Fusing Tabular Features and Deep Learning for Fetal Heart Rate Analysis: A Clinically Interpretable Model for Fetal Compromise Detection

Lochana Mendis, *Student Member, IEEE*, Debjyoti Karmakar, Marimuthu Palaniswami, *Fellow, IEEE*, Fiona Brownfoot*, and Emerson Keenan*, *Member, IEEE*

***Abstract*—*Objective:* Cardiotocography (CTG) is commonly used to monitor fetal heart rate (FHR) and assess fetal well-being during labor. However, its effectiveness in reducing adverse outcomes remains limited due to low sensitivity and high false-positive rates. This study aims to develop an interpretable deep learning model that fuses FHR time series with tabular clinical features to improve prediction of fetal compromise (umbilical artery pH $<$ 7.05). *Methods:* We introduce Fusion ResNet, a novel architecture combining residual convolutional networks for FHR signal processing with a parallel neural network for tabular features. The model was trained and internally validated on a private dataset of 9,887 FHR recordings. External validation was performed on the open-access CTU-UHB dataset comprising 552 recordings. Model interpretability was evaluated using Shapley Additive Explanations (SHAP) and Gradient-Weighted Class Activation Mapping (Grad-CAM). *Results:* Fusion ResNet achieved a mean area under the ROC curve (AUC) of 0.77 during internal cross-validation and a state-of-the-art AUC of 0.84 on the CTU-UHB dataset, outperforming existing deep learning approaches. SHAP analysis identified key clinical features contributing to predictions, while Grad-CAM highlighted salient FHR patterns linked to fetal compromise. *Conclusion:* The proposed model enhances predictive accuracy while providing clinically meaningful explanations, enabling more transparent and reliable CTG interpretation. *Significance:* This work demonstrates the potential of interpretable deep learning to improve fetal monitoring by integrating multimodal data, supporting timely and informed decision-making in obstetric care.**

***Index Terms*—Cardiotocography, Deep Learning, Electronic Health Records, Explainability, Fetal Heart Rate**

L. Mendis and M. Palaniswami are with the Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville 3010 Victoria, Australia.

F. Brownfoot and D. Karmakar are with the Obstetric Diagnostics and Therapeutics Group, Department of Obstetrics and Gynaecology, The University of Melbourne, Heidelberg 3084 Victoria, Australia.

E. Keenan is with the Obstetric Diagnostics and Therapeutics Group, Department of Obstetrics and Gynaecology, The University of Melbourne, Heidelberg 3084 Victoria, Australia and with the Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville 3010 Victoria, Australia.

* denotes equal authorship.

Corresponding author: Lochana Mendis (lochana.mendis@ieee.org)

I. Introduction

INTRAPARTUM stillbirths account for 45% of stillbirths worldwide [1]. A leading cause of intrapartum stillbirths is fetal asphyxia. It is caused by impaired placental blood gas exchange during labor leading to reduced oxygen supply to fetal tissues followed by acidosis compromising fetal wellbeing [2]. Causes of fetal asphyxia may include maternal hypo-tension, poor placental perfusion, and umbilical cord compressions [2]. The compensatory fetal responses to oxygen deprivation are reflected as abnormal changes in the fetal heart rate (FHR). Clinicians monitor the FHR to detect potential events that may compromise fetal well-being, allowing timely intervention through a cesarean section or instrumental delivery to prevent adverse outcomes.

The most widely used technique for FHR monitoring is cardiotocography (CTG), which simultaneously records the FHR and uterine activity [3]. These signals are visually evaluated based on consensus guidelines for intrapartum fetal monitoring such as the International Federation of Gynecology and Obstetrics guidelines (FIGO) [4]. Therefore, CTG interpretation is subjective and marked by significant intra- and interobserver variability [5]. Despite more than five decades of widespread use since its introduction in the 1960s, CTG has yet to show a significant impact in reducing neonatal mortality or long-term neurological impairment [4], [6]. However, its association with increased cesarean section rates has raised ongoing concerns [4], [6]. Furthermore, in clinical practice, the true positive rate (TPR) for detecting fetal compromise is low, typically ranging from 31% to 48%, with a false positive rate (FPR) between 16% and 21% [7], [8].

Recent research has proposed data-driven systems using machine learning (ML) and deep learning (DL) methods for FHR evaluation [7], [9]–[12] aiming to address the challenges of observer variability, low sensitivity, and high false positive rates associated with visual CTG interpretation. The ML classifiers commonly used for feature-based FHR evaluation are support vector machines (SVM) [8], [9] and logistic regression [13], [14]. Primarily, these ML methods use classical FHR features (e.g.: baseline, variability, accelerations, and decelerations) and/or human-crafted new features using feature selection and feature extraction [12]. Among these new features, decelerative capacity from phase-rectified signal averaging (DC_{PRSA}), median absolute deviation of FHR (MAD_{dtrd}), β_0 parameter representing the FHR baseline, and the Hurst parameter (H) have shown strong classification performance for fetal compromise detection [9], [15]. In contrast, DL methods learn directly from the raw FHR signals and may evaluate novel features that do not have a clear physiological meaning [16]. These methods often use feedforward neural networks such as convolutional neural networks (CNNs) [7], [10], [17], and recurrent neural networks such as long short-term memory (LSTM) networks. Existing works have shown that CNNs outperformed LSTMs as well as other ML methods in fetal compromise detection [10], [18], [19]. The main challenges for these DL methods are the limited availability of CTGs for training and interpretability [10], [16]. The current state-of-the-art performance reported on the public CTU-UHB intrapartum CTG dataset is by a multimodal CNN (MCNN) using the raw CTG signals along with quality features calculated based on the percentage of signal loss present in the signal. It was trained on over 35,000 CTGs and achieved an area under the receiver operating characteristic curve (AUC) of 0.81 by evaluating the last 60 minutes of CTG regardless of labor.

