ITC 1/55
Information Technology and Control
Vol. 55 / No. 1/ 2026
pp. 204-221
DOI 10.5755/j01.itc.55.1.42087

Multiclass Fetal Abnormality Detection Using Ensemble Deep Learning Techniques

Received 2025/07/01 | Accepted after revision 2025/10/05

HOW TO CITE: Ramya, R., Krishnamoorthi, M. (2026). Multiclass Fetal Abnormality Detection Using Ensemble Deep Learning Techniques. *Information Technology and Control*, 55(1), 204-221. https://doi.org/10.5755/j01.itc.55.1.42087

Multiclass Fetal Abnormality Detection Using Ensemble Deep Learning Techniques

R. Ramya*, M. Krishnamoorthi

Department of Information Technology, Dr. N. G. P. Institute of Technology, Coimbatore, 641048, India; e-mail: ramyaramakrishnan4.r@outlook.com

Corresponding author: ramyaramakrishnan4.r@outlook.com

Classifying fetal cardiotocography data is essential in the efficient prenatal risk assessment due to its potential for identifying errors or abnormalities during pregnancy. Traditional fetal heart rate (FHR) analysis frameworks, which unfortunately still rely on manual interpretation of results, subsequently lead to the inefficient use of human resources and sometimes require more time for abnormality detection. With the implementation of Machine Learning (ML) algorithms, automatic analysis and early detection of abnormalities are now possible. The model's performance is directly influenced by the retrieval of features and the optimal management of class imbalance in the dataset. In this regard, we introduce a feature-based innovative strategy for multi-class classification in fetal cardiograph datasets based on feature importance analysis. The proposed model utilizes Random Forest (RF) for feature extraction, which employs two distinct target importance analyses: 1. class imbalance, and 2. class weights. In Phase 1, an artificial neural network and an improved TabNet model were utilised for classifying three classes: Normal, Suspect, and Pathology (NSP), with SMOTE balancing. In Phase 2, we identify the features of classes that contribute to NSP classification, and we consider nine additional features based on class weight for various cardiotography features, such as baseline, ASTV, ALTV, etc. In Phase 2, NSP classification is performed by including class 1-9 features (A, B, C, D, E, AD, DE...) and assigning class weights. Using our proposed ensemble deep learning model, the accuracy of prediction is improved. The RF model retrieves primary features from the fetal cardiograph, and complex relationships among these features enhance the representation of information. The next step is the classification stage, which applies an attention-based deep learning model, TabNet. Due to the nature of the TABNET model in handling tabular data, it can selectively focus on relevant features while ensuring explainability. The proposed model is evaluated using different performance metrics for two novel feature importance analyses. The RF+TabNet+LSTM achieves a maximum accuracy of 97% with SMOTE in NSP target classification (phase 1), while including Class weight in class1-9 features, the model

achieves classification accuracy of 92% (phase II) and proves the importance of features contributing to prediction and classification. All code and the curated dataset for Multiclass Fetal Abnormality Detection are available at https://github.com/rrramyaresea/Multiclass-Fetal-Abnormality-Detection, enabling the reproducibility of our findings.

KEYWORDS: Fetal health assessment, ECG signals, Cardiotography, Multiple multi-class classification, TABNET model.

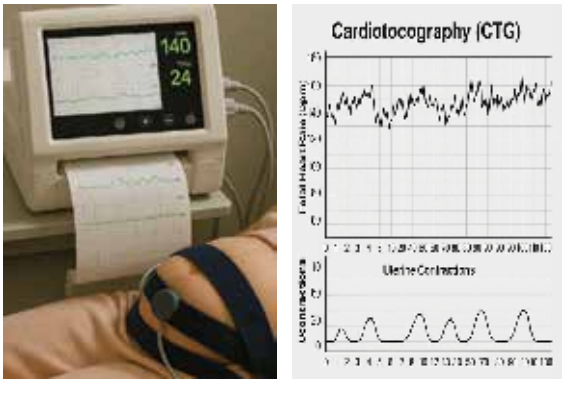
1. Introduction

Cardiotocography (CTG) refers to the visual representation of fetal heart rate (FHR) [15] and uterine contractions, which is monitored by the principle of the fetal neurologic system through its afferent and efferent networks [16]. The cardiotocograph shows a continuous electronic record of the fetal heart rate obtained, which is indicated by an ultrasound transducer placed on the mother's abdomen, and where the second transducer is placed on the mother's abdomen over the uterus and fundus to record [1]. CTG monitoring should never be observed as a substitute for clinical observation or as an excuse for leaving the mother unattended during labour, where Unexpected complications may occur during labour, even in patients with prior signs of risk [23]. Over one-third of maternal deaths and severe life conditions, approximately half of them stillbirths, and a quarter of neonatal deaths result from complications during labour. According to an InterAgency group report, an estimated 340600 stillbirths occurred in the Indian population of 1.4 billion in 2019 [24].

The CTG machine, also known as an Electronic Fetal Monitor, is shown in Figure 1. Two Transducers (either External or Internal) are Capable of performing an ultrasound recording of fetal heart rate. A Tocodynamometer (Toco) is used to measure uterine contractions. The results are either printed or displayed digitally, with the upper portion indicating Fetal Heart Rate (FHR) and the lower, indicating Uterine Contraction (UC). Transducers are placed around the patient's midsection with elastic bands, which help maintain the position. The ultrasound probe is placed on the mother's belly, on the child's back, or chest to record the child's heartbeats via Doppler ultrasound. The Toco Sensor is placed on the top of the uterus to measure pressure changes during contractions. Modification or adaptation of these procedures can allow monitoring for the duration of labour. CTG monitors the extended time between 20 and 40 minutes and identifies the cause for suspicion of being outside the normal range.

In fetal ECG analysis, the use of cardiotocography combined with deep learning is an advanced technique that enhances the accuracy and efficiency of fetal health assessments. It is a prevalent method for monitoring fetal well-being by analysing FHR and uterine contraction signals [28]. Deep learning offers a promising solution by automating the analysis and providing more reliable results, where those responses explore various deep learning approaches in fetal ECG analysis. It introduces a hybrid neural network that combines a multilayer perceptron (MLP) and a Convolutional Neural Network (CNN) to classify healthy and pathological features. This model processes both quantitative and image representations of FHR signals. AlexNet with Support Vector Machine (SVM), which classifies CTG recordings into standard and suspected categories, emphasises reducing computational time and making it suitable for clinical settings. CTG is widely used in obstetrics

Figure 1
Sample feat heart rate, uterine contractions measured in CTG.



for assessing fetal health, but its interpretation can be challenging due to the complexity of FHR signals.

Feature importance in clinical data is a challenging issue to address. Features such as uterine contractions, accelerations, long-term variability, short-term variability, fetal movements, and rhythm changes over time are recorded by a gynaecologist to detect abnormalities. In this research, based on these feature patterns, standard, suspect, and pathological cases are reported. Additionally, based on these feature patterns, classes 1-9 are reported in categorical form. Traditional research [19] using the CTG dataset utilises the target variable of NSP (class 10) and evaluates its performance within a machine learning environment. Additionally, a significant limitation of machine learning models is discussed, including the issues of overfitting and the need for complex models to process features efficiently. This limitation is overcome by using a lightweight TabNet model with attention functions to handle tabular data and predict fetal health accurately.

To mitigate the class imbalance issue in the dataset, steps like feature scaling and other resampling methods are applied to promote the equitable training of the model. Overfitting is mitigated by integrating early stopping during training to improve generalisation. The model is tested using a dataset consisting of fetal cardiography signals, and the classification accuracy shows significant improvement compared to the previous approaches. The model's performance highlights the effectiveness of using this system in monitoring fetal health, considering its potential as an efficient clinical device. This method has the potential to improve the automated classification of fetal heart rates. It could be adapted to other classification fields in medical data where efficient feature extraction and imbalanced classes are present.

Challenges and Limitations

- Despite the advancements, challenges such as data imbalance and the need for large, diverse datasets remain. Techniques like SMOTE for oversampling have been used to address these issues, improving model sensitivity and specificity
- The complexity of CTG signals and the variability in clinical interpretations necessitate continuous refinement of deep learning models to ensure their reliability and generalizability across different clinical settings [15].

The contribution of the research is as follows,

1. The feature importance learned from the RF model. The RF model is implemented after SMOTE balancing and the assignment of minority class weights.
2. Further deep learning models, RF, TABNET+LSTM, are designed for the efficient classification outcome.
3. Multiclass is classified based on the efficiency of deep learning models with 10 10-fold cross-validation approach. Fold 6 gives the best results as discussed in the results section.

