



Article

Multimodal Deep Learning for Predicting Adverse Birth Outcomes Based on Early Labour Data

Daniel Asfaw ^{1,2,*}, Ivan Jordanov ¹ , Lawrence Impey ², Ana Namburete ³, Raymond Lee ⁴ and Antoniya Georgieva ² 

¹ School of Computing, University of Portsmouth, Portsmouth PO1 3HE, UK

² Nuffield Department of Women's & Reproductive Health, University of Oxford, Oxford OX1 2JD, UK; antoniya.georgieva@wrh.ox.ac.uk (A.G.)

³ Department of Computer Science, University of Oxford, Oxford OX1 3QG, UK

⁴ Faculty of Technology, University of Portsmouth, Portsmouth PO1 2UP, UK

* Correspondence: daniel.asfaw@port.ac.uk

Abstract: Cardiotocography (CTG) is a widely used technique to monitor fetal heart rate (FHR) during labour and assess the health of the baby. However, visual interpretation of CTG signals is subjective and prone to error. Automated methods that mimic clinical guidelines have been developed, but they failed to improve detection of abnormal traces. This study aims to classify CTGs with and without severe compromise at birth using routinely collected CTGs from 51,449 births at term from the first 20 min of FHR recordings. Three 1D-CNN and LSTM based architectures are compared. We also transform the FHR signal into 2D images using time-frequency representation with a spectrogram and scalogram analysis, and subsequently, the 2D images are analysed using a 2D-CNNs. In the proposed multi-modal architecture, the 2D-CNN and the 1D-CNN-LSTM are connected in parallel. The models are evaluated in terms of partial area under the curve (PAUC) between 0–10% false-positive rate; and sensitivity at 95% specificity. The 1D-CNN-LSTM parallel architecture outperformed the other models, achieving a PAUC of 0.20 and sensitivity of 20% at 95% specificity. Our future work will focus on improving the classification performance by employing a larger dataset, analysing longer FHR traces, and incorporating clinical risk factors.

Keywords: CTG; FHR; deep learning; CNN; LSTM



Citation: Asfaw, D.; Jordanov, I.; Impey, L.; Namburete, A.; Lee, R.; Georgieva, A. Multimodal Deep Learning for Predicting Adverse Birth Outcomes Based on Early Labour Data. *Bioengineering* **2023**, *10*, 730. <https://doi.org/10.3390/bioengineering10060730>

Academic Editor: Mario Petretta

Received: 2 May 2023

Revised: 29 May 2023

Accepted: 7 June 2023

Published: 19 June 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Cardiotocography (CTG) is a continuous and simultaneous measurement of fetal heart rate (FHR) and maternal uterine contraction signals. CTG is commonly performed during or preceding labour to assess fetal wellbeing and reduce its mortality and morbidity [1]. Interpretation of the CTG patterns requires assessing the FHR baseline, variability, accelerations, and decelerations by a trained clinician. However, due to the complexity of CTG signals, visual interpretation is often challenging and imprecise [2], leading to miss diagnoses [3,4]. In the United Kingdom (UK), each year between 2015 to 2018, on average 125 intrapartum still births, 154 neonatal deaths, and 854 severe injuries were registered [5]. These adverse outcomes frequently lead to litigation. In England, in 2020/2021, over £4.1 billion was spent on settling obstetric claims, 59% of which were clinical negligence payments [6]. Enhancing the accuracy of CTG interpretation has the potential to enable clinicians to intervene earlier, thereby potentially preventing some of these adverse outcomes. This, in turn, can alleviate the substantial financial burden on the healthcare system. Globally in 2019, an estimated 2 million babies were stillborn [7]. Most these adverse outcomes that occurred during intrapartum periods are potentially preventable with CTG monitoring and appropriate interventions.

CTG remains at the center of the decision-making process in intrapartum fetal monitoring despite its limitations, as there is no other technology or method that has been shown

to be as effective in assessing fetal well-being during labour. CTG is typically performed at maternity admission units/triage wards or after admission to the labour ward and in some countries like Sweden it is routinely performed, while in others, such as the UK, CTG is not recommended for low-risk births (about 40% of all) [8,9]. Under some circumstances, the initial 20–30 min CTG recording is called ‘admission CTG’ [10], and its role is controversial: some studies show that it may increase the incidence of unnecessary caesarean sections, especially in ‘low risk’ pregnancies [10], while others report its benefit in the decision to perform caesarean delivery when administered routinely to all births [8]. This lack of consensus can be attributed to the imprecision of current clinical guidelines and the poor sensitivity and specificity of the available tools to interpret CTG patterns. During the evaluation of 27,000 high-risk births, our team noted significant differences in the first-hour CTG features (extracted using objective computerized methods) and clinical risk factors between births with severe compromise and those without severe compromise [11].

Several automated methods have been proposed to address the subjective visual interpretation of CTG recordings [12–15]. Research efforts have been devoted to developing techniques that can automatically detect characteristics of the CTG signal [16–19]. These studies focus on detecting or quantifying abnormal patterns by mimicking what clinical rules and guidelines suggest or what experts do during their visual assessment. Such methods are commercially available but have not shown clinical benefit in randomised clinical trials [20,21] and have not been widely adopted. Other studies have used advanced signal processing techniques to extract multiple features from the CTG signal alone or in conjunction with clinical risk factors and apply machine learning approaches, including hierarchical Dirichlet process mixture models [22], logistic regression [23], neural networks [24], support vector machines [14,25], random forests [15], Bayesian classifier [26], and XGBoost [15,27], to find patterns from the extracted features. One of the key issues with the conventional machine learning models is that they require careful design of feature extractors that transform the input CTG signal into compact representations or feature vectors [28].

More recent studies on computerised CTG analysis apply modern deep learning (DL) techniques [12,29–31]. Convolutional Neural Networks (CNN) are among the most notable DL approaches that learn automatically abstract hierarchical representations directly from the input data using multiple hidden layers [28]. CNN models have been extensively applied in various medical data analysis tasks involving image and time series analysis [32,33]. Long Short-Term Memory (LSTM) networks are another class of DL models suitable for analysing sequential (time series) data [34]. Most previous works incorporating deep learning methods for the CTG analysis have used primarily 1D-CNN based networks. For example, Comert et al. [12], and Baghel et al. [31] analysed the last 90 min CTG data of the CTU-UHB open-access dataset [35], using 1D CNNs. Zhao et al. [12] reported better performance on the same dataset when implementing 2D CNNs by firstly transforming the CTG signal into 2D images using recurrence plots. More recently, Liu et al. [36] proposed attention based CNN-LSTM network with features of discrete wavelet transformation to analyse the CTU-UHB dataset. However, these studies have been evaluated mostly on small datasets (40–160 abnormal birth outcomes), which are prone to large within-class variability and between-class similarity, resulting in a model with poor generalisation performance. Furthermore, although the proposed deep learning models are valid, due to absence of a distinct holdout (testing) sets in the CTU-UHB dataset and the less rigorous evaluation approach adopted (using the whole dataset for the cross validation instead of only on the training set), the results should be interpreted with caution. On a much larger dataset, Petrozziello et al. [37] and Mohannad et al. [38] evaluated the performance of CNNs to classify CTGs. Petrozziello et al. [37] achieved promising performance using multimodal 1D CNNs to predict cord acidemia from the last hour CTG signal. Mohannad et al. [38] applied a multi-input CNN network to analyse CTG plots of the initial 30 min of the last 50 min before delivery and gestational age to predict foetuses with a low Apgar score.

Overall, previous studies on analysis of CTG using deep learning approaches focused on the last hour recordings mostly using small datasets.

In this study, we present three deep learning models for prediction of birth outcome using FHR traces recorded around the onset of labour in both: the time domain, implementing a combination of 1D CNNs and LSTMs; and in the frequency domain, employing a 2D CNNs [33,39]. The models are trained to classify new-borns with and without severe compromise at birth. To our knowledge, this study represents a pioneering effort in the application of deep learning techniques for analyzing CTG traces during early labour. Given the absence of published results from computer-based methods, we conducted a comparison of our findings with the existing standards of clinical care. We hypothesise that DL methods trained with a large clinical dataset of CTGs from around the onset of labour could hold the potential to ultimately assist clinicians in identifying fetuses who are already compromised or are vulnerable at labour onset and may thus be at high risk for further injury during labour.

2. Materials and Methods

2.1. Data and Pre-Processing

2.1.1. Description of the Dataset

This was a retrospective cohort study of infants delivered at the John Radcliffe Hospital in Oxford, UK, using a clinical data collection system between 1993 and 2012. The study received ethical approval from the Newcastle & North Tyneside 1 Research Ethics Committee, Reference 11/NE0044 (data before 2008), and from the South Central Ethics Committee, Reference 13/SC/0153 (for data after 2008). Informed consent by the participants was not required.

