consistent with prior findings on frequency localization characteristics [43]. While this configuration is supported by literature, a more comprehensive empirical evaluation using alternative wavelets and parameter settings could provide deeper insights into its optimality. Sampling frequency was 4 Hz with a signal length of 7200 and voices per octave of 12. Each signal was individually transformed into a scalogram using Continuous Wavelet Transform (CWT) with its respective β value, ensuring optimal time–frequency resolution for both modalities while preserving essential information in both low and high-frequency components.

CWT can be used to decompose one-dimensional (1D) signals into two-dimensional (2D) [44]. The 1D signal is decomposed into wavelet coefficients, capturing similarities at different scales and positions. This produces a 2D scalogram, where time and frequency are represented, and color indicates coefficient amplitude. Scalograms enhance feature and frequency visualization, revealing details not apparent in 1D representations [44].

3.4 Dual-channel RGB encoding

This RGB representation not only preserves the distinct characteristics of both signals but also spatially separates them into dedicated channels, reducing feature overlap and enhancing the network's ability to learn discriminative patterns from each modality. Additionally, encoding them in RGB format enables seamless integration with pre-trained convolutional based architectures such as SE-ResNet50, which are optimized for three-channel image inputs. The resulting H×W×3 image was resized to 224×224×3 for model input.

3.5 Squeeze-and-excitation residual network

Squeeze-and-Excitation ResNet50 (SE-ResNet50) is a variant of the ResNet50 architecture enhanced with Squeeze-and-Excitation (SE) blocks, as proposed by Hu et al. [45]. While ResNet50 employs residual learning to address vanishing gradients and degradation issues in deep convolutional networks [46], SE blocks introduce a channel-wise attention mechanism that adaptively recalibrates feature responses. This mechanism allows the network to emphasize informative features and suppress less relevant ones, thereby improving representational power without significantly increasing computational cost. SE-ResNet50 has been shown to enhance performance in various computer vision tasks by combining the benefits of deep residual learning and channel-wise attention.

The input to the model is a scalogram image with dimensions 224×224×3. The model is trained using the Adam optimizer with a learning rate of 0.001. For multiclass classification into three categories (normal, suspicious, pathological), the loss function is based on categorical cross-entropy. To address class imbalance, we apply focal loss, which modifies cross-entropy to focus more on hard-to-classify samples, together with class weighting to reduce bias toward the majority classes. To overcome class imbalance, focal loss and class weight are used. Focal loss is a variant of the cross entropy loss function designed to focus on examples that are difficult to classify [47]. In class imbalance, models tend to focus on majority classes, neglecting harder minority cases. Focal loss mitigates this by down-weighting easy examples and emphasizing difficult ones, improving learning for minority classes. Equation (7) provides the focal loss formula [47]:

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{7}$$

Where p_t is the correct prediction probability for a particular class, α_t is a weighing factor for minority classes (e.g., giving more weight to minority classes), and γ is a parameter that controls how much attention is focused on a difficult instance (the larger the γ, the greater the focus on the difficult example). Focal loss helps the model learn examples from more difficult minority classes by giving more weight to the errors from those examples. In this study, we addressed class imbalance by setting the focal loss parameter α_t according to the computed class weights, while also experimenting with alternative α_t values to assess their impact. The focusing parameter γ was set at 2, as this value has been reported to perform well in various scenarios [48]. Class weighting assigns greater importance to minority classes within the loss calculation, thereby improving the model's sensitivity to underrepresented categories [49].

Class weight is a technique that involves giving more weight to a minority class in a standard loss function such as cross entropy loss [49]. This approach adjusts weights for sparse classes, ensuring the model treats minority class predictions as equally important. Class weights are computed using equation (8) [49].

$$\omega_i = \frac{N}{n_i} \tag{8}$$

Where ω_i is the weight for the i class, N is the total number of samples, and n_i is the number of samples in class i. Less frequent classes receive higher weights to balance the loss function, ensuring the model considers minority classes.

To further improve performance, dropout layers were added to the SE-ResNet50 architecture, and early stopping with a patience of 10 epochs was applied to prevent overfitting. Figure 3 illustrates the architecture of SE-ResNet50 proposed in this study.

Model performance evaluation involves a set of metrics that can measure certain aspects of the model's performance. Some common evaluation metrics used are accuracy, precision, sensitivity, and F1-score [50]. Accuracy shown in equation (9), measures the proportion of the number of correct predictions to the total number of samples [50].

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{9}$$

TP is the true positive (number of correct positive predictions) and FP is the false positive. TN is true negative (number of correct negative predictions), FP is false positive (number of false positive predictions), and FN is false negative (number of false negative predictions). The precision shown in equation (10) measures the proportion of positive predictions that are truly positive.

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{10}$$

Sensitivity measures the model's ability to identify all positive samples. The formula is in (11) [50]:

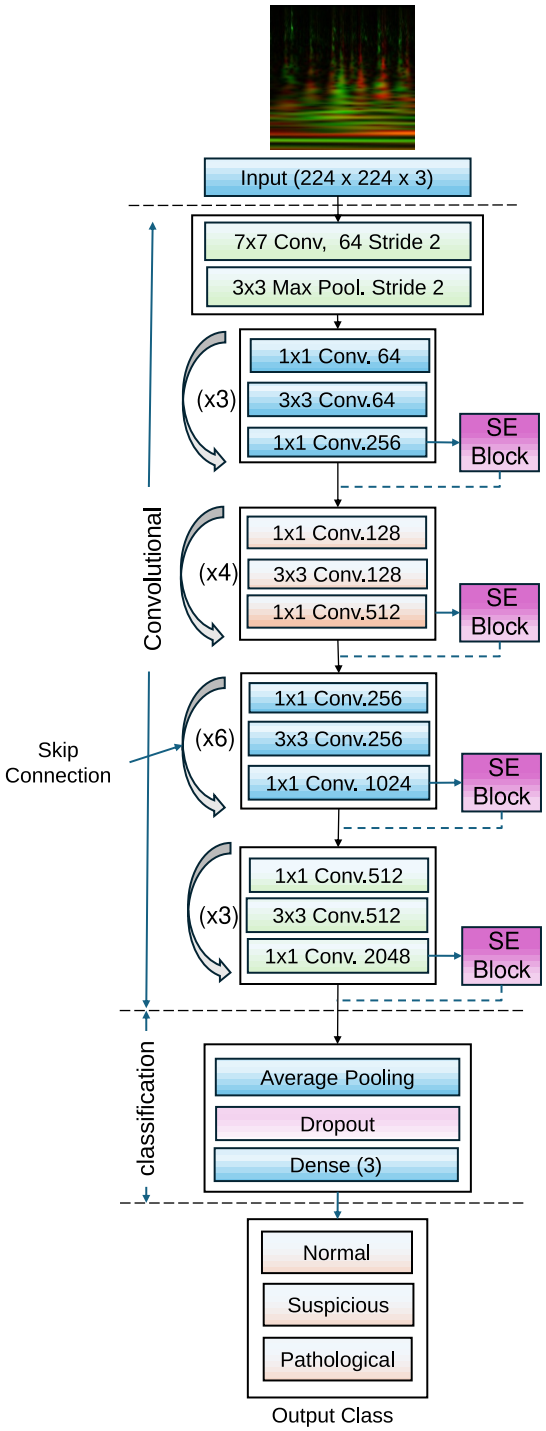

Figure 3: SE-ResNet50 architecture

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

F1-score is the harmonic means of precision and sensitivity, providing a balance between the two. The formula is shown in (12) [50]:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \tag{12}$$

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is a widely used performance metric that evaluates a model's ability to distinguish between classes by measuring the area under the ROC curve. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various classification thresholds. The AUC is computed using the trapezoidal rule as follows (13)(14) [50]:

$$\text{AUC} = \sum_{i=1}^{n-1} (\text{FPR}_{i+1} - \text{FPR}_i) \times \frac{\text{TPR}_{i+1} + \text{TPR}_i}{2} \tag{13}$$

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{14}$$

For multiclass evaluation, we adopted a one-vs-rest approach for each class (normal, suspicious, pathological), computing precision, recall (sensitivity), F1-score, and AUC individually, then averaging them equally to obtain macro-averaged metrics.