2. Related Works

This section explores previous work on fetal health classification. The potential of deep learning models [23] to improve fetal health classification from Cardiotocography (CTG) data reaches an accuracy of 93%. However, data quality presents challenges, including issues with overfitting and underfitting models. Lack of transparency and ethical issues, such as bias and data privacy, also indicate difficulties in integrating and accepting ML models in clinical processes. It identifies the possibility of improving CTG classification for fetal hypoxia prediction and supports the clinical decision. The requirement of interpretable ML models is limited to class imbalance. The Mixed-Data Type Approach [16] proposes a hybrid neural architecture that integrates quantitative parameters with feature representations obtained from images of FHR signals. A Multi-Layer Perceptron (MLP) and a Convolutional Neural Network (CNN) were employed on the image data to distinguish between healthy and pathological fetuses with an accuracy of 80.1% on 14,000 CTG tracings.

LW-FHRNet, a lightweight model [6] employing a cross-channel interactive attention mechanism, Model 3: Cross-channel interaction. The warm embrace of comprehension beckons as we note that this model transforms one-dimensional FHR signals into two-dimensional wavelet packet coefficient matrices, achieving an overwhelming classification gain of 95.24% while minimising model shards to 0.33 M. This approach utilises the time-frequency representation [29] of FHR signals with Morse wavelets and a ResNet 50 model, which is already trained on other

datasets. This yielded highly accurate classification results of 98.7% and 96.1% for FHR data occurring at different stages of labour.

Models such as Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbours (K-NN) [28] demonstrate good accuracy in identifying pathological states, although their effectiveness may vary in other classes. Further research employs more diverse datasets to enhance model robustness and clinical relevance. Convolutional neural networks (CNN) are utilised in real-time classification of fetal status via a computer server and mobile application, achieving high sensitivity and specificity. The mobile application using 1D CNN reports a 93% harmonic mean, while the computer server employing 2D CNN achieves 98.7%. However, the limitations are contingent on the hospital's dataset, which restricts the dataset size to mitigate the risk of overfitting. This limitation emphasises fetal heart rate and uterine contractions as predictive factors. Subsequent validation against a multivariable dataset and other clinical factors enhances generalisability. A recent study introduces a 1D Cycle GAN-based model for reconstructing non-invasive fetal ECG (fetal ECG, or fecg) signals from maternal ECG (maternal ECG, or mecg), demonstrating high accuracy in diagnosing fetal heart abnormalities.

An AI-driven decision support system [7] that analyses two-dimensional videos from fetal echocardiograms conducted in the first trimester to identify critical cardiac structures. This system is designed to provide remote second opinions to sonographers and can be integrated into ultrasound machines for live evaluation during procedures. Siamese CNNs (SCNNS) [22] are best suited for tasks such as ECG classification with limited datasets because they can learn to distinguish pairs of input samples. The SCNN model achieved an accuracy of up to 95% on publicly available datasets, utilising the hold-out validation technique and a mean AUC score of 89% over 7-fold cross-validation in classifying heartbeats from 12-lead ECGs, surpassing several previous benchmarks and demonstrating its robustness.

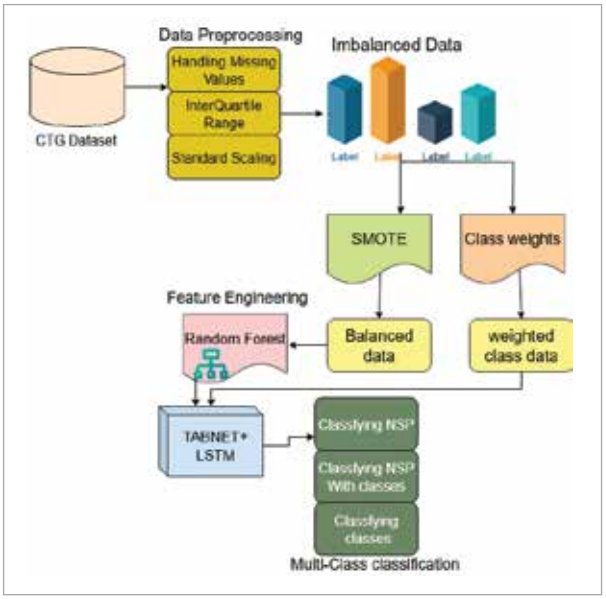
Machine learning workflow [14] for FHR deceleration classification using four techniques: MLP, RF, NB, and SLR. The work highlights feature selection as a critical step, leveraging a newly introduced fuzzy logic mechanism that achieved the highest classification accuracy of 97.94% with MLP, significantly outperforming 63.92% with RF using clinician-annotated data. A hybrid method of classification of fetal electrocardiogram (fECG) [24] using multimodal data fusion with deep learning techniques. This method aims to enhance the accuracy and efficiency of detecting heart-related issues in fetuses, which is crucial for effective prenatal care. The application of multimodal data integration with deep learning models addresses the limitations of older techniques that heavily rely on manual feature extraction and single-modality data. This novel approach is backed by several studies, which underscore the impact of deep learning and multimodal fusion on ECG classification.

The framework developed a Machine Learning-based Congenital Heart Disease Prediction Method (ML-CHDPM) [11] for this study, aimed at detecting and identifying R-peaks of the fetal ECG signal directly from a 12-channel abdominal composite signal recorded noninvasively from 70 pregnant women, both healthy and with health conditions, ensuring no records of fetal abnormalities were present. R-peak detection of fetal ECG was performed using the proposed model based on a recurrent neural network architecture. The framework's performance was evaluated through subject-dependent (5-fold cross-validation) and independent (leave-one-subject-out) tests, achieving average accuracies of 94.2% and 88.8%, respectively. Machine Learning (ML) and Deep Learning (DL) techniques [3] are transforming fetal health diagnosis and monitoring through a new wave of research focused on fECG classification. The application of ML and DL [4] addresses the challenges associated with the noisy and intricate nature of FCG signals, while enhancing the reliability and accuracy of health evaluations related to the fetus. The intention is to mitigate fetal mortality rates and enhance outcomes for neonates.

3. Proposed Methodology

This study sensitively concentrates on the correlation between all features and multiple techniques to evaluate the multi-class classification. We focus on both the classes 1-9 and NSP for improving model generalizability by different feature balancing techniques like SMOTE and Class weight assignment.

Figure 2
Overall architecture of the proposed research.



Fetal ECG data-based health classification using machine learning and deep learning models empowers medical services for gynaecologists. Though fetal heart rate is viable for monitoring, we need an AI model to examine the heartbeat condition. Figure 2 illustrates the overall architecture of the research. Initially, data is preprocessed with scaling, label encoders, and SMOTE to balance the dataset. Now, balanced datasets are processes that select the best feature engineering. Finally, 21 features are selected to process with deep learning models. ANN and Tabnet with LSTM are trained and tested for efficient outcomes. This model gives multiple outcomes by classifying two types of multi-class targets.

3.1. Dataset Description

The Cardiotocography dataset from the UCI Machine Learning Repository contains data related to fetal heart rate (FHR) and uterine contraction (UC) measurements, which are used in obstetrics to monitor fetal health. The dataset can be downloaded from https://archive.ics.uci.edu/dataset/193/cardiotocography. For fetal health analysis, this dataset was prepared by S. M. Guven Kaya, Department of Obstetrics and Gynaecology, Medical School of Dokuz Eylul University, Turkey. This multivariate dataset comprises 2,126 fetal cardiotocograms with 23 feature attributes. There are three health classification reports: Normal (N), Suspect (S) and Pathological

Table 1
Feature Description.

Fefeatures	Dedescription	Fefeatures	Dedescription
FileName	CTG examination	**ASTV**	Percentage of time with abnormal short-term variability (SisPorto)
Date	of the examination	**mSTV**	mean value of short-term variability (SisPorto)
B	start instant	**ALTV**	Percentage of time with abnormal long-term variability (SisPorto)
E	end instant	**mLTV**	mean value of long-term variability (SisPorto)
LBE	Baseline value (medical expert)	**DL**	light decelerations
LB	baseline value (SisPorto)	**DS**	severe decelerations
AC	accelerations (SisPorto)	**DP**	prolongued decelerations
FM	fetal movement (SisPorto)	**DR**	repetitive decelerations
UC	uterine contractions (SisPorto)	**Width**	histogram width
Min	low freq. of the histogram	**Nmax**	number of histogram peaks
Max	high freq. of the histogram		
CLASS	Class code (1 to 10) for classes A to SUSP	**NSP**	Normal=1; Suspect=2; Pathologic=3

(P). In addition, they have ten classes on predicting waveform-based fetal activity as shown in table 1.

3.2. Data Preprocessing

Feature selection is the most critical parts of the preprocessing steps. The CTG dataset, values of LB, AC, FM, UC, DL, DS, and DP are chosen since they directly relate to the output variable. In addition, mean, median (middle value after sorting the data), variance, mode (most frequently repeated value) and tendency of the histogram are deemed necessary. Mean and Variance are defined using Equations (3)-(4).

$$Mean\ \mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

$$Variance\ \sigma^2 = \frac{1}{n}\ \sum_{i=1}^{n}(x_i - \ \mu)^2. \tag{2}$$

Other features, such as file names and timestamps, are irrelevant and therefore omitted to enhance model performance by reducing dimensionality.