The clinical protocol has been to administer intrapartum CTG only to pregnancies deemed at ‘high-risk’. From them, 51,449 CTG tracings include births at gestation ≥ 36 weeks, longer than 1 h, and have no second stage trace in the first hour (Figure 1). In these records, 452 are births with a severe compromise—a composite outcome of intrapartum stillbirth, neonatal death, neonatal encephalopathy, seizures, and resuscitation followed by over 48 h in the neonatal intensive care unit. The rest of the cohort samples are labelled as no severe compromise.

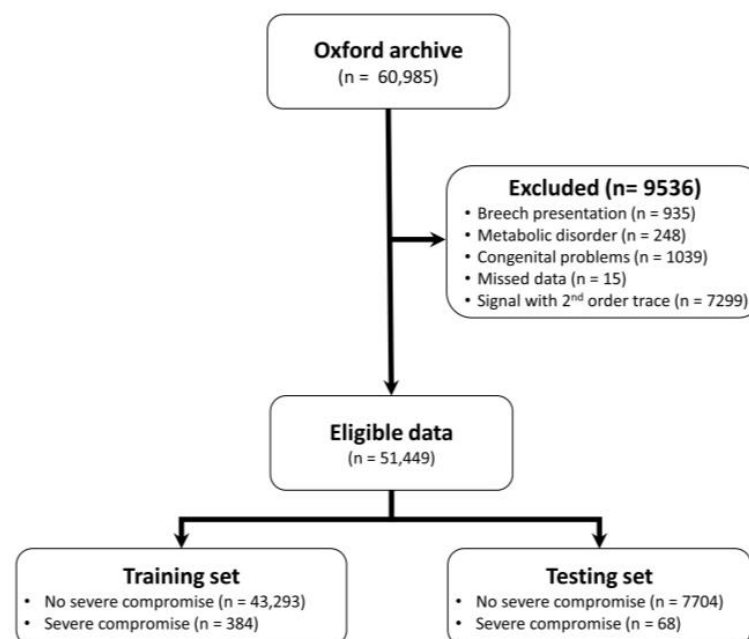


Figure 1. Data flow chart.

2.1.2. Pre-Processing

The CTG datasets are obtained with standard fetal monitors at a 4 Hz sampling rate for the FHR and 2 Hz for the uterine contraction signals. Consistently with our prior work [37], the FHR signals are down sampled to 0.25 Hz. We apply a two-stage pre-processing procedure to deal with the noise and missingness: signal cleaning; and gap imputation.

2.1.3. Signal Cleaning

A bespoke algorithm is applied to remove artefacts from the CTG signal, for example, erroneous maternal heart rate capture and extreme outliers (FHR measurements >230 or <50 beats per minute (bpm)). The start time of the signal is adjusted to ensure adequate signal quality: the CTG is analysed using a sliding 5-min window (with a one-minute stride) to ensure that the signal loss is less than 50% in the first five minutes. The extracted cleaner 20-min FHR tracing has less than 50% signal loss, i.e., any sample with signal loss greater than the threshold is discarded. The cleaning process reduced the signal loss of the no severe compromise group (mean, \pm std) from 26.9% (26.7, 27.2%) to 12.3% (12.2, 12.4%); and for the severe cases from 29.1% (26.6%, 31.5%) to 13.4% (12.3, 14.5%).

2.1.4. Gap Imputation

Signal noise and loss are common in CTG tracings, resulting in both short (few seconds) and long (many minutes) gaps in the signal [40]. Following efficient noise removal, reliable gap imputation is an important task in data analysis and pre-processing phases, which is expected to improve the classifier performance at the later learning stage. There are a number of techniques proposed for inferring and imputing the gaps in FHR signals recorded by CTG, including Linear interpolation [37], Cubic spline interpolation [41], Sparse representation with dictionaries [42], Gaussian processes (GP), and others [43,44]. We compared the performance of the Linear, GP, and Autoregressive (AR) imputation techniques (example shown in Figure 2) based on their effect on the performance of the CNN, which is evaluated on the testing set. Since the AR consistently outperformed the Linear and GP gap imputation techniques, and achieved the highest accuracy the CNN's classification accuracy, we used it to impute the gaps in the FHR signals of our dataset (results of the comparison between the gap imputation techniques is reported in [45]).

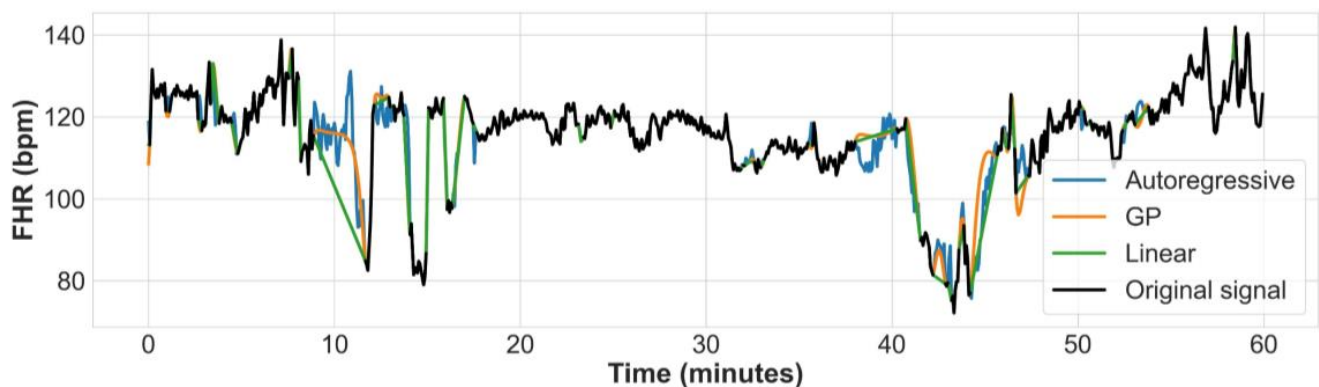


Figure 2. Example of an FHR signal from the dataset where gaps are imputed using Autoregressive, GP based, and Linear interpolation methods.

2.1.5. Transformation of the FHR Signal to a 2D Image

The raw FHR signals (each sample 20 min long) are transformed into time-frequency using Fourier and Wavelet transforms, and the resulting images are analysed using the well-established 2D-CNNs. The spectrogram of Short-Time Fourier Transform (STFT) represents the normalised, squared magnitude of short-time Fourier transform coefficients [46]. To convert the input 1D FHR signal using spectrograms, the time domain signals are divided into shorter segments (windows), and Fourier transform is computed for each segment to obtain the frequencies. We used a 1 Hz FHR signal in the spectrogram and scalogram

(Wavelet transform) analysis since it produced better accuracy in our preliminary experiment. The spectrogram is the STFT of each short signal segment, computed by sliding the window with a constant stride and an overlap through the entire record. In this work, we investigate the effect of different window strides and overlapping sizes on the classifier's performance. The FHR signals are converted into spectrogram images by applying STFT given with (1).

$$X(n, \omega) = \sum_{m=0}^{L-1} x[m]w[m-n]e^{-j2\pi m\omega/L} \quad (1)$$

where $x[m]$ is the input FHR signal, $w[m]$ is the window, and L is the window length. $X(n, \omega)$ STFT of a windowed data centred at time point n and the log values of $X(n, \omega)$ are represented as spectrogram (128×128) images. Since the window length L is a hyperparameter, we investigated the effect of different window sizes on the classification performance. In our initial experiments the 128×128 spectrogram images lead to better classifier performance than 64×64 or 256×256 image sizes. Thus, we used the 128×128 image size for the rest of our analysis.

In addition to the STFT spectrograms, we also investigated the wavelet scalograms. A scalogram is a time-frequency representation with a wavelet basis instead of sinusoidal functions. Like the Fourier spectrogram, the scalogram analyses compute the coefficients using sliding windows called wavelets, i.e., the input signal is multiplied with the wavelet at different time locations. The process is repeated by increasing the scale of the wavelet (also known as the mother wavelet). This dilation and contraction operation captures long- and short-time events from the input, where the dilated wavelet is sensitive to long-time events, and the contracted wavelet to short-time events [47]. The wavelet transform of a signal $x(t)$ is defined as the integration of the $x(t)$ with the shifted or scaled shapes from a mother wavelet $\psi_{a,b}(t)$ as shown in (2).

$$WT_x(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \psi\left(\frac{t-b}{a}\right) dt \quad (2)$$

where a is a scale parameter, b is a translation parameter, and $\psi(t)$ is the mother wavelet function. By using different scale factors of the wavelet transform, WT_x computes wavelet coefficients of the signal at different scales. The absolute values of these continuous coefficients define the scalogram (in our case, represented as 128×128 image). The choice of the mother wavelet is important as the time-frequency analysis represents the match between the wavelet and the FHR signal. Here, we compare the Gaussian of order 8 ('gaus'), Morlet ('morl'), Shannon ('shan'), and Mexican Hat ('mexh') wavelets. [47].