Prior evidence demonstrates that adverse outcomes are influenced by other clinical risk factors, including meconium-stained fluid, gestational diabetes, hypertension, and fetal growth restriction [14], [16]. Several studies have shown that including such factors, which are typically captured in electronic health records (EHR), alongside FHR features could improve fetal compromise detection [13], [14]. A recent retrospective study combined two FHR features and two clinical risk factors into a computerized CTG system (OxSys 1.5) and showed increased sensitivity for fetal compromise detection and reduced the false-positive rate [7], [11]. Another recent multicohort study investigated the impact of signal loss in CTG signals on perinatal asphyxia and found a strong association and concluded that integrating measures of signal loss into fetal monitoring algorithms may improve decision-making [20].

More recent studies have further advanced computerized CTG interpretation through transformer-based and multimodal deep learning architectures. For example, 1D-U-Net–based models that jointly analyze fetal heart rate and uterine activity signals have demonstrated improved accuracy and reliability in detecting abnormal CTG patterns [21]. Similarly, transformer-based frameworks, such as PatchCTG, have demonstrated robust performance in identifying high-risk cases during antepartum monitoring [22]. The integration of large language models into CTG interpretation has also been explored, highlighting the potential of natural language–augmented decision support to enhance interpretability and clinical confidence [23]. Furthermore, DeepCTG® 2.0, a CNN-based model validated across multiple clinical centers, demonstrated strong predictive capability for detecting neonatal acidemia during labor, with AUC values ranging from 0.74 to 0.83. The study also emphasized that incorporating relevant clinical variables could further enhance model performance by accounting for risk factors not evident in CTG signals [24].

Despite the existing literature on fetal compromise detection, no study has investigated the integration of these tabular features with DL models that use raw FHR as input and explored their interpretability. This study proposes a novel deep learning approach that builds on a previously established processing workflow and generalizable model to integrate tabular features with raw FHR time series for improved fetal compromise detection, with performance validated across two international CTG cohorts [10]. Furthermore, the model explainability is investigated to understand the models' decision-making process in both a global manner and based on individual case explanations to assist with clinical adoption.

The primary contributions of this study are as follows:

- Propose a novel deep learning fusion model to integrate tabular clinical features and classical FHR features with raw time-series FHR signals to enhance fetal compromise detection.
- Assess performance of traditional feature-based methods and deep learning fusion models on both internal and external validation data.
- Integrate two explainability AI techniques, SHAP and Grad-CAM, to improve the clinical interpretability of the model predictions and demonstrate their alignment with clinical knowledge.

II. Materials and Method

A. Data acquisition

This work uses retrospective FHR records and electronic health records acquired from two international cohorts: Australia and the Czech Republic. The first cohort consists of FHR records acquired from laboring women of gestation more than 36 weeks who had singleton deliveries between January 2010 and December 2021 at the Mercy Hospital for Women, Heidelberg, Melbourne, Australia. The Mercy Health Human Research Ethics Committee of the hospital approved the data extraction (Approval Number 2020-077). The FHR signal data were extracted from the Philips IntelliSpace Perinatal™ information system with technical support from Philips. The maternal, fetal, and neonatal information related to each delivery was extracted from the Birthing Outcomes System (BOS) electronic medical record software. This study only selected the FHR recordings ending within 1 hour of birth and had a cord blood pH evaluation within 40 minutes of birth resulting in 9,887 records. Hereafter, this dataset is referred to as "MHW-pH". For a detailed data selection process, we refer the reader to our prior study [10].

The second cohort was acquired between April 2010 and August 2012 at the obstetrics ward of the University Hospital in Brno, Czech Republic. The data extraction was approved by the Institutional Review Board of University Hospital Brno and all women have signed informed consent. It consists of 552 FHR records linked with electronic health records including cord pH and clinical features [25]. These records are available for public access in the PhysioNet database repository (https://physionet.org/content/ctu-uhb-ctgdb/1.0.0/).

B. Data preprocessing, splitting, and labeling

In this study, the last 60 minutes of the FHR signal regardless of the stage of labor were analyzed following prior works [7], [10], [17]. The FHR signal was preprocessed by removing artefacts and linearly interpolating short signal gaps of less than 15 seconds. Signal gaps longer than 15 seconds were considered as zero values for input to further data processing. The 15-second threshold for FHR gap interpolation was selected based on empirical evidence from prior fetal heart rate studies [12], [26]–[28], which demonstrated that interpolating short gaps maintains signal integrity, whereas longer imputations introduce distortion. Then, the FHR signal was smoothed down to 1 Hz to remove redundant beats and improve computational complexity while maintaining performance as determined by prior work [12]. The missing values in the tabular features were minimal ($<0.4\%$) and were inputted using the mean of the respective feature. The UC signals were not used in the study due to their poor quality, and no improvement in fetal compromise detection performance was observed in a previous study [12]. All FHR recordings of both datasets were included irrespective of their signal loss to reflect real clinical behavior.

The MHW-pH data was split into training and validation datasets using 5-fold cross-validation. The CTU-UHB dataset was used as an external testing dataset. The MinMax scaler normalization was applied for the FHR time series signals on both datasets. The tabular features were standardized with a zero mean and unit variance.

As the outcome-defining criterion, this study uses the arterial cord blood pH. It measures the acidity of fetal blood just after birth which is used as an objective biochemical indicator for identifying potentially compromised babies. This criterion has been widely used by existing studies for computerized fetal compromise detection to label their data. Predominantly, births with pH < 7.05 were labeled as the compromised class [7], [10], [12], [19]. This study uses the same threshold to label the datasets into Normal and Compromised classes following prior work for fair comparisons.

C. Tabular features

This study analyzed twelve tabular features composed of electronic health records (EHR), enhanced FHR features used in classical machine learning, and features related to signal quality. Specifically, six EHR features—Parity, Gestation, Maternal Age, Diabetes, Hypertension, and Meconium were selected as they were available in both cohorts. The first three features were used as real values and the remaining three as boolean values. In practice, these features are readily available for clinicians as clinical data for decision-making before delivery.