4 Results

This section presents the findings from each stage of the research methodology that has been conducted.

4.1 Data selection and annotation analysis results

In the CTG-UHB dataset, annotations are assigned as follows: 1 (normal), 2 (suspicious), 3 (pathological), and -1 (uninterpretable). Table 4 presents the data distribution from the majority voting of nine experts across three evaluation steps. The sample count varies at each step, reflecting differences in expert agreement and data interpretation. The large number of uninterpretable cases in evaluation step 3 indicates that more cases were considered ambiguous or difficult to categorize. Since this study only requires normal, suspicious, and pathological labels, the uninterpretable category (-1) is ignored. A correlation matrix is calculated for each step, as shown in Figure 4 for step 1, excluding the uninterpretable label.

Table 4: Data distribution based on majority votes from nine experts

Label	Criteria	CTG Evaluation		
		Step 1	Step 2	Step 3
1	Normal	296	229	127
2	Suspicious	220	251	153
3	Pathological	30	68	57
-1	Uninterpretable	6	4	215

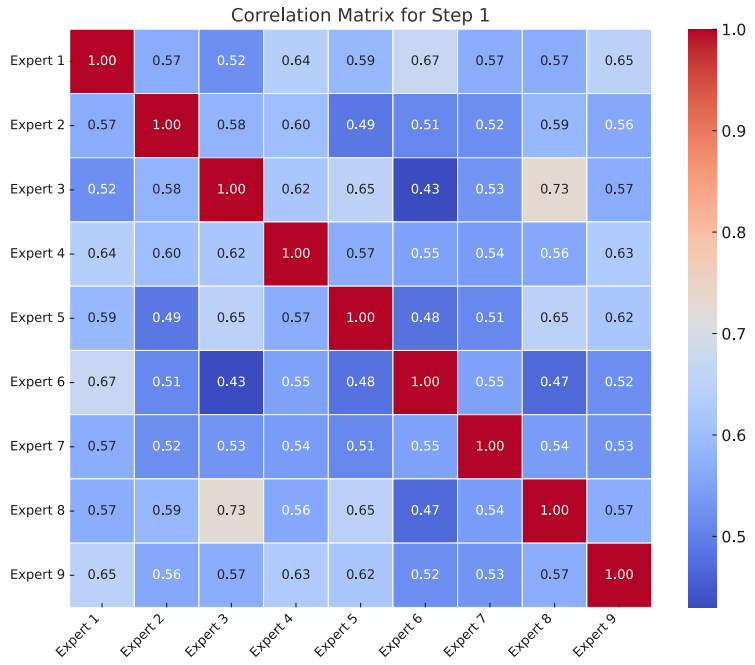

Figure 4: Expert correlation matrix step 1 excluding uninterpretable label (-1)

To assess the consistency of expert annotations, we evaluated inter-rater agreement using both average pairwise correlation and Fleiss' kappa across the three annotation steps. The average correlation values were 0.62 (Step 1), 0.59 (Step 2), and 0.54 (Step 3), indicating a moderate degree of similarity in experts' rating patterns. However, Fleiss' kappa scores were notably lower, with values of 0.3 (fair agreement), 0.22 (slight agreement), and 0.185 (slight agreement), respectively. This discrepancy arises because correlation measures the similarity in score trends between experts, while kappa evaluates exact label matches, adjusted for chance agreement. For this reason, CTG data from step 1 will be used in this study because it has a high correlation between experts that shows strong agreement, which means that the labels on this data are more consistent and reliable. Stable training data and minimal noise are essential for machine learning models because it makes it easier for models to find accurate patterns. The labels assigned to the data in evaluation step 1 tend to be uniform. This reduces the risk of ambiguity in the training data, so the model can learn from clearer patterns and not be affected by inconsistent labels.

4.2 Data preprocessing and feature extraction results

Both FHR and UC signals, in all categories (normal, suspicious, and pathological), Both FHR and UC signals, in all categories (normal, suspicious, and pathological), were processed through data cleaning process. Figure 5 shows one of the results of the data cleaning stage in the suspicious category. It shows how the signals are efficiently refined by the data cleaning process, which eliminates artifacts and irregularities that can obstruct precise analysis.

Following data cleaning, augmentation was performed to address class imbalance among the normal, suspicious, and pathological categories. After applying the time-domain augmentation techniques consisting of cyclic temporal shifting, Gaussian noise injection, and segmented Gaussian noise injection, the dataset size increased from 522 to 1,452 samples. The final distribution consisted of 636 normal, 606 suspicious, and 210 pathological cases. This process substantially reduced class imbalance, with the pathological class, originally the smallest, experiencing the largest relative growth due to the application of additional shift intervals and multiple noise levels.

After the dataset was cleaned and augmented, each FHR and UC signal was individually transformed into scalogram images using Continuous Wavelet Transform (CWT) with different β parameters (β = 50 for FHR and β = 100 for UC) to optimize time–frequency resolution for each signal type. Examples of the resulting scalograms for both signals are shown in Figure 6. These two scalograms were then combined into an RGB-encoded image by assigning the FHR scalogram to the red channel, the UC scalogram to the green channel, and leaving the blue channel empty, thereby preserving the modality-specific features while enabling simultaneous processing by the deep learning model. An example of the resulting RGB-encoded representation is shown in Figure 7. The resulting RGB-encoded image was then resized to 224×224×3 before being fed into the model.