3.2.1. Categorical Feature Encoding

Categorical features like neonatal state prediction (NSP) require encoding into numbers for modelling, which can be done using one-hot encoding based on the variable and model. Let

$$x = NSP\ \in\ \{1,2,3\}\ \rightarrow One - hot\ encoded\ vector\ :\ \left\{\begin{matrix}(1,0,0), if\ x = 1\\ (0,1,0), if\ x = 2\\ (0,0,1), if\ x = 3\end{matrix}\right\} \tag{3}$$

3.4. Handling Class Imbalance

3.4.1. SMOTE

Handling and managing feature imbalance and scaling is essential, especially while constructing a machine learning model. To overcome this, this research deployed the Synthetic Minority Over-Sampling Technique (SMOTE) to build artificial samples for the minority class. SMOTE works by interpolating between a sample and its nearest neighbours using Equation (6).

$$X_{new} = x_i.\ \lambda.\ (x_{nn} - x_i),\ \lambda\epsilon[0,1]. \tag{4}$$

Figure 4 shows the before-and-after results of the SMOTE operation. With features having different scales, standardisation and normalization become vital. The above is achieved by applying StandardScaler(), which sets the mean to zero and the standard deviation to one. The dataset is then split using train_test_split(), achieving the desired 80/20 split for the training and testing datasets. Most datasets have class imbalance, which is often exacerbated in classification tasks. To address this, the research employed the Synthetic Minority Over-sampling Technique (SMOTE) to generate artificial samples for the minority class, thereby improving class balance. Executing these preprocessing methods stepwise resulted in a dataset of higher quality and accuracy for training a machine learning model.

3.4.2. Class Weights

we train the model with class 1–9 classification with class weights. The model predicts fetal health under classes 1–10, where every class coincides to distinct ECG signal characteristics. Class 3 has a weight of 3.85, depicting it is less frequent but more essential, while Class 1 has a minimum weight of 0.43, as shown in Figure 5 for describing it is more frequent. The model utilizes these weights to handle class imbalance efficiently. After classifying the 1–10 classes, the model completes into NSP classes such as Normal, Suspect, and Pathological. The class weight assigned features are processed with proposed improved TABNET+LSTM model.

3.3. Feature Selection Phase I and II

Initially, phase I extraction SMOTE balanced data is used by RF. The RF selected features are processed using proposed TabNet+LSTM to predict fetal abnormalities. In feature extraction, the Random Forest can be employed to determine the significance of every feature by predicting the target variable. This is an impurity in holding decision trees. In our research, the Random Forest algorithm will use 100 decision trees while retaining all other parameters. This way, the model can be trained without needing any extra tuning. To control for the model's performance not being overly dependent on a particular data split, and to check for generalisation to truly unseen data, Stratified K-Fold Cross-Validation is employed. This form of cross-validation also splits the

dataset into five folds, known as a five-fold cross-validation, but each fold must have an equal class distribution as the entire dataset. Stratified sampling enhances performance estimation by reinforcing the proportion of classes present in each fold, which leads to more reliable performance estimates.

Then, based on feature importance and correlation analysis, classes 1–9 are utilised for exact abnormality prediction. Class 1–9 is categorical data used to determine the fetal state in the uterus. By including classes 1–9 with the target class NSP, the model performance is evaluated in phase II.

Addressing outliers is crucial to maintaining data credibility. An extreme value that's too far beyond the accepted range is something that the Interquartile Range (IQR) method is designed to detect using masking so that the model can work reliably within defined boundaries, as defined by Equation (5).

$$\begin{aligned} IQR &= Q3 - Q1 \\ Lower\ bound &= Q1 - 1.5 \times IQR \\ Upper\ bound &= Q3 - 1.5 \times IQR. \end{aligned} \tag{5}$$

Any value outside this range is marked as an outlier and tends to be masked or removed.

Phase II extarction uses class weights. Based on class weights the proposed model learns the feature correlation and learns the abnormalities.

3.5. Proposed-TabNet Architecture

The architecture of the TabNet model is a deep learning framework tailored for handling tabular data. The TabNet architecture works with CTG tabular data, which can be processed and predictions made by examining complex features. It accomplishes this by learning a differentiable decision tree-like structure that is optimal for CTG datasets with numerous numerical and categorical features.

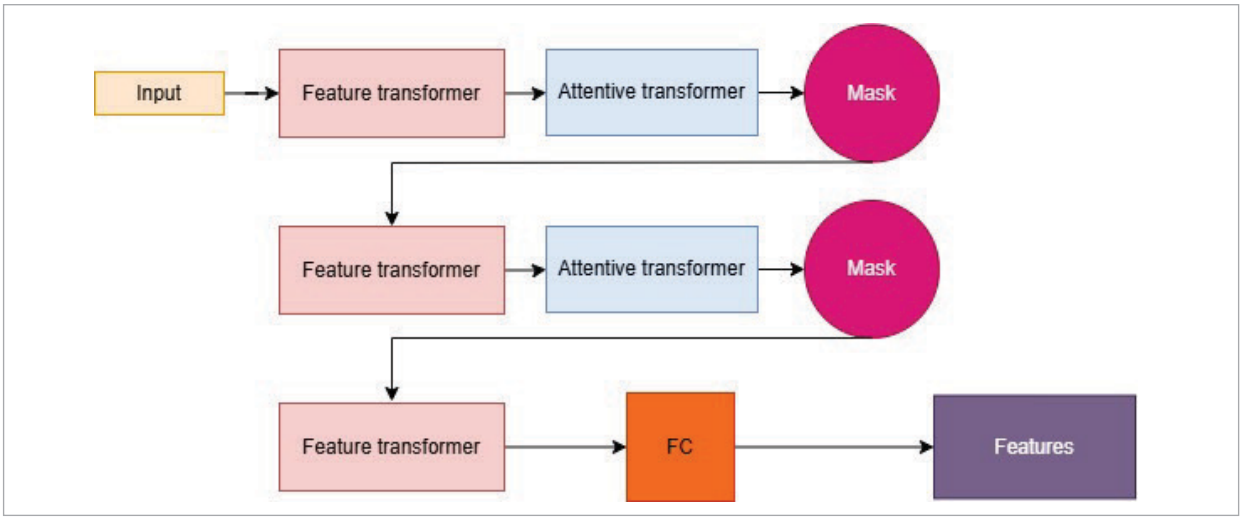
Embedding Generator (TabNetEmbeddings)

The input is first passed through embedding generators, which convert categorical features into dense vectors. Let $x_c \in R^d$ be a nominal feature and let $E_c \in R^{|V_c| \times \epsilon}$ be the embedding matrix where $|V_c|$ refers to the vocabulary size and e is the embedding dimension. The embedded feature is defined as $\tilde{x}_c = E_c[x_c]$. The CTG dataset consists of both numerical and categorical features, like the fetal heart rate, alongside the neonatal state prediction; hence, embedding generators will utilise the categorical features. Batch Normalization BN is performed on embeddings to impose form on the data defined by Equation (6).

$$BN(x) = \gamma \cdot \frac{x-\mu}{\sqrt{\sigma^2+\varepsilon}} + \beta, \tag{6}$$

where μ and σ^2 refers to the batch mean and variance, and γ, β are learnable parameters. A data split is done in this step into suitable components which

Figure 3
Tabnet architecture.



can be treated separately. This guarantees the tabular data is split in an orderly manner before being sent to more advanced processing layers. Figure 3 shows the TABNET architecture.

Tabnet Encoder

The feature transformer takes care of the raw data. A feature transformer consists of the following sub-parts:

a **Gated Linear Unit (GLU) Layers:** Each GLU layer transforms the input as mentioned in Equation (7).

$$GLU(x) = (Wx + b).\sigma\left(W_g x + b_g\right), \tag{7}$$

where W, b are learnable parameters for the linear path, W_g *and* b_g are for the gate, σ is the sigmoid function and $\cdot$ is the element-wise multiplication. In our CTG data, it is essential to capture complex interactions between features, as this involves modelling interactions between fetal heart rate-related features like LB, AC, FM, etc. Thus, GLU non-linear activation functions were added to the design.

b **Shared Layers:** A set of layers shared across all features to ensure that the model captures a general representation of the data before specific transformations. Let h_i^8 represent the output from the shared layers for feature i, defined using Equation (8).

$$h_i^8 = f_{shared}(x_i). \tag{8}$$

These layers ensure a general representation across all features.

c **Specific Layers:** $h_i^p = f_{specifici}(h_i^8)$ These layers enable the model to capture relationships between features. In the CTG dataset, this could be the learning patterns for different fetal health states.

d **Attentive Transformer:** Like transformers, the attentive transformer helps focus on relevant features for a particular decision. This is particularly useful for CTG data, where certain features are more predictive or informative of the neonatal state. The attention scores are calculated using Equation (9).

$$M^{(t)} = Sparsemax(P^{(t)}.h^{(t-1)}). \tag{9}$$

e **Linear Layers:** After attention is applied to the feature set, linear layers are applied using Equation (10).

$$z = Wx + b. \tag{10}$$

In this context, a transformation like $R^8 \rightarrow R^{22}$ is done, $z = W_{8\times22}x + b$.

f **Softmax:** The softmax function is used to get a probability distribution over the output classes defined using Equation (11).

$$\hat{y} = Software(W_3 z). \tag{11}$$

For the CTG dataset, this would mean predicting the neonatal state as normal, suspect, or pathological.

g **Final Mapping:** After the attention process, the features undergo a final mapping with a linear projection of $(8 \rightarrow 3)$, which is described using Equation (12).

$$Z_{final} = W_f x_{attended} + b_f, \quad W_f \in R^{8\times3} \tag{12}$$

Here, the data's dimensionality is aligned to the number of output classes, which for the CTG dataset is 3.