2.2. Deep Learning Models

2.2.1. Data Augmentation

The dataset used in this study is substantially imbalanced: there are fewer positive samples ($n = 384$) than the negative ones ($n = 43,293$). The class imbalance in the dataset can make the learning process very challenging for any classifier and usually leads to poor prediction performance [48]. In our initial model training, we observed signs of overfitting. To mitigate this problem, we augmented the data of the severe compromise class, which we expected to act as a regulariser and improve the generalisation performance of the classifier. Several data augmentation methods have been proposed for time series signals, such as flipping the signal, adding noise, masking the segment of the signal [49]. We developed a simple data augmentation approach, tailored to our specific task, which involved extracting additional 20-min FHR segments from the first 1-h FHR data with 50% overlap, thereby increasing the size of positive samples by a factor of 4. Only the 20-min segments with less than 50% signal loss were augmented. The positive instances are then further oversampled by a factor of 2, which led to their overall increase by a factor of 8 in the training dataset. In CTG analysis, the under-sampling of the negative samples is a common practice [12,30],

but in our experiments, these techniques did not improve the generalisation performance of the trained models.

2.2.2. Deep Learning Architectures

The proposed architectures primarily constitute 1D-CNN and LSTMs. Three variations are investigated:

- (i) 5-layer 1D-CNN network, in which the encoder is composed of five 1D-CNN layers and two fully connected (FC) layers;
- (ii) CNN-LSTM sequential architecture, in which the network has 5-layer 1D-CNN followed by 2-layer LSTM component and two FC layers;
- (iii) 5-layer CNN-LSTM parallel architecture, in which a 5-layer 1D-CNN and two-layer LSTM networks are connected in parallel, followed by 2 FC layers.

The architectures of the three models are shown in Figure 3. We also employed a 2D-CNN of the FHR signal to analyse the spectrograms and scalograms. The network architecture comprises of 2D-CNNs (with ReLU activation), along with five residual blocks, followed by an average pooling layer. The residual blocks are a 2D-CNN with skip connections, a widely used network architecture for various image recognition tasks (28). Each residual block in the skip connection has a 2D-CNN component, followed by a sequence of batch normalisation, dropout, and max pooling operations (Figure 4). In addition, we investigated the effect of different kernel sizes on classification performance as well.

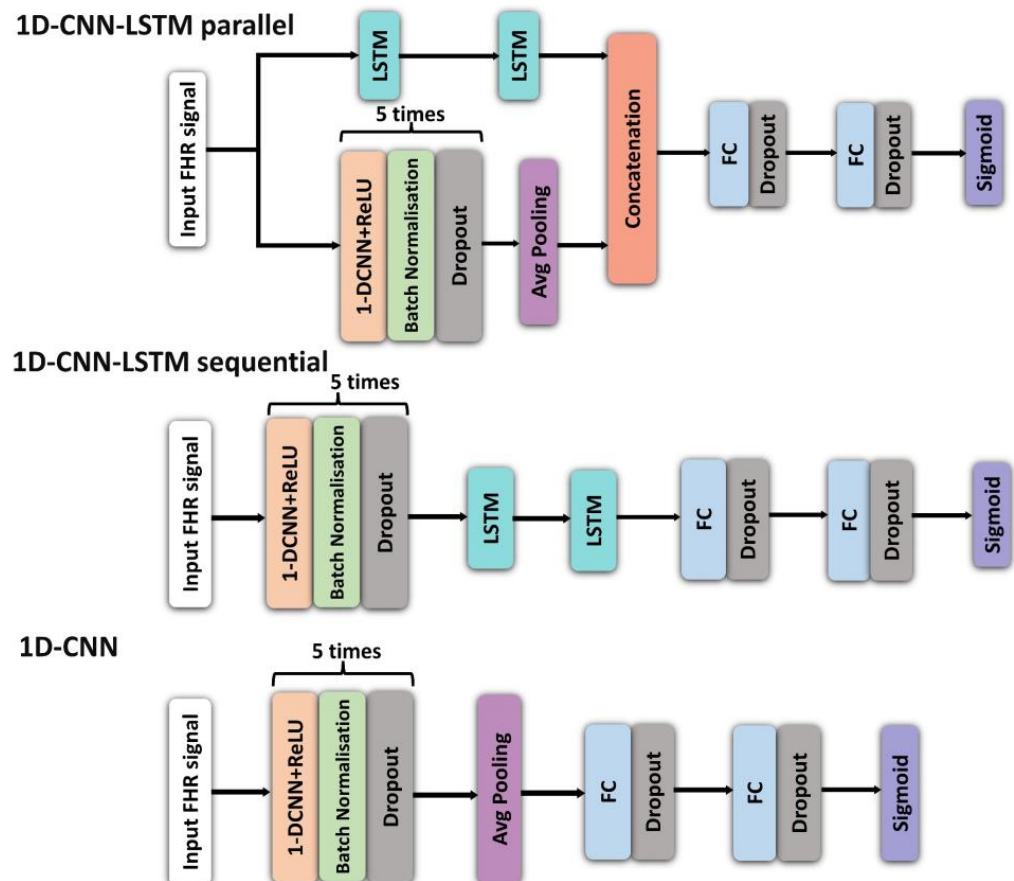


Figure 3. Architectures of: 1D-CNN-LSTM parallel (top); 1D-CNN-LSTM sequential (middle); and 1D-CNN (bottom) models (hyperparameters of these models are provided in Table 1).

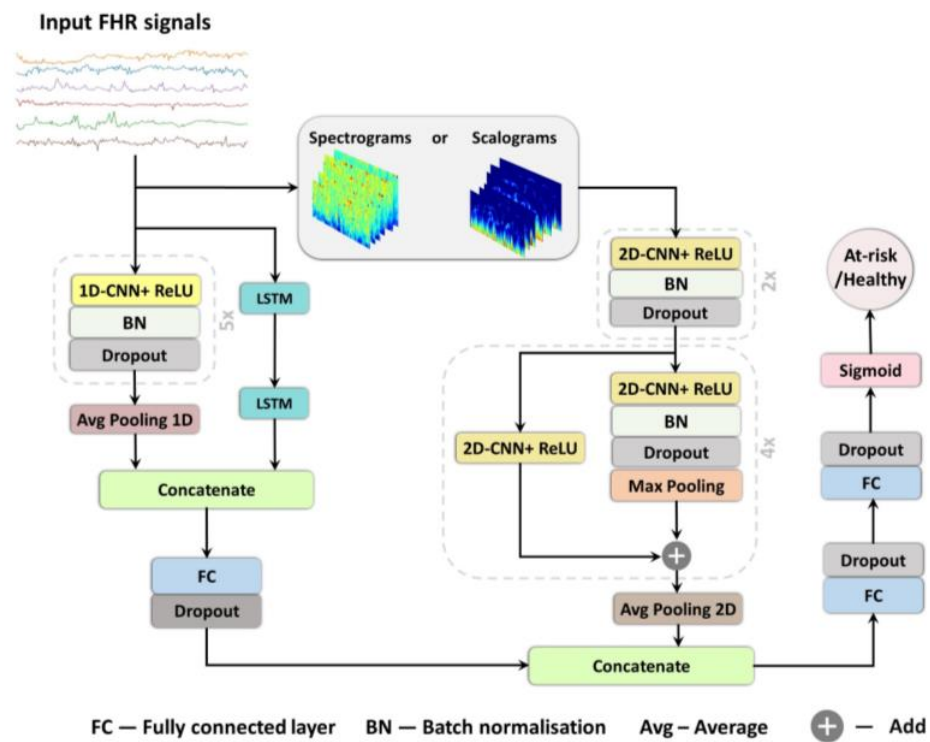


Figure 4. The proposed combination of 1D-CNN-LSTM and 2D-CNN architectures, used for the FHR signal classification.

Table 1. Optimal model hyperparameters tuned for the parallel 1D-CNN-LSTM architecture. Note that the number of filters is for each of the five layers respectively.