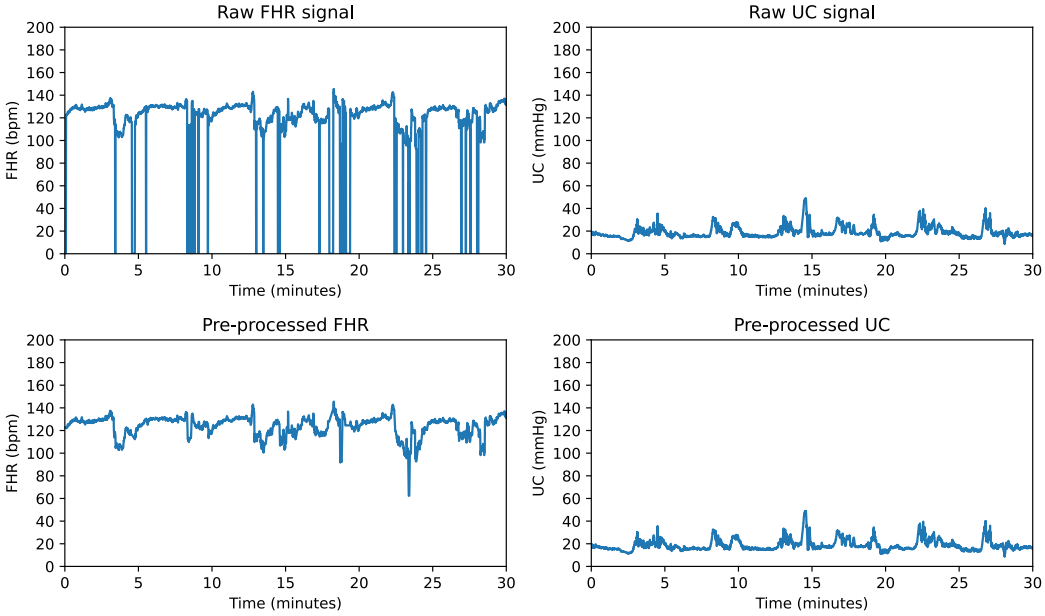

Figure 5: Example of cleaned FHR and UC signals from the suspicious class

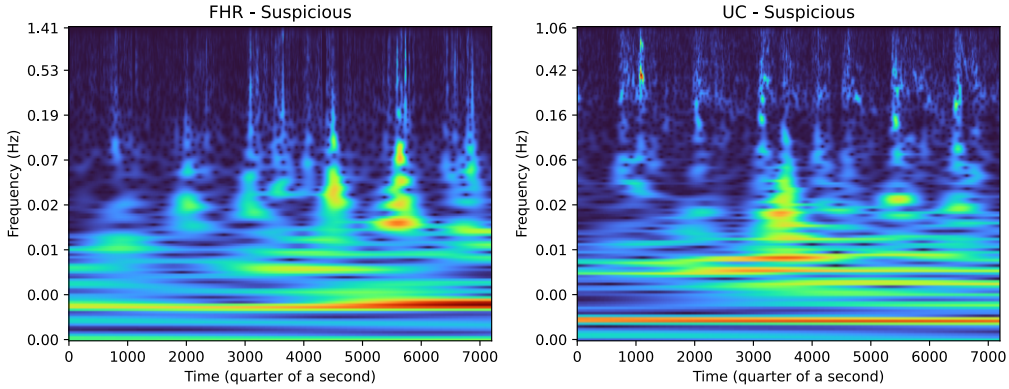

Figure 6: Scalogram representation (a) FHR signal (b) UC signal

4.3 Model training and evaluation results

The experiments were carried out on Google Colab Pro with an NVIDIA GPU (16 GB VRAM) and 62.8 GB RAM, using Python 3.10, PyTorch 2.2.2, and Torchvision 0.17.2. All runs were configured with deterministic settings (random seed = 42) for reproducibility. In the first experiment, the proposed SE-ResNet50 model enhanced with dropout layers was evaluated using stratified five-fold cross-validation. Performance metrics including accuracy, precision, recall, F1-score, and AUC were calculated for each fold, with the mean values reported to assess the model's overall effectiveness as shown in Table 5 and Figure 8.

Across the five folds, the model achieved an average F1-score of 0.6015 ± 0.0718, with Fold 1 yielding the highest F1-score of 0.7073. This fold was then evaluated on the independent test set, achieving an F1-score of 0.7267 and an AUC of 0.7489. Figure 9 and Figure 10 present the per-class performance metrics and the confusion matrix of the final test set, respectively, while Table 6 summarizes the precision, recall, and F1-score for each class.

The performance of the SE-ResNet-50 model was compared with baseline models using CTG evaluation step 1 data, with the same number of epochs (100), early stopping (patience = 10), dropout rate (0.6), and learning rate (0.001), all initialized with ImageNet-pretrained weights. Table 7 presents the mean ± standard deviation results from five-fold cross-validation for all models. The overall performance on the final test set is illustrated in Figure 11, while Figures 12 and 13 show the relationship between model complexity (number of parameters) and final test set performance, as well as the trade-off between inference time and final test set F1-score.

Table 5: Performance of the proposed model across five folds

Fold	Accuracy	Precision	Recall	F1	AUC	
1	0.6821	0.6495	0.7795	0.7073	0.7963	
2	0.5375	0.5284	0.5258	0.5142	0.6700	
3	0.5857	0.6175	0.6034	0.5966	0.7266	
4	0.6091	0.6132	0.6089	0.5926	0.7302	
5	0.6301	0.6510	0.6206	0.6268	0.8216	
Mean ± Std	0.6089 ± 0.0535	0.6119 ± 0.0499	0.6276 ± 0.0928	0.6075 ± 0.0696	0.7489 ± 0.0604	

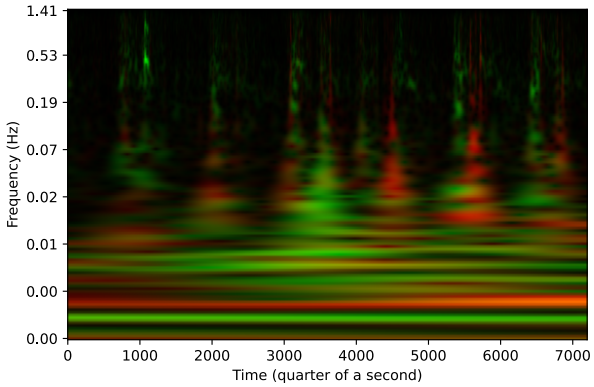

Figure 7: RGB-encoded representation

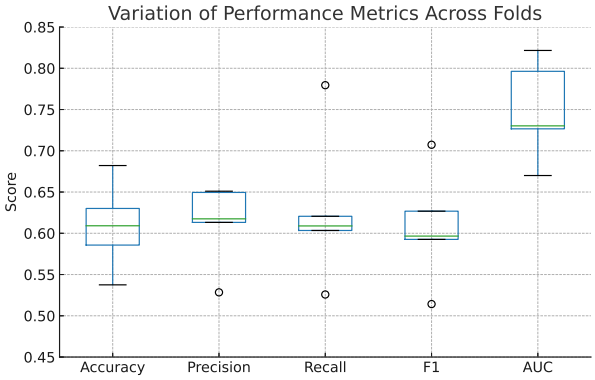

Figure 8: Variation of performance metrics across folds

Table 6: Per-class precision, recall, and F1-score for the final test set

Class	**Precision**	**Recall**	**F1-score**
Normal	0.5833	0.6364	0.6087
Suspicious	0.6000	0.5455	0.5714
Pathologic	1.0000	1.0000	1.0000

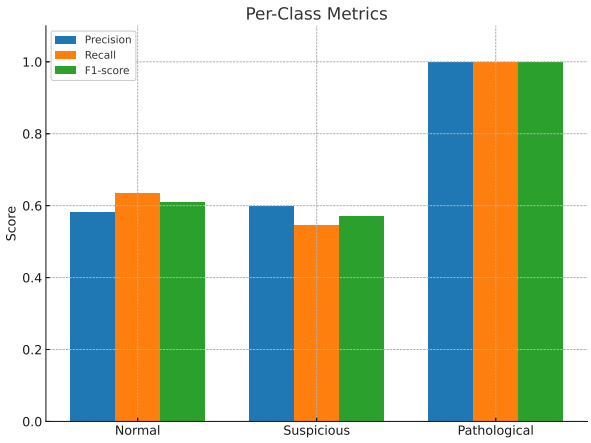

Figure 9: Per-class performance metrics

To evaluate the effect of key hyperparameters on model performance, an ablation study was conducted by varying the learning rate, batch size, and dropout rate while keeping other training settings constant. The learning rates tested were 0.001 and 0.0001, batch sizes were set to 8, 16, and 32, and dropout rates were tested at 0.3, 0.5, and 0.6. All experiments used the same number of epochs with early stopping (patience = 10) to ensure a fair comparison. Table 8 summarizes the 5-fold cross-validation mean ± standard deviation results for F1-score and AUC, with the best results highlighted in bold.

