



A Hybrid LSTM–Random Forest Framework for Explainable Fetal Health Classification Using Cardiotocography Signals

P Sowjanya¹, Prof. D Lalitha Bhaskari²

^{1,2} Department of Computer Science and Systems Engineering,
Andhra University College of Engineering(A), Visakhapatnam, Andhrapradesh-530003

How to Cite this Article:

Sowjanya, P. (2026). A Hybrid LSTM–Random Forest Framework for Explainable Fetal Health Classification Using Cardiotocography Signals. International Journal of Creative and Open Research in Engineering and Management, 2(9), 1-9.
<https://doi.org/10.55041/ijcope.v2i9.172>

License:

This article is published under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

© The Author(s). Published by International Journal of Creative and Open Research in Engineering and Management.



<https://doi.org/10.55041/ijcope.v2i9.172>

Abstract: Cardiotocography (CTG) is an established method for monitoring fetal well-being through measurements associated with fetal heart rate and uterine contractions. Automated classification of CTG records into Normal, Suspect, and Pathological categories can support clinicians by providing rapid and consistent decision-support information. Conventional machine-learning techniques, particularly Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost), have demonstrated strong performance on publicly available CTG datasets. Deep-learning models such as Long Short-Term Memory (LSTM) networks have also been investigated because of their ability to learn nonlinear representations. This paper proposes a hybrid LSTM–Random Forest framework in which an LSTM network performs representation learning and the learned representation is subsequently classified using Random Forest. The framework is designed for the UCI Cardiotocography dataset, which contains 2,126 CTG records and 21 automatically extracted features with three fetal-state classes. A leakage-resistant experimental protocol is proposed using stratified partitioning, training-only preprocessing, class-imbalance handling, and multiple evaluation metrics including accuracy, balanced accuracy, precision, recall, macro-F1, Matthews correlation coefficient, and ROC-AUC. Explainability is incorporated using SHAP. Published literature provides benchmarks against which the proposed framework can be evaluated. Reported results include RF accuracies of 94.90% and 97.20% under different experimental protocols, XGBoost accuracy of

95.27%, LSTM accuracy of 93.4%, and hybrid/stacked approaches reaching 95.39% and 98.9% accuracy in different studies.

Keywords: Cardiotocography, fetal health classification, fetal heart rate, uterine contractions, Random Forest, Long Short-Term Memory, XGBoost, hybrid machine learning.

INTRODUCTION

Fetal health assessment is an important component of antenatal and intrapartum care. Cardiotocography (CTG) is routinely used to observe fetal heart-rate behavior and uterine contraction activity. Abnormal patterns may indicate fetal compromise and can require additional clinical assessment or intervention. Because manual CTG interpretation can be complex and subject to inter-observer variation, machine-learning techniques have increasingly been investigated as decision-support tools.

Machine-learning studies have demonstrated that conventional classifiers can achieve strong results on this dataset. Hoodbhoy *et al.* investigated several machine-learning methods for fetal-risk prediction and reported approximately 93% overall test accuracy using XGBoost, with particularly strong sensitivity for pathological cases [2]. Rahmayanti *et al.* compared Artificial Neural Networks, LSTM, XGBoost, SVM, KNN, LightGBM,



and Random Forest and reported strong performance from several conventional machine-learning algorithms [3].

Hybrid approaches have also attracted attention. Feng *et al.* developed a stacked ensemble using SVM, XGBoost, and Random Forest as base learners and a backpropagation network as the meta-learner. Their public-dataset experiment achieved 95.39% accuracy and an average F1-score of 92.49% using ten-fold cross-validation [5]. More recently, Ahmed and Mahmoud evaluated multiple machine-learning and deep-learning models and reported 93.4% accuracy for LSTM and 98.9% accuracy for their stacking classifier, with the stacking model obtaining 99.0% precision, 98.6% recall, 99.3% F1-score, and 99.8% ROC-AUC [6].

This study therefore proposes a hybrid LSTM–Random Forest framework. The LSTM is used as a nonlinear representation-learning component, while RF performs the final multiclass classification. The proposed methodology also incorporates class-imbalance-aware evaluation and SHAP-based explainability.

The principal contributions of this study are:

1. Development of a hybrid LSTM–Random Forest architecture for three-class CTG fetal-health classification.
2. Definition of a leakage-resistant preprocessing and evaluation protocol.
3. Comparative evaluation against SVM, RF, XGBoost, and standalone LSTM baselines.
4. Use of accuracy, balanced accuracy, precision, recall, macro-F1, MCC, and ROC-AUC for comprehensive evaluation.
5. Incorporation of SHAP-based model interpretation.
6. Comparison with verified numerical results reported in previous CTG studies.