Hyperparameter	Range	Optimal Value
Batch size	16, 32, 64, 128, 256, 512, 1024	512
Number of CNN layers	4, 5, 6, 7, 8	5
kernel size	3, 5, 7, 10, 15	3
Number of CNN filters	8, 16, 32, 64, 128	16, 32, 64, 64, 16
Optimization functions	SGD, RMSprop, Adam, Nadam	Adam
LSTM units	8, 16, 32, 64	16

The combined 1D-CNN-LSTM and 2D-CNN deep learning network architecture is a two-channel network: one channel to analyse the input 1D-FHR signal using 1D-CNN-LSTM parallel topology, while the other channel uses a 2D-CNN with skip connections (Figure 4) to extract spectral features from the sample. This makes the architecture capable of analysing and capturing the signal’s temporal and spectral characteristics. The input FHR signal (1D signal) and the corresponding spectrogram or scalogram (2D) are simultaneously fed to the respective channels. Subsequently, the output of the two channels is concatenated and fed to an FC layer.

The input to each of the 1D-CNN-LSTM based models is prepared as (B, T, F), where B, T and F represent the batch size, the length of times slices, and the signal dimension, respectively. In a 20-min FHR signal sampled at 0.25 Hz, there are 300 time steps (15 per minute, $T = 300$, $F = 1$) in each CTG sample. Similarly, the input to the 2D-CNN network is arranged as (B, H, W, C), where B is the number of samples or the batch size, while W (128), H (128) and C (3) are the width, height and number of channels of the image respectively. A Sigmoid activation function is used at the last layer of all the networks to obtain the class probability prediction of each sample.

2.2.3. Training Procedure

We split the data randomly into training (85%) and test (15%) sets, while preserving the class ratio in each subset. Ten-fold cross-validation was performed using the training dataset, in which 90% of the samples are used for training the model and the remaining 10% for validation. During each fold, the model is trained for a maximum of 400 epochs with early stopping to mitigate overfitting (with a window size of 50 epochs) by monitoring the partial area under the receiver characteristic curve (PAUC). After the optimal model parameters are obtained, the PAUC is evaluated on the test set. We report the average performance of the ten models evaluated on the test set.

Since our class distribution is unbalanced, weighted binary cross-entropy loss is used for training, where the weight is based on the inverse class frequency, i.e., during training, the loss function penalises more (by factor of 14) the misclassification of severe compromised cases. As mentioned above, the network is trained for 400 epochs, using Adam optimiser with an initial learning rate of 0.001, decayed by a factor of 2 every 50 epochs. The batch size is set to 128 and batch normalisation is used with default parameters as recommended in [50]. Furthermore, we used dropout with a probability of 0.3 in all layers and early stopping based on the PAUC. The tuned hyperparameter values, including the CNN module's filter size, the number of filters used in each CNN layer, and the unit number of LSTM modules, are summarised in Table 1. The model is implemented using TensorFlow on an NVIDIA GTX 2080 Ti 12GB GPU machine.

To ensure that every feature in the data has the same level of importance, features are standardized using the z-score:

$$z_i = (x_i - \mu) / \sigma$$

where x_i is the original value of sample i in the dataset; z_i is the normalized value, μ is the mean, and σ is the standard deviation.

3. Results

3.1. Performance Metrics

The performance of the models is evaluated using the partial area under the receiver operating characteristic curve (PAUC) and the true positive rate (TPR = TP/(TP + FN)) at a 5% false-positive rate (FPR). The PAUC is the AUC between 0 and 10% false-positive rates (FPR = FP/(FP + TN)). The metrics are selected to assess the accuracy of the model only at a very low specificity, i.e., to detect adverse birth outcomes as accurately as possible, while minimising the rate of false positives. It is crucial to minimise unnecessary interventions, particularly early in the labour. ROC curve is a commonly accepted graphical plot that shows the performance of a binary classifier for all classification thresholds. It is based on TPR and FPR values, which are calculated from true positive (TP), false positive (FP), true negative (TN), and false-negative (FN) values. The AUC measures the area underneath the entire ROC curve from (0, 0) to (1, 1). We also considered the Precision, Recall and F1-score values of the three 1D-CNN and LSTM based models.

3.2. Hyperparameter Tuning

Optimal model hyperparameters are tuned using the Bayesian optimisation (BO) and Hyperband (HB), (BOHB) optimiser [51]. The selected values for the batch size, the number of filters in CNN each layer, the kernel size, the number of layers, and the optimisation functions are all summarised in Table 1.

3.3. Performance of the Proposed Models

The performance of the three 1D-CNN and LSTM based models is shown in Figure 5. The 1D-CNN-LSTM parallel architecture achieved higher Sensitivity and PAUC on the testing set than the 1D-CNN-LSTM sequential architecture. The difference between PAUC values of the 1D-CNN-LSTM parallel and 1D-CNN-LSTM sequential models are statistically significant (Mann-Whitney U tests, two-tailed, $U = 10$, $p = 0.002$). All other differences

between the three models in terms of PAUC and Sensitivity at 0.95 specificity values are not statistically significant. The comparison of the three models in terms of Precision, Recall, and F1-score is shown in Table 2. The best performing 1D-CNN and LSTM based models (from the 10 models trained using a 10-fold cross-validation) are shown in Figure 6.

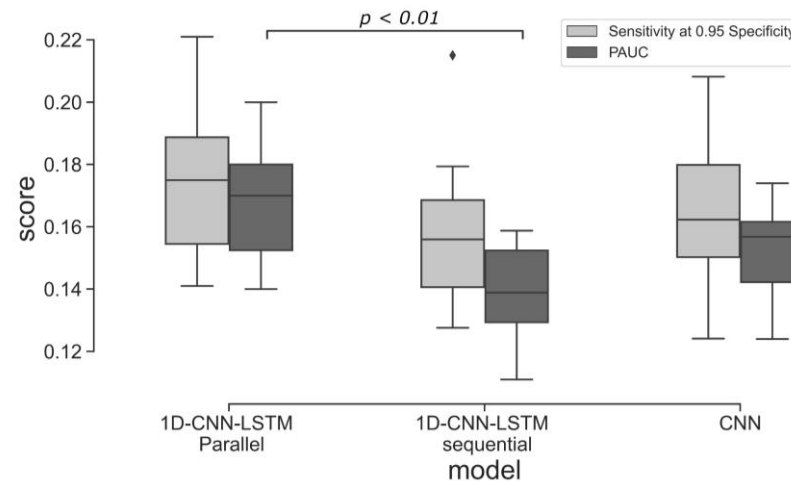


Figure 5. Classification performance (the ten models trained using 10-fold cross-validation evaluated on the test set) of the three 1D-CNN and LSTM based architectures.

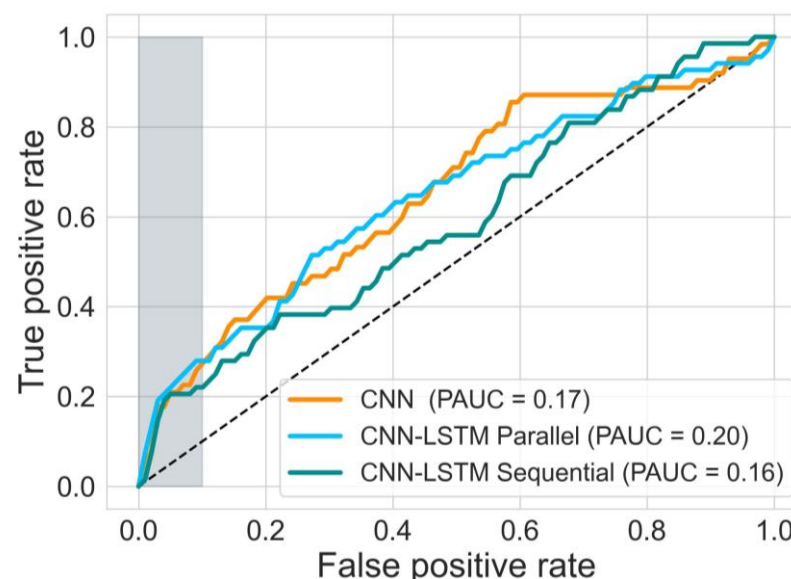


Figure 6. ROC of the best performing 1D-models from the 10-fold cross-validation, evaluated on the test set (68 severe compromises and 7704 samples without severe compromise).

Table 2. Classification performance (mean of the 10-fold cross-validation on the test set) of the 1D CNN and LSTM based architectures.