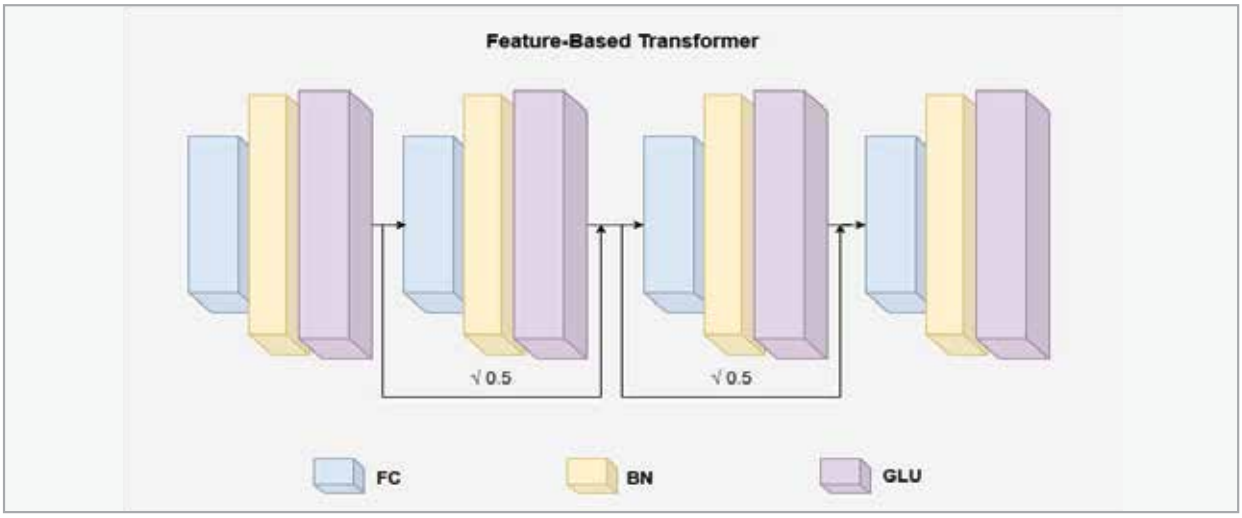
h The model mainly focuses on the most relevant features for every decision. The attention mechanism is sparse, focusing only on a subset of features at every step, which aids in handling high-dimensional datasets, such as CTG data. The attention mechanism can be expressed using Equation (18).

$$a_t = Attention(x_t, h_t), \tag{13}$$

where a_t is the attention vector at step t, which shows the relevanc e of each feature in the input. x_t is the input feature vector at step t. h_t is the hidden state vector from the above layer.

As shown in Figure 4, the input features and CTG dataset components LB, AC, FM, UC, etc. undergo processing by the embedding generator. This phase is vital for the CTG dataset's continuous and categorical features. Initially, input features undergo processing at the Feature Transformer, where the output will be $h^{(t)} = GLU\left(BN(W_x^{(t)})\right)$, where the model simultaneously encodes shared and specific repre-

Figure 4
Layers of Feature transformer in TABNET.



sentations. The GLU layers capture the non-linear patterns, while the shared and specific layers ensure that the model adopts general and domain-specific patterns, such as recognizing FHR patterns, which are abnormal for health. The attention mechanism enables the model to concentrate on relevant features for classification. Some features within the CTG dataset, such as FM, UC, fetal heart rate, and long-term variability, are more closely associated with neonatal health; therefore, attention prioritizes these traits when making decisions. Following the attention layers, the output undergoes a last linear mapping step, which shrinks the dimensionality to three, the number of classes in the CTG dataset [25-27]. These are normal, suspect, and pathological.

A softmax function then translates these outputs into probability distributions, which prevail in classification. The final decision of the model can be expressed using Equation (14).

$$\hat{y} = Software(W_3, DecisionLayerOutput), \quad (14)$$

where W_3 is the weight matrix in the final decision layer.The Softmax maps the output of the decision layer to probabilities suitable for classification. In binary classification predicting fetal health, whether regular or abnormal, the output $\hat{y}$ will be a single probability value with 0 and 1. Typically, 0.5 is used to determine the class:

$$\hat{y} = \begin{cases} 1 \; if \; P(class = 1) > 0.5 \\ 0 \qquad\qquad otherwise \end{cases}, \quad (15)$$

where P(class=1) is the probability of the fetus being in distress and abnormal health. The backbone model can operate on raw tabular data, which is simpler than other models that may require heavy text preprocessing, feature extraction, or encoding work. The attention-based interpretability included in TabNet helps explain the feature focus for each output, which assists in fetal monitoring features associated with expectant mother healthcare. The model processes numerical and categorical data proficiently, which is beneficial because CTG datasets encompass FHR, LB, and UC features that differ significantly in type and scale.

The TabNet architecture is highly optimised for tabular datasets, such as the CTG dataset. It efficiently manages the data and prioritises essential features, making it ideal for neonatal state prognosis in cardiotocography. Finally, LSTM is used for temporal feature processing with a wide advantage for classifying the multi classes. It can be expressed as $h_t,c_t=LSTM(x_t, h_{t-1},c_{t-1})$.This ensures the model captures temporal dynamics from time-series signals like FHR traces. Figure 5 shows the TABNET attention architecture, and Table 2 describes the hyperparameter values of the TABNET model.

Figure 5
Layers of Attention mechanism in TABNET.

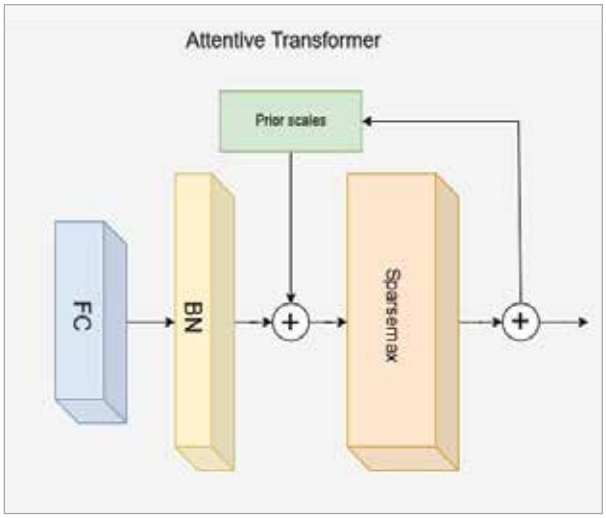


Table 2
Hyperparameter values of the TABNET model.

Hyperparameter	Proposed Value
n_d (decision step number)	8
n_a (attention head number)	8
n_steps (attention steps)	3
Gamma (sparsity coefficient)	1.3
lambda_sparse(L1 regularization coefficient)	1e-5
Optimizer	Adam
learning_rate	0.02
batch_size	128
max_epoch	50
early_stopping	True
Patience	5
dropout_rate	0.5
feature_mask_type	"sparsemax"
virtual_batch_size	128
num_virtual_batches	2
activation_fn	"relu"
LSTM LAYERS	Basic model

4. Experiments and Results

4.1. System Requirements

The experiments utilised a machine equipped with a Ryzen 5 5600H processor and 16GB of RAM, thereby enhancing the dataset's computational capabilities and facilitating model training. The RAM and processor are further beneficial for training multiple machine learning models, as they require high RAM and memory simultaneously. The experiment was run on Scikit-learn on Windows, along with Python 3.8, described as one of the best languages to use with machine learning and its associated libraries. Scikit-learn is used for complex machine learning operations, particularly model building, data partitioning, and feature extraction. TensorFlow and PyTorch methods were used to compute and train it with TabNet models. These frameworks allowed excellent robustness and flexibility in scaling both models. Table 3 below presents various performance metrics for different classification modes. Results were recorded using multiple classification models for reproducibility, which was confirmed by fixing random seeds (*torch.manual_seed*(42)).