II. RELATED WORK

A. Machine Learning for CTG Classification

The structured nature of CTG-derived features makes the dataset suitable for conventional machine-learning algorithms. Random Forest is particularly attractive because it can model nonlinear relationships and interactions without requiring strong assumptions regarding feature distributions. SVM has also been widely applied to CTG classification because of its ability to construct nonlinear decision boundaries through kernel functions. XGBoost uses sequential gradient-boosted decision trees and has demonstrated strong performance in structured classification tasks.

A recent benchmark study reported test accuracy values of 93.01% for SVM, 94.90% for RF, and 95.27% for XGBoost under its particular preprocessing and evaluation protocol [7]. These values should not be interpreted as universal performance levels because different studies use different data splits, class-balancing methods, feature-selection procedures, and optimization strategies. Hardalaç *et al.* demonstrated the effect of feature engineering and optimization by obtaining 97.20% accuracy using RF [4]. This indicates that the preprocessing and optimization pipeline can be as important as the choice of classifier.

B. LSTM-Based CTG Classification

LSTM is a recurrent neural-network architecture designed to learn representations from ordered input sequences through gated memory mechanisms. LSTM has been applied to CTG-related classification because fetal heart-rate behavior is inherently temporal.

However, the standard UCI Cardiotocography dataset contains 21 automatically extracted features rather than the original continuous FHR and UC time series [1]. Therefore, applying LSTM to the 21 features represents feature-level representation learning rather than direct raw-signal temporal modeling.



Ahmed and Mahmoud reported 93.4% accuracy for their LSTM model, while their overall stacking approach obtained 98.9% accuracy [6]. These results provide useful published benchmarks for the present study.

C. Hybrid and Ensemble CTG Models

Hybrid approaches attempt to combine the strengths of multiple learning algorithms.

Feng *et al.* proposed a stacked ensemble consisting of SVM, XGBoost, and RF base learners with a BP meta-learner. The model achieved 95.39% accuracy and 92.49% average F1-score on the public dataset using ten-fold cross-validation[4] [5][6]. The same study reported that the individual RF model achieved 89.49% accuracy, while SVM and XGBoost achieved 93.15% and 93.14%, respectively, under its ten-fold cross-validation protocol. The proposed stacked strategy achieved 95.39% accuracy[4] [5].

RESEARCH GAP

Despite the considerable progress in automated CTG classification, three research gaps remain.

First, many studies emphasize overall accuracy despite the class imbalance present in the fetal-health dataset. A model can obtain high overall accuracy while performing relatively poorly on the Suspect or Pathological classes.

Second, conventional machine-learning models and deep-learning models are frequently investigated independently. There is comparatively less emphasis on using a neural network specifically as a representation-learning stage followed by a robust tabular classifier such as RF. Third, there is a methodological distinction between raw CTG sequence learning and learning from automatically extracted CTG features. The UCI dataset represents the latter. Consequently, a careful description is required when LSTM is used with this dataset.

The proposed study addresses these issues by using the LSTM as a feature-representation module, RF as the final classifier, and multiple class-sensitive evaluation metrics.

IV. DATASET

A. UCI Cardiotocography Dataset

The proposed study uses the UCI Cardiotocography dataset [1]. The dataset contains:

- 2,126 CTG records.
- 21 numerical CTG-derived features.
- Three fetal-state classes.
- No missing values according to the UCI repository.
- Expert-derived consensus labels.

The three target classes are:

Class	Label	Description
Normal	0	Normal fetal state
Suspect	1	Suspected fetal compromise
Pathological	2	Pathological fetal state

The dataset features include variables associated with baseline fetal heart rate, accelerations, fetal movements, uterine contractions, decelerations, short-term variability, long-term variability, and related CTG measurements [1].



B. Dataset Limitation

The dataset should not be described as raw CTG waveform data. The UCI repository explicitly states that the 2,126 cardiotocograms were automatically processed and diagnostic features were measured [1].

Therefore, the LSTM in the proposed framework learns nonlinear representations of CTG-derived features rather than directly modeling continuous fetal-heart-rate sequences.