As FHR features, four enhanced FIGO features—median absolute deviation of FHR after baseline subtraction (MAD_{dtrd}) [9] quantifying the average depth of FHR, the intercept of the linear regression model for the evolution of baseline along time (β_0) [9], Hurst parameter (H) [9] quantifying the variability of FHR and the decelerative capacity of FHR calculated from the phase rectified signal averaging method (DC_{PRSA}) [15] quantifying the average downward movement of FHR. These four FHR features are identified as top features in prior works for automated fetal compromised detection using classical machine learning methods [9], [11]. This study utilized the original implementation available in their respective studies to calculate these features. The FHR features were calculated on the last 60 minutes of the 4 Hz FHR signal. To include the impact of signal loss inherent to FHR monitoring methodologies, this study used two features related to signal quality—the ratio of zeros in the FHR signal (SigLoss) and the mode of FHR acquisition (AqMode). The AqMode feature is used as a categorical variable coded with 1, 2, and 3 representing Doppler ultrasound-based FHR acquisition, direct fetal electrocardiogram using fetal scalp clip, and use of both respectively.

D. Feature ranking and machine learning approach

The tabular features were first ranked using the minimum Redundancy Maximum Relevance (mRMR) feature selection algorithm as shown in Fig. 1a. The mRMR algorithm selects the most relevant features to the class labels while minimizing the redundancy among the selected features [29]. All the ranked tabular features were analyzed using classical machine learning methods in the literature for fetal compromise detection. Specifically, this study compared the area under the receiver operating characteristic curve (AUC) performance for classifying fetal compromise using two support vector machine (SVM) methods already used for fetal compromise detection—classical linear SVM [30], and Sparse linear SVM with ℓ_1-norm as penalty and squared Hinge loss function (instead of standard ℓ_2-norm as penalty and Hinge loss) [9].

Each SVM method was evaluated in two experiments as shown in Fig. 1b. First, these features were evaluated using a stratified 5-fold cross-validation (CV) on the MHW-pH dataset. Then, the trained model on each fold was externally tested on the CTU-UHB dataset. During training, the class weights were adjusted inversely proportional to class frequencies in the input data to tackle the class imbalance. In all analyses, the regularization parameter C, which controls the trade-off between misclassification rate and data sparsity in SVM classifiers was optimized using a grid search with a 5-fold CV.

To compute the performance, each feature was first used in a univariate evaluation. Then the ranked features were iteratively evaluated with the SVM methods by adding one feature at a time to perform a multivariate analysis.

E. Fusion of tabular features and deep learning

This study proposes a novel deep learning model for fetal compromise detection that fuses tabular features with time series FHR, building upon the previously established ResNet-based architecture and optimum processing workflow [10]. An overview of the Fusion ResNet model architecture with two inputs (FHR time series and tabular features) is shown in Fig. 2a. The top five ranked tabular features from the mRMR algorithm evaluated on the MHW-pH dataset were selected for the deep learning approach.

The input to the ResNet branch is the 60-minute FHR time series (3600 values sampled at 1Hz). The ResNet model

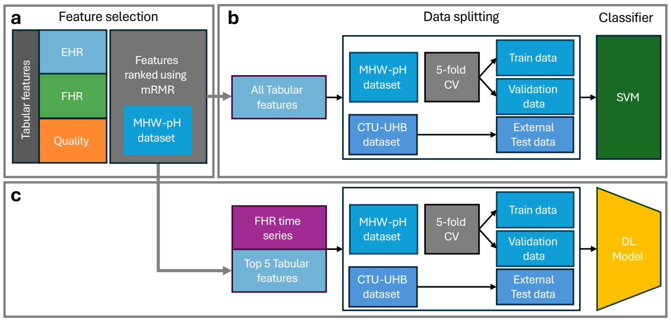

Fig. 1. Panel (a) shows the feature ranking of EHR (n=6), FHR (n=4), and Quality (n=2) features using the minimum redundancy maximum relevance (mRMR) feature selection algorithm. Panel (b) illustrates the data splitting and performance evaluation of all ranked tabular features with the SVM classifier (Classical SVM and Sparse SVM). Panel (c) illustrates the use of the top five tabular features ranked by using the mRMR method on the MHW-pH training dataset, along with the FHR time-series signal for deep learning evaluation. In both evaluations, the MHW-pH dataset was used for the 5-fold CV, while CTU-UHB was used as the external testing data.

consists of three residually connected 1D convolutional blocks. The output of the last convolution block is passed through a global average pooling layer (GAP) that averages the temporal dimension to a single value resulting in a feature vector equal to the number of filters of the last convolutional layer (1 x 128 in this case). More details of the ReNet model architecture are given in our prior work [10]. The tabular features are processed through a second branch comprising two fully connected (FC) hidden layers with output dimensions of 10 and 128, respectively. A dropout layer is included between the FC layers to reduce the risk of overfitting. The first FC layer uses a rectified linear unit (ReLU) activation function, while the second employs a sigmoid activation function to constrain the output range. The feature vectors from both branches are combined via a fusion operator and subsequently passed through a final output layer with a sigmoid activation function to generate the predicted class probabilities. Three fusion operators were evaluated: element-wise multiplication, addition, and concatenation.

In training and evaluating this proposed Fusion ResNet model, we used a stratified 5-fold CV on the MHW-pH dataset and used CTU-UHB as the external testing dataset as shown in Fig. 1c. The cases with pH that belong in the range, pH $\geq$ 7.05 and pH $<$ 7.15 (intermediate cases), were removed from the training data to minimize potential errors in the labels as it is not well established in the literature whether this group consists of compromised events or not [7], [10], [12]. However, the intermediate cases were not removed from the testing datasets as they are present in reality. During the training process, class weights based on inverse class frequency were used to tackle the class imbalance. This ensured heavy penalization for misclassifying minority cases during training. The binary cross entropy was used as the loss function and Adam as the optimizer. All models were trained for a maximum of 400 epochs with a batch size of 16 and an initial learning rate of 0.0001. Early stopping based on validation loss was used to dynamically stop the training.