4.3. Phase 1 Result Analysis

Among the classifiers analyzed, Random Forest performed best at 93% accuracy because of its strong non-linear relation handling, and overfitting reduction from ensemble averaging. Logistic Regression and K-Nearest Neighbor (KNN) followed closely with 89% and 90% accuracy respectively, demonstrating their effectiveness albeit with simplistic approaches to feature interactions.

Gradient Boosting, an ensemble method, also reached 90% accuracy, showcasing the iterative improvement offered by boosting in refining predictions. In contrast, Decision Tree and Support Vector Classifier (SVC) suffered with 85% and 81% accuracy respectively, due to over-sensitivity to feature scaling, class imbalance, and overfitting. Furthermore, combining SMOTE (Synthetic Minority Oversampling Technique) with Random Forest brought accuracy to 95.5%, demonstrating the impact class imbalance has on medical datasets. This table 3 illustrates that while model selection is essential, preprocessing like resampling is just as crucial in improving accuracy.

Table 3
Multi-class classification labels.

A:	Class 1	calm sleep	DE:	Class 7	decelerative pattern (vagal stimulation)
B:	Class 2	REM sleep	LD:	Class 8	largely decelerative pattern
C:	Class 3	calm vigilance	FS:	Class 9	flat-sinusoidal pattern (pathological state)
D:	Class 4	active vigilance	SUSP:	Class 10	suspect pattern
SH:	Class 5	shift pattern (A or Susp with shifts)	10 CLASSES	(class1-10)	Class code (1 to 10) for classes A to SUSP
AD:	Class 6	accelerative/de-celerative pattern (stress situation)	NSP (3 classes)	3 classes=NSP	Normal=1; Suspect=2; Pathologic=3

Table 4
State of art [19] comparison of machine learning models and importance of class imbalance handling.

Classifiers	Accuracy
Random Forest [19]	93%
Logistic regression [19]	89%
KNN [19]	90%
Gradient boosting [19]	90%
Decision Tree [19]	85%
Support vector classifier [19]	81%
SMOTE+TABNET+RF	**95.5%**

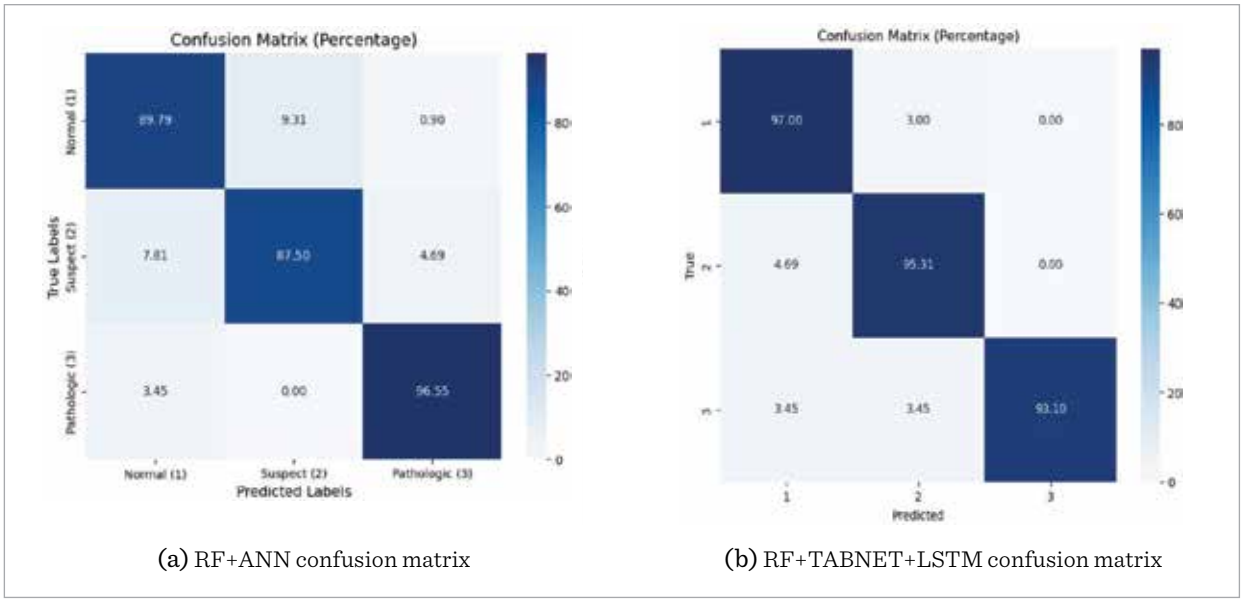
Even though Random Forest (RF) classifier functions independently, it boasts 95.53% accuracy and 95% precision boasting an F1-score of 93% which indicates that the model is conteh with the data after being processed by SMOTE, suggesting the model's conti. It is worth noting these figures demonstrate an SPF in the RESULTS SECTION, implying it is in the lowest tier of accuracy graduates with solid views on the goals they aim. However, combining RF with advanced architectures like Artificial Neural Networks (RF+ANN) and TabNet (RF+TabNet) poses substantial challenges.

An even more preferable result comes with RF+LSTM which increases accuracy to 91.76% while also achieving a reasonably good balance between precision and recall (93% and 92%) which indicates that conducting classification using LSTM after temporal sequence modeling is advantageous. Importantly, the best performance is attributed to the ensemble of RF+TabNet+LSTM, which attains an astounding 97% accuracy and 96% F1 score, outperforming all other methods by a noticeable margin. This demonstrates that the combination of feature-aware learning TabNet and sequence modeling LSTM, integrated with a

Table 5
Phase 1 Outcome after smote processing and classification

Criteria	Random Forest	RF+ ANN	RF+ TabNet	RF+ LSTM	RF+ TabNet+ LSTM
Accuracy	95.53%	86.85%	86.85%	91.76%	97%
Precision	95%	90%	89%	93%	95%
Recall	91%	87%	89%	92%	94%
F1 – score	93%	88%	88%	92%	96%
Computation cost (param)	0.1 M	1.5 M	5M	3M	3M

Figure 6
Confusion matrix of the proposed model.



(a) RF+ANN confusion matrix

(b) RF+TABNET+LSTM confusion matrix

Figure 7
Classification report heatmap for RF+TABNET.

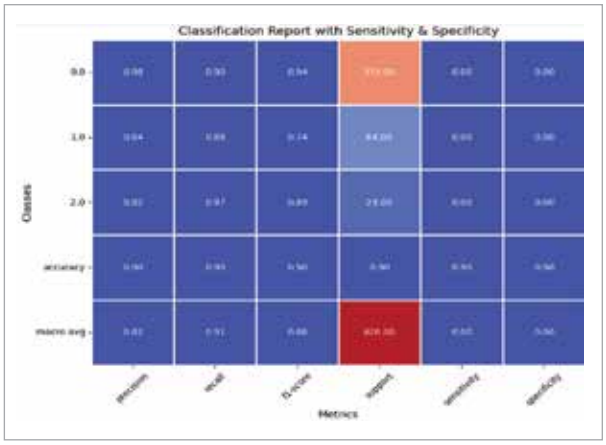
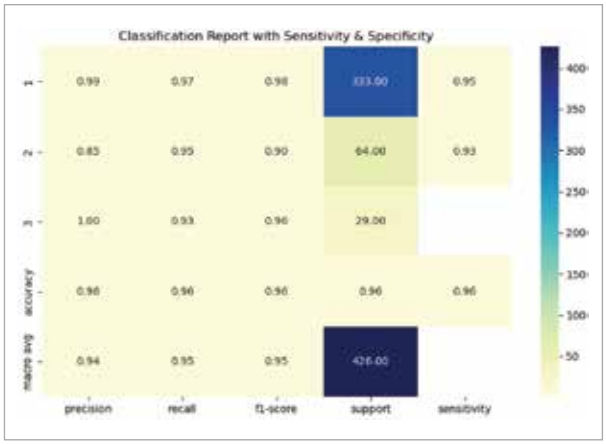


Figure 8
Performance metrics of RF+TABNET+LSTM.



baseline random forest, provides significant improvements in generalization and sensitivity to minority class patterns after applying SMOTE.

As noted, Table 5 clearly demonstrates the importance of leveraging complementary learning approaches. RF+TabNet+LSTM emerged as the optimal model, owing to the best reported results of balanced precision and recall, alongside high predictive accuracy and resilience to commonly encountered challenges in classification scenarios with unequal class distributions.

The confusion matrix presented in Figure 6 demonstrates Random forests with ANN and Random forests with TabNet+LSTM. The performance of Figure 6(b) shows high accurate prediction of NSP using ensemble model technique.

The classification report of the TabNet model, shown in Figures 7-8 above, gives precision, recall, F1-score, support, and sensitivity for each of the three classes (NSP): 1, 2, and 3. Furthermore, the matrix contains the macro average and model accuracy, thus enabling more informed value judgment concerning performance.