Model	AUC	Precision	Recall	F1-Score
1D-CNN	0.66	0.014	0.65	0.029
1D-CNN-LSTM sequential	0.61	0.012	0.49	0.025
1D-CNN-LSTM parallel	0.68	0.017	0.60	0.034

Table 3 shows the performance of the 2D-CNN and the multimodal architecture (combined 1D-CNN-LSTM and 2D-CNN). The 2D-CNN using scalogram analysis produced slightly better results compared to the case of using the spectrograms. However, the performance of the 2D-CNN (alone) and the multimodal architecture was inferior compared to the 1D-CNN-LSTM parallel architecture. This indicates that using the raw FHR (temporal representation) as input can lead to a better classification performance than the case relying on the time-frequency representations.

Table 3. Classification performance (mean of the 10-fold cross-validation on the test set) of the 2D CNN alone and when combined with 1D CNN-LSTM parallel architectures.

Method		PAUC	Sensitivity at 0.95 Specificity	AUC	Precision	Recall	F1-Score
Spectrogram	Only 2D CNN	0.12	0.13	0.59	0.010	0.48	0.019
	Combined with 1D-CNN-LSTM	0.15	0.14	0.60	0.011	0.45	0.021
Scalogram	Only 2D CNN	0.13	0.14	0.60	0.013	0.54	0.025
	Combined with 1D-CNN-LSTM	0.16	0.15	0.61	0.017	0.59	0.033

The performance metrics when varying the window size of the spectrograms and the kernel sizes of the 2D CNN model are given in Table 4. Three different kernel sizes are considered for the comparison: 3×3 ; 5×5 ; and 7×7 . The rest of the hyperparameters, such as, batch size, number of layers, and number of filters are selected using cross-validation. The highest PAUC and the sensitivity performance at a Specificity of 0.95 is achieved using a window size (overlapping size) of 64 (32) with a 3×3 kernel. This indicates that the two classes can be better separated when the input FHR signals' frequency content is analysed using a window size of about 1 min. Nevertheless, the performance of the spectrogram analysis is inferior compared to the 1D-CNN and LSTM based models.

Table 4. Impact of the window length on the classification performance of the spectrograms. The results are performances of the best models from the 10-fold cross-validation, evaluated on the test set (the best classification performances are given in bold).

Window Length (Overlap Size)	Kernel Size	PAUC	Sensitivity at 0.95 Specificity
32 (16)	$3 \times 3/5 \times 5/7 \times 7$	0.09/0.09/0.09	0.09/0.10/0.12
64 (32)	$3 \times 3/5 \times 5/7 \times 7$	0.11 /0.07/0.09	0.13 /0.12/0.13
128 (64)	$3 \times 3/5 \times 5/7 \times 7$	0.11/0.11/0.09	0.10/0.12/0.10
256 (128)	$3 \times 3/5 \times 5/7 \times 7$	0.09/0.09/0.08	0.10/0.13/0.13

Table 5 demonstrates the performance metrics and the influence of the different kernel sizes on the scalograms generated using a variety of wavelet functions. The proposed 2D-CNN achieved the highest classification results using a 3×3 kernel on scalograms generated with Mexican hat wavelet functions. This indicates a greater similarity between the input FHR signal and the Mexican hat, making it a preferable wavelet. The scalograms analysis showed slightly better separation between the two classes than the spectrograms. However, the wavelets' performances were inferior to those of the 1D-CNN and LSTM based models.

Table 5. Classification performance of the different wavelet functions and kernels sizes. The results are performances of the best models from the 10-fold cross-validation, evaluated on the test set (the best classification performances are given in bold).

Wavelet Function	Kernel Size	PAUC	Sensitivity at 0.95 Specificity
<i>Mexican hat</i> wavelet (mexh)	$3 \times 3/5 \times 5/7 \times 7$	0.13 /0.11/0.10	0.14 /0.09/0.09
<i>Morlet</i> wavelet (morl)	$3 \times 3/5 \times 5/7 \times 7$	0.11/0.11/0.11	0.12/0.12/0.11
<i>Shannon</i> wavelet (shan)	$3 \times 3/5 \times 5/7 \times 7$	0.12/0.09/0.08	0.14/0.12/0.13
<i>Gaussian</i> wavelet (gaus8)	$3 \times 3/5 \times 5/7 \times 7$	0.11/0.08/0.09	0.14/0.09/0.09

In the combined 1D and 2D architectures, the 2D CNN models based on spectrograms and scalograms were aggregated with the best performing 1D model (the 1D-CNN-LSTM parallel architecture). In this analysis, the spectrograms were computed with Fourier transform using a window size of 64, while the scalograms were obtained implementing the Mexican hat wavelets. The scalograms achieved slightly better results than the spectrograms, but the combined model's overall performance was inferior to the 1D-CNN-LSTM parallel construct, indicating that the time-frequency representation provides not enough useful content in separating the two classes. Conversely, the temporal features extracted using the 1D-CNN-LSTM parallel architecture separated better the two classes than when the features were extracted from the time-frequency representation using the 2D-CNN architecture.

3.4. Comparison with Clinical Practice and OxSys

We compared the 1D-CNN-LSTM parallel model to OxSys 1.5 (3) and clinical benchmark (11). OxSys uses two FHR features and two clinical risk factors to analyze the entire FHR trace with a 15-min sliding window. While in clinical practice, clinicians consider not only the findings of CTG interpretation but also consider clinical risk factors when making the diagnosis. The TPR of detecting severe adverse outcomes by the *Clinical Practice*, OxSys 1.5, and our model is presented in Table 6. The TPR in clinical practice reported in [3], which is defined as the number of emergency deliveries, is based on a clinical decision for “presumed fetal compromise” as a proportion of the total number of babies with compromise (5/162) within 2-h of start of CTG recording. The FPR is the number of emergency deliveries, based on a clinical decision for “presumed fetal compromise”, where there was no compromise as a proportion of the total number of normal cases (108/27,652). The results show that the OxSys 1.5 which is based on the entire CTG and clinical risk factors achieved highest sensitivity.

Table 6. Performance comparison between 1D-CNN-LSTM model, OxSys 1.5, and emergency deliveries in clinical practice for fetal distress cases.

Method	Sample Size		TPR	FPR
	Severe Compromise	No Severe Compromise		
OxSys 1.5 (entire CTG)	187	22,603	43.32%	16.45%
Clinical practice (first 2-h of CTG)	167	27,927	3.08%	0.39%
1D-CNN-LSTM (first 20 min)	68	7703	35.29%	16.16%

We also compared the sensitivity between our 1D CNN and LSTM-based models and clinical practice, focusing on a similar FPR value of 0.4%. The results, presented in Figure 7, indicate that the sensitivity of the 1D CNN-LSTM parallel model falls slightly below the optimal sensitivity achieved in clinical practice (2.4% vs. 3.1%). However, it is important to

interpret these findings with caution, as the results of *Clinical Practice* are based on clinical risk factors and the initial 2-h CTG recording, whereas our analysis is solely based on the initial 20 min of FHR recording.

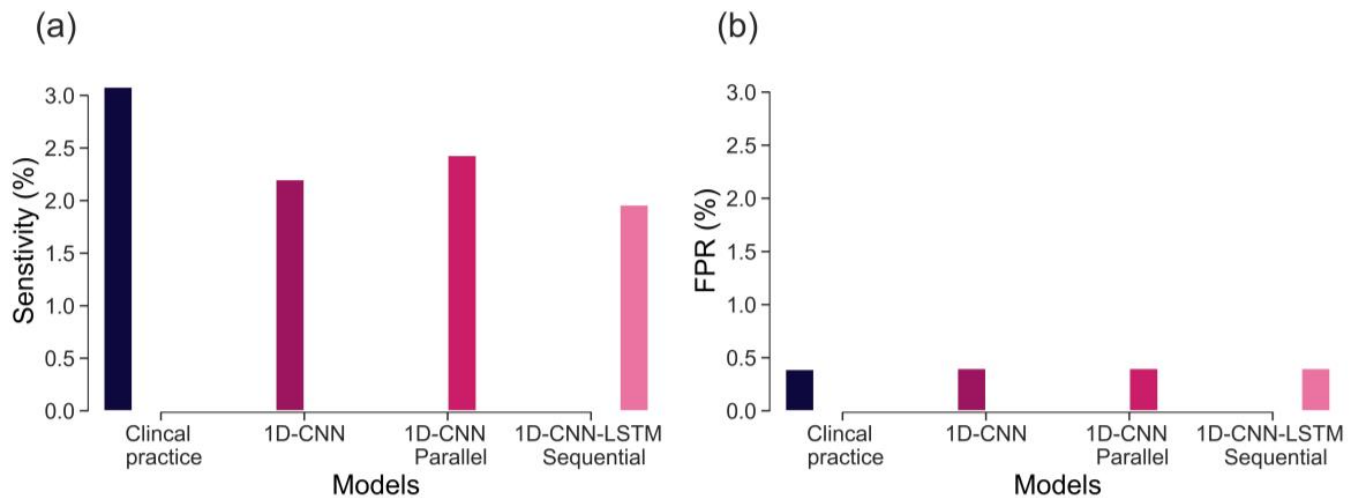


Figure 7. Performance comparison between 1D-CNN, 1D-CNN-LSTM parallel, 1D-CNN-LSTM sequential, OxSys 1.5, and *Clinical Practice* (emergency deliveries in clinical practice for fetal distress cases within two hours of admission to labour ward) in terms of: (a) Sensitivity (TPR); and (b) FPR.