PROPOSED METHODOLOGY

A. Overall Architecture

The proposed system consists of the following stages:

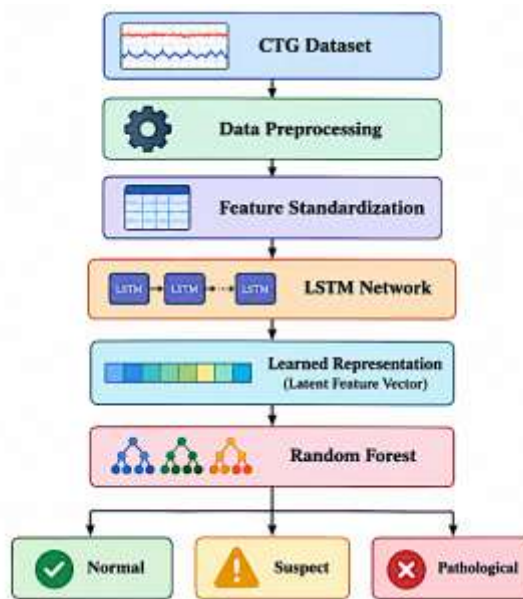


Fig. 1. Overall Architecture of the Proposed LSTM–Random Forest Framework for Fetal Health Classification Using CTG Data.

B. Data Preprocessing

The preprocessing pipeline consists of:

1. Loading the UCI CTG dataset.
2. Separating input features and target labels.
3. Checking for duplicates and inconsistent records.
4. Performing stratified train-test splitting.
5. Fitting normalization parameters exclusively on the training set.
6. Applying the same transformation to the validation and test sets.
7. Addressing class imbalance using training-only class weighting or resampling.
8. Reshaping the feature representation for LSTM input.

No test-set information is permitted to influence preprocessing or hyperparameter optimization.



LSTM REPRESENTATION LEARNING

A. LSTM Representation Learning

An LSTM network is employed as the representation-learning component of the proposed framework. The gated architecture of LSTM enables the network to transform the input feature sequence into a nonlinear latent representation. For an input vector x_t at time step t , the LSTM computes the forget gate, input gate, candidate cell state, cell state, output gate, and hidden state as follows:

$$\begin{aligned} f_t &= \sigma(W_f[h_{t-1}, x_t] + b_f) \\ i_t &= \sigma(W_i[h_{t-1}, x_t] + b_i) \\ \tilde{C}_t &= \tanh(W_C[h_{t-1}, x_t] + b_C) \\ C_t &= f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \\ o_t &= \sigma(W_o[h_{t-1}, x_t] + b_o) \\ h_t &= o_t \odot \tanh(C_t) \end{aligned}$$

where f_t , i_t , and o_t denote the forget, input, and output gates, respectively; C_t represents the cell state; h_t denotes the hidden state; W_f , W_i , W_C , and W_o are trainable weight matrices; b_f , b_i , b_C , and b_o are bias vectors; $\sigma(\cdot)$ denotes the sigmoid activation function; and $\odot$ represents element-wise multiplication.

The final hidden state is extracted as the learned latent representation:

$$z = h_T$$

where T represents the final input position and z denotes the learned latent feature vector generated by the LSTM.

RANDOM FOREST CLASSIFIER

Random Forest is used as the final classifier because of its ability to model nonlinear relationships in structured feature spaces.

For an ensemble of MM decision trees, the Random Forest classifier determines the final class through majority voting among the individual tree predictions

The principal hyperparameters include:

- Number of trees.
- Maximum tree depth.
- Minimum samples per split.
- Minimum samples per leaf.
- Number of features considered at each split.
- Class weighting.

Hyperparameter optimization should be performed only on the training set through stratified cross-validation.

BASELINE MODELS

The following baseline models should be evaluated:

1. Support Vector Machine.
2. Random Forest.



3. XGBoost.
4. Standalone LSTM.
5. Proposed LSTM–RF.

Where computational resources permit, Logistic Regression and LightGBM can also be included.

The same train-test split should be used for all models to provide a fair experimental comparison.

EXPERIMENTAL PROTOCOL

A. Data Splitting

The recommended primary protocol is:

- 80% training.
- 20% independent testing.
- Five-fold stratified cross-validation within the training set.

The independent test set is not used for model selection.

B. Class Imbalance

The dataset is imbalanced, making accuracy alone insufficient.

The proposed evaluation therefore reports:

- Accuracy.
- Balanced accuracy.
- Macro precision.
- Macro recall.
- Macro-F1.
- MCC.
- ROC-AUC.
- Per-class precision.
- Per-class recall.
- Per-class F1.