An ablation study compared cross-entropy and focal loss with various class weight settings, using a fixed learning rate of 0.001, dropout rate of 0.6, batch size of 32, and training for up to 100 epochs with early stopping (patience = 10). Focal loss parameters (γ and α) were varied as shown in Table 9, which reports 5-fold cross-validation mean ± standard deviation for F1-score and AUC, with best results in bold.

Grad-CAM interpretability analysis was performed to highlight the regions most influential in the model predictions. Figure 14 presents the Grad-CAM visualizations of representative scalogram images from the normal, suspicious, and pathological classes using the complete scalogram for each case. To provide a more detailed view, Figure 15 shows the Grad-CAM visualization of a representative suspicious case with separated FHR and UC channels, enabling observation of class-specific attention patterns for each signal type.

Table 7: Mean ± std cross-validation performance of deep learning models

Model	Accuracy	Precision	Recall	F1-Score	AUC
CNN	0.5418 ± 0.0478	0.5248 ± 0.1031	0.4881 ± 0.0673	0.4895 ± 0.0786	0.7026 ± 0.0449
EfficientNetB0	0.5472 ± 0.0202	0.5953 ± 0.0744	0.5155 ± 0.0354	0.5129 ± 0.0198	0.6880 ± 0.0638
MobileNetV2	**0.6313 ± 0.0936**	**0.6509 ± 0.0660**	0.5781 ± 0.0635	0.5933 ± 0.0623	0.7314 ± 0.1332
ResNet18	0.5800 ± 0.0274	0.5713 ± 0.1012	0.5607 ± 0.0943	0.5335 ± 0.0798	0.7347 ± 0.0497
ResNet50	0.6044 ± 0.0459	0.6405 ± 0.0704	0.5749 ± 0.0520	0.5913 ± 0.0576	0.7185 ± 0.0299
SE-ResNet50 (Proposed)	0.5888 ± 0.0511	0.6276 ± 0.0691	**0.6040 ± 0.0847**	**0.6015 ± 0.0718**	**0.7530 ± 0.0423**

Table 8: Comparison of hyperparameter tuning results (5-fold mean ± std) for F1-score and AUC

Learning Rate	Batch Size	Dropout	F1-score	AUC
0.001	32	0.6	**0.6542 ± 0.0629**	0.7366 ± 0.0769
	32	0.5	0.6209 ± 0.0440	0.7251 ± 0.0834
	32	0.3	0.5794 ± 0.0475	0.7116 ± 0.0930
0.0001	32	0.6	0.4605 ± 0.0774	0.6704 ± 0.0573
0.001	16	0.6	0.6360 ± 0.0523	**0.7902 ± 0.0213**
	8	0.6	0.6177 ± 0.0252	0.7512 ± 0.0455

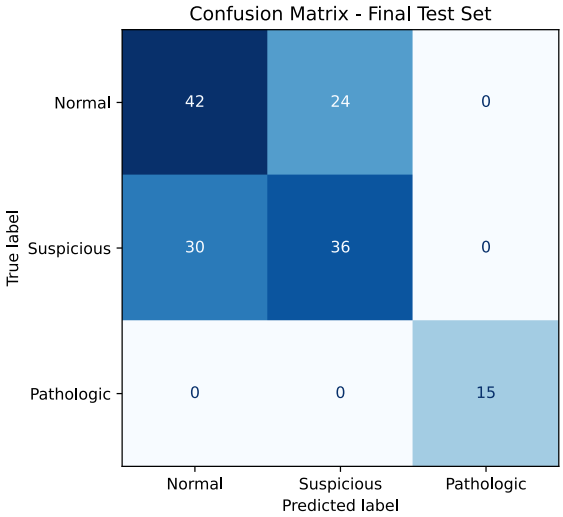

Figure 10: Confusion matrix

5 Discussion

This section discusses the experimental results, beginning with an evaluation of annotation consistency through Pearson correlation and Fleiss' kappa, followed by the performance analysis of the proposed model across cross-validation folds and the independent test set. The results are further interpreted in comparison with baseline models, along with per-class performance, to provide insights into the strengths and limitations of the approach.

The quality of expert annotations was assessed using both Pearson correlation and Fleiss' kappa across three annotation steps. The average pairwise Pearson correlation values were 0.62, 0.59, and 0.54 for Step 1, Step 2, and Step 3, respectively, indicating moderate similarity in scoring trends among experts, with Step 1 showing the highest consistency. However, Fleiss' kappa values were considerably lower. The scores were 0.3 (fair agreement), 0.22 (slight agreement), and 0.185 (slight agreement), revealing that exact label agreement, adjusted for chance, was limited, particularly in Step 2 and Step 3. This lower inter-rater reliability is consistent with the model's final test performance (Figure 9) and the confusion matrix (Figure 10), where the normal and suspicious classes achieved F1-scores of only 0.6087 and 0.5714, respectively. These results suggest that the ambiguity in expert annotations for these categories likely contributed to the model's reduced classification performance.

To assess the stability of performance metrics across folds, we calculated the coefficient of variation (CV), defined as the ratio of the standard deviation to the mean, expressed as a percentage. Based on Table 5, AUC (CV = 8.07%), precision (CV = 8.15%), and accuracy (CV = 8.79%) had low variation, indicating stable discriminative capability and consistent predictive accuracy. In contrast, F1-score (CV = 11.46%) and recall (CV = 14.79%) showed moderate variation, with recall exhibiting the largest fluctuation, particularly due to a notable drop in Fold 2. The boxplot further illustrates this pattern: Precision and AUC have compact distributions with minimal spread, while Recall and F1 display wider interquartile ranges and several low outliers. This variability is likely linked to limited generalization on certain data subsets, potentially influenced by label inconsistencies. Conducting an error analysis of high-confidence misclassifications could help detect and correct such labeling issues, while enhancing feature representation or applying additional regularization may improve robustness and stabilize recall and F1 performance