For Class 1 (True Label 1), the RF+TABNET+LSTM model achieved optimal effectiveness, as shown by the measurements of precision (0.99), recall (0.99), F1-score (0.98), and support 333, which reflects a large sample size in the dataset for this category. Sensitivity was 0.95. These scores mean that the model has a high probability of successfully identifying Class 1—effectively classifying most instances with a slight chance of error. The model performs well for Class 2 (True Label 2), although its accuracy is slightly lower, with a precision and recall of 0.85 and 0.95, respectively. Support sits at 64, significantly lower than Class 1, while the F1-score here is 0.90. Sensitivity is 0.93, meaning that while being competent in identifying Class 2, the model mistakenly associates some Class 2 instances with other classes, especially with Class 1.

For Class 3 (True Label 3), the model exhibits an even stronger performance with perfect precision (1.00) and a recall of 0.93, yielding an F1-score of 0.96. The support for this class is 29, and the sensitivity value is 0.96, which reflects the model's excellent ability to identify captures Class 3 with small misclassifications. The macro average precision, recall, and F1 score are uniformly at 0.94, 0.95, and 0.95, respectively. These values represent strong performance across all classes. The model's accuracy, at 0.96, demonstrates the model's dependability in classifying the classes correctly. The TabNet model achieves remarkable results in all three classes, with the highest performance reserved for Class 3 and strong performance for Classes 1 and 2. The model robustly distinguishes the classes with minimal misclassification and balanced sensitivity across all classes. The macro-average values indicate the overall effectiveness of the model.

4.5. Analysis of Accuracy and Loss

The accuracy and loss curve for the RF+TABNET+LSTM model in Figure 9 initially shows overfitting issues. However, we employed patience and regularisation to reduce this problem. The validation accuracy reaches 90% after 35 epochs, which is higher. At this 30th to 35th epoch, we can find that overfitting is wholly reduced. This shows that our model performs well after the 35th epoch. Loss curves tend to be reduced and stagnant after the 30th epoch

4.6. Classification Phase II Report of the NSP with 10 CLASSES with Class Weights

In this section, we train the model with class 1-9 classification with weights. The model predicts fetal health under classes 1–10, where every class corresponds to distinct ECG signal characteristics. For example, Class 3 weights 3.85, indicating that it is less frequent but more essential, while Class 1 has a minimum weight of 0.43, as shown in Figure 10, indicating that it is more frequent. The model utilises

Figure 9
RF+TABNET+LSTM accuracy and loss curve.

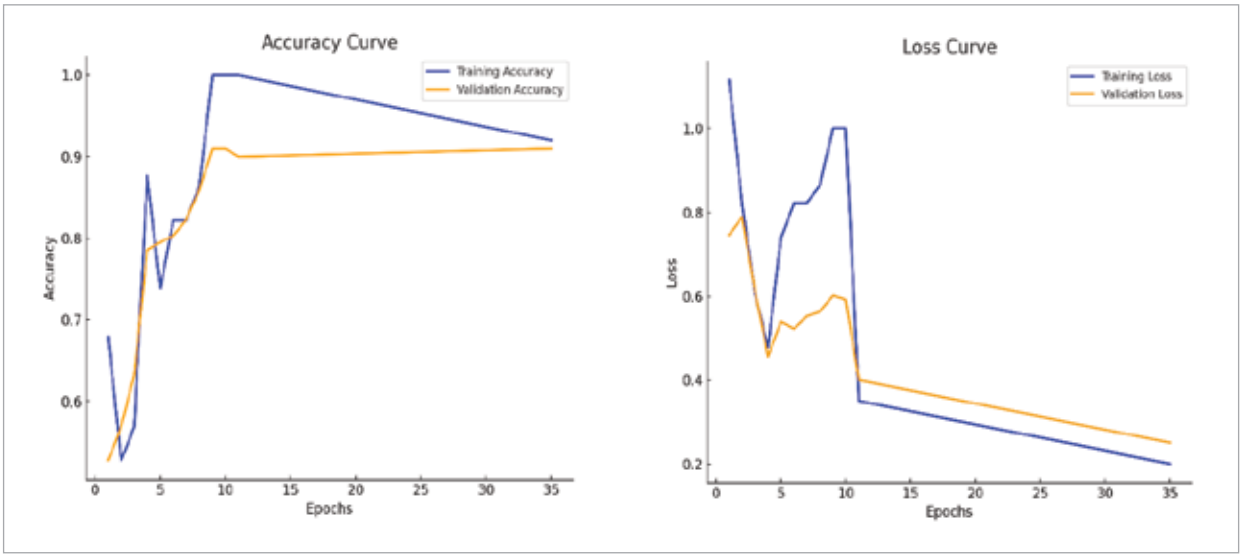


Figure 10
Class weight assignment in histogram.

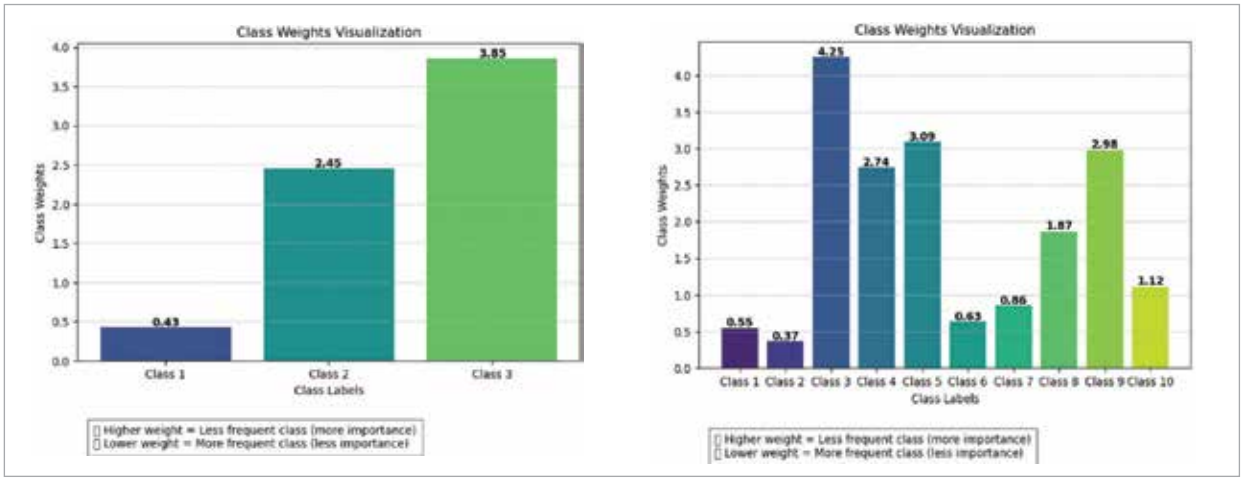


Figure 11
Classification report in heatmap for RF+;TABNET with class weight on NSP.

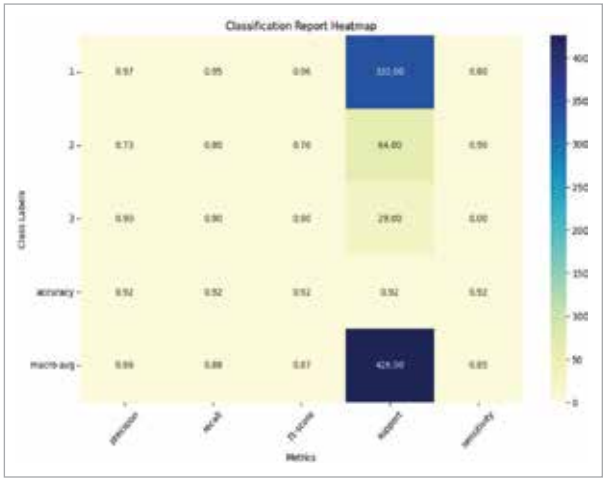
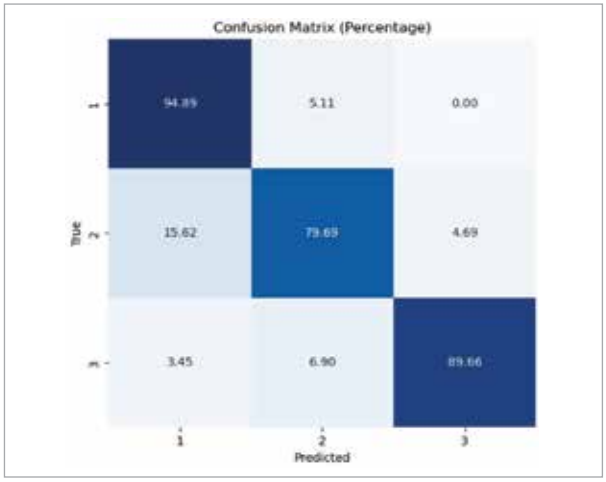


Figure 12
Confusion matrix of RF+TABNET with class weights and target NSP.



these weights to handle class imbalance efficiently. After classifying the 1–10 classes, the model is completed into NSP classes such as Normal, Suspect, and Pathological.