This study also investigated the use of a pre-trained ResNet branch compared to keeping it trainable during the training process (Fig. 2b). Both trainable and pre-trained deep learning models were independently trained and evaluated by progressively adding tabular features to the model, one at a time, according to their feature importance ranking. The performance of only using the FHR time series branch was also evaluated. As performance metrics, the true positive rate (TPR) at 5%, 10%, 15%, and 20% false positive rate (FPR) and AUC were calculated. The mean and standard deviation of these performance metrics for 5-fold CV on MHW-pH and external test on CTU-UHB for each experiment were reported. The best-performing model was selected based on the 5-fold CV performance on the MHW-pH dataset. All experiments were carried out on an Intel® Xeon® Gold 6326 CPU at 2.90 GHz with an Nvidia A100 GPU and 16 GB of RAM. The models were implemented using the TensorFlow 2.0 framework.

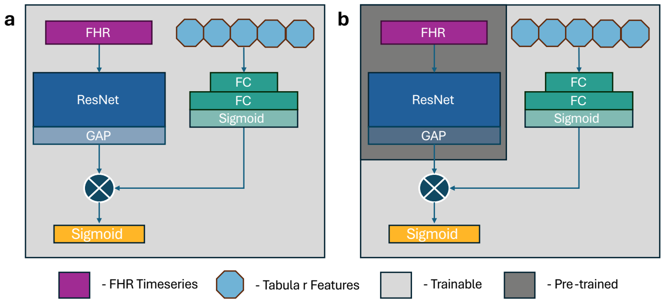

Fig. 2. Fusion ResNet model architecture with two branches, where time series FHR uses the existing ResNet model and tabular features use a dense model in a separate branch. Outputs from both branches are fused via a fusion operator and passed through the final sigmoid layer. Three fusion operators were evaluated: element-wise multiplication, addition, and concatenation. (a) and (b) shows the trainable regions for the trainable model and the pre-trained model, respectively.

F. Model interpretability

To explain the model predictions of the proposed Fusion ResNet model, this study used Grad-CAM (Gradient weighted Class Activation Mapping) [31] and SHAP (Shapely Additive exPlanations) [32] methods widely used for explainable artificial intelligence. Both methods focus on calculating feature attribution scores assigned to each input of the model to quantify their significance to the prediction made by the model.

The Grad-CAM is used in this study to visualize the regions of the FHR time series that contribute to the model prediction. It uses the channel-wise mean of the gradient information flowing into the last convolutional layer of the ResNet model as weights to calculate a weighted sum of the feature maps in the last convolutional layer. This is then passed through a ReLU function to focus on the positive contributions producing a coarse localization map that highlights the important regions of FHR time series influencing the model prediction.

The SHAP values are used in this study to visualize the contributions of tabular features and FHR time series as a whole for the model prediction. To compute the SHAP values, we use SHAP DeepExplainer [32] from the SHAP package. The SHAP values ensure a fair distribution of importance among the input features based on their contributions to the

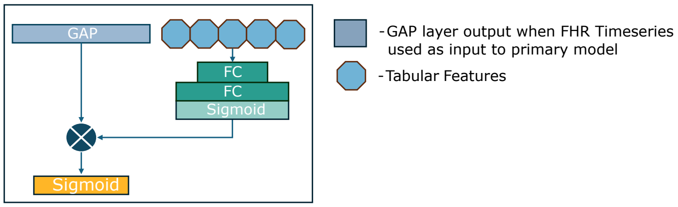
GAP
FC
FC
Sigmoid
Sigmoid
-GAP layer output when FHR Timeseries used as input to primary model
-Tabular Features

Fig. 3. Secondary fused deep learning model used with SHAP DeepExplainer. All corresponding layer weights were replicated from the trained primary model.

model output. Applying SHAP directly to the FHR time series would produce 3600 SHAP values which will be hard to interpret and aggregate to compare with the SHAP values from tabular features. Therefore, as illustrated in Fig. 3, we created a secondary model with the GAP layer (128 values) from the primary model as input for the FHR time series branch. All layers of the secondary model were replicated with the corresponding layer weights from the primary model. The 128 GAP values are a compressed representation of the time series, and computing SHAP on these values enables more meaningful and computationally efficient insights. To generate a single SHAP value representing the FHR time series, we aggregated the 128 SHAP values from the GAP layer using summation. The MHW-pH data was used as the background data for the SHAP DeepExplainer and was tested with the CTU-UHB dataset.

G. Statistical testing

The Mann-Whitney U test was used to test the statistical significance of the tabular features for the Normal and Compromised groups. The same test was used to test the significant differences between the AUC performance of the Fusion ResNet model on CTU-UHB for all tabular feature pairs including the FHR time series. The Pearson correlation testing was used to compute the significance of the correlation among the tabular features.

III. Results

A. Data distribution and correlation

This study evaluated automated fetal compromise detection using the last 60 minutes of the FHR time series in labor from two international cohorts from Australia (MHW-pH, n=9,887) and the Czech Republic (CTU-UHB, n=552). The participant characteristics of the two cohorts are shown in Table I. Cord pH measurements were used as the outcome criterion to divide the data into Compromised (pH < 7.05) and Normal classes (pH ≥ 7.05). This resulted in 40 (7.2%) and 111 (1.1%) cases for the compromised class for CTU-UHB and MHW-pH datasets, respectively.

In addition to the FHR time series, twelve tabular features including six electronic health record features, four enhanced FHR features, and two quality features were included in the analysis. The distributions of twelve features for the two classes in both datasets are shown in Fig. 4a. A significant difference in the parity, MAD_{dtrd} (median absolute deviation of the detrended FHR), DC_{PRSA} (decelerative capacity of FHR), and SigLoss (ratio of signal loss in FHR) features was observed in the Normal and Compromised classes in both datasets. Additionally, maternal age showed a significant difference in the two classes for the MHW-pH dataset, while β_0 (intercept of the FHR baseline evolution model) was significant for the CTU-UHB dataset. Furthermore, to examine the relationship among the selected features in each dataset, the Pearson correlation was computed and shown as a heatmap in Fig. 4b.