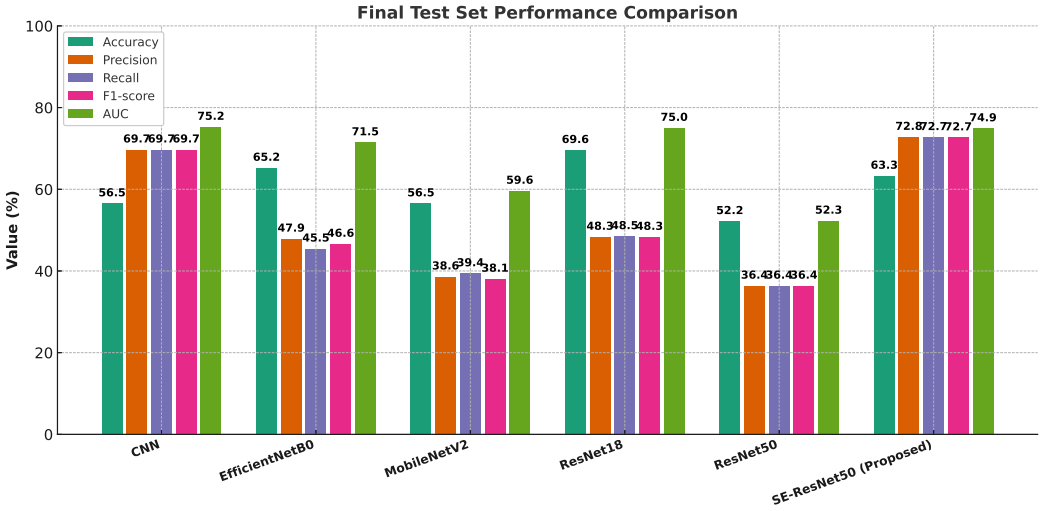

Figure 11: Final test set performance comparison

Table 9: Ablation study results for focal loss and class weight configurations (5-fold mean ± std)

Model	Focal Loss	Class Weight	Gamma	Alpha	F1-score	AUC
1	False	False	–	–	**0.6542 ± 0.0629**	0.7366 ± 0.0769
2	False	True	–	–	0.6506 ± 0.0382	0.7909 ± 0.0197
3	True	True	2	[1.0, 1.0, 1.0]	0.6015 ± 0.0718	0.7530 ± 0.0423
4	True	False	2	[0.8, 2.2, 1.4]	0.5902 ± 0.0910	0.7281 ± 0.0343
5	True	False	2	[0.8, 1.2, 2.0]	0.6296 ± 0.0638	0.7486 ± 0.0323
6	True	False	1	[0.8, 1.0, 2.0]	0.6163 ± 0.0983	**0.7669 ± 0.0691**

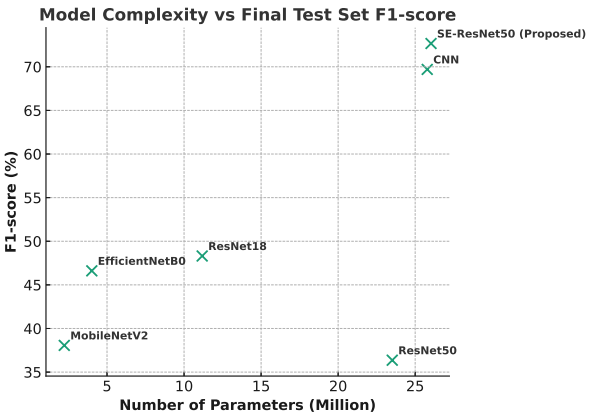

Figure 12: Parameter vs performance

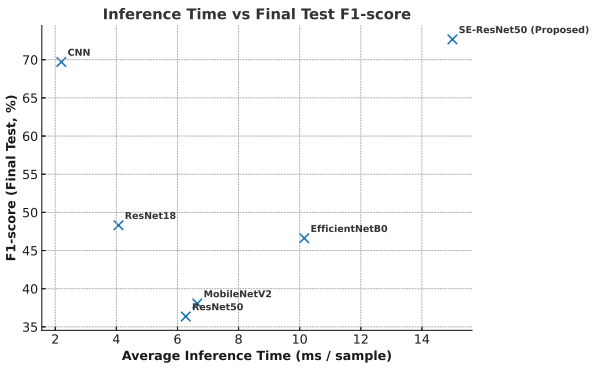

Figure 13: Inference time vs performance

across folds.

In the final test set, the model achieved perfect precision, recall, and F1-score for the pathological class, without misclassifications. Performance was lower for normal (precision 0.5833, recall 0.6364) and suspicious (precision 0.60, recall 0.5455), with most errors occurring between these two classes. This pattern suggests overlapping features or unclear decision boundaries, indicating the need for improved feature representation or targeted data augmentation.

Compared to previous studies, the proposed model achieved lower overall accuracy (0.6327) and macro-average F1-score (0.7267) than high-performing methods such as Liang et al. (accuracy= 0.9515, F1-score=0.9520) [34] but was comparable to Ogasawara et al. (F1-score=0.67) [35] and Saini et al. (accuracy 0.70) [36]. Although prior work primarily addressed binary classification (normal vs. abnormal), our model tackled a more challeng-

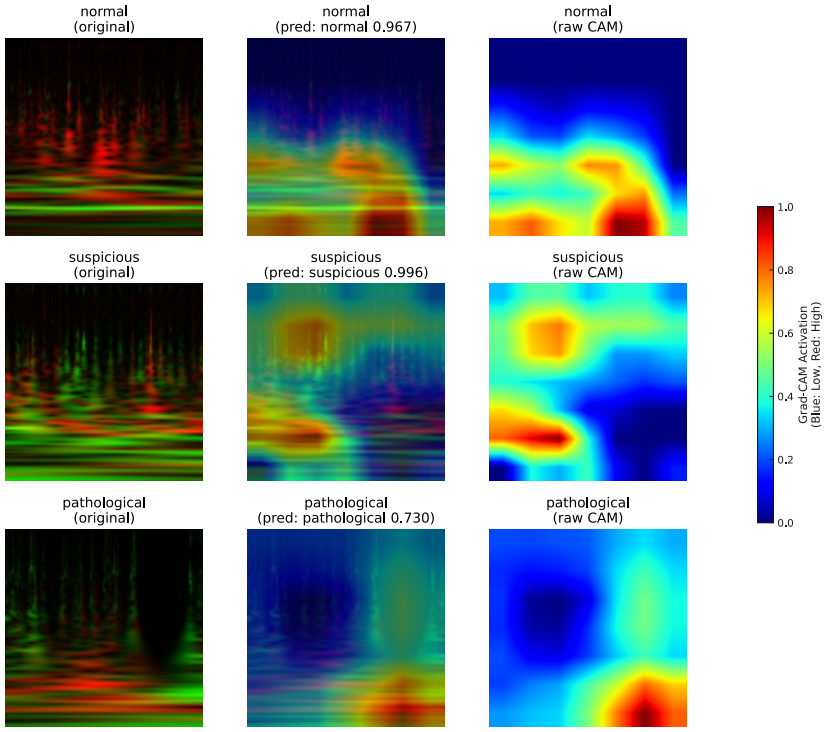

Figure 14: Grad-CAM visualizations of representative normal, suspicious, and pathological scalograms (complete images)
.