Figure 11 shows the classification report of the proposed RF+TABNET with class weights and target NSP. Precision for class 1 (normal) is 97% which is higher and class 3 (pathology) which is 90%. The suspect class 2 have more deviation is classification task. The accuracy of model achieves 90% is classifying the fetal condition. Figure 12 shows the confusion matrix with the proposed classification task performing better.

4.7. Evaluation of 10-class Classification with NSP Feature

After merging the 'NSP' feature with other existing features in CTG and classifying the CLASS, class weights were utilised to handle imbalance. Minority class Class 3 acquired the highest weight of 4.25, showing its rarity and higher significance. Conversely, Class 2 had the lowest weight of 0.37, indicating a superior frequency and minimal impact. Other eminent weights comprise Class 5 with 3.09, Class 9 with 2.98, and Class 1 with 0.55. These weights directed the model to prioritise less frequent but crucial fetal conditions precisely.

Figure 13
Confusion matrix with 10 classes using RF+TABNET in classification.

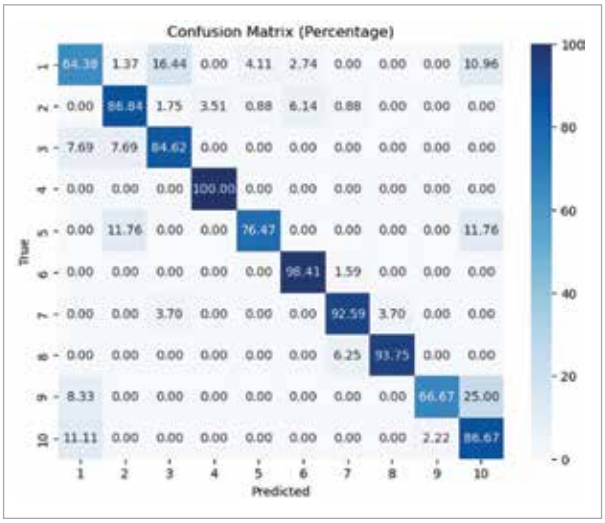
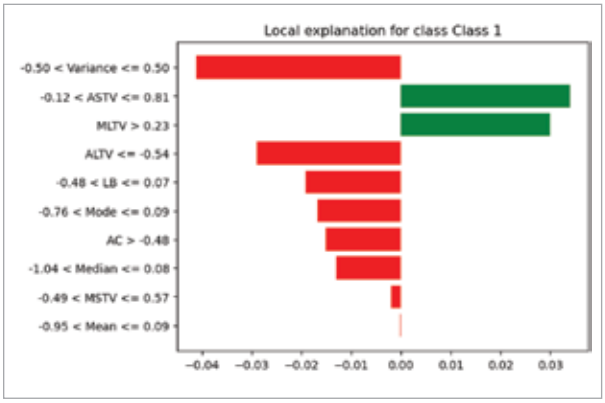


Figure 13 shows the confusion matrix for 10 classes. Prediction is 100% for 4th class and 98.4% for 6th class, which has better prediction results. The minimum prediction accuracy is achieved in the 9th class and should be considered for class balancing in future research.

According to the performance metrics, the TABNET model achieves 95% accuracy with both LSTM and RF, as illustrated in Figure 17. Other performance metrics also work better than the existing model.

Figure 15
LIME for CLASS 1 classification.



The training decreases rapidly in the beginning epochs and continues improving gradually throughout the model training process, as shown in Figure 14. The loss begins to stabilise during the later stages of training. It achieves a value of around 0.4 to 0.5, indicating the model has most likely converged in minimizing the error for the training data. Training accuracy is initially low but improves sharply with the data, reaching approximately 0.9 (90%) by the end of training. This indicates that the model performs reasonably well on the training data as true positives. Validation accuracy also improves to 0.8 (80%).

Explainable AI helps to understand the transparency of backbox decision-making, as given in Figure 15.

Figure 14
LOSS and ACCURACY Curve for predicting fetal health using RF+ TABNET+LSTM with various class weights.

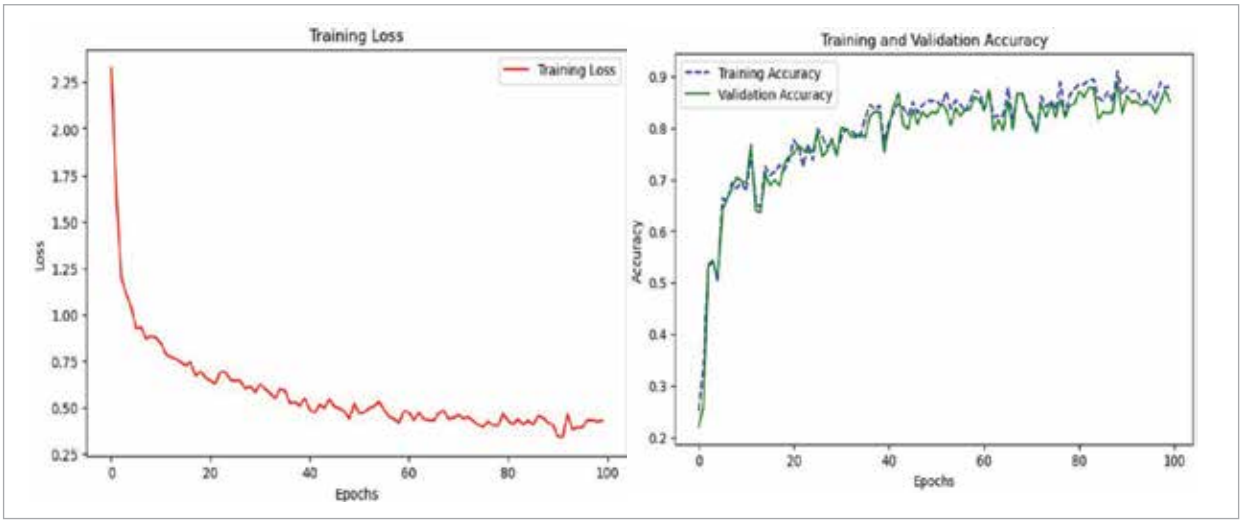


Table 6
Existing state-of-the-art works based on performance.

Model	Accuracy
Random forest [19]	93%
Gradient boosting [19]	90%
Discriminant Analysis [8]	83%
Decision tree [8]	86%
SVM [5]	84%
NavieBayes	83.7%
Proposed RF+TABNET+LSTM	**97%**

LIME shows that ASTV and MLTV are the strongest positive drivers for this classification, while other features like Variance, ALTV, and LB pull the prediction away from Class 1. Likewise, every class has certain features to consider for its true value. Table 6 shows a state-of-the-art comparison.

5. Conclusion

The study evaluates different machine learning techniques, including the principles of Random Forest (RF), TabNet, and Long Short-Term Memory (LSTM) networks, to assess their efficiency in predicting fetal health using the CTG dataset in a multi-class classification setting. The findings indicate that model performance is maximised with the hybrid RF-TABNET-LSTM architecture with SMOTE validation, achieving a 97.42% accuracy rate in three-class classification tasks. Performance was also enhanced with the addition of ANN, achieving higher precision and recall across multiple classes due to its advanced modelling capabilities and integration of LSTM for handling temporal dependencies. The assessment of multi-class classification models aimed at predicting NSP (Normal, Suspect, Pathologic) alongside 10 distinct health statuses of the fetus demonstrates that the TabNet + LSTM model was the most accurate and precise, yielding the best recall, and showcasing a robust capability to manage the complexity of the data. This study demonstrates the need for model optimisation through class weight adjustments, transfer learning, and other methods for effective classification, particularly when the amount of labelled data available is small. This study highlights the importance of data balancing over weight balancing, as our model achieves high performance after the SMOTE process, rather than using class weights.