PUBLISHED LITERATURE BENCHMARK

The following values are **TABLE I: Published CTG Classification Results**

Study	Model	Dataset/Protocol	Accuracy	Other reported metric
Hoodbhoy <i>et al.</i> [2]	XGBoost	CTG test set	~93.0%	Pathological sensitivity >92%
Rahmayanti <i>et al.</i> [3]	Multiple ML/DL	UCI CTG	89–99% range across scenarios/models	—
Feng <i>et al.</i> [5]	RF	Public CTG, 10-fold CV	89.49%	F1 = 77.18%
Feng <i>et al.</i> [5]	SVM	Public CTG, 10-fold CV	93.15%	F1 = 87.82%
Feng <i>et al.</i> [5]	XGBoost	Public CTG, 10-fold CV	93.14%	F1 = 87.94%



Feng <i>et al.</i> [5]	Stacked SVM-XGB-RF-BP	Public CTG, 10-fold CV	95.39%	F1 = 92.49%
Akmal <i>et al.</i> [8]	Autoencoder-RFE-RF	UCI CTG	96.62%	—
Hardalaç <i>et al.</i> [4]	Optimized RF	UCI CTG	97.20%	MCC = 0.919
Ahmed and Mahmoud [6]	LSTM	CTG	93.4%	F1 = 89.3%
Ahmed and Mahmoud [6]	Stacking	CTG	98.9%	F1 = 99.3%; AUC = 99.8%
Proposed study	LSTM–RF	UCI CTG	To be experimentally determined	—

The published RF result of 97.20% was obtained using feature elimination and hyperparameter optimization [4]. The 95.39% stacked result was obtained by Feng *et al.* using ten-fold cross-validation [5]. The 98.9% stacking result was reported by Ahmed and Mahmoud under their specified preprocessing and modeling scenarios [6].

These differences demonstrate that reported performance depends strongly on experimental methodology.

Ahmed and Mahmoud reported a substantially higher 98.9% accuracy for their stacking classifier and 93.4% for LSTM [6]. The difference illustrates the importance of experimental protocol, preprocessing, feature engineering, and ensemble construction.

COMPARISON WITH PREVIOUS HYBRID APPROACHES

The proposed LSTM–RF model belongs to a broader class of hybrid CTG systems.

Feng *et al.* used a stacked SVM-XGBoost-RF architecture with a BP meta-learner and achieved 95.39% accuracy on the public dataset [5]. Akmal *et al.* combined autoencoder feature extraction, recursive feature elimination, Bayesian optimization, and RF and reported 96.62% accuracy [8]. Hardalaç *et al.* further demonstrated the value of feature elimination and hyperparameter optimization, obtaining 97.20% accuracy using RF [4]. Ahmed and Mahmoud reported 98.9% accuracy for a stacking classifier combining multiple machine-learning and deep-learning models [6]. Therefore, the contribution of the proposed LSTM–RF system should not be framed simply as an attempt to achieve the highest possible accuracy. Its scientific contribution should instead be evaluated in terms of:

- representation learning,
- minority-class performance,
- model robustness,
- interpretability,
- computational complexity,
- reproducibility, and
- generalization.

FUTURE WORK

Future research should investigate the following directions.

A. Raw CTG Sequence Learning: Future work should obtain raw FHR and UC sequences and investigate true temporal architectures including:

- LSTM.



- BiLSTM.
- CNN-LSTM.
- Temporal convolutional networks.
- Transformers.

B. Multimodal Learning: CTG information could be combined with maternal and pregnancy-related information such as:

- maternal age,
- gestational age,
- parity,
- pregnancy complications,
- obstetric history.

C. External Validation: The proposed system should be evaluated on independent datasets collected from different hospitals and populations.

D. Clinical Outcome Prediction: Future research should investigate clinically meaningful endpoints such as neonatal acidemia and adverse neonatal outcomes when appropriate ground-truth data are available.

E. Explainable AI: Additional explainability techniques such as Integrated Gradients, attention visualization, counterfactual explanations, and permutation importance could complement SHAP.

F. Prospective Evaluation: Before clinical deployment, prospective studies should evaluate:

- calibration,
- robustness,
- subgroup performance,
- false-negative rates,
- clinician interaction,
- workflow integration,
- clinical outcomes.

CONCLUSION

This paper presented a hybrid LSTM–Random Forest framework for fetal-health classification from cardiotocography-derived features. The framework uses LSTM as a nonlinear representation-learning stage and Random Forest as the final classifier.