B. Univariate and multivariate feature analysis using classical machine learning

The predictive capability of each tabular feature was individually evaluated using two support vector machine (SVM) models to determine their effectiveness in fetal compromise detection. The AUC performances on the univariate and multivariate feature evaluation using SVM models on 5-fold CV on the MHW-pH dataset and external testing on the CTU-UHB dataset are shown in Fig. 5. The rank of the features for multivariate analysis was determined by the minimum redundancy maximum relevance (mRMR) method on the MHW-pH dataset.

In the univariate analysis using a 5-fold CV on the MHW-pH dataset, the MAD_{dtrd} feature showed the best performance (AUC of 0.66) closely followed by SigLoss (AUC=0.60), DC_{PRSA} (AUC=0.59), and Parity (AUC=0.59) on the Sparse SVM model. Similarly, the MAD_{dtrd} feature achieved the best performance (AUC=0.76) when externally tested on CTU-UHB followed by features SigLoss (AUC=0.71) and DC_{PRSA} (AUC=0.66) using the same Sparse SVM model. In both 5-fold CV and external testing, the Sparse SVM model showed superior performance compared to the Classical SVM model in the majority of the features.

In the multivariate analysis using a 5-fold CV on the MHW-pH dataset, the AUC performance on both SVM models showed an upward trend initially for the first three features (β_0, Parity, and MAD_{dtrd}), then when more features were added Sparse SVM performance remained marginally stable while Classical SVM performance showed a slight reduction in performance. Both models showed the best AUC performance of 0.69 with the first three features. In contrast, the external test on CTU-UHB showed an initial drop in the performance of both SVM models when the Parity feature was combined with β_0 and then increased sharply with the addition of the MAD_{dtrd} feature and remained relatively stable beyond three features. The AUC performances of the first three features when tested externally on CTU-UHB were 0.75 and 0.74 for Sparse SVM and Classical SVM, respectively.

C. Fusing FHR time series with tabular features using deep learning

The top five ranked features based on the MHW-pH dataset were used to integrate with the time-series FHR data using a deep learning approach to evaluate their performance in fetal compromise detection. The proposed fusion deep learning model (Fusion ResNet) was trained and evaluated in two ways — first with the time series FHR branch as trainable and second with the time series FHR branch as pre-trained. The 5-fold CV performance on the MHW-pH dataset and the external test

TABLE I
PARTICIPANT CHARACTERISTICS OF THE CZECH REPUBLIC AND AUSTRALIAN COHORTS. STD = STANDARD DEVIATION, IQR = INTERQUARTILE RANGE.

Characteristics	**CTU-UHB (n=552)**				**MHW-pH (n=9887)**			
	Mean (STD)	**Median(IQR)**	**Min**	**Max**	**Mean(STD)**	**Median(IQR)**	**Min**	**Max**
Gestation (weeks)	40(1.13)	40(39-41)	37	43	39.45(1.25)	39.5(38.5-40.3)	36.10	43.10
Maternal Age (years)	29.67(4.54)	30(27-33)	18	46	31.95(4.48)	32(29.05-35)	18	58.1
pH	7.23(0.11)	7.25(7.17-7.3)	6.85	7.47	7.25(0.08)	7.26(7.2-7.3)	6.00	7.52
	n(%)				**n(%)**			
Parity	0: 376(68.1) 1: 140(25.4) 2: 29(5.2) ≥3: 7(1.3)				0: 6491(65.7) 1: 2460(24.9) 2: 697(7.0) ≥3: 239(2.4)			
Diabetes	37(6.7)				1333(13.5)			
Hypertension	44(8.0)				276(2.8)			
Meconium	64(11.6)				1832(18.5)			

