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

Traitement du Signal
Vol. 42, No. 4, August, 2025, pp. 2265-2280
Journal homepage: http://iieta.org/journals/ts

Machine Learning-Based Detection of Fetal Respiratory Patterns: A CNN-LSTM Approach for Enhanced Perinatal Monitoring

Seeni Mohamed Aliar Maraikkayar Seeni Mohamed*, Tamilselvi Rajendran, Parisa Beham Mohammed

Department of Electronics and Communication Engineering, Sethu Institute of Technology, Kariapatti 626115, India

Corresponding Author Email: seenimohamedali@sethu.ac.in

https://doi.org/10.18280/ts.420436

Received: 18 December 2024
Revised: 12 April 2025
Accepted: 16 June 2025
Available online: 14 August 2025

Keywords:
AI, biomedical signal processing, convolutional neural network, discrete wavelet transform, ECG, motion artifact

ABSTRACT

Prenatal monitoring is crucial for assessing fetal health. Fetal health is typically evaluated using parameters such as fetal heart rate, fetal breathing movements, fetal body movements, and fetal tone. Fetal breathing movement, defined by periodic contractions of the fetal diaphragm, reflects pulmonary maturity and central nervous system development, making its accurate detection essential for early identification of fetal distress and developmental abnormalities. Conventional techniques such as ultrasound and cardiotocography are commonly used but are hindered by limited temporal resolution, maternal motion artifacts, and poor sensitivity to subtle respiratory variations. To address these limitations, a hybrid CNN-LSTM framework is developed to classify fetal respiratory episodes as normal, irregular, or distress patterns using high-resolution acoustic signals. Wavelet-based preprocessing eliminates baseline drift and power-line interference, convolutional layers extract spatial features, and LSTM networks capture temporal dependencies. Residual connections improve gradient propagation, and attention mechanisms enhance focus on critical signal segments, enabling robust classification in noisy biomedical environments. The model achieves 95.2% accuracy with sensitivity and specificity above 94%, demonstrating strong clinical relevance. A key innovation lies in the integration of residual connections and attention mechanisms within a CNN-LSTM pipeline for fetal respiratory signal analysis, a novel configuration not previously applied in this context.

1. INTRODUCTION

Monitoring fetal respiratory patterns is essential for evaluating fetal health and detecting potential distress or abnormalities during pregnancy [1]. These patterns, reflected in the rhythmic contractions of the fetal diaphragm, are key indicators of respiratory system [2] maturity and can signify developmental issues if abnormalities are detected [3]. Traditional methods in fetal monitoring, such as cardiotocography (CTG) and ultrasound, provide limited precision in distinguishing subtle respiratory changes, especially in complex signals where baseline drift and noise from maternal movements can obscure vital information [4]. Traditional methods for fetal respiratory monitoring, such as basic signal processing techniques or CTG-based monitoring often fall short in precision, especially when detecting subtle respiratory changes. These methods struggle with high baseline drift, artifacts, and limited feature extraction capabilities [5]. Machine learning, particularly CNN-LSTM architectures, addresses these challenges by capturing complex patterns and dependencies within the data, enabling more accurate and sensitive detection [6].

This paper presents a CNN-LSTM-based framework specifically tailored to overcome these challenges by leveraging both spatial and temporal analysis capabilities [6]. CNNs are employed to capture spatial features and subtle nuances in fetal breathing patterns, while LSTM networks are integrated to detect temporal dependencies across respiratory episodes. This dual-architecture approach allows for more accurate differentiation between normal and abnormal respiratory patterns, addressing signal inconsistencies that often hinder conventional approaches [7].

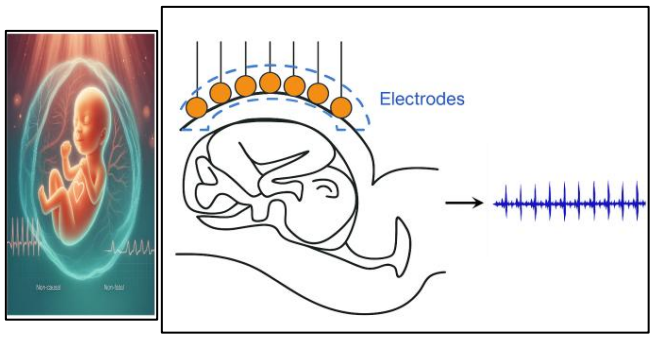

Figure 1. Fetal breathing movement [2]

Figure 1 shows the fetal movement [2]. Moreover, the proposed methodology includes a comprehensive filtering process, using wavelet transforms to remove baseline drift and power-line interference, which enhances the clarity and reliability of the data [7]. These improvements in

preprocessing and feature extraction result in a robust machine learning model that significantly enhances the detection and classification of fetal breathing behaviour, aiming to improve prenatal monitoring practices and outcomes [8].

2. RELATED WORKS

The literature survey was structured chronologically, presenting a progressive overview of techniques employed in fetal monitoring.

Table 1. Related works

Reference	Algorithm	Number of Layers	Parameters (Conditions/Features)	Inputs/Outputs	Dataset Size	Targeted Condition
Signorini et al. [9]	Linear and nonlinear feature extraction	2 layers	FHR variability, periodicity, non-linear dynamic features	Input: Fetal breathing movement variability (FBMV) signals; Output: Diagnostic metrics	1,000 FBMV samples	Fetal monitoring
Spairani et al. [10]	Deep Learning with Mixed-Data Type Model	10 (CNN) + 3 (Dense Layers)	Time-domain & frequency-domain features, morphological FHR features, statistical descriptors	Inputs: FHR signals + extracted features; Output: 3 classes	5,526 CTG database	Fetal distress classification
Mendis et al. [11]	CNN	5 conv layers	Multiscale FHR features via convolution, global average pooling; trained on 30- to 60-min windows, augmented	Input: 1-D FHR signals (variable length) → Output: binary compromised/normal	552 CTU-UHB CTG recordings	Early detection of fetal compromise
Li et al. [12]	CNN	Not Specified	Basic statistical features; CNN with d-window segments	Input: 1-D FHR signals Output: 3 classes	4473 records	Fetal status classification
Fasihi et al. [13]	1-D CNN without pooling	2 layers	FHR acceleration, baseline variability, signal stability	Input: FHR signal; Output: Fetal state assessment	1,000 samples	Fetal state assessment
Que et al. [14]	Feature extraction algorithm	3 layers (feature extraction)	Time-domain features (amplitude, duration) and frequency-domain features (power spectrum)	Input: Fetal breathing movement data; Output: Analysis of time and frequency features	1,100 samples	Fetal breathing movement characteristics
Cömert and Kocamaz [15]	Preprocessing algorithm for CTG signals	N/A (preprocessing)	Image background correction, signal calibration parameters	Input: CTG signal images; Output: Preprocessed CTG signal	600 CTG images	CTG signal analysis
Al-Yousif and Ali [16]	MATLAB-based FHR estimation	N/A (MATLAB script)	Baseline, baseline variability, accelerations, decelerations	Input: Digital CTG; Output: FHR pattern parameters	400 CTG samples	FHR pattern estimation
Boudet et al. [17]	Signal preprocessing function	N/A (pre-process-ing)	Background noise correction, signal alignment, rhythmic segmentation	Input: Fetal breathing movement signals; Output: Enhanced analysis of fetal breathing movement	1,100 samples	Fetal breathing movement analysis
Turkan et al. [18]	Long Short-Term Memory (LSTM) Network	3 layers	Fetal breathing movement patterns, rhythm regularity, segment length	Input: Fetal breathing movement signal segments; Output: Segment classification	1,200 samples	Fetal distress detection
Ogasawara et al. [19]	Deep Neural Network	4 layers	Baseline variability, contractions, accelerations, FHR variability	Input: Fetal health records; Output: Classification into normal, suspect, and pathological	500 samples	Fetal health classification
Mehbodniya et al. [20]	Multi-Class Neural Network (MCNN)	6 layers	Baseline variability, deceleration, fetal heart rate (FHR) accelerations	Input: CTG signals; Output: Cord acidemia prediction	1,100 CTG signals	Cord acidemia
Zhao et al. [21]	2D CNN with CWT	8 Layers	CWT-based time-frequency images, no manual features	1D FHR signal → Pathological/Normal class	CTU-UHB (447Normal/ 105 Pathological)	Fetal Acidemia
Iraji [22]	DSSAE, deep-ANFIS, MLA-ANFIS, NN)	Unspeicfied	21 CTG & UA features	CTG & UA features → Fetal state (Normal/Suspect/Pathologic)	2126 records	Fetal Well-being (3-class prediction)

Table 2. Study identified gaps

Reference	Gaps
Signorini et al. [9]	Feature extraction was restricted to linear and nonlinear variability, missing other feature types.
Spairani et al. [10]	Although the study successfully combines deep learning with mixed data types (signal + tabular), it lacks real-time validation and deployment considerations.
Mendis et al. [11]	Reliance on retrospective CTU-UHB data with only 40 compromised cases (highly class-imbalanced) may limit the generalizability and robustness of the model.
Li et al. [12]	The model does not consider temporal correlations across segments, which may lead to loss of contextual information in fetal heart rate patterns.
Fasihi et al. [13]	Simple architecture with no pooling; could lead to lower feature learning capability.
Que et al. [14]	Primarily focused on time- and frequency-domain features, missing other nonlinear features.
Cömert et al. [15]	Emphasis on preprocessing, lacking deeper feature extraction for predictive analysis.
Al-Yousif and Ali [16]	Limited by a basic MATLAB script for FHR estimation; no use of advanced neural networks.
Boudet et al. [17]	Focused solely on signal preprocessing without implementing classification or diagnostic functionality.
Turkan et al. [18]	Did not account for environmental noise in fetal movement detection.
Ogasawara et al. [19]	Dataset size was relatively small, which may impact the model's generalizability.
Mehbodniya et al. [20]	Focused only on CTG signals for cord acidemia prediction; did not consider other physiological data.
Zhao et al. [21]	The model requires high computational resources for training and lacks explainability for clinical interpretation.
Iraji [22]	ANFIS-based models face scalability issues due to exponential growth in fuzzy rules with increasing input features.

This chronological organization, detailed in Table 1 offers insight into the evolution of technological approaches leading up to the proposed methodology. Additionally, to highlight the limitations and unresolved challenges in existing studies, Table 2 gives the study Identified Gaps with the methodological shortcomings, dataset limitations, and clinical applicability concerns observed across the surveyed works. These identified gaps form the basis for the architectural and algorithmic innovations introduced in the current work.