The literature demonstrates that conventional machine-learning approaches are highly competitive for CTG classification. Published results include approximately 93% accuracy for XGBoost in an earlier study, 94.90% for RF and 95.27% for XGBoost in a recent benchmark, 96.62% for an optimized autoencoder-RFE-RF framework, and 97.20% for a feature-elimination and hyperparameter-optimized RF approach [2], [4], [7], [8]. Hybrid approaches have also demonstrated strong performance, including 95.39% accuracy for a stacked SVM-XGBoost-RF-BP model and 98.9% for a more recent stacking framework [5], [6].

The proposed LSTM–RF framework therefore provides a structured methodology for investigating whether neural representation learning can complement Random Forest classification. The final experimental evaluation should compare RF, SVM, XGBoost, LSTM, and LSTM–RF under an identical leakage-resistant protocol and should emphasize balanced accuracy, macro-F1, MCC, ROC-AUC, and Pathological-class recall in addition to overall accuracy.



A key methodological consideration is that the UCI CTG dataset contains automatically extracted CTG features rather than raw continuous CTG signals. Consequently, the proposed LSTM should be interpreted as a feature-representation learner rather than a raw temporal-signal model. Future work using raw FHR and UC sequences could provide a stronger test of the temporal modeling capabilities of LSTM and other sequence-learning architectures.

REFERENCES

- [1] D. Campos and J. Bernardes, “Cardiotocography,” UCI Machine Learning Repository, Dataset 193, 2010, doi: 10.24432/C51S4N.
- [2] Z. Hoodbhoy, M. Noman, A. Shafique, A. Nasim, D. Chowdhury, and B. Hasan, “Use of machine learning algorithms for prediction of fetal risk using cardiotocographic data,” *Int. J. Appl. Basic Med. Res.*, vol. 9, no. 4, pp. 226–230, 2019, doi: 10.4103/ijabmr.IJABMR_370_18.
- [3] N. A. Rahmayanti, H. N. Pradani, M. R. Pahlawan, and R. A. Vinarti, “Comparison of machine learning algorithms to classify fetal health using cardiotocogram data,” *Procedia Comput. Sci.*, vol. 197, pp. 162–171, 2022, doi: 10.1016/j.procs.2021.12.130.
- [4] F. Hardalaç, H. Akmal, K. Ayturan, U. R. Acharya, and R.-S. Tan, “A pragmatic approach to fetal monitoring via cardiotocography using feature elimination and hyperparameter optimization,” *Interdiscip. Sci.: Comput. Life Sci.*, vol. 16, pp. 882–906, 2024, doi: 10.1007/s12539-024-00647-6.
- [5] J. Feng, J. Liang, Z. Qiang, Y. Hao, X. Li, L. Li, Q. Chen, G. Liu, H. Wei, *et al.*, “A hybrid stacked ensemble and Kernel SHAP-based model for intelligent cardiotocography classification and interpretability,” *BMC Med. Inform. Decis. Mak.*, vol. 23, Art. no. 273, 2023, doi: 10.1186/s12911-023-02378-y.
- [6] S. S. Ahmed and N. M. Mahmoud, “Early detection of fetal health status based on cardiotocography using artificial intelligence,” *Neural Comput. Appl.*, vol. 37, pp. 16753–16779, 2025, doi: 10.1007/s00521-025-11343-x.
- [7] C. Puri, K. T. V. Reddy, and P. M. Gote, “Machine learning-based fetal health prediction and development of smart web application,” *Frontiers in Artificial Intelligence*, 2026.
- [8] H. Akmal, F. Hardalaç, and K. Ayturan, “A fetal well-being diagnostic method based on cardiotocographic morphological pattern utilizing autoencoder and recursive feature elimination,” *Diagnostics*, vol. 13, no. 11, Art. no. 1931, 2023, doi: 10.3390/diagnostics13111931.
- [9] J. Breiman, “Random forests,” *Machine Learning*, vol. 45, pp. 5–32, 2001.
- [10] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [11] L. Breiman, “Bagging predictors,” *Machine Learning*, vol. 24, pp. 123–140, 1996.
- [12] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794.
- [13] S. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.



[14] S. Lundberg *et al.*, “From local explanations to global understanding with explainable AI for trees,” *Nature Machine Intelligence*, vol. 2, pp. 56–67, 2020.

[15] Prachi, P. Sabherwal, M. Agrawal, and A. Sharma, “Precision unveiled in unborn: A cutting-edge hybrid machine learning approach for fetal health state classification,” *Cardiovascular Engineering and Technology*, vol. 16, no. 6, pp. 641–662, 2025, doi: 10.1007/s13239-025-00800-2.