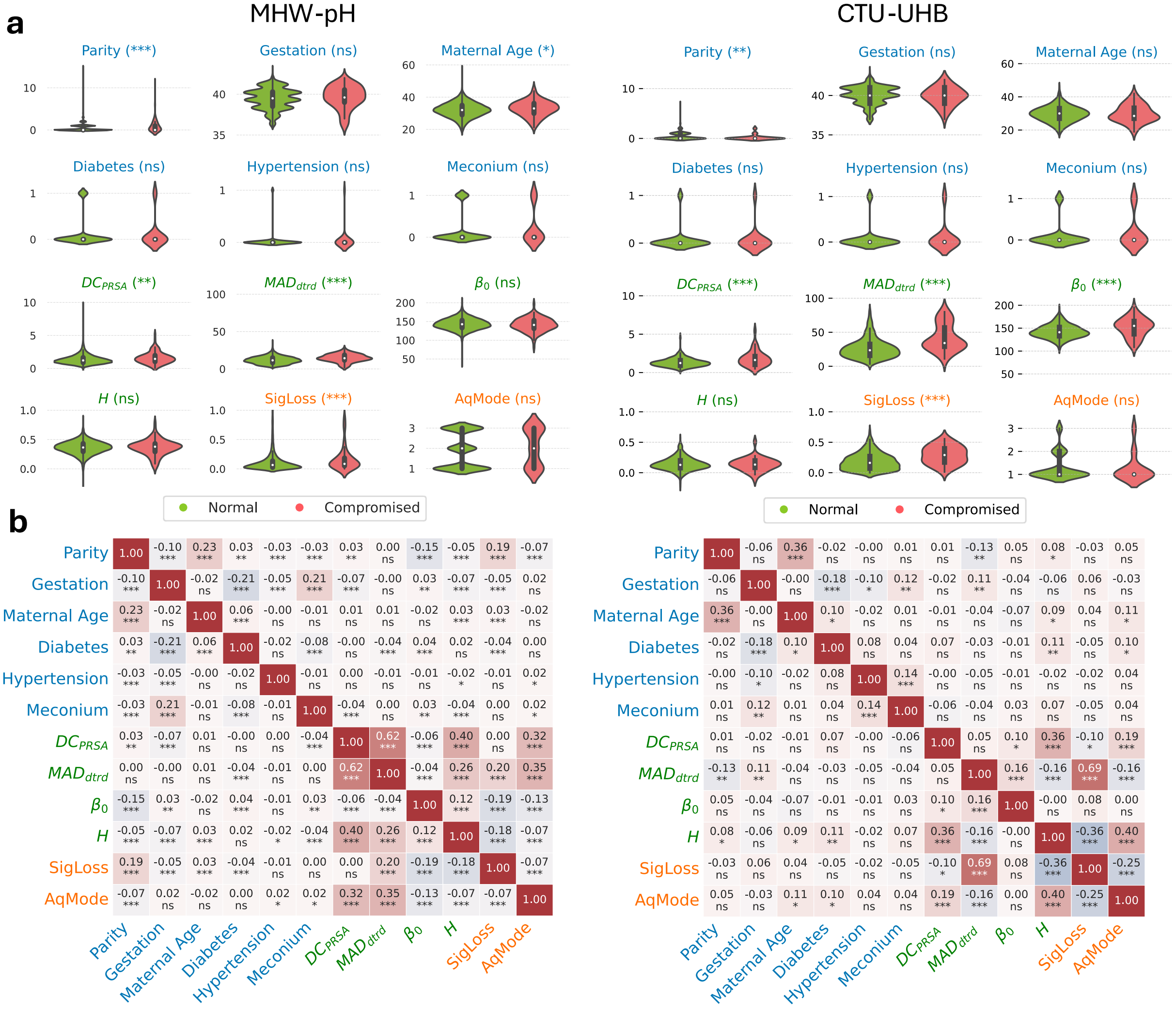

Fig. 4. Top row (a) Violin plots illustrate the distribution of features for the two classes: Compromised (pH $<$ 7.05) and Normal, in two international cohorts. Bottom row (b) Heatmap of Pearson correlation among tabular features consisting of Electronic Health Records (EHR) in blue, enhanced FHR features in green, and Quality features in orange. Statistical significance is indicated by asterisk: ns (p$\geq$0.05), * (p$<$0.05), ** (p$<$0.01), and *** (p$<$0.001).

performance on CTU-UHB for the trainable and pre-trained models are given in Table II and Table III, respectively.

As observed in Table II, the 5-fold CV and external test results demonstrated that combining tabular features with the FHR time series, using the Fusion ResNet model with a trainable FHR branch, outperformed the model that relied

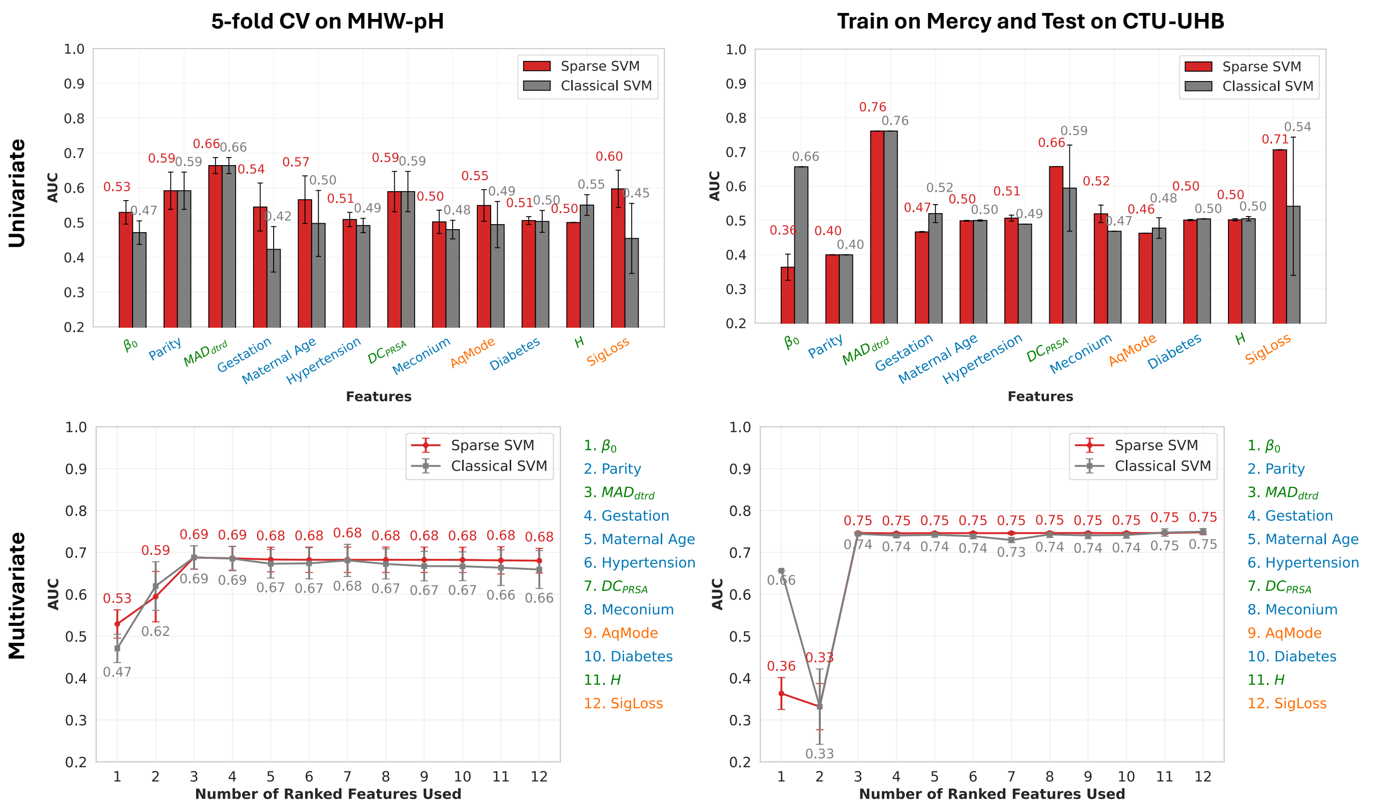

Fig. 5. Univariate and multivariate mean AUC performance of the twelve tabular features evaluated with two types of SVM models using 5-fold cross-validation on the MHW-pH dataset (left column) and when trained on MHW-pH and tested on CTU-UHB (right column). The features were ranked using the minimum redundancy maximum relevance (mRMR) method on the MHW-pH dataset for multivariate evaluation. The same feature order was used for univariate illustrations.