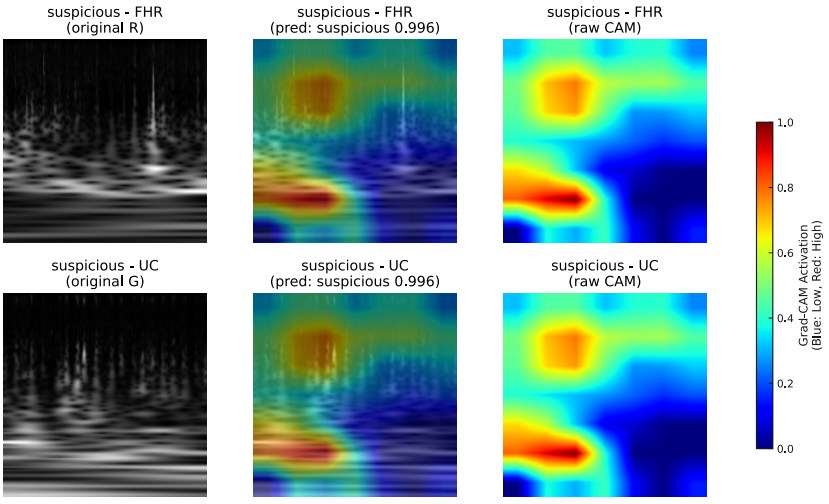

Figure 15: Grad-CAM visualizations for suspicious scalograms (separated channels: FHR and UC)

ing three-class problem (normal, suspicious, pathological) and demonstrated a key strength by achieving perfect precision, recall, and F1 score for the pathological class, highlighting its robustness in identifying high-risk cases.

The proposed SE-ResNet50 outperformed all baseline models in the final test set, achieving the highest F1-score (0.727), recall (0.727), precision (0.728), and AUC (0.749), demonstrating balanced performance across metrics. Although its accuracy (0.633) was slightly lower than CNN and MobileNetV2, it offered a better generalization for the three-class classification problem. The complexity analysis of the model showed that SE-ResNet50 delivered the best

performance without significantly increasing the parameter count compared to ResNet50. From the model complexity vs. the the final test F1 score graph, SE-ResNet50 achieved the highest F1-score (0.727) with ∼24 million parameters, slightly higher than CNN (0.697) but significantly outperforming ResNet50 despite having the same complexity. This indicates that the addition of Squeeze-and-Excitation blocks effectively enhances feature representation without substantially increasing the parameter load. Although it recorded the longest inference time (15 ms/sample), this trade-off is justified by its substantially higher classification performance, making it suitable for applications prioritizing detection accuracy over speed.

We also conducted a statistical analysis to evaluate whether the performance improvements of the proposed SE-ResNet50 over baseline models were statistically significant. The analysis used the F1-scores obtained from each fold of the 5-fold cross-validation for SE-ResNet50 and each baseline model. Two tests were performed: the paired t-test, which assumes normality in the differences between paired samples, and the Wilcoxon signed-rank test, a non-parametric alternative that does not require the normality assumption. The results showed that SE-ResNet50 achieved the highest mean F1-score (0.6015) among all models. The paired t-test indicated that the improvements over CNN (mean F1 = 0.4895, p = 0.0036) and EfficientNetB0 (mean F1 = 0.5129, p = 0.0428) were statistically significant at the 0.05 level. However, the Wilcoxon test did not confirm significance for these comparisons ($p > 0.05$), suggesting that the differences may not be robust under non-parametric assumptions. Comparisons with MobileNetV2, ResNet18 and ResNet50 yielded p values greater than 0.05 in both tests, indicating that there were no statistically significant differences in those cases. These findings suggest that SE-ResNet50 offers a statistically supported improvement over CNN and EfficientNetB0 in terms of F1-score, while its advantage over the other per-class, although numerically higher, was not statistically significant given the current sample size.

The ablation study results in Table 9 indicate that the use of focal loss and class weight yields varying effects on model performance. Statistical analysis using one-way ANOVA revealed no significant differences in F1-score across all model configurations ($p = 0.8496$), indicating that neither class weight nor focal loss consistently improved the balance between precision and recall compared to the baseline. In contrast, AUC differences were statistically significant ($p = 0.0389$), and post-hoc pairwise comparisons with Bonferroni correction identified a significant improvement for the configuration without focal loss but with class weight (Model 2) over the focal loss configuration with gamma = 2, alpha = [0.8, 1.2, 2.0] (Model 5, $p < 0.05$). This suggests that class weight is more effective in enhancing the model's discriminative ability than focal loss in this dataset.

Overall, the results indicate that class weight provides a statistically supported benefit for improving AUC, while focal loss alone can achieve competitive AUC with tuned parameters but does not surpass the best class weight configuration. For F1-score, neither method yields statistically significant improvements, suggesting that alternative optimization strategies may be required when this metric is the primary objective.

In the complete RGB Grad-CAM visualizations, distinct attention patterns were observed across the normal, suspicious, and pathological classes. For the normal class, high confidence predictions were associated with well-localized warm regions concentrated in the lower central portion of the scalogram. These hotspots likely represent stable segments of the fetal heart rate (FHR) signal and consistent uterine contraction (UC) patterns without abnormal variability. The focused activation suggests that the model identifies characteristic steady-state features that are typical of normal CTG recordings.

In the suspicious class, the activation maps showed broader but still localized warm areas, particularly in temporal regions where contraction peaks aligned with subtle changes in the FHR baseline. These regions correspond to clinical patterns such as mild or intermittent decelerations, which may not meet the pathological threshold but still warrant closer monitoring. The model's attention in this class indicates recognition of moderate deviations from normal patterns that can signal potential fetal distress. As an additional observation, visual inspection revealed that in some cases, the suspicious class shared overlapping activation regions with the normal class, particularly in the lower central scalogram areas corresponding to steady FHR segments. This overlap may contribute to the model's misclassification between these two classes, as mild deviations in suspicious cases can resemble normal patterns in both spatial location and intensity of activations.

For the pathological class, the Grad-CAM heatmaps displayed more diffuse and widespread activations across the scalogram, with concentrated warm regions in the lower central area overlapping contraction periods and pronounced FHR fluctuations. Such activation patterns are consistent with severe decelerations, abnormal variability, or prolonged recovery times following contractions, which align with clinical definitions of pathological CTG. The broader distribution of activations suggests that the model considers multiple abnormal signal segments when forming its decision.

The single-channel Grad-CAM visualizations offer a clearer interpretation of the model's attention by isolating the contribution of each physiological signal. This approach allows domain experts to verify whether the features emphasized by the model align with established clinical knowledge for each signal type. For the FHR channel, the visualization reveals how the model responds to baseline stability, variability, and decelerations without interference from other signals. For the UC channel, it highlights the timing, frequency, and intensity of contractions as perceived by the model. By separating these channels, it becomes possible to determine whether the model's decision

is predominantly influenced by FHR patterns, UC activity, or a combination of both.

In the suspicious case, the FHR channel Grad-CAM displayed concentrated warm regions on segments with subtle baseline shifts and mild decelerations. These features are consistent with early signs of potential fetal compromise, even though they may not fulfill the criteria for pathological classification. The UC channel Grad-CAM showed activations primarily on contraction peaks, indicating that the model incorporates the temporal context of uterine activity when interpreting changes in FHR.

When compared to the complete RGB Grad-CAM for the same case, the single-channel visualizations make the source of the model's attention more explicit. The RGB visualization showed broader activation in regions where contraction peaks coincided with mild FHR changes, suggesting that the model leverages multi-signal interactions. The single-channel analysis confirmed that each signal contains distinct features relevant to the prediction, while the combined RGB image captures their integration into a clinically meaningful temporal relationship.

Grad-CAM visualizations showed that the model attends to physiologically meaningful regions, with RGB maps capturing multi-signal interactions and single-channel maps revealing distinct contributions from each signal. These results confirm that the model integrates relevant temporal relationships for prediction. However, the classification of suspicious cases remains challenging due to overlapping patterns with normal cases, which may limit the separation of decision boundaries. The proposed SE-ResNet-50, which already incorporates channel-wise attention through Squeeze-and-Excitation blocks, improved feature representation but did not fully resolve the overlap, indicating that additional spatial or hybrid attention mechanisms may be needed to better isolate features unique to suspicious cases.