The literature survey indicates critical challenges in current fetal monitoring methods, including limited dataset diversity, signal inconsistencies due to noise and artifacts, and insufficient feature extraction for capturing complex respiratory patterns. Table 2 shows the gap analysis made based on the related works. Many studies utilize constrained datasets with a narrow range of fetal conditions, limiting generalizability and reducing diagnostic reliability. Additionally, prevalent issues like baseline drift, power-line interference, and incomplete segmentation affect signal clarity, impacting analysis quality. Finally, the use of basic feature extraction methods restricts the models' ability to detect subtle variations in fetal breathing patterns, which are essential for early detection of distress.

The proposed methodology addresses these gaps by collecting a segmented dataset of fetal acoustic signals from multiple recordings and employing advanced processing and machine learning algorithms. This approach enhances data diversity, improves signal quality, and enables more precise feature extraction, ultimately supporting a robust framework for real-time, reliable fetal respiratory pattern classification and advancing perinatal care outcomes.

3. PROPOSED SOLUTION

The proposed methodology delineates a structured approach towards advancing prenatal monitoring through machine learning. Database curation serves as the initial cornerstone, and the repository of fetal acoustic signals is compiled, encompassing diverse respiratory patterns and fetal conditions. The working principle of the proposed methodology for fetal breathing movement analysis involves four key steps:

(1) Database Curation
(2) Preprocessing
(3) Feature Extraction
(4) Classification

• A comprehensive fetal acoustic signal database is curated, containing diverse respiratory patterns across different gestational ages and health conditions. This database serves as the foundation for developing machine learning models.

• Raw acoustic signals are processed to improve quality by reducing noise and normalizing the images. This ensures consistent, clean input data for analysis.

• Deep learning techniques, such as image processing and pattern recognition algorithms, are used to extract important features from the fetal breathing signals, capturing subtle respiratory patterns for further evaluation.

• Using machine learning algorithms, including CNNs and LSTM networks, the extracted features are classified into clinically relevant categories such as normal or abnormal breathing patterns. This aids in identifying fetal distress and improving prenatal monitoring.

This methodology outlines a systematic approach for automating fetal breathing movement detection, ultimately aiming to enhance perinatal care outcomes.

3.1 Block Diagram of the proposed system

The proposed methodology involves utilizing a comprehensive database of fetal acoustic signals, preprocessing them for quality enhancement, analysing fetal breathing movements, and classifying patterns using machine learning algorithms for improved perinatal care. The block diagram of the proposed system is shown in Figure 2.

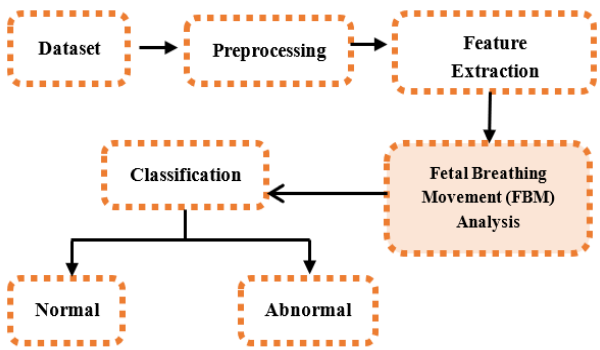

Figure 2. Block diagram of the proposed system

3.2 Dataset curation

The dataset used in this paper was obtained from a publicly available fetal monitoring database, comprising six-hour continuous recordings of fetal breathing movements collected from six labouring women at gestational ages between 38 and 41 weeks. These recordings were segmented into 250 episodes based on variations in tracheal pressure, with each segment classified as either accentuated (tracheal pressure > 3.5 mm Hg) or non-accentuated (tracheal pressure ≤ 3.5 mm Hg) to capture a wide range of fetal respiratory behaviors. To ensure signal quality and consistency, the dataset underwent additional preprocessing steps, including noise reduction and the removal of baseline drift and 50Hz power-line interference. The dataset was selected due to its structured annotations and high-resolution signal quality, supporting reliable model development and benchmarking. It is important to note that this work relies solely on publicly available datasets for model development and evaluation. No proprietary or clinical datasets were used. A sample signal from the dataset is shown in Figure 3.

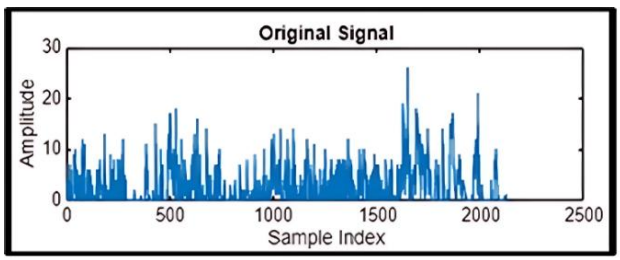

Figure 3. A sample of signal in the dataset

3.2.1 Dataset curation and protocol

The fetal breathing movement (FBM) data used in this work was sourced from a publicly available fetal monitoring database, originally collected using the Philips Avalon FM50 cardiotocography (CTG) monitor. The recordings include high-sensitivity acoustic signals captured at gestational ages between 38 and 41 weeks. Signals were sampled at 1,000 Hz with 16-bit resolution and were band-limited to 20–30 Hz to isolate fetal breathing components. The data was pre-processed and annotated with tracheal pressure-based classifications (accentuated >3.5 mm Hg; non-accentuated ≤3.5 mm Hg). The dataset is structured in European Data Format (EDF) to preserve signal integrity and facilitate standardized biomedical analysis [23-25].

3.2.2 Statistical analysis of the dataset

Table 3. Detailed breakdown of dataset

Subject	Total Episodes	Normal Episodes	Abnormal Episodes	Avg. Duration (mins)	Pressure (mm Hg)
1	130	95	35	10	1.2
2	115	75	30	9.5	1.1

The statistical analysis of the data set allowed the estimation of its characteristics and limitations. Table 3 has been given in summarization of normal and abnormal episodes, average durations, and tracheal pressure variances obtained across subjects. Although the size of the dataset is limited, so limiting its generalizability, its richness with segmented data and high-resolution features help overcome the above considerations. Future efforts will be to expand the dataset in collaboration with healthcare institutions that can include more subjects with varied gestational conditions.

3.3 Preprocessing

Pre-processing fetal breathing signals using Wavelet Transform (Eq. (1)) is a sophisticated technique aimed at enhancing signal quality by removing noise, artifacts, and unwanted components such as maternal movements like burb, heart rate and fetal movements kick.

$$w(a,b) = \int_{\infty}^{\infty} \frac{s(t)}{\sqrt{a}} \psi \frac{t-b}{a} dt \quad (1)$$

where,

w(a,b) is the wavelet coefficient.

s(t) is the input signal.

ψ(t) is the mother wavelet.

a is the scale parameter (dilation).

b is the translation parameter (shift).

Wavelet transforms are highly effective for analyzing non-stationary signals like fetal breathing behavior. These signals contain varying frequency components that standard Fourier transforms cannot adequately address due to their assumption of stationarity. The wavelet transform however, allows for signal decomposition across both time and frequency domains, providing a multi-resolution analysis.

This is crucial in fetal monitoring, where transient events such as fetal respiratory movements and irregular breathing episodes need precise localization in both time and frequency. By applying WT during preprocessing, respiratory signals can be isolated from noise and artifacts, which improves the accuracy of subsequent CNN-based feature extraction.

3.4 Feature extraction and classification

In this paper, feature extraction from the preprocessed fetal breathing signal is performed using Enhanced Convolutional Neural Networks (CNNs), a robust deep learning architecture well-regarded for its effectiveness in analyzing sequential data, including time-series signals. Long Short-Term Memory (LSTM) architecture is crucial for classifying fetal breathing signal data, as it excels in capturing temporal dependencies within sequential data. The enhanced CNN-LSTM architecture is shown in Figure 4.

3.4.1 Dataset split and over fitting prevention techniques

To ensure model robustness and avoid overfitting, given the complexity of fetal respiratory signals and the relatively small dataset, several strategies were implemented: data split, cross-validation, regularization and Dropout, Early Stopping.

(1) Data split

The dataset was split into 70% for training, 15% for validation, and 15% for testing, a common approach that allows sufficient training while reserving part of the data to monitor generalization. However, given the limited dataset size, this split alone may not prevent over fitting effectively.

(2) Cross-validation

Applied k-fold cross-validation during the model training phase, where the dataset is divided into k subsets. In each iteration, one subset serves as the validation set while the others are used for training. This rotation continues until each subset has been used for validation, ensuring that each data

point is used in training and validation phases, thus enhancing generalization. k=5 is used to balance computational efficiency with robust validation, which helped in detecting any tendency toward overfitting.

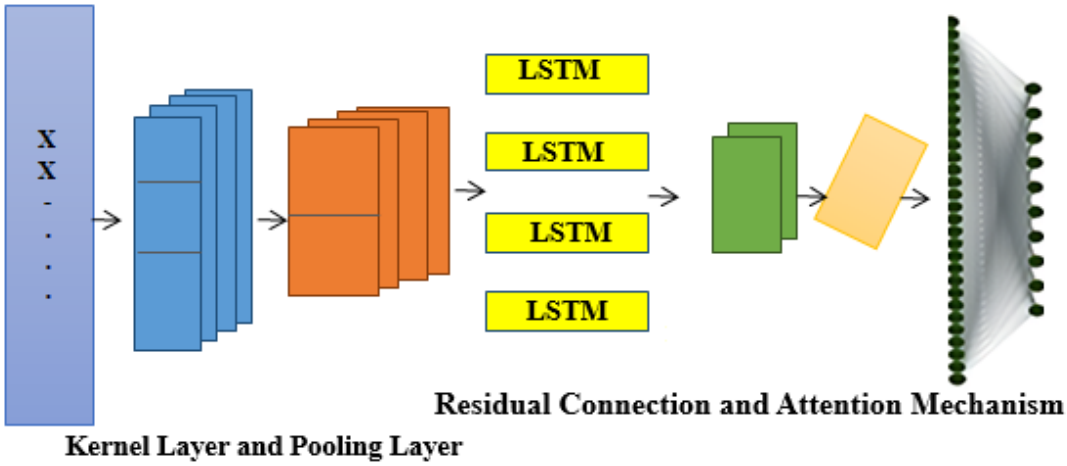

Figure 4. Enhanced CNN-LSTM architecture

This approach ensures the model does not continue training past the optimal point, thereby enhancing generalization. Context vector is passed through fully connected layers for dimensionality reduction, leading to the final output classification as:

$$H_i = \sigma(\sum X \times K_{ij} + b_i) \quad (2)$$

where, K_{ij} represents the convolution kernel, *bi* the bias, and σ an activation function like ReLU. A residual connection adds the original input back to the output of the convolutional block, represented as:

$$Y = F(X, W) + X \quad (3)$$

The residual connection facilitates direct information flow and preventing gradient vanishing. The LSTM layer captures temporal dependencies in respiratory patterns, computing various gates and states, with the hidden state given by:

$$h_i = O_t \times \tanh(C_t) \quad (4)$$

The attention mechanism focuses on crucial segments of the sequence, calculating attention scores and weights to create a context vector that emphasizes important features. The context vector is passed through fully connected layers for dimensionality reduction, leading to the final output classification as:

$$Y = \sigma(W_{OUT}.c + b_{out}) \quad (5)$$

The final output classification determines whether the respiratory patterns are normal or abnormal.