References

1. Alkhodari, M., Vartanian, O. Deep Learning Identifies Cardiac Coupling Between Mother and Fetus During Gestation. Frontiers in Cardiovascular Medicine, 2022, 9, 926965. DOI: https://doi.org/10.3389/fcvm.2022.926965
2. Almadani, M. M., Alkhodari, M., Ghosh, S. K., Hadjileontiadis, L., Khandoker, A. Extraction of Fetal Heartbeat Locations in Abdominal Phonocardiograms Using Deep Attention Transformer. Computers in Biology and Medicine, 2025, 189, 110002. https://doi.org/10.1016/j.compbiomed.2025.110002
3. Barnova, K., Martinek, R., Vilimkova Kahankova, R., Jaros, R., Snasel, V., Mirjalili, S. Artificial Intelligence and Machine Learning in Electronic Fetal Monitoring. Archives of Computational Methods in Engineering, 2024, 31(5), 2557-2588. https://doi.org/10.1007/s11831-023-10055-6
4. Chiou, N., Young-Lin, N., Kelly, C., Cattiau, J., Tiyasirichokchai, T., Diack, A., Koyejo, S., Heller, K., Asiedu, M. Development and Evaluation of Deep Learning Models for Cardiotocography Interpretation. NPJ Women's Health, 2025, 3(1), 21. https://doi.org/10.1038/s44294-025-00068-w
5. Cömert, Z., Kocamaz, A.F. Evaluation of Fetal Distress Diagnosis During Delivery Stages Based on Linear and Nonlinear Features of Fetal Heart Rate for Neural Network Community. International Journal of Computer Applications, 2016, 156(4), 26-31. https://doi.org/10.5120/ijca2016912417

6. Das, S., Obaidullah, S. M., Mahmud, M., Kaiser, M. S., Roy, K., Saha, C. K., Goswami, K. A Machine Learning Pipeline to Classify Fetal Heart Rate Deceleration with Optimal Feature Set. Scientific Reports, 2023, 13(1). DOI: https://doi.org/10.1038/s41598-023-27707-z
7. Francis, F., Luz, S., Wu, H., Stock, S. J., Townsend, R. Machine Learning on Cardiotocography Data to Classify Fetal Outcomes: A Scoping Review. Computers in Biology and Medicine, 2024, 108220. https://doi.org/10.1016/j.compbiomed.2024.108220
8. Fuentealba Ortiz, P.F. Automatic Fetal Distress Assessment During Labor Based on Modal and Parametrical Analysis of the Cardiotocographic Recording (Doctoral Dissertation), 2020.
9. Kanna, S. R., Shajin, F. H., Rajesh, P., Mannepalli, K. A Multi-Branch Multi-Scale Convolutional Neural Network Using Automatic Detection of Fetal Arrhythmia. Signal, Image and Video Processing, 2024, 18, 87-96. https://doi.org/10.1007/s11760-024-03133-0
10. Lin, Z., Liu, X., Wang, N., Li, R., Liu, Q., Ma, J., Wang, L., Wang, Y., Hong, S. Deep Learning with Information Fusion and Model Interpretation for Long-Term Prenatal Fetal Heart Rate Data. NPJ Women's Health, 2024, 2(1), 31. https://doi.org/10.1038/s44294-024-00033-z
11. Liu, Z., Si, L., Shi, S., Li, J., Zhu, J., Lee, W. H., Wang, G. Classification of Three Anesthesia Stages Based on Near-Infrared Spectroscopy Signals. IEEE Journal of Biomedical and Health Informatics, 2024, 28(9), 5270-5279. https://doi.org/10.1109/JBHI.2024.3409163
12. Lovers, A., Ugwumadu, A., Georgieva, A. Cardiotocography and Clinical Risk Factors in Early Term Labor: A Retrospective Cohort Study Using Computerized Analysis with Oxford System. Frontiers in Pediatrics, 2022, 10, 784439. DOI: https://doi.org/10.3389/fped.2022.784439
13. Mendis, L., Palaniswami, M., Keenan, E., Brownfoot, F. Rapid Detection of Fetal Compromise Using Input Length Invariant Deep Learning on Fetal Heart Rate Signals. Scientific Reports, 2024, 14(1), 12615.https://doi.org/10.1038/s41598-024-63108-6
14. Ou, J., Li, N., He, H., He, J., Zhang, L., Jiang, N. Detecting Muscle Fatigue Among Community-Dwelling Senior Adults with Shape Features of the Probability Density Function of sEMG. Journal of NeuroEngineering and Rehabilitation, 2024, 21(1), 196. https://doi.org/10.1186/s12984-024-01497-5
15. Pan, H., Li, Z., Fu, Y., Qin, X., Hu, J. Reconstructing Visual Stimulus Representation from EEG Signals Based on Deep Visual Representation Model. IEEE Transactions on Human-Machine Systems, 2024, 54(6), 711-722. https://doi.org/10.1109/THMS.2024.3407875
16. Pan, H., Tong, S., Song, H., Chu, X. A Miner Mental State Evaluation Scheme with Decision Level Fusion Based on Multidomain EEG Information. IEEE Transactions on Human-Machine Systems, 2025, 55(2), 289-299. https://doi.org/10.1109/THMS.2025.3538162
17. Pradipta, G. A., Ayu, P. D. W., Liandana, M., Hostiadi, D. P. Enhanced Fetal Arrhythmia Classification by Non-Invasive ECG Using Cross-Domain Feature and Spatial Differences Windows Information. IEEE Access, 2025. https://doi.org/10.1109/ACCESS.2025.3526586
18. Pruksanusak, N., Chainarong, N., Boripan, S., Geater, A. Comparison of the Predictive Ability for Perinatal Acidemia in Neonates Between the NICHD 3-Tier FHR System Combined with Clinical Risk Factors and the Fetal Reserve Index. PLOS ONE, 2022. https://doi.org/10.1371/journal.pone.0276451
19. Salini, Y., Mohanty, S. N., Ramesh, J. V. N., Yang, M., Chalapathi, M. M. V. Cardiotocography Data Analysis for Fetal Health Classification Using Machine Learning Models. IEEE Access, 2024, 12, 26005-26022. https://doi.org/10.1109/ACCESS.2024.3364755
20. Shekhawat, D., Chaudhary, D., Kumar, A., Kalwar, A., Mishra, N., Sharma, D. Binarized Spiking Neural Network Optimized with Momentum Search Algorithm for Fetal Arrhythmia Detection and Classification from ECG Signals. Biomedical Signal Processing and Control, 2024, 89, 105713. https://doi.org/10.1016/j.bspc.2023.105713

21. Ungureanu, A., Marcu, A. S., Patru, C. L., Ruican, D., Nagy, R., Stoean, R., Stoean, C., Iliescu, D. G. Learning Deep Architectures for the Interpretation of First-Trimester Fetal Echocardiography (LIFE): A Study Protocol for Developing an Automated Intelligent Decision Support System for Early Fetal Echocardiography. BMC Pregnancy and Childbirth, 2023, 23(1). DOI: https://doi.org/10.1186/s12884-022-05204-x

22. Vadivu, M.S., Kavithaa, G. Fetal QRS Complexes Detection Using Deep Learning Technique. Journal of Electrical Engineering & Technology, 2024, 19(3), 1909-1918. https://doi.org/10.1007/s42835-023-01682-x

23. Vasconcellos, M.E., Ferreira, B.G., Leandro, J.S., Neto, B.F., Cordeiro, F.R., Cestari, I.A., Gutierrez, M.A., Sobrinho, A., Cordeiro, T.D. Siamese Convolutional Neural Network for Heartbeat Classification Using Limited 12-Lead ECG Datasets. IEEE Access, 2023, 11, 5365-5376. https://doi.org/10.1109/ACCESS.2023.3236189

24. Yin, J., Qiao, Z., Han, L., Zhang, X. EEG-Based Emotion Recognition with Autoencoder Feature Fusion and MSC-TimesNet Model. Computer Methods in Biomechanics and Biomedical Engineering, 2025, 1-18. https://doi.org/10.1080/10255842.2025.2477801

25. Zheng, Q., Saponara, S., Tian, X., Yu, Z., Elhanashi, A., Yu, R. A Real-Time Constellation Image Classification Method of Wireless Communication Signals Based on the Lightweight Network MobileViT. Cognitive Neurodynamics, 2024, 18(2), 659-671. https://doi.org/10.1007/s11571-023-10015-7

26. Zheng, Q., Tian, X., Yu, Z., Ding, Y., Elhanashi, A., Saponara, S., Kpalma, K. MobileRaT: A Lightweight Radio Transformer Method for Automatic Modulation Classification in Drone Communication Systems. Drones, 2023, 7(10), 596. https://doi.org/10.3390/drones7100596

27. Zheng, Q., Tian, X., Yu, Z., Yang, M., Elhanashi, A., Saponara, S. Robust Automatic Modulation Classification Using an Asymmetric Trilinear Attention Net with a Noisy Activation Function. Engineering Applications of Artificial Intelligence, 2025, 141, 109861. https://doi.org/10.1016/j.engappai.2024.109861

28. Zhu, C. Computational Intelligence-Based Classification System for the Diagnosis of Memory Impairment in Psychoactive Substance Users. Journal of Cloud Computing, 2024, 13(1), 119. https://doi.org/10.1186/s13677-024-00675-z

29. Ziani, S. Enhancing Fetal Electrocardiogram Classification: A Hybrid Approach Incorporating Multimodal Data Fusion and Advanced Deep Learning Models. Multimedia Tools and Applications, 2024, 83(18), 55011-55051. https://doi.org/10.1007/s11042-023-17305-6

30. Ziani, S., Rizal, A., Ziani, Y. Refining Fetal Electrocardiogram Classification: A Hybrid Approach with Multimodal Data Fusion and Advanced Deep Learning. In International Conference on Connected Objects and Artificial Intelligence, Cham: Springer Nature Switzerland, May 2024, 378-388. https://doi.org/10.1007/978-3-031-70411-6_57