Although the model demonstrated strong performance, particularly in detecting pathological cases, there are still areas for improvement. First, the selection of Continuous Wavelet Transform (CWT) parameters, was based on literature rather than exhaustive empirical tuning, leaving room for parameter optimization in future studies. Second, label quality in CTG datasets can be inconsistent due to subjective interpretation, which may contribute to misclassification. Addressing this issue could involve multi-expert consensus labeling or incorporating external datasets to enhance generalizability. Future work could also explore semi-supervised learning or label-noise–robust learning approaches to better handle label variability, as well as investigate advanced feature representation methods to refine classification boundaries and improve the model's ability to distinguish among all fetal state categories

From a clinical perspective, the model could support prenatal care workflows by providing real-time decision support in identifying suspicious or pathological patterns, enabling earlier and more targeted interventions. Such integration has the potential to improve decision-making efficiency for obstetricians and midwives, ultimately enhancing maternal and fetal outcomes.

6 Conclusion

The proposed SE-ResNet-50 model applied to scalogram representations of cardiotocographic signals achieved strong performance in multiclass fetal state classification, particularly in detecting pathological cases under imbalanced conditions. These results support the study's hypothesis that incorporating Squeeze-and-Excitation mechanisms improves classification performance by leveraging channel interdependencies and time–frequency features. The best configuration reached an F1-score of 0.7267 and an AUC of 0.7489 on the independent test set.

Future work will focus on improving classification in borderline suspicious cases, strengthening generalizability through larger and more diverse datasets, and exploring integration into clinical workflows to provide real-time decision support for earlier and more targeted interventions.

Acknowledgement

We sincerely thank Dr. Ari Waluyo, Sp.OG, for his valuable insights and feedback, which greatly contributed to the development of this research. We also appreciate the support from the Faculty of Engineering, Universitas Indonesia, through the Seed Funding research grant program (NKB-2614/UN2.F4.D/PPM.00.00/2023).

References

[1] W. H. Organization, *Trends in maternal mortality 2000 to 2020: estimates by WHO, UNICEF, UNFPA, World Bank Group and UNDESA/Population Division*. World Health Organization, 2023.

[2] W. H. Organization, *Improving maternal and newborn health and survival and reducing stillbirth: progress report 2023*. World Health Organization, 2023.

[3] T. Ermias Geltore and D. Laloto Anore, "The Impact of Antenatal Care in Maternal and Perinatal Health," in *Empowering Midwives and Obstetric Nurses*, IntechOpen, 2021, ch. 8, https://doi.org/10.5772/intechopen.98668.

[4] Z. Anwar et al., "Association of Intrapartum CTG with Fetomaternal Outcome," *Pakistan Journal of Medical & Health Sciences*, vol. 16, no. 03, pp. 1045–1045, 2022, https://doi.org/10.53350/pjmhs22164666.

[5] S. T. Nabipour Hosseini, F. Abbasalizadeh, S. Abbasalizadeh, S. Mousavi, and P. Amiri, "A comparative study of CTG monitoring one hour before labor in infants born with and without asphyxia," *BMC Pregnancy and Childbirth*, vol. 23,

no. 1, p. 758, 2023, https://doi.org/10.1186/s12884-023-06040-3.

[6] Z. Alfirevic, G. M. Gyte, A. Cuthbert, and D. Devane, "Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour," *Cochrane Database of Systematic Reviews*, no. 2, 2017. https://doi.org/10.1002/14651858.CD006066.pub3.

[7] A. Sharma, Ed., *Labour Room Emergencies*. Springer, 2020.

[8] O. Nzelu, E. Chandraharan, and S. Pereira, "Human factors: the dirty dozen in CTG misinterpretation," *Global Journal of Reproductive Medicine*, vol. 6, no. 2, p. 555683, 2018. https://doi.org/10.19080/GJORM.2018.06.555683.

[9] Y. Zhang, Q. Zhou, and X. Li, "The Advent of a New Era of Antenatal Cardiotocography," *Maternal-Fetal Medicine*, vol. 4, no. 2, pp. 93–94, 2022. https://doi.org/10.1097/FM9.0000000000000144.

[10] P. Garcia-Canadilla, S. Sanchez-Martinez, F. Crispi, and B. Bijnens, "Machine Learning in Fetal Cardiology: What to Expect," *Fetal Diagnosis and Therapy*, vol. 47, no. 5, pp. 363–372, 2020. https://doi.org/10.1159/000505021.

[11] G. Hever, L. Cohen, M. F. O'Connor, I. Matot, B. Lerner, and Y. Bitan, "Machine learning applied to multi-sensor information to reduce false alarm rate in the ICU," *Journal of Clinical Monitoring and Computing*, vol. 34, no. 2, pp. 339–352, Apr. 2020. https://doi.org/10.1007/s10877-019-00307-x.

[12] W. H. Organization, *WHO Recommendations Non-clinical Interventions to Reduce Unnecessary Caesarean Sections*. World Health Organization, 2018.

[13] J. O'Heney, S. McAllister, M. Maresh, and M. Blott, "Fetal monitoring in labour: summary and update of NICE guidance," *BMJ*, vol. 379, p. o2854, Dec. 2022. https://doi.org/10.1136/bmj.o2854.

[14] E. Chandraharan, "Updated NICE Cardiotocograph (CTG) guideline: Is it suspicious or pathological," *Journal of Clinical Medicine and Surgery*, vol. 3, no. 2, p. 1129, 2023.

[15] T. M. Kadarina, Basari, and D. Gunawan, "ML-Based Interpretation of Cardiotocography Data: Current State and Future Research," in *2023 International Conference of Computer Science and Information Technology (ICOSNIKOM)*, 10–11 Nov. 2023, pp. 1–6. https://doi.org/10.1109/ICoSNIKOM60230.2023.10364517.

[16] 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, Apr. 2024. https://doi.org/10.1016/j.compbiomed.2024.108220.

[17] P. Fergus, C. Chalmers, C. C. Montanez, D. Reilly, P. Lisboa, and B. Pineles, "Modelling Segmented Cardiotocography Time-Series Signals Using One-Dimensional Convolutional Neural Networks for the Early Detection of Abnormal Birth Outcomes," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2021. https://doi.org/10.1109/TETCI.2020.3020061.

[18] D. Panda, D. Panda, S. R. Dash, and S. Parida, "Extreme learning machines with feature selection using GA for effective prediction of fetal heart disease: A novel approach," *Informatica (Slovenia)*, 2021. https://doi.org/10.31449/INF.V45I3.3223.

[19] P. Riskyana Dewi Intan, M. A. Anwar Ma'sum, N. Alfiany, W. Jatmiko, A. Kekalih, and A. Bustamam, "Ensemble learning versus deep learning for Hypoxia detection in CTG signal," in *2019 International Workshop on Big Data and Information Security (IWBIS)*, 2019. https://doi.org/10.1109/IWBIS.2019.8935796.