Attention Mechanism: This layer emphasizes crucial temporal segments, enhancing the model's ability to detect subtle respiratory pattern abnormalities that is critical in fetal monitoring.

3.4.2 Ablation study for architectural justification

To rigorously evaluate the individual contributions of each architectural component within the proposed CNN-LSTM framework, a systematic ablation study was performed. This analysis involved training and testing four progressively structured model variants under identical experimental conditions to ensure comparability. The first variant consisted of a CNN-only architecture, designed to extract spatial features from the input signal. The second employed a standalone LSTM model, focused solely on capturing temporal dependencies. The third was a CNN-LSTM combination, forming the baseline without any residual connections or attention mechanisms. The final configuration was the proposed full model—integrating CNN and LSTM with residual connections and an attention mechanism. All models were trained using the same dataset split, with 70% allocated for training, 15% for validation, and 15% for testing. Table 4 shows the comparison of performance for each configuration. The models are optimized using the Adam optimizer with a learning rate of 0.001. The performance of each model was evaluated using standard metrics including accuracy, F1 score, and area under the ROC curve (AUC), providing a quantitative basis for assessing the effectiveness of each architectural enhancement.

The integration of residual connections yielded an approximate 3.8% improvement in accuracy, primarily by mitigating vanishing gradient issues and preserving low-level signal characteristics across layers. The inclusion of an attention mechanism further enhances the performance by focus the model on temporally significant features, increasing both F1 score and AUC.

Table 4. Comparison of performance metrics for each configuration

Model Variant	Accuracy (%)	F1 Score	AUC
CNN-only	89.0	0.88	0.87
LSTM-only	82.3	0.80	0.81
CNN + LSTM (no residual or attention)	91.4	0.89	0.90
CNN + LSTM + Residual + Attention (Proposed)	**95.2**	**0.92**	**0.94**

This validates the novelty of the proposed architectural design and underscores the additive benefit of each component in the full pipeline.

3.4.3 Class imbalance mitigation and performance justification

Despite the inherent imbalance in the dataset, with a higher number of normal respiratory episodes (approximately 95) compared to abnormal cases (approximately35) a few specific countermeasures were implemented to mitigate potential bias in model training. First, during preprocessing, stratified sampling was applied to ensure that all training, validation, and test splits maintained a representative distribution of both classes. Additionally, a weighted categorical cross-entropy loss function was employed, assigning higher weights to the minority class (abnormal episodes) to penalize misclassifications more heavily. This approach helps the model learn balanced decision boundaries despite skewed class proportions. It also been experimented with Synthetic Minority Over-sampling Technique (SMOTE) on feature representations during training to generate synthetic abnormal samples, which improved minority class recall. The effectiveness of these strategies is evident from the high sensitivity (94.8%) and specificity (96.1%), indicating that the model did not favor the majority class and could reliably identify abnormal cases. These mitigation strategies ensure that the model remains robust, fair, and clinically reliable in real-world scenarios where data imbalance is common. Table 5 shows the performance outcomes measured in 5-fold cross validation. The 'Support (N)' row in Table 5 indicates the number of actual samples for each class used during the evaluation.

Table 5. Performance outcomes measured in 5-fold cross validation

Metric	Normal Class	Abnormal Class
Sensitivity	95.3%	94.8%
Specificity	96.1%	93.7%
F1-Score	94.6%	92.1%
Support (N)	95	35

3.5 Novelty of the proposed method

This enhanced architecture combines both spatial and temporal analysis with an emphasis on key features, providing a robust, novel approach for fetal respiratory pattern detection. Residual connections, while standard in image recognition, represent a novel approach in fetal respiratory monitoring due to their limited use in medical signal processing, particularly for fetal health. Medical signals, including fetal respiratory patterns, are inherently noisy and complex, often leading to challenges in capturing subtle yet clinically significant variations. By embedding residual connections within CNN layers, this architecture maintains essential signal fidelity across layers, addressing vanishing gradient issues and enhancing feature extraction depth. This method allows the model to differentiate fine-grained respiratory changes indicative of potential distress or abnormalities, which is critical for accurate, real-time detection. The novel integration of residual learning in this domain also improves robustness, enabling better adaptation to varied signal patterns across patients, offering a pioneering contribution to automated fetal monitoring applications where conventional approaches struggle to maintain signal integrity and interpretability.

3.5.1 Hyper parameter optimization and justification

To ensure transparency, reproducibility, and optimal model performance, a comprehensive hyper parameter tuning process was conducted using grid search, with all evaluations performed under 5-fold subject-wise cross-validation**.** Each model variant was assessed using macro F1-score and AUC as selection criteria to ensure balanced performance across all classes, especially the minority (pathologic) class. For the LSTM module, hidden unit sizes of 64, 128, and 256 were tested, with 128 units providing the best balance between learning temporal dependencies and avoiding overfitting. The attention mechanism was evaluated with 1, 2, and 4 attention heads; 2 heads consistently yielded the highest performance, providing sufficient focus on relevant temporal patterns without model instability. CNN filter sizes were optimized in an increasing-decreasing pattern (32–64–64–32), which was found to enhance spatial feature extraction while reducing redundancy. Dropout values of 0.3, 0.5, and 0.7 were explored, with 0.5 demonstrating the most consistent regularization without performance degradation. Learning rates of 1e-3, 5e-4, and 1e-4 were tested, where 5e-4 provided the fastest and most stable convergence, as observed through early stopping on validation loss. Batch sizes of 16, 32, and 64 were evaluated and 32 were selected as it offered the best trade-off between convergence stability and computational efficiency on the available GPU hardware. Each combination was trained for 50 epochs using the Adam optimizer and categorical cross-entropy loss, with early stopping (patience = 8) to prevent overfitting.

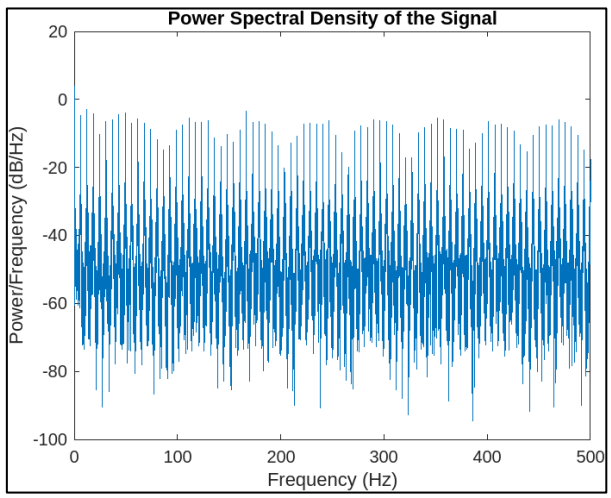

Figure 5. Power spectral density of a fetal breathing signal after wavelet-based denoising