TABLE II

PERFORMANCE OF THE DEEP LEARNING MODELS WITH TRAINABLE FHR BRANCH GIVEN AS MEAN ± STANDARD DEVIATION. 5-FOLD CV ON MHW-PH AND EACH FOLD MODEL EXTERNALLY TESTED ON CTU-UHB. FINAL MODEL SELECTED BASED ON 5-FOLD CV PERFORMANCE HIGHLIGHTED IN BOLD. FHR = FHR TIME SERIES.

Features	5-fold CV (MHW-pH)					External Test (CTU-UHB)				
	TPR (%)				AUC	TPR (%)				AUC
	At 5% FPR	At 10% FPR	At 15% FPR	At 20% FPR		At 5% FPR	At 10% FPR	At 15% FPR	At 20% FPR	
FHR	26 ± 8	40 ± 7	47 ± 7	50 ± 8	0.73 ± 0.04	33 ± 5	43 ± 6	54 ± 4	62 ± 0	0.80 ± 0.02
FHR + β_0	26 ± 9	39 ± 7	46 ± 6	50 ± 6	0.72 ± 0.02	43 ± 4	50 ± 6	57 ± 6	65 ± 5	0.82 ± 0.04
FHR + β_0 + Parity	27 ± 6	41 ± 9	47 ± 7	54 ± 9	0.76 ± 0.04	33 ± 6	44 ± 4	51 ± 2	61 ± 2	0.79 ± 0.02
FHR + β_0 + Parity + MAD_{dtrd}	26 ± 7	41 ± 7	50 ± 5	60 ± 5	0.76 ± 0.04	48 ± 10	64 ± 4	70 ± 4	74 ± 2	0.85 ± 0.02
FHR + β_0 + Parity + MAD_{dtrd} + Gestation	**28 ± 6**	**40 ± 6**	**50 ± 7**	**55 ± 7**	**0.77 ± 0.03**	**45 ± 3**	**59 ± 8**	**67 ± 5**	**72 ± 4**	**0.84 ± 0.02**
FHR + β_0 + Parity + MAD_{dtrd} + Gestation + Maternal Age	22 ± 6	41 ± 7	50 ± 6	56 ± 5	0.75 ± 0.03	48 ± 6	62 ± 6	70 ± 6	74 ± 4	0.85 ± 0.02

solely on FHR time series data across most tabular feature combinations. Furthermore, all deep learning models outperformed the classical machine learning model performance in 5-fold CV on MHW-pH and external CTU-UHB evaluations. The mean AUC 5-fold CV and external test performance of the Fusion ResNet model with only FHR time series as input (without fusing tabular features) were 0.73 and 0.80 respectively. The 5-fold CV performance increased when more tabular features were added to the model until the fourth feature and marginally dropped when the Maternal age feature was added. External evaluation on the CTU-UHB dataset showed an increase in AUC performance with the addition of the β_0 feature. However, a notable drop in performance occurred when the Parity feature was included. The subsequent addition of the next three features resulted in relatively stable AUC values, with no significant differences among them, but they were statistically significant compared to the model using only the FHR time series and model with only β_0 and Parity tabular features, as illustrated in Fig. 6.

The highest mean 5-fold CV performance (AUC=0.77) as shown in Table II was achieved by the model that fused the FHR time series data with the top four tabular features (β_0,

TABLE III

PERFORMANCE OF THE DEEP LEARNING MODELS WITH PRE-TRAINED FHR BRANCH GIVEN AS MEAN ± STANDARD DEVIATION. 5-FOLD CV ON MHW-PH AND EACH FOLD MODEL EXTERNALLY TESTED ON CTU-UHB. FHR = FHR TIME SERIES

Features	5-fold CV (MHW-pH)					External Test (CTU-UHB)				
	TPR (%)				AUC	TPR (%)				AUC
	At 5% FPR	At 10% FPR	At 15% FPR	At 20% FPR		At 5% FPR	At 10% FPR	At 15% FPR	At 20% FPR	
FHR	26 ± 8	40 ± 7	47 ± 7	50 ± 8	0.73 ± 0.04	33 ± 5	43 ± 6	54 ± 4	62 ± 0	0.80 ± 0.02
FHR + β_0	22 ± 5	36 ± 4	43 ± 3	50 ± 3	0.73 ± 0.04	33 ± 3	44 ± 5	54 ± 4	64 ± 4	0.81 ± 0.01
FHR + β_0 + Parity	23 ± 6	38 ± 8	45 ± 4	51 ± 6	0.74 ± 0.03	29 ± 4	42 ± 8	56 ± 7	64 ± 7	0.79 ± 0.04
FHR + β_0 + Parity + MAD_{dtrd}	22 ± 7	35 ± 6	47 ± 2	55 ± 3	0.75 ± 0.03	35 ± 6	51 ± 7	60 ± 7	68 ± 6	0.81 ± 0.05
FHR + β_0 + Parity + MAD_{dtrd} + Gestation	24 ± 10	37 ± 5	49 ± 4	56 ± 4	0.74 ± 0.03	34 ± 4	49 ± 10	58 ± 8	66 ± 9	0.79 ± 0.08
FHR + β_0 + Parity + MAD_{dtrd} + Gestation + Maternal Age	24 ± 9	37 ± 4	48 ± 6	54 ± 6	0.73 ± 0.04	39 ± 6	54 ± 6	61 ± 7	69 ± 1	0.83 ± 0.01