[20] X. Kang et al., "Prediction of Delivery Mode from Fetal Heart Rate and Electronic Medical Records Using Machine Learning," in *2022 Computing in Cardiology (CinC)*, 2022, vol. 498, pp. 1–4. https://doi.org/10.22489/CinC.2022.116.

[21] H. Allahem and S. Sampalli, "Automated labour detection framework to monitor pregnant women with a high risk of premature labour using machine learning and deep learning," *Informatics in Medicine Unlocked*, vol. 28, p. 100771, 2022. https://doi.org/10.1016/j.imu.2021.100771.

[22] D. Campos and J. Bernardes, "Cardiotocography," *UCI Machine Learning Repository*, 2010. https://doi.org/10.24432/C51S4N.

[23] B. Amin, M. Gamal, A. A., I. M. El-Henawy, and K. Mahfouz, "Classifying Cardiotocography Data based on Rough Neural Network," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 8, 2019. https://doi.org/10.14569/ijacsa.2019.0100846.

[24] M. Chen and Z. Yin, "Classification of cardiotocography based on the apriori algorithm and multi-model ensemble classifier," *Frontiers in Cell and Developmental Biology*, vol. 10, p. 888859, May 2022. https://doi.org/10.3389/fcell.2022.888859.

[25] A. K. Pradhan, J. K. Rout, A. B. Maharana, B. K. Balabantaray, and N. K. Ray, "A Machine Learning Approach for the Prediction of Fetal Health using CTG," in *2021 19th OITS International Conference on Information Technology (OCIT)*, 2021. https://doi.org/10.1109/OCIT53463.2021.00056.

[26] R. Chinnaiyan and D. Stalin Alex, "Early Analysis and Prediction of Fetal Abnormalities Using Machine Learning Classifiers," in *2021 2nd International Conference on Smart Electronics and Communication (ICOSEC)*, 2021. https://doi.org/10.1109/ICOSEC51865.2021.9591828.

[27] M. T. Alam et al., "Comparative Analysis of Different Efficient Machine Learning Methods for Fetal Health Classification," *Applied Bionics and Biomechanics*, vol. 2022, p. 6321884, 2022. https://doi.org/10.1155/2022/6321884.

[28] A. Ilham, I. N.Istiqomah, A. T. A. Nugroho, and S. P. Hadi, Fetal Health Risk Classification using Important Feature Selection and CART Model on Cardiotocography Data,"*Informatica*, vol. 49, no. 1, pp. 1–12, 2025. https://doi.org/10.31449/inf.v49i1.5658.

[29] Y. D. Daydulo, B. L. Thamineni, H. K. Dasari, and G. T. Aboye, "Deep learning based fetal distress detection from time frequency representation of cardiotocogram signal using Morse wavelet: research study," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, p. 329, Dec. 2022. https://doi.org/10.1186/s12911-022-02068-1.

[30] M. A. Ma'sum, P. R. D. Intan, W. Jatmiko, A. A. Krisnadhi, N. A. Setiawan, and I. M. A. D. Suarjaya, "Improving Deep Learning Classifier for Fetus Hypoxia Detection in Cardiotocography Signal," in *2019 International Workshop on Big Data and Information Security (IWBIS)*, 2019, pp. 51–56. https://doi.org/10.1109/IWBIS.2019.8935835.

[31] S. Magesh and P. S. Rajakumar, "Ensemble feature extraction-based prediction of fetal arrhythmia using cardiotocographic signals," *Measurement: Sensors*, 2023. https://doi.org/10.1016/j.measen.2022.100631.

[32] K. Bhogal, "Focus on cardiotocography: Intrapartum monitoring of uterine contractions," *British Journal of Midwifery*, vol. 25, no. 8, pp. 491–497, 2017. https://doi.org/10.12968/bjom.2017.25.8.491.

[33] R. Zeng, Y. Lu, S. Long, C. Wang, and J. Bai, "Cardiotocography signal abnormality classification using time-frequency features and Ensemble Cost-sensitive SVM classifier," *Computers in Biology and Medicine*, vol. 130, p. 104218, Mar. 2021. https://doi.org/10.1016/j.compbiomed.2021.104218.

[34] H. Liang and Y. Lu, "A CNN-RNN unified framework for intrapartum cardiotocograph classification," *Computer Methods and Programs in Biomedicine*, vol. 229, p. 107300, Feb. 2023. https://doi.org/10.1016/j.cmpb.2022.107300.

[35] J. Ogasawara et al., "Deep neural network-based classification of cardiotocograms outperformed conventional algorithms," *Scientific Reports*, vol. 11, no. 1, p. 13367, Jun. 2021. https://doi.org/10.1038/s41598-021-92805-9.

[36] H. D. Singh, M. Saini, and J. Kaur, "Fetal distress classification with deep convolutional neural network," *Current Women's Health Reviews*, vol. 17, no. 1, pp. 60–73, 2021.

[37] V. Chudacek et al., "Open access intrapartum CTG database," *BMC Pregnancy and Childbirth*, vol. 14, p. 16, Jan. 2014. https://doi.org/10.1186/1471-2393-14-16.

[38] L. Hruban et al., "Agreement on intrapartum cardiotocogram recordings between expert obstetricians," *Journal of Evaluation in Clinical Practice*, vol. 21, no. 4, pp. 694–702, 2015.

[39] A. M. Davani, M. Díaz, and V. Prabhakaran, "Dealing with disagreements: Looking beyond the majority vote in subjective annotations," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 92–110, 2022.

[40] S. Das, S. M. Obaidullah, K. C. Santosh, K. Roy, and C. K. Saha, "Cardiotocograph-based labor stage classification from uterine contraction pressure during ante-partum and intra-partum period: a fuzzy theoretic approach," *Health Information Science and Systems*, vol. 8, no. 1, p. 16, Dec. 2020. https://doi.org/10.1007/s13755-020-00107-7.

[41] S. Mallat, *A Wavelet Tour of Signal Processing*. Elsevier, 1999.

[42] N. Al Bassam, V. Ramachandran, and S. Eratt Parameswaran, "Wavelet Theory and Application in Communication and Signal Processing," in *Wavelet Theory*, M. Somayeh, Ed. Rijeka: IntechOpen, 2021, ch. 3. https://doi.org/10.5772/intechopen.95047.

[43] J. M. Lilly and S. C. Olhede, "Higher-order properties of analytic wavelets," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 146–160, 2008. https://doi.org/10.1109/TSP.2008.2007607.

[44] Y. H. Byeon, S. B. Pan, and K. C. Kwak, "Intelligent Deep Models Based on Scalograms of Electrocardiogram Signals for Biometrics," *Sensors (Basel)*, vol. 19, no. 4, p. 935, Feb. 2019. https://doi.org/10.3390/s19040935.

[45] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020. https://doi.org/10.1109/TPAMI.2019.2913372.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

[47] T.-Y. Ross and G. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 2980–2988, 2018. https://doi.org/10.1109/TPAMI.2018.2858826.

[48] X. Li, C. Lv, W. Wang, G. Li, L. Yang, and J. Yang, "Generalized focal loss: Towards efficient representation learning for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3139–3153, 2022. https://doi.org/10.1109/TPAMI.2022.3180392.

[49] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009. https://doi.org/10.1109/TKDE.2008.239.

[50] A. Zheng, *Evaluating Machine Learning Models: A Beginner's Guide to Key Concepts and Pitfalls*. O'Reilly Media, 2015.