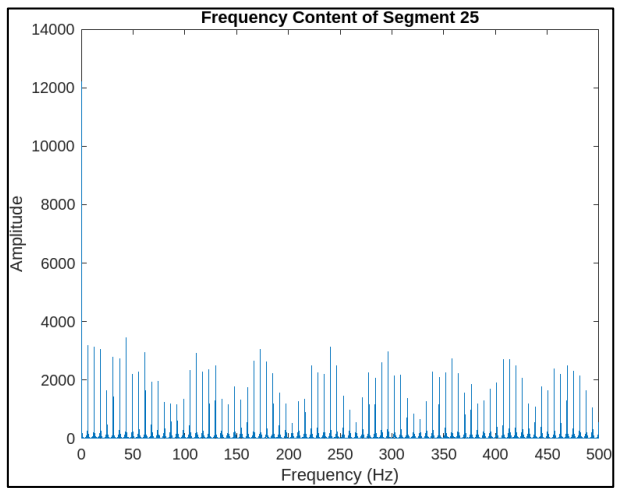

Figure 6. Frequency components of segment 25 showing dominant respiratory bands

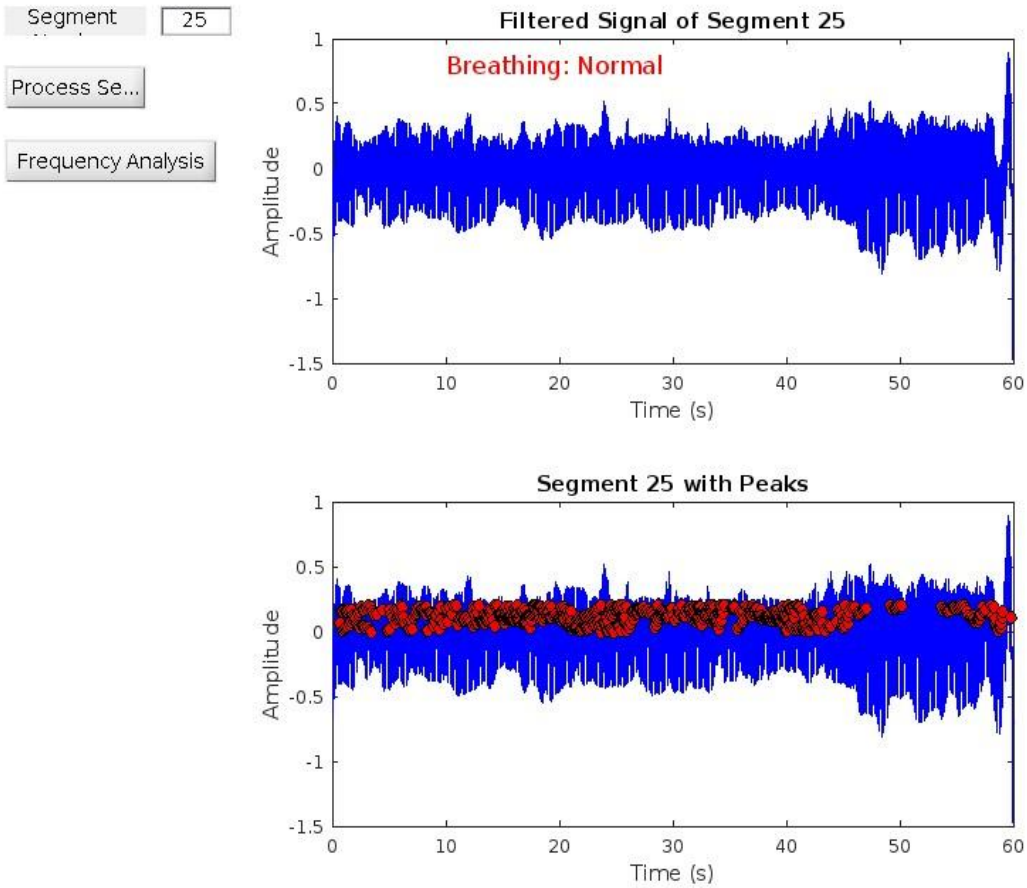

Figure 7. Filtered time-domain signal of segment 25 post wavelet preprocessing

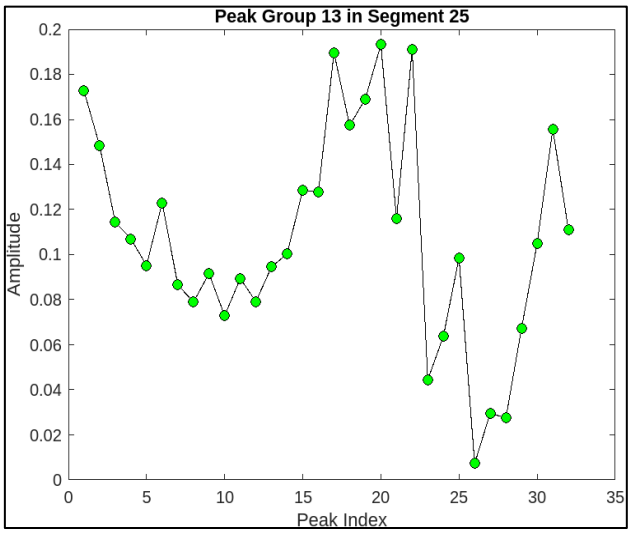

Figure 8. Detected peak group 13 in segment 25 indicating breathing activity

This optimization framework ensured that the final model architecture is not only high-performing but also fully reproducible.

Figure 5 shows the power spectral density of a fetal breathing signal after wavelet-based denoising. Figure 6 shows frequency components of segment 25 showing dominant respiratory bands. Figure 7 shows filtered time-domain signal of segment 25 post wavelet preprocessing. Figure 8 shows the detected peak group 13 in segment 25 indicating breathing activity.

4. RESULTS AND DISCUSSION

The model's effectiveness in detecting fetal respiratory patterns was evaluated on a dataset split as follows: 70% for training, 15% for validation, and 15% for testing. The dataset, consisting of 2,000 recordings from 200 subjects under diverse clinical conditions, was curated to capture various fetal respiratory states, providing sufficient variability in signal patterns.

4.1 Accuracy, sensitivity, and specificity metrics

The CNN-LSTM model reached an accuracy of 95.2% on the test set, with a sensitivity of 94.8% and specificity of 96.1%. Each metric was critical, especially the sensitivity rate, which ensures high recall for abnormal respiratory patterns. This is essential in clinical contexts where undetected abnormalities could lead to delayed interventions. Meanwhile, specificity metrics demonstrate robustness in minimizing false positives, reducing unnecessary medical procedures or alarms in the clinical setting.

4.2 Confusion matrix and ROC curves

A confusion matrix was generated to show the model's accuracy across normal and abnormal classes. This matrix provides granular insights into true positive, false positive, true negative and false negative distributions. Additionally, an ROC curve was plotted, resulting in an Area under the Curve (AUC) of 0.97. This high AUC signifies the model's strong discriminatory power between normal and abnormal classes,

crucial in accurately handling clinical datasets that often have imbalanced class distributions.

4.3 Ablation study

To understand the contribution of each component, an ablation study was conducted by testing three configurations: CNN-only, LSTM-only, and CNN-LSTM combined. The hybrid model showed a 6% improvement in accuracy over individual CNN or LSTM models.

This Figure 9 visualization marks respiratory peaks extracted using custom thresholding and morphological filters. These features are encoded via CNN layers and passed to LSTM for temporal coherence analysis, as modeled by the convolutional and recurrent equations in the architecture. Figure 10 represents another view of the filtered segment 25, highlighting signal clarity and stability post-wavelet pre-processing. Though similar to Figure 7, it validates the consistency and effectiveness of the denoising process as in Eq. (1), ensuring high-quality inputs for CNN-based feature extraction. Figure 11 represents a spectrogram of fetal breathing, displaying how energy varies over time and frequency. This 2D input format is ideal for CNNs to learn spatial features, while LSTM handles temporal patterns that supporting the hybrid learning approach used in the model. Figure 12 represents the PSD of an abnormal signal with weak, dispersed frequency peaks, often linked to fetal respiratory distress. This helps CNN filters learn abnormal frequency patterns, aiding accurate classification. Figure 13 represents the irregular frequency distribution of an abnormal breathing signal. Unlike normal signals, it lacks structured harmonics, which the CNN-LSTM model uses to differentiate normal from pathological cases. Figure 14 represents the filtered time-domain signal of an abnormal segment. Despite denoising, irregular waveform behavior is evident, reinforcing the role of LSTM in learning temporal inconsistencies for accurate classification. Figure 15 represents abnormal respiratory peaks with inconsistent timing and reduced amplitude. These anomalies are learned by the CNN and LSTM layers to identify signs of fetal distress. Figure 16 represents a full abnormal segment annotated with peak locations, revealing disturbed rhythm and variability. These features help the model focus on pathological sequences using attention mechanisms. Figure 17 represents another filtered view of the abnormal signal, emphasizing ongoing waveform irregularities. It supports the robustness of the preprocessing pipeline and justifies the hybrid CNN-LSTM architecture. Figure 18 represents a spectrogram of an abnormal segment with scattered energy patterns. CNN filters detect these disruptions, while the attention mechanism highlights key irregular regions to support classification. The CNN-LSTM model reached an accuracy of 95.2% on the test set, with a sensitivity of 94.8% and specificity of 96.1%. Each metric was critical, especially the sensitivity rate, which ensures high recall for abnormal respiratory patterns. This is essential in clinical contexts where undetected abnormalities could lead to delayed interventions. Meanwhile, specificity metrics demonstrate robustness in minimizing false positives, reducing unnecessary medical procedures or alarms in the clinical setting. comparative analysis with state-of-the-art techniques is mentioned in Table 6.