3.5. Effect of the Pre-Processing on the Model's Output

We investigated the relationship between the model's output and the location and/or magnitude of signal loss within CTG segments. Table 7 shows Spearman's correlation coefficient between the model predictions (range between 0 and 1) and the percentage of signal loss, number of gaps in the signal, the longest gap length, and its location. The signal loss summary statistics are computed before the gap imputation. The results show weak correlation between the model's outputs and the signal loss summary statistics, indicating that our model predictions are largely independent of the magnitude and the location of signal loss.

Table 7. Spearman correlation between the 1D-CNN-LSTM network predictions and signal loss (testing set data).

Data (n = 7772)	Signal Loss	Number of Gaps	Longest Gap Length	Location of Longest Gap
Severe compromise	0.10	0.05	0.12	0.07
No severe compromise	0.14	0.09	0.13	0.07

3.6. Post-Hoc Analysis of the Models' Prediction

We utilized concept attribution as a technique to enhance the explainability of our deep learning model [52]. Concept attribution seeks to identify the key features or concepts in the input data that exert the greatest influence on the model's decision-making. In this case, we investigated whether the predictions generated by the deep learning model were related to the clinical features of FHR used for evaluating initial FHR traces. In the initial stages of labour, the FHR baseline and variability play a crucial role in the interpretation of CTG (11). Table 8 shows the relationship between the predictions of the model (1D-CNN-LSTM parallel architecture) and the standard clinical features of the FHR, such as FHR baseline and FHR short-term variability (STV). The testing set samples are divided into three groups, based on the quartiles of the predicted values from the 1D-CNN-LSTM network: low (≤ 25 percentile), medium (> 25 and < 75 percentile), and high (≥ 75 percentile). The samples with clinically low STV ($STV \leq 3$ ms) are more likely to have high DL predictions signifying increased risk (39.1% vs. 12.1%), which is in line with current clinical understanding that

diminished STV is a significant risk for fetal compromise. On the other hand, the model's predictions do not appear to be associated with the FHR baseline.

Table 8. Relationship between the model's predicted outputs, the FHR STV and Baseline values (Spearman's correlation coefficient (number of samples from the testing set)).

DL Predicted Values	STV ≤ 3	STV > 3	Baseline ≥ 150	Baseline < 150
Low	0.12 (99)	0.27 (1796)	0.24 (1613)	0.36 (282)
Medium	0.49 (401)	0.50 (3360)	0.53 (3564)	0.25 (195)
High	0.39 (321)	0.23 (1561)	1580)	0.39 (301)

4. Discussion

This study investigates the potential of implementing different deep learning architectures for predicting births with and without severe compromise outcomes, using the first 20 min of the FHR signals of more than 51,000 births, recorded as per clinical practice in a UK hospital during 1993–2012. From the designed and proposed DL architectures, the 1D-CNN-LSTM parallel topology model achieved superior classification performance compared to the other two developed models: 2D-CNN; and the multimodal architectures (combined 1D-CNN-LSTM and 2D-CNN). The suboptimal performance of the 2D-CNNs could be attributed to the lack of informative features in the time-frequency representations of the FHR signal. The post-hoc analysis also indicates that the performance of the 1D-CNN-LSTM model is not biased to signal loss, and its predictions are related to a degree to low STV values, which aligns with the clinical expectation of what is important in the initial FHR [11].

The sensitivity of CTG, based on initial hour in detecting severely compromised births, is not well established and there is limited evidence available. Lovers et al. [11] reported that the sensitivity of admission CTG is approximately 3.1% at 0.4% FPR (Figure 7). Our best model, the parallel 1D-CNN-LSTM model, achieved a slightly reduced sensitivity of 2.4% at a 0.4% FPR. This outcome is promising, particularly considering that clinical sensitivity in practice relies on evaluating the initial 2-h CTG data and incorporates various clinical risk factors. Nonetheless, the findings imply that our model has the potential to serve as a valuable aid for clinicians in identifying fetal distress and averting adverse birth outcomes.

Previous studies have also explored the potential of data-driven approaches in detecting abnormalities in CTG traces, focusing on the last hour CTG recording. For instance, Petrozziello et al., [37], demonstrated that a 1D CNNs model that employs more than 35,000 CTGs of the last hour recording can achieve higher TPR in predicting birth acidemia (pH < 7.05) than the clinical diagnosis (53% vs. 31% at about 15% FPR). Other studies (12, 30, 31), using a much smaller dataset and pH < 7.15 as an abnormal outcome, also implemented 1D CNNs to classify FHR signals. However, these works have focused on detecting birth acidemia based on the last hour CTG recording. When similar outcome groups are investigated (as in this work: with and without severe compromise), the OxSys 1.5 (3), achieved slightly higher TPR (43% at 14% FPR) than the clinical diagnosis (35% at 16% FPR) and our 1D-CNN-LSTM model (35% at 16% FPR) on a dataset of more than 22,000 CTGs. Nevertheless, this relatively higher accuracy results from analysing the entire FHR trace (in our case, it is based on the first 20 min only).

The main contributions of our work are: proposing and implementing DL models, based on uniquely large and detailed dataset allowing their successful training, validation, and testing; the clinically relevant definition of a rare severe compromise; and the focus on the first 20 min of the FHR, seeking an early warning for those fetuses that are unlikely to sustain the stress of labour due to pre-existing vulnerability. We capped the false positive rate to 5% and achieved sensitivity of 20%—an encouraging result given that most infants who would sustain severe compromise, are expected to do so later in labour. Also, given the fact that the false positive rate cannot be precisely defined, due to the routine nature of

our data, which includes high rates of clinical intervention and censoring the data. The achieved performance, as compared to the clinical benchmark (as shown in Figure 7), is highly encouraging. This is particularly noteworthy because predicting adverse outcomes in clinical practice relies not only on CTG patterns but also on various risk factors such as abnormal fetal growth, antepartum hemorrhage, prolonged rupture of membranes, and meconium staining of the amniotic fluid [9]. Consequently, a model that provides an objective assessment of CTG, without imposing a significant computational burden (with trace prediction in under a second with the ready trained model), can serve as an integral part of a clinical decision support tool. By doing so, it could contribute to optimizing the allocation of clinical resources, allowing clinicians to focus on other crucial responsibilities. Finally, in our future work, we expect to further improve the model accuracy by incorporating clinical risk factors into the analysis.

Some limitations of our approach are also worth noting. While data augmentation has demonstrated some effectiveness in addressing label imbalance and reducing model overfitting, it is important to consider the potential risk of amplifying label noise within the dataset. Thus, it is necessary to acknowledge that data augmentation alone may not offer a complete solution to the challenges associated with learning from imbalanced datasets. This limitation is evident in the variance of the ten cross-validated models, as depicted in Figure 5. Future work should address this by employing other techniques, perhaps creating synthetic data using generative adversarial networks [53]. The classification performance should also be improved by analysing longer traces and incorporating uterine contraction signals and clinical risk factors (such as fetal gestation and maternal co-morbidities) into the model. Finally, our approach does not explain which segment of the input leads to a particular prediction. Therefore, future work will consider applying an attention layer to provide better explainability [54].

5. Conclusions

We developed and evaluated three different deep neural network architectures to classify a 20-min FHR segment, recorded around the onset of labour, to investigate their potential in providing very early warning and triage women in labour into high or low-risk groups for further monitoring and/or review. We achieved superior performance using the proposed 1D-CNN-LSTM parallel architecture: the best model achieved a sensitivity of 20% at 95% specificity. The results are clinically encouraging, considering the fact that the majority of the compromised babies are not expected to have demonstrated problems, and if any, they are challenging to detect at the onset of labour. It is important to note that there is also room to improve the model classification performance by analysing the entire FHR trace and incorporating clinical risk factors.