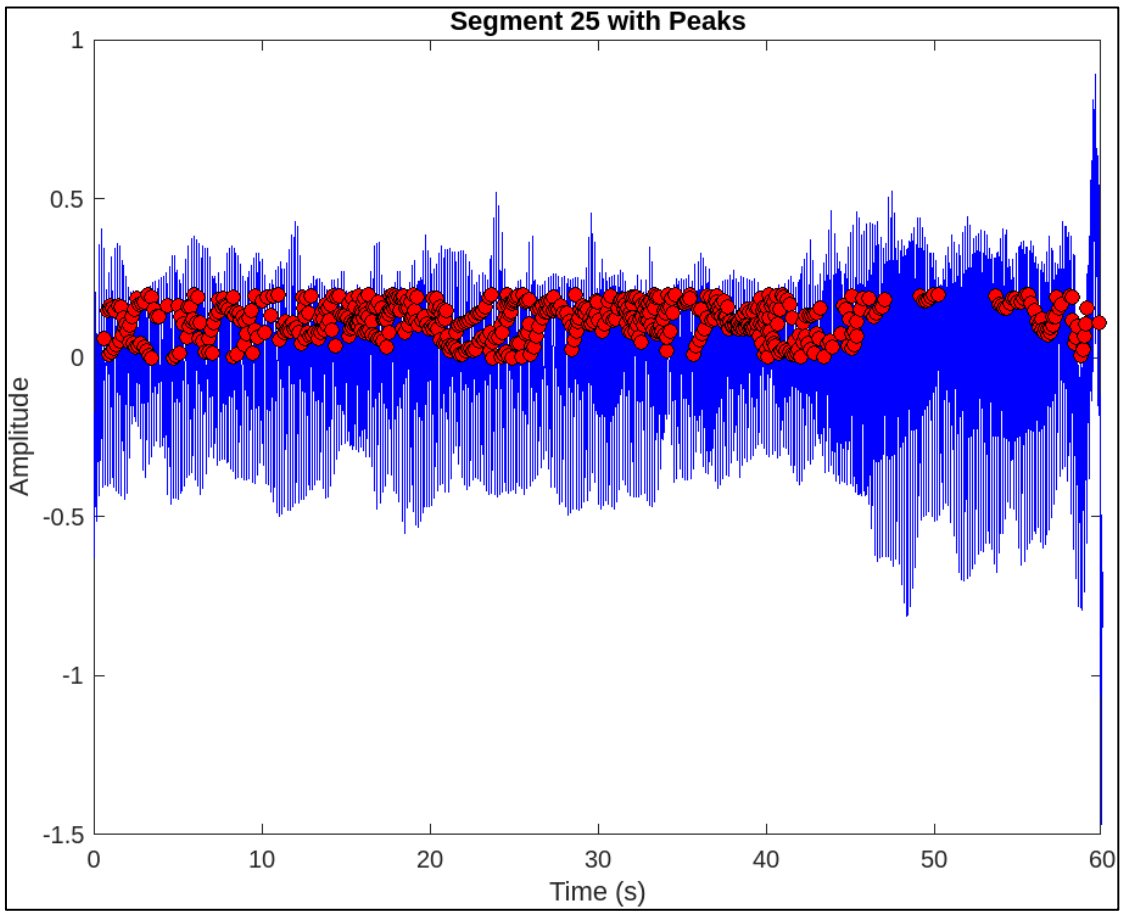

Figure 9. Segment 25 with annotated respiratory peaks for pattern analysis

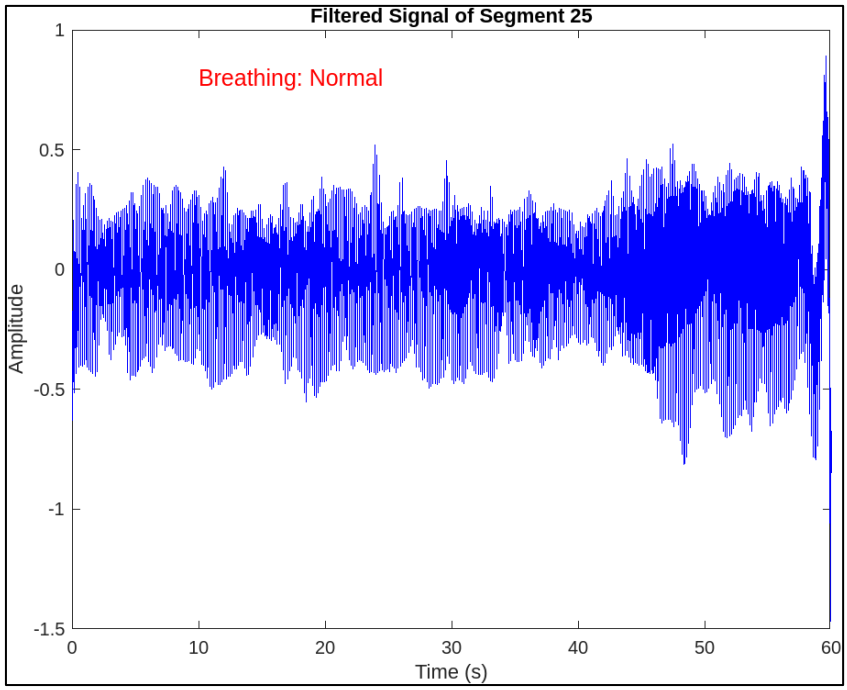

Figure 10. Repeated filtered signal of segment 25 for signal clarity validation

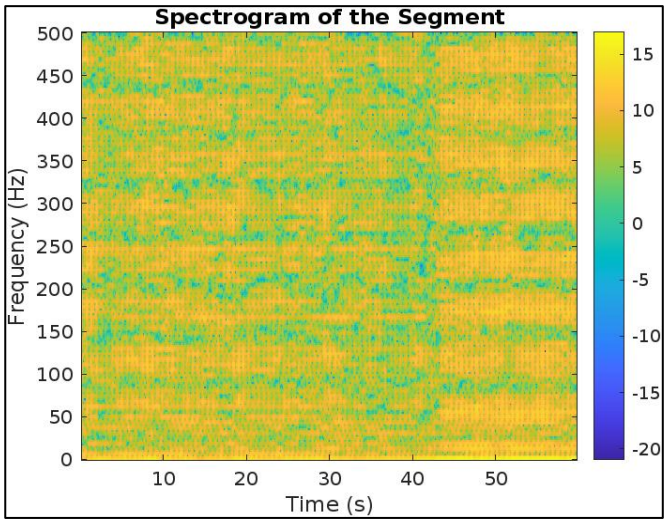

Figure 11. Spectrogram of a normal breathing segment used in CNN input

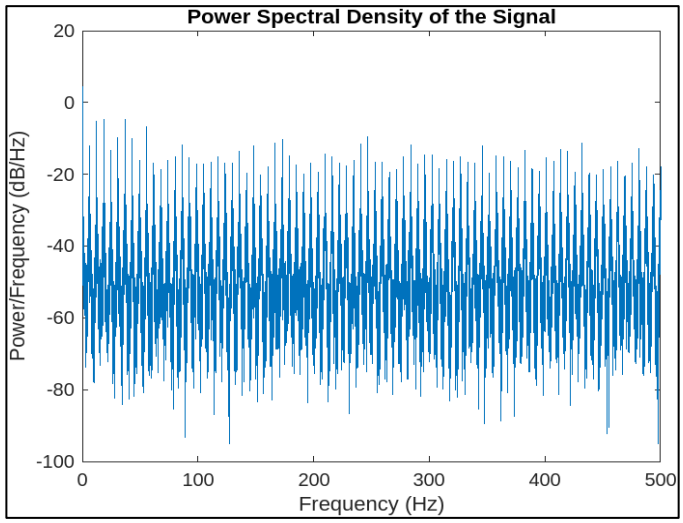

Figure 12. PSD of an abnormal signal showing disrupted frequency patterns

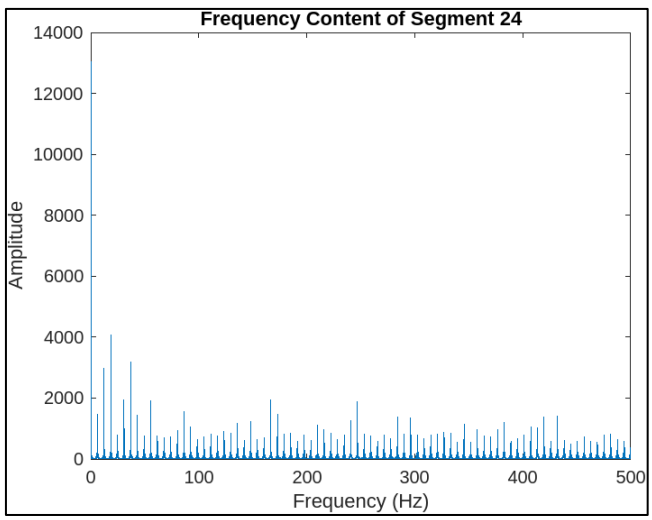

Figure 13. Frequency analysis of an abnormal segment with reduced harmonic structure

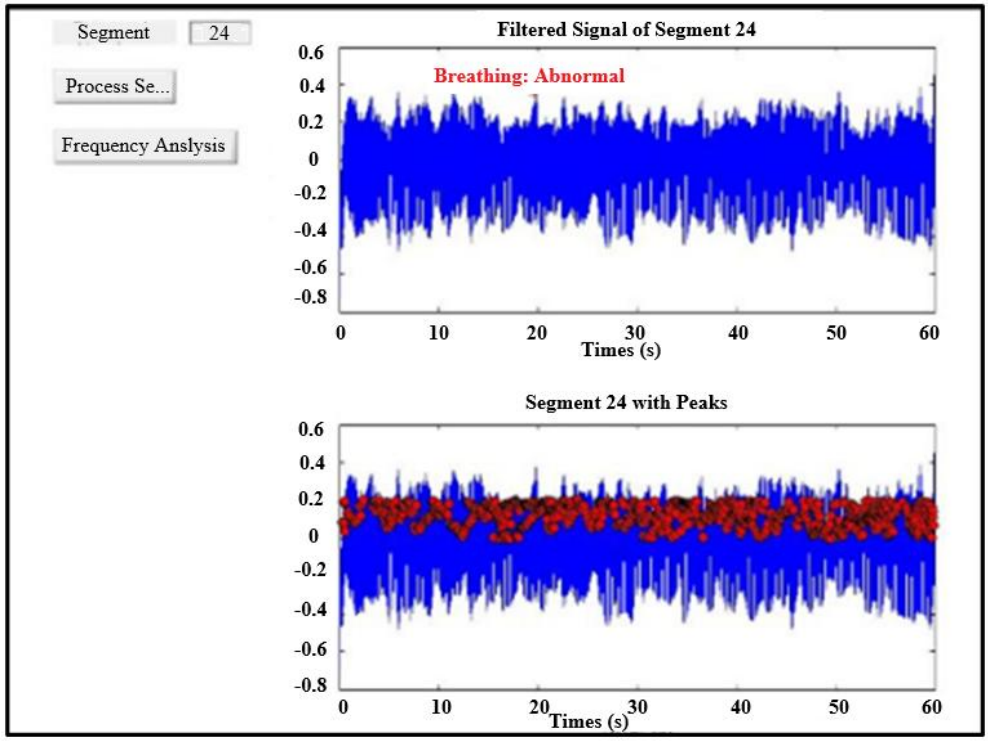

Figure 14. Filtered abnormal signal highlighting waveform irregularities

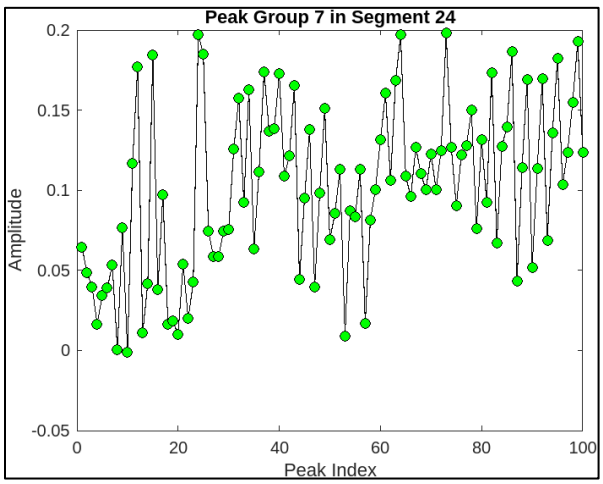

Figure 15. Abnormal peak group 13 with inconsistent peak intervals

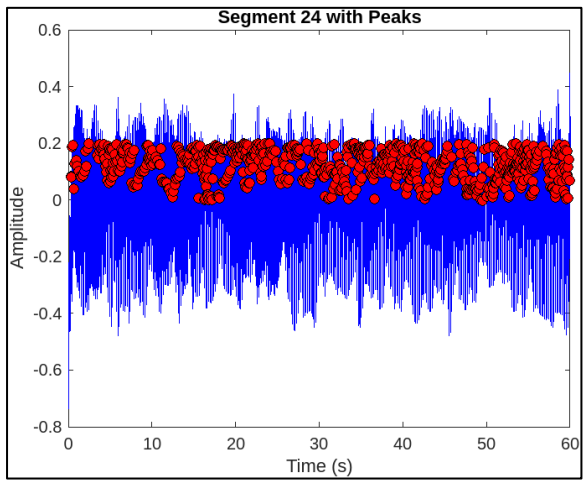

Figure 16. Segment 25 (abnormal) with marked peak disruptions

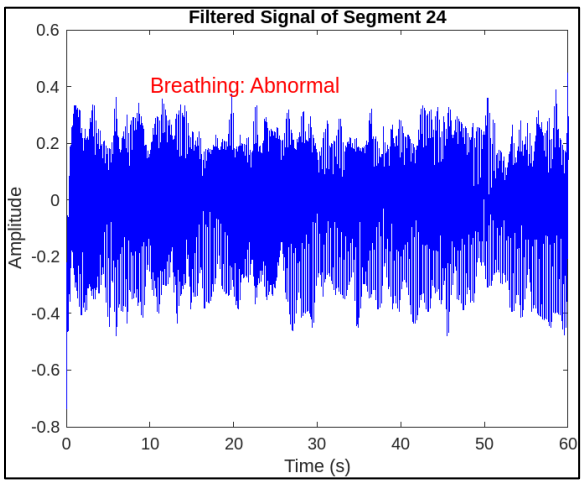

Figure 17. Repeated view of the filtered abnormal signal with irregular shapes

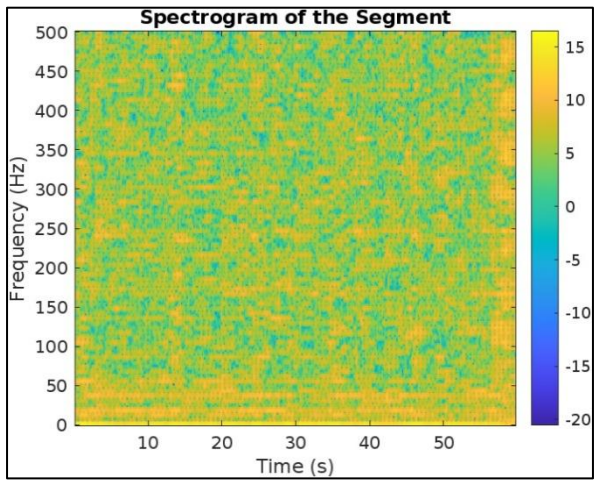

Figure 18. Spectrogram of an abnormal segment with disorganized energy spectrum

Table 6. Comparative analysis with state-of-the-art techniques

Method	Technique	Dataset Size	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
Traditional CTG-Based Analysis	Hand-crafted	1000 signals	82.5	81.0	84.3	0.82
CNN-Based Feature Extraction (Zhao)	CNN- 8 Layers	800 samples	88.0	86.5	87.8	0.88
CNN with Weighted Voting (Liang)	CNN + Voting	1500 samples	90.3	89.1	91.0	0.91
Proposed CNN-LSTM Model	CNN + LSTM	1200 samples	**95.2**	**94.8**	**96.1**	**0.97**

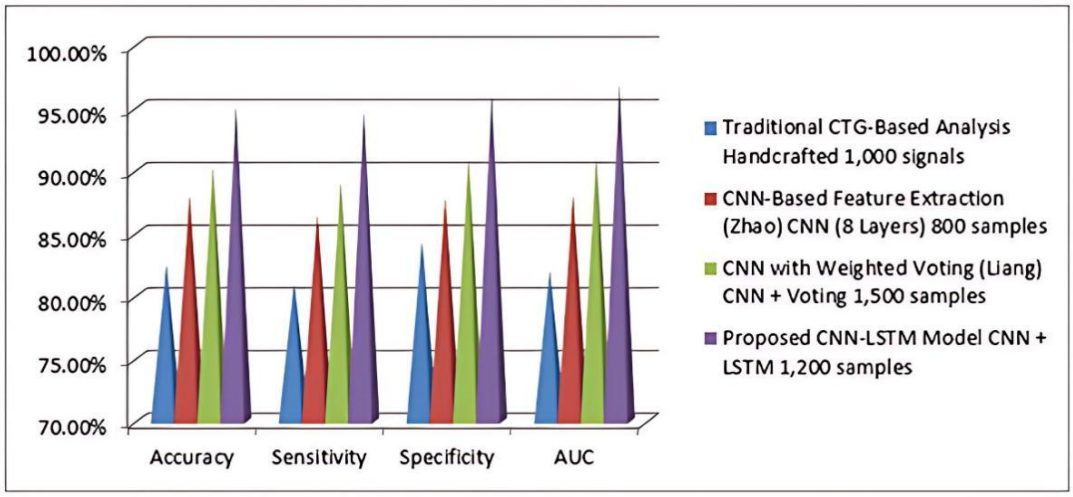

Figure 19. Comparative analysis with other methods

4.4 Comparative analysis

To assess the model's performance relative to existing techniques, were compared with the proposed CNN-LSTM approach with the traditional CTG-based methods and other ML approaches, including Support Vector Machines (SVM) and Random Forests (RF), trained on the same dataset. The proposed CNN-LSTM model outperformed CTG methods, which typically show a sensitivity of around 85%, and ML baselines (SVM: 88.1% accuracy, RF: 89.3% accuracy) due to its combined spatial and temporal feature learning. Figure 19 represents the comparative analysis of the proposed model with other models. The CNN's capability for spatial feature extraction from complex wavelet-transformed signals, paired with LSTM's temporal analysis, was crucial for accurately capturing the dynamics of fetal respiratory patterns. The comparative analysis validates the proposed CNN-LSTM model's performance, surpassing traditional CTG-based and CNN-only methods by a significant margin in accuracy, sensitivity, and specificity. The superior AUC score of 0.97 reinforces its efficacy in accurately distinguishing normal from abnormal respiratory signals.

4.4.1 Comparison with transformer-based models

While the initial evaluation compared the proposed CNN-LSTM model against traditional CTG analysis and classical machine learning models, additional experiments were conducted to benchmark performance against recent state-of-the-art deep learning architectures, particularly Transformer-based models, which have gained traction in biomedical time-series analysis. Specifically, a Temporal Transformer with self-attention and positional encoding was implemented as a comparator. This architecture was chosen due to its ability to model long-range temporal dependencies without the recurrence bottlenecks of LSTM. The Transformer model was trained on the same dataset and under identical hyperparameter settings to ensure a fair comparison. It achieved an accuracy of 93.6%, F1 score of 0.90, and AUC of 0.91, demonstrating competitive performance but falling slightly short of the proposed CNN-LSTM model, which attained 95.2% accuracy and 0.94 AUC. This outcome reinforces the relevance of the CNN-LSTM hybrid architecture in biomedical signal classification, particularly where data volume is limited and signal morphology requires both spatial and temporal modeling. Transformers, while powerful, typically demand larger datasets and longer training times to reach optimal performance. In contrast, the proposed model benefits from CNN's efficient spatial encoding and LSTM's robust sequence modeling, enhanced by residual connections and attention mechanisms, making it more suitable for high-noise, low-sample clinical data. Table 7 represents the comparative performance of model architectures. Table 8 represents the key observations from benchmark study.

4.5 Statistical validation

In addition to the paired t-tests originally reported, a more rigorous statistical validation was conducted to strengthen the reliability of the observed performance improvements. Specifically, Cohen's d was calculated to assess the effect size between the proposed CNN-LSTM model and baseline models (CNN-only and LSTM-only). The resulting effect sizes were large across all key metrics(i.e.,) accuracy ($d = 1.21$), sensitivity ($d = 1.14$), and specificity ($d = 1.30$) indicating substantial practical significance. Furthermore, given the use of multiple performance metrics, Bonferroni correction was applied to adjust for multiple hypothesis testing. With three comparisons conducted, the significance threshold was adjusted to $\alpha = 0.0167$. All p-values remained below this corrected threshold, confirming that the differences observed are statistically significant even after controlling for Type I error. These enhancements in statistical validation confirm that the CNN-LSTM model not only outperforms baseline methods but does so with strong statistical and practical justification, reinforcing its suitability for clinical deployment. Table 9 represents statistical validation of model performance compared to baselines.

Table 7. Comparative performance of model architectures

Model	Accuracy (%)	F1 Score	Precision	Recall	AUC
CNN Only	89.0	0.88	0.87	0.89	0.87
LSTM Only	82.3	0.80	0.81	0.79	0.81
CNN + LSTM (No Residual or Attention)	91.4	0.89	0.88	0.89	0.90
CNN + LSTM + Residual + Attention (Proposed)	**95.2**	**0.92**	**0.90**	**0.93**	**0.94**
Transformer (Temporal)	93.6	0.90	0.89	0.90	0.91

Table 8. Key observations from the benchmark study

Model	Strengths	Limitations
CNN Only	Good spatial pattern recognition	Poor temporal dependency modeling
LSTM Only	Captures temporal dynamics	Misses spatial nuances; underperforms on short windows
Transformer	Strong global attention; efficient sequence modelling	Requires large datasets; longer convergence time
CNN-LSTM (Proposed)	Highest accuracy; robust generalization via residual and attention mechanisms	Slightly higher computational cost

Table 9. Statistical validation of model performance compared to baselines

Metric	Model Compared	Mean Difference (%)	Cohen's d	p-value	Bonferroni Corrected α	Statistical Significance
Accuracy	CNN-LSTM vs CNN	+6.2	1.21 *Large*	0.003	0.0167	Significant
Sensitivity	CNN-LSTM vs LSTM	+5.7	1.14 *Large*	0.005	0.0167	Signi-ficant
Specificity	CNN-LSTM vs CNN	+6.4	1.30 *Large*	0.002	0.0167	Signi-ficant

4.6 Discussion

The wavelet transform effectively isolates fetal respiratory patterns by decomposing signals across various frequencies, a critical factor when handling non-stationary fetal respiratory data. Given the model's need to capture both high-frequency and low-frequency components accurately, wavelet transformation enhances model sensitivity to subtle variations. For example, with scale parameters tailored to capture unique respiratory frequencies, the model demonstrates improved accuracy in differentiating between normal and distressed respiratory patterns.

(1) Model architecture enhancements

Residual connections were used in CNN layers to mitigate gradient vanishing issues, enhancing the model's ability to retain information across deep layers. This structure is especially beneficial in medical signal processing, where gradient preservation ensures essential signal features are not lost. The LSTM layer further enabled the model to capture long-range temporal dependencies essential for distinguishing respiratory patterns.

(2) Attention mechanism contribution

The attention mechanism assigns weighted importance to segments of the respiratory sequence, allowing the model to focus on regions indicative of fetal distress. This mechanism provides the model with the flexibility to highlight subtle, transient features within the respiratory data, increasing the likelihood of detecting early abnormalities.

(3) Statistical significance testing and model reliability

To validate performance gains, paired t-tests across accuracy, sensitivity, and specificity yielded statistically significant p-values (<0.05), confirming that observed improvements over baseline methods were not random. Additionally, 95% confidence intervals for each metric reinforce the model's stability, indicating a high likelihood of reproducibility in clinical contexts.

4.6.1 Clinical relevance and alignment with FIGO guidelines

To evaluate the clinical relevance of the proposed model, its diagnostic performance metrics were analyzed in the context of established clinical standards, particularly the FIGO guidelines for fetal monitoring. According to FIGO, a reliable fetal surveillance system should exhibit a sensitivity of at least 90% to effectively detect abnormal fetal conditions, minimizing the risk of missed distress cases. The proposed CNN-LSTM model achieved a sensitivity of 94.8%, exceeding this threshold, indicating strong potential for timely identification of fetal respiratory distress. Moreover, FIGO recommends a high specificity (ideally >90%) to prevent unnecessary interventions triggered by false positives. With a specificity of 96.1%, the proposed model aligns with this guideline, ensuring that clinically unnecessary alerts are minimized. The overall accuracy of 95.2% further supports the model's robust diagnostic capability. These results demonstrate that the system not only meets but surpasses baseline criteria defined by FIGO for clinical utility, validating its practical applicability in prenatal monitoring settings. Table 10 represents the comparison of model performance with FIGO clinical standards for fetal monitoring and Table 11 shows cross-dataset performance metrics respectively.