Although the proposed DL architecture achieved encouraging results on the holdout test set, it could also be tested on an external dataset to further validate its generalisation performance. In addition, the investigated models do lack explainability, and future work will incorporate an attention mechanism that will introduce it.

Author Contributions: Conceptualization, D.A., I.J. and A.G.; methodology, D.A., I.J., L.I., A.N., R.L. and A.G.; data curation, A.G.; Data-Analysis: D.A., I.J., R.L. and A.G.; writing—original draft preparation, D.A., I.J. and A.G.; writing—review and editing: D.A., I.J., L.I., A.N., R.L. and A.G.; supervision, A.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC Grant number EP/V002511/1).

Institutional Review Board Statement: The study received ethical approval from the Newcastle & North Tyneside 1 Research Ethics Committee, Reference 11/NE0044 (data before 2008), and from the South Central Ethics Committee, Reference 13/SC/0153 (for data after 2008).

Informed Consent Statement: Informed consent by the participants was not required.

Data Availability Statement: The data that support the findings of this study are available from the corresponding authors upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Alfirevic, Z.; Gyte, G.M.; Cuthbert, A.; Devane, D. Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour. *Cochrane database Syst. Rev.* **2017**, 2, CD006066. [\[CrossRef\]](#)
- Farquhar, C.M.; Armstrong, S.; Masson, V.; Thompson, J.M.D.; Sadler, L. Clinician Identification of Birth Asphyxia Using Intrapartum Cardiotocography Among Neonates with and Without Encephalopathy in New Zealand. *JAMA Netw. Open* **2020**, 3, e1921363. [\[CrossRef\]](#) [\[PubMed\]](#)
- Georgieva, A.; Redman, C.W.G.; Papageorgiou, A.T. Computerized data-driven interpretation of the intrapartum cardiotocogram: A cohort study. *Acta Obstet. Gynecol. Scand.* **2017**, 96, 883–891. [\[CrossRef\]](#)
- Alfirevic, Z.; Devane, D.; Gyte, G.M. Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour. In *Cochrane Database of Systematic Reviews*; The Cochrane Collaboration, Ed.; John Wiley & Sons, Ltd.: Chichester, UK, 2013; p. CD006066. [\[CrossRef\]](#)
- Draper, E.; Gallimore, I.; Smith, L.; Fenton, A.; Kurinczuk, J.; Smith, P. *Maternal, Newborn and Infant Clinical Outcome Review Programme MBRACE-UK Perinatal Mortality Surveillance Report*; Infant Mortality and Morbidity Studies, Department of Health Sciences, University of Leicester: Leicester, UK, 2020.
- Resolution, N.H.S. Annual report and accounts 2020/21'. 2021. Available online: <https://resolution.nhs.uk/wp-content/uploads/2021/07/Annual-report-and-accounts-2020-2021-WEB-1.pdf> (accessed on 18 March 2023).
- Hug, L.; You, D.; Blencowe, H.; Mishra, A.; Wang, Z.; Fix, M.J.; Wakefield, J.; Moran, A.C.; Gaigbe-Togbe, V.; Suzuki, E.; et al. Global, regional, and national estimates and trends in stillbirths from 2000 to 2019: A systematic assessment. *Lancet* **2021**, 398, 10302. [\[CrossRef\]](#)
- Parts, L.; Holzmann, M.; Norman, M.; Lindqvist, P.G. Admission cardiotocography: A hospital based validation study. *Eur. J. Obstet. Gynecol. Reprod. Biol.* **2018**, 229, 26–31. [\[CrossRef\]](#) [\[PubMed\]](#)
- Devane, D.; Lalor, J.G.; Daly, S.; McGuire, W.; Cuthbert, A.; Smith, V. Cardiotocography versus intermittent auscultation of fetal heart on admission to labour ward for assessment of fetal wellbeing. *Cochrane Database Syst. Rev.* **2017**, 1, CD005122. [\[CrossRef\]](#) [\[PubMed\]](#)
- Blix, E. The admission CTG: Is there any evidence for still using the test? *Acta Obstet. Gynecol. Scand.* **2013**, 92, 613–619. [\[CrossRef\]](#)
- Lovers, A.A.K.; Ugwumadu, A.; Georgieva, A. Cardiotocography and Clinical Risk Factors in Early Term Labor: A Retrospective Cohort Study Using Computerized Analysis with Oxford System. *Front. Pediatr.* **2022**, 10, 784439. [\[CrossRef\]](#) [\[PubMed\]](#)
- Zhao, Z.; Zhang, Y.; Comert, Z.; Deng, Y. Computer-Aided Diagnosis System of Fetal Hypoxia Incorporating Recurrence Plot with Convolutional Neural Network. *Front. Physiol.* **2019**, 10, 255. [\[CrossRef\]](#)
- Huang, M.-L.; Hsu, Y.-Y. Fetal distress prediction using discriminant analysis, decision tree, and artificial neural network. *JBiSE* **2012**, 5, 526–533. [\[CrossRef\]](#)
- Czabanski, R.; Jezewski, J.; Matonia, A.; Jezewski, M. Computerized analysis of fetal heart rate signals as the predictor of neonatal acidemia. *Expert Syst. Appl.* **2012**, 39, 11846–11860. [\[CrossRef\]](#)
- Chen, T.; Guo, A.; Chen, Q.; Quan, B.; Liu, G.; Li, L.; Hong, J.; Wei, H.; Zhifeng, H. Intelligent classification of antepartum cardiotocography model based on deep forest. *Biomed. Signal Process. Control* **2021**, 67, 102555. [\[CrossRef\]](#)
- Ayres-de-Campos, D.; Rei, M.; Nunes, I.; Sousa, P.; Bernardes, J. SisPorto 4.0—Computer analysis following the 2015 FIGO Guidelines for intrapartum fetal monitoring. *J. Matern. Fetal. Neonatal. Med.* **2017**, 30, 62–67. [\[CrossRef\]](#) [\[PubMed\]](#)
- Cömert, Z.; Kocamaz, A.F. Open-access software for analysis of fetal heart rate signals. *Biomed. Signal Process. Control* **2018**, 45, 98–108. [\[CrossRef\]](#)
- Romano, M.; Bifulco, P.; Ruffo, M.; Improta, G.; Clemente, F.; Cesarelli, M. Software for computerised analysis of cardiotocographic traces. *Comput. Methods Programs Biomed.* **2016**, 124, 121–137. [\[CrossRef\]](#)
- de l'Aulnoit, A.H.; Boudet, S.; Demailly, R.; Delgranche, A.; Génin, M.; Peyrodie, L.; Beuscart, R.; de l'Aulnoit, D.H. Automated fetal heart rate analysis for baseline determination and acceleration/deceleration detection: A comparison of 11 methods versus expert consensus. *Biomed. Signal Process. Control* **2019**, 49, 113–123. [\[CrossRef\]](#)
- The INFANT Collaborative Group; Brocklehurst, P.; Field, D.; Greene, K.; Juszczak, E.; Keith, R.; Kenyon, S.; Linsell, L.; Mabey, C.; Newburn, M.; et al. Computerised interpretation of fetal heart rate during labour (INFANT): A randomised controlled trial. *Lancet* **2017**, 389, 1719–1729. [\[CrossRef\]](#)
- Nunes, I.; Ayres-de-Campos, D.; Ugwumadu, A.; Amin, P.; Banfield, P.; Nicoll, A.; Cunningham, S.; Sousa, P.; Costa-Santos, C.; Bernardes, J. Central Fetal Monitoring with and Without Computer Analysis: A Randomized Controlled Trial. *Obstet. Gynecol.* **2017**, 129, 83–90. [\[CrossRef\]](#)
- Yu, K.; Quirk, J.G.; Djuric, P.M. Fetal heart rate analysis by hierarchical dirichlet process mixture models. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 709–713. [\[CrossRef\]](#)