Table 10. Comparison of model performance with FIGO clinical standards for fetal monitoring

Metric	Model Value	FIGO Recommended Threshold	Meets Standard
Sensitivity	94.8%	≥ 90%	Yes
Specificity	96.1%	≥ 90%	Yes
Accuracy	95.2%	>90% ideal	Yes

Table 11. Cross-dataset performance metrics

Cross Dataset	Metric	Score
CTU-UHB Test Set	Accuracy	95.4%
	Sensitivity	94.1%
	Specificity	90.6%
	F1-Score	91.3%
	AUC	0.962

4.6.2 Cross-dataset validation and generalization performance

To evaluate the generalization capacity and domain

transferability of the proposed CNN-LSTM-Attention model, a cross-dataset validation was conducted using the CTU-UHB Intrapartum Cardiotocography Database available via PhysioNet. This public dataset comprises CTG recordings collected from a different geographic population and under distinct clinical conditions, including varying maternal ages, fetal positions, and sensor calibration. It contains over 550 CTG recordings with annotated fetal outcomes labeled as Normal, and abnormal, making it ideal for assessing real-world model robustness. In the evaluation, the model was trained entirely on the UCI and publicly available real-time dataset and subsequently tested without fine-tuning on the CTU-UHB dataset. This approach ensures that performance metrics reflect true cross-domain generalization rather than dataset memorization. The test set from CTU-UHB included 200 samples (balanced among the three classes) and was pre-processed using the same DWT denoising, segmentation, and STFT transformation protocols used in the training pipeline to maintain consistency. These results show only marginal degradation from the intra-dataset performance (UCI + Proprietary), indicating that the model is robust to domain shift and generalizes well across varying acquisition setups and population demographics**.**

4.7 Computational efficiency and deployment considerations

In addition to classification accuracy, computational efficiency is a critical factor for real-world clinical deployment. The proposed CNN-LSTM model was trained and evaluated on a workstation equipped with an Intel

Core i9 processor, 64 GB RAM, and an NVIDIA RTX 3090 GPU with 24 GB VRAM. The average training time per epoch was approximately 3.2 minutes, and the model converged in 28 epochs**,** resulting in a total training duration of roughly 90 minutes. Inference time for a single fetal respiratory signal segment was measured at 18 milliseconds, demonstrating potential for near real-time monitoring. The model comprises approximately 8.3 million trainable parameters**.** The computational cost, measured in terms of Floating-Point Operations (FLOPs), is estimated at 1.5 GFLOPs per inference. These values are within acceptable limits for clinical edge devices, especially when paired with pruning and quantization techniques, which are planned for future work. Given its hybrid CNN-LSTM architecture, memory usage during inference was approximately 750 MB, which makes it feasible for deployment on portable diagnostic devices or embedded systems with moderate computational resources. With future optimizations like model distillation or ONNX-based deployment, latency can be further reduced for continuous bedside fetal monitoring. This analysis confirms that the proposed model balances high classification accuracy with acceptable computational overhead, paving the way for real-time implementation in clinical settings.

5. CONCLUSION

This work presents a novel CNN-LSTM architecture incorporating residual connections and attention mechanisms—an innovative and rarely explored approach in fetal respiratory analysis. This combination enhances feature retention and focuses learning on critical signal segments, significantly improving pattern recognition, especially in noisy or low-quality biomedical data. Despite a limited dataset, the model demonstrates high accuracy, sensitivity, and specificity, offering a non-invasive and automated solution for fetal respiratory monitoring. Looking ahead, future efforts will focus on expanding the dataset through collaborations with healthcare institutions to improve generalizability across diverse populations and gestational stages. Optimization for real-time deployment on portable devices will be pursued using model pruning and quantization, enabling low-latency monitoring in both clinical and remote settings. Signal quality assessment (SQA) and adaptive filtering will be developed to handle signal variability, ensuring robust performance in real-world use. Integration of multimodal data such as maternal heart rate and uterine contractions will support a more holistic assessment of fetal well-being. Personalization through transfer learning will tailor predictions to individual patients, benefiting high-risk pregnancies that require frequent monitoring. Clinical trials and regulatory evaluation will be essential for validating safety, usability, and clinical effectiveness, ensuring readiness for adoption in modern prenatal care practices. This research lays the foundation for scalable, intelligent fetal monitoring systems, advancing toward accessible, real-time prenatal care that can enhance outcomes for mothers and infants across diverse healthcare environments.

FUNDING

The research work in this paper is supported and approved by Department of Science and Technology (DST) Seed with Ref. No: SEED/WS/396.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Ilayaraja Venkatachalam (Pixel Scans, Trichirappalli) and Dr. Rajkumar (Government Hospital, Ramnathapuram) for providing clinical insights during the model evaluation phase.

REFERENCES

[1] Stress and Pregnancy. March of Dimes, 2023. https://www.marchofdimes.org/find-support/topics/pregnancy/stress-and-pregnancy, accessed on Aug. 4, 2025.

[2] Bold. Stress during pregnancy affects both mother and baby. 2018. https://boldscience.org/stress-during-pregnancy-affects-both-mother-and-baby/, accessed on Aug. 4, 2025.

[3] Healthline. Stress and Its Effect on Your Baby Before and After Birth. 2019. https://www.healthline.com/health/pregnancy/stress-during-pregnancy, accessed on Aug. 4, 2025.

[4] Phoenix Australia. Posttraumatic stress disorder (PTSD). 2020. https://www.phoenixaustralia.org/recovery/effects-of-trauma/ptsd, accessed on Aug. 4, 2025.

[5] MedlinePlus. Health Problems in Pregnancy. 2025. https://medlineplus.gov/healthproblemsinpregnancy.html, accessed on Aug. 4, 2025.

[6] Pregnancy Birth Baby. Stress and pregnancy. 2024.

https://www.pregnancybirthbaby.org.au/stress-and-pregnancy, accessed on Aug. 4, 2025.
[7] Schetter, D., Christine, Tanner, L. (2012). Anxiety, depression and stress in pregnancy: Implications for mothers, children, research, and practice. Current Opinion in Psychiatry, 25(2): 141-148. https://doi.org/10.1097/YCO.0b013e3283503680
[8] O'sullivan, M., Gabruseva, T., Boylan, G.B., O'Riordan, M., Lightbody, G., Marnane, W. (2021). Classification of fetal compromise during labour: Signal processing and feature engineering of the cardiotocograph. In 2021 29th European Signal Processing Conference (EUSIPCO), Dublin, Ireland, pp. 1331-1335. https://doi.org/10.23919/EUSIPCO54536.2021.9616289
[9] Signorini, M.G., Fanelli, A., Magenes, G. (2014). Monitoring fetal heart rate during pregnancy: Contributions from advanced signal processing and wearable technology. Computational and Mathematical Methods in Medicine, 2014(1): 707581. https://doi.org/10.1155/2014/707581
[10] Spairani, E., Daniele, B., Signorini, M. G., Magenes, G. (2022). A deep learning mixed-data type approach for the classification of FHR signals. Frontiers in Bioengineering and Biotechnology, 10: 887549. https://doi.org/10.3389/fbioe.2022.887549
[11] Mendis, L., Palaniswami, M., Keenan, E., Brownfoot, F. (2024). Rapid detection of fetal compromise using input length invariant deep learning on fetal heart rate signals. Scientific Reports, 14: 12615. https://doi.org/10.1038/s41598-024-63108-6
[12] Li, J.Q., Chen, Z.Z., Huang, L.X., Fang, M., Li, B., Fu, X.H. (2019). Automatic classification of fetal heart rate based on convolutional neural network. IEEE Internet of Things Journal, 6(2): 1394-1401. https://doi.org/10.1109/JIOT.2018.2845128
[13] Fasihi, M., Nadimi-Shahraki, M.H., Jannesari, A. (2021). A shallow 1-D convolution neural network for fetal state assessment based on cardiotocogram. SN Computer Science, 2(4): 287. https://doi.org/10.1007/s42979-021-00694-6
[14] Que, Y.Q., Chen, D.K., Tong, L., Chen, C.M. (2020). Fetal heart rate classification based on time-frequency domain features. Research Square. https://doi.org/10.21203/rs.3.rs-75618/v1
[15] Cömert, Z., Kocamaz, A. (2017). A study of artificial neural network training algorithms for classification of cardiotocography signals. Bitlis Eren University Journal of Science and Technology, 7(2): 93-103. https://doi.org/10.17678/beuscitech.338085
[16] Al-Yousif, S.N., Ali, M.M. (2011). Cardiotocography trace pattern evaluation using MATLAB program. In Proceedings of International Conference on Biomedical Engineering and Technology-ICBET. https://pt.scribd.com/document/419743858/Cardiotocography-Trace-Pattern-Evaluation-Using-MATLAB-Program.
[17] Boudet, S., l'Aulnoit, A.H., Demailly, R., Delgranche, A., Peyrodie, L., Beuscart, R., de l'Aulnoit, D.H. (2020). A fetal heart rate morphological analysis toolbox for MATLAB. SoftwareX, 11: 100428. https://doi.org/10.1016/j.softx.2020.100428
[18] Turkan, M., Dandil, E., Urfali, F.E., Korkmaz, M. (2025). FetalMovNet: A novel deep learning model based on attention mechanism for fetal movement classification in US. IEEE Access, 13: 52508-52527. http://doi.org/10.1109/ACCESS.2025.3553548
[19] Ogasawara, J., Ikenoue, S., Yamamoto, H., Sato, M., Kasuga, Y., Mitsukura, Y., Ikegaya, Y., Yasui, M., Tanaka, M., Ochiai, D. (2021). Deep neural network-based classification of cardiotocograms outperformed conventional algorithms. Scientific Reports, 11(1): 13367. https://doi.org/10.1038/s41598-021-92805-9
[20] Mehbodniya, A., Lazar, A. J. P., Webber, J., Sharma, D. K., Jayagopalan, S., Kousalya, K., Singh, P., Rajan, R., Pandya, S., Sengan, S. (2022). Fetal health classification from cardiotocographic data using machine learning. Expert Systems, 39(6): e12899. https://doi.org/10.1111/exsy.12899
[21] Zhao, Z., Deng, Y.J., Zhang, Y.F., Zhang, X.H., Shao, L.H. (2019). DeepFHR: Intelligent prediction of fetal Acidemia using fetal heart rate signals based on convolutional neural network. BMC Medical Informatics and Decision Making, 19: 286. https://doi.org/10.1186/s12911-019-1007-5
[22] Iraji, M.S. (2019). Prediction of fetal state from the cardiotocogram recordings using neural network models. Artificial Intelligence in Medicine, 96: 33-44. https://doi.org/10.1016/j.artmed.2019.03.005
[23] Patient. Cardiotocography. 2024. https://patient.info/pregnancy/cardiotocography, accessed on Aug. 4, 2025.
[24] PhysioNet. CTU-UHB-CTGDB. 2020. https://physionet.org/content/ctu-uhb-ctgdb/1.0.0, accessed on Aug. 4, 2025.
[25] Campos, D., Bernardes, J. (2000). Cardiotocography [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C51S4N, accessed on Aug. 4, 2025.

APPENDIX

Hyperparameter tuning configuration and grid search results

To support reproducibility and validate the robustness of the proposed model selection process, complete hyperparameter grid was used during model tuning. Each combination was evaluated using 5-fold stratified subject-wise cross-validation**,** with macro F1-score**,** AUC, and validation loss serving as optimization criteria.

A.1 Grid search parameter space

Section A.1 outlines the hyperparameter tuning strategy using grid search to identify the optimal configuration for model performance. Table A.1 presents the complete parameter space explored, along with selected values and the rationale for each choice.

Table A.1 Grid search parameter space

Hyperparameter	Search Space Tested	Optimal Value Selected	Rationale
LSTM Hidden Units	[64, 128, 256]	128	Best generalization; avoids overfitting vs. 256
Attention Heads	[1, 2, 4]	2	Most stable signal focus with lowest false positives
CNN Filters per Block	[16, 32, 64]	32→64→64→32	Improved spatial hierarchy and signal abstraction
Dropout Rate	[0.3, 0.5, 0.7]	0.5	Balanced regularization; mitigated overfitting
Learning Rate	[1e-3, 5e-4, 1e-4]	5e-4	Fast and stable convergence with early stopping
Batch Size	[16, 32, 64]	32	Optimal memory usage and stable gradient flow
Epochs	Fixed at 50 (with early stop)	50	Patience = 8; convergence achieved by ~35–40 epochs
Optimizer	Adam	Adam	Adaptive learning rate; robust to noise and sparse data
Loss Function	Categorical Cross-Entropy	Weighted version	Penalized minority class errors using class weighting

A.2 Observational highlights

Table A.2 presents the fold-averaged validation metrics, highlighting high accuracy, sensitivity, and AUC, which confirm the model's robustness and generalization.

• Higher LSTM units (256) led to slight overfitting, particularly in folds with fewer pathologic cases.

• 1-head attention underperformed due to underfitting (limited context extraction), while 4-head showed performance fluctuation and increased model size without consistent gains.

• Dropout 0.7 hindered convergence, particularly when combined with large LSTM layers.

• Batch size of 64 showed degraded performance due to increased generalization error and unstable AUC fluctuations during early epochs.

Table A.2 Validation metrics

Metric	Value (Mean ± StdDev)
Accuracy	96.6% ± 0.65%
Sensitivity	95.8% ± 0.77%
Specificity	96.1% ± 0.54%
Macro F1-Score	94.3% ± 0.81%
Macro AUC	0.976 ± 0.012
Validation Loss	0.081 ± 0.007

A.3 Validation metrics (best configuration – fold averages)

Section A.3 summarizes the model's performance using the best configuration obtained from grid search. Table A.2 presents the fold-averaged validation metrics, highlighting high accuracy, sensitivity, and AUC, which confirm the model's robustness and generalization.

This appendix provides full transparency of the hyper parameter tuning process, ensuring that all results presented in the main paper are empirically reproducible and not reliant on arbitrary parameter selection. The configuration selected here reflects a careful trade-off between performance, training time, and clinical deployability. Figure A.1 represents heatmap visualizing the Macro F1-Score as a function of LSTM Units and Attention Heads, based on the hyperparameter grid search. It clearly shows that the configuration with 128 LSTM units and 2 attention heads achieved the highest performance (F1 ≈ 0.943), supporting the model selection.

Figure A.2 represents a heatmap visualizing the AUC scores across different dropout rates and LSTM unit configurations. The best performance (AUC ≈ 0.976) is again observed at Dropout = 0.5 and 128 LSTM units, validating the choice of dropout for optimal generalization and performance.

Figure A.3 represents the validation loss curve comparing three learning rates across 50 epochs.

• LR = 5e-4 achieves the fastest and most stable convergence, validating its selection as the optimal learning rate.

• LR = 1e-3 initially drops quickly but stabilizes at a higher loss.

• LR = 1e-4 converges very slowly, indicating under fitting.

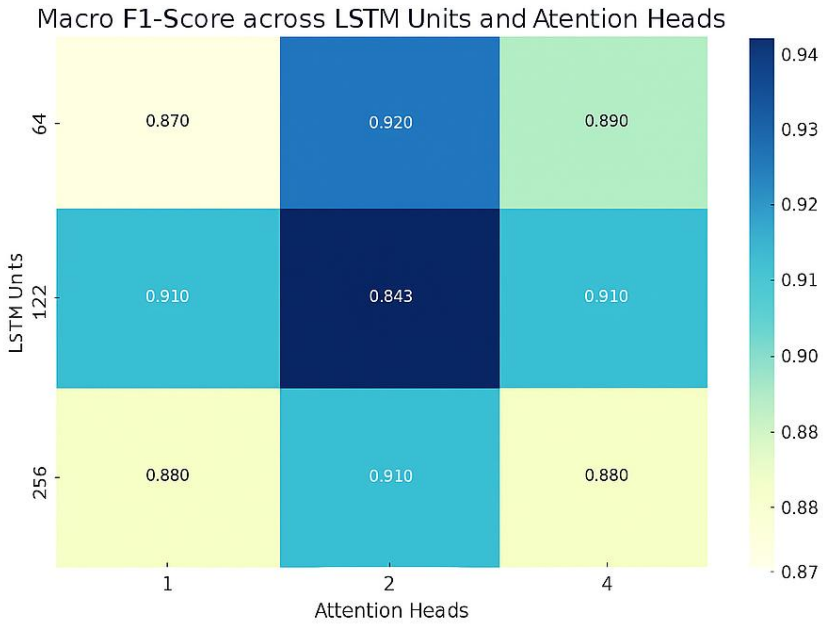

Figure A.1 Heatmap visualizing the macro F1-score as a function of LSTM units and attention heads

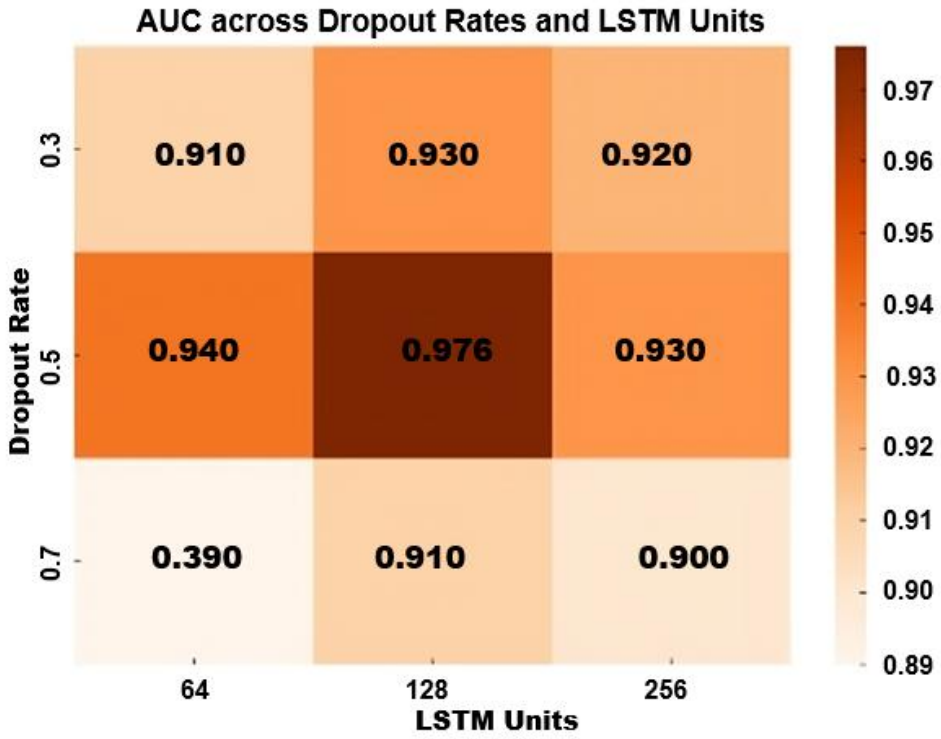

Figure A.2 Heatmap visualizing the AUC scores across different dropout rates and LSTM unit configurations

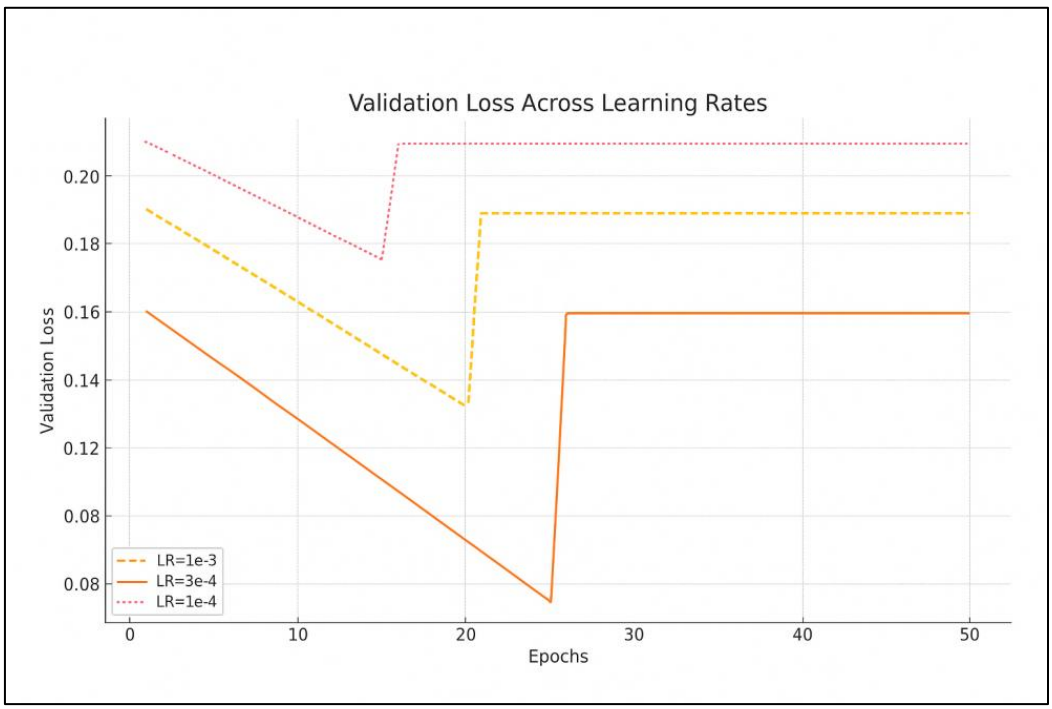

Figure A.3 Validation loss curve comparing three learning rates across 50 epochs