23. De l'Aulnoit, A.H.; Génin, M.; Boudet, S.; Demailly, R.; Ternynck, C.; Babykina, G.; de l'Aulnoit, D.H.; Beuscart, R. Use of automated fetal heart rate analysis to identify risk factors for umbilical cord acidosis at birth. *Comput. Biol. Med.* **2019**, *115*, 103525. [\[CrossRef\]](#)
24. Georgieva, A.; Payne, S.J.; Moulden, M.; Redman, C.W.G. Artificial neural networks applied to fetal monitoring in labour. *Neural. Comput. Applic.* **2013**, *22*, 85–93. [\[CrossRef\]](#)
25. Cömert, Z.; Kocamaz, A.F.; Subha, V. Prognostic model based on image-based time-frequency features and genetic algorithm for fetal hypoxia assessment. *Comput. Biol. Med.* **2018**, *99*, 85–97. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Dash, S.; Quirk, J.G.; Djuric, P.M. Fetal Heart Rate Classification Using Generative Models. *IEEE Trans. Biomed. Eng.* **2014**, *61*, 2796–2805. [\[CrossRef\]](#)
27. Hoodbhoy, Z.; Noman, M.; Shafique, A.; Nasim, A.; Chowdhury, D.; Hasan, B. Use of machine learning algorithms for prediction of fetal risk using cardiotocographic data. *Int. J. App. Basic Med. Res.* **2019**, *9*, 226. [\[CrossRef\]](#)
28. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [\[CrossRef\]](#)
29. Petrozziello, A.; Jordanov, I.; Papageorgiou, T.A.; Redman, W.G.C.; Georgieva, A. Deep Learning for Continuous Electronic Fetal Monitoring in Labor. In Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 18–21 July 2018; pp. 5866–5869. [\[CrossRef\]](#)
30. Ogasawara, J.; Ikenoue, S.; Yamamoto, H.; Sato, M.; Kasuga, Y.; Mitsukura, Y.; Ikegaya, Y.; Yasui, M.; Tanaka, M.; Ochiai, D. Deep neural network-based classification of cardiotocograms outperformed conventional algorithms. *Sci. Rep.* **2021**, *11*, 13367. [\[CrossRef\]](#)
31. Baghel, N.; Burget, R.; Dutta, M.K. 1D-FHRNet: Automatic Diagnosis of Fetal Acidosis from Fetal Heart Rate Signals. *Biomed. Signal Process. Control* **2022**, *71*, 102794. [\[CrossRef\]](#)
32. Hannun, A.Y.; Rajpurkar, P.; Haghpanahi, M.; Tison, G.H.; Bourn, C.; Turakhia, M.P.; Ng, A.Y. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Med.* **2019**, *25*, 65–69. [\[CrossRef\]](#)
33. Wang, T.; Lu, C.; Sun, Y.; Yang, M.; Liu, C.; Ou, C. Automatic ECG Classification Using Continuous Wavelet Transform and Convolutional Neural Network. *Entropy* **2021**, *23*, 119. [\[CrossRef\]](#)
34. Liao, Z.; Song, Y.; Ren, S.; Song, X.; Fan, X.; Liao, Z. VOC-DL: Deep learning prediction model for COVID-19 based on VOC virus variants. *Comput. Methods Programs Biomed.* **2022**, *224*, 106981. [\[CrossRef\]](#) [\[PubMed\]](#)
35. Chudáček, V.; Spilka, J.; Burša, M.; Janků, P.; Hruban, L.; Huptych, M.; Lhotská, L. Open access intrapartum CTG database. *BMC Pregnancy Childbirth* **2014**, *14*, 16. [\[CrossRef\]](#)
36. Liu, M.; Lu, Y.; Long, S.; Bai, J.; Lian, W. An attention-based CNN-BiLSTM hybrid neural network enhanced with features of discrete wavelet transformation for fetal acidosis classification. *Expert Syst. Appl.* **2021**, *186*, 115714. [\[CrossRef\]](#)
37. Petrozziello, A.; Redman, C.W.G.; Papageorgiou, A.T.; Jordanov, I.; Georgieva, A. Multimodal Convolutional Neural Networks to Detect Fetal Compromise During Labor and Delivery. *IEEE Access* **2019**, *7*, 112026–112036. [\[CrossRef\]](#)
38. Mohannad, A.; Shibata, C.; Miyata, K.; Imamura, T.; Miyamoto, S.; Fukunishi, H.; Kameda, H. Predicting high risk birth from real large-scale cardiotocographic data using multi-input convolutional neural networks. *NOLTA* **2021**, *12*, 399–411. [\[CrossRef\]](#)
39. Arpitha, Y.; Madhumathi, G.L.; Balaji, N. Spectrogram analysis of ECG signal and classification efficiency using MFCC feature extraction technique. *J. Ambient. Intell. Humaniz. Comput.* **2021**, *13*, 757–767. [\[CrossRef\]](#)
40. Hamelmann, P.; Vullings, R.; Kolen, A.F.; van Laar, J.O.E.H.; Tortoli, P.; Misch, M. Doppler Ultrasound Technology for Fetal Heart Rate Monitoring: A Review. *IEEE Trans. Ultrason. Ferroelect. Freq. Control.* **2020**, *67*, 226–238. [\[CrossRef\]](#) [\[PubMed\]](#)
41. Spilka, J.; Georgoulas, G.; Karvelis, P.; Chudáček, V.; Stylios, C.D.; Lhotská, L. Discriminating Normal from “Abnormal” Pregnancy Cases Using an Automated FHR Evaluation Method’. In *Artificial Intelligence: Methods and Applications*; Lecture Notes in Computer Science; Likas, A., Blekas, K., Kalles, D., Eds.; Springer International Publishing: Cham, Switzerland, 2014; Volume 8445, pp. 521–531. [\[CrossRef\]](#)
42. Barzideh, F.; Urdal, J.; Hussein, K.; Engan, K.; Skretting, K.; Mdoe, P.; Kamala, B.; Brunner, S. Estimation of Missing Data in Fetal Heart Rate Signals Using Shift-Invariant Dictionary. In Proceedings of the 2018 26th European Signal Processing Conference (EUSIPCO), Rome, Italy, 3–7 September 2018; pp. 762–766. [\[CrossRef\]](#)
43. Feng, G.; Quirk, J.G.; Djuric, P.M. Recovery of missing samples in fetal heart rate recordings with Gaussian processes. In Proceedings of the 2017 25th European Signal Processing Conference (EUSIPCO), Kos, Greece, 28 August–2 September 2017; pp. 261–265. [\[CrossRef\]](#)
44. Feng, G.; Quirk, J.G.; Heiselman, C.; Djuric, P.M. Estimation of Consecutively Missed Samples in Fetal Heart Rate Recordings. In Proceedings of the 2020 28th European Signal Processing Conference (EUSIPCO), Amsterdam, The Netherlands, 18–21 January 2021; pp. 1080–1084. [\[CrossRef\]](#)
45. Asfaw, D.; Jordanov, I.; Impey, L.; Namburete, A.; Lee, R.; Georgieva, A. Fetal Heart Rate Classification with Convolutional Neural Networks and the Effect of Gap Imputation on Their Performance. In *International Conference on Machine Learning, Optimization, and Data Science, Certosa di Pontignano, Italy, 19–22 September 2022*; Springer Nature: Cham, Switzerland, 2022; pp. 459–469.
46. Şengür, A.; Guo, Y.; Akbulut, Y. Time–frequency texture descriptors of EEG signals for efficient detection of epileptic seizure. *Brain Inf.* **2016**, *3*, 101–108. [\[CrossRef\]](#) [\[PubMed\]](#)
47. Verstraete, D.; Ferrada, A.; Droguett, E.L.; Meruane, V.; Modarres, M. Deep Learning Enabled Fault Diagnosis Using Time-Frequency Image Analysis of Rolling Element Bearings. *Shock. Vib.* **2017**, *2017*, 5067651. [\[CrossRef\]](#)

48. Megahed, F.M.; Chen, Y.J.; Megahed, A.; Ong, Y.; Altman, N.; Krzywinski, M. The class imbalance problem. *Nat. Methods*. **2021**, *18*, 1270–1272. [[CrossRef](#)]
49. Iwana, B.K.; Uchida, S. An Empirical Survey of Data Augmentation for Time Series Classification with Neural Networks. *arXiv* **2007**, arXiv:2007.15951v4. [[CrossRef](#)]
50. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv* **2015**, arXiv:1502.03167.
51. Falkner, S.; Klein, A.; Hutter, F. BOHB: Robust and Efficient Hyperparameter Optimization at Scale. *arXiv* **2021**, arXiv:1807.01774.
52. Graziani, M.G.; Andrearczyk, V.; Marchand-Maillet, S.; Müller, H. Concept attribution: Explaining CNN decisions to physicians. *Comput. Biol. Med.* **2020**, *123*, 103865. [[CrossRef](#)]
53. Ramponi, G.; Protopapas, P.; Brambilla, M.; Janssen, R. ‘T-CGAN: Conditional Generative Adversarial Network for Data Augmentation in Noisy Time Series with Irregular Sampling. *arXiv* **2019**, arXiv:1811.08295.
54. Mousavi, S.; Afghah, F.; Razi, A.; Acharya, U.R. ECGNET: Learning where to attend for detection of atrial fibrillation with deep visual attention. In Proceedings of the 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), Chicago, IL, USA, 19–22 May 2019; pp. 1–4.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.