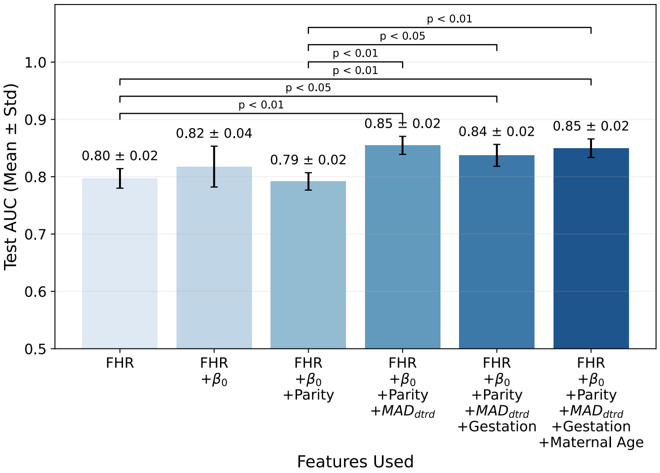

Fig. 6. The comparison of AUC performance in using the Fusion ResNet model with only FHR time series, and the Fusion ResNet model with a trainable FHR branch and a combination of tabular features, tested on the CTU-UHB dataset. A p-value < 0.05 is considered statistically significant. FHR = FHR time series.

Parity, MAD_{dtrd}, and Gestation). The corresponding mean AUC performance on the test CTU-UHB dataset was 0.84 (67-72% TPR at 15-20% FPR). Furthermore, the performance of using a trainable FHR branch, as shown in Table II, achieved superior performance compared to using a pre-trained FHR branch, as shown in Table III, across all evaluated feature combinations.

D. Comparison of Fusion Operators

A comparison study was performed to evaluate the impact of different fusion operators on model performance while keeping all other architectural and training components identical. Three fusion mechanisms were compared: (1) element-wise multiplication, (2) addition, and (3) concatenation. For all variants, the FHR and EHR branches each produced 128-dimensional latent representations before fusion, and the same learning schedules and regularization parameters were applied to ensure a fair comparison.

The results presented in Table IV show that element-wise multiplication achieved the highest internal validation AUC (0.77) and strong external performance (AUC = 0.84) on the CTU-UHB dataset. The addition and concatenation operators produced comparable but slightly lower internal validation AUCs; however, the concatenation approach achieved superior AUC on the CTU-UHB external validation dataset, at the cost of a slight increase in parameter count.

E. Explaining deep learning model predictions

This study computed SHAP values of the inputs to the Fusion ResNet model to investigate their contribution to the model output. The SHAP values indicate the amount of impact of individual features on the model output. The SHAP beeswarm plot shown in Fig. 7 demonstrates the FHR time series data is the most dominant global feature influencing the model predictions on the CTU-UHB external testing dataset. The MAD_{dtrd} was the next most influential feature, followed by β_0, Maternal age, and Parity. The Gestation feature is the least influential.

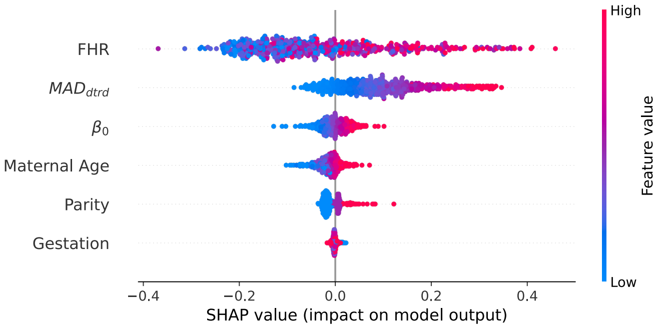

Fig. 7. Beeswarm summary plot illustrating the distribution of SHAP values of tabular features computed for the CTU-UHB dataset indicating the global impact of each feature on the model output.

The positive SHAP values of FHR were generally observed to be associated with high FHR latent space values (3600 FHR values were compressed into a latent space of 128 values and summed to generate a single value, see Method section) while negative SHAP FHR values were observed to be associated with low FHR latent space values. Additionally, high feature values of the MAD_{dtrd}, β_0, Maternal age, and Parity features showed a clear trend of raising the likelihood of the model prediction towards classification of fetal compromise.

TABLE IV
COMPARISON OF THE SELECTED MODEL PERFORMANCE WITH DIFFERENT FUSION OPERATORS.

Fusion Operator	**AUC (MHW-pH, CV)**	**%TPR at 15% FPR (MHW-pH, CV)**	**AUC (CTU-UHB)**	**%TPR at 15% FPR (CTU-UHB)**	**Params (K)**
Multiplication	**0.77 ± 0.03**	50 ± 7	0.84 ± 0.02	67 ± 5	505.5
Addition	0.74 ± 0.04	47 ± 7	0.79 ± 0.08	57 ± 15	505.5
Concatenation	0.75 ± 0.02	53 ± 2	0.85 ± 0.01	70 ± 4	505.7

In addition to SHAP, this study also utilized the gradient-weighted class activation mapping (Grad-CAM) to investigate the FHR time series regions that strongly impact the model output. The Grad-CAMs along with the SHAP local bar plots of four random cases of the CTU-UHB dataset reflecting true positive, false positive, false negative, and true negative predictions of the Fusion ResNet model fused with the top five features are shown in Fig. 8.

As illustrated, the true positive case shows that the most impactful feature for its prediction was the FHR time series and the Grad-CAM shows strong activation in a deceleration region of FHR. Higher values of MAD_{dtrd}, β_0, and maternal age, reflected by positive SHAP values, contributed positively to the likelihood of predicting fetal compromise, whereas being nulliparous, indicated by a negative SHAP value, was associated with a lower likelihood of being classified as a compromised case. In the true negative case, the FHR is still the dominant feature negatively affecting the model prediction, thus reducing the likelihood of fetal compromise. The Grad-CAM also aligns with FHR SHAP values as no activated region corresponding to fetal compromise was observed. Low MAD_{dtrd} values and moderate β_0 showed negative SHAP values, reducing the likelihood of fetal compromise. However, high Gestation showed a negative SHAP value impacting the model output negatively.

For the false positive case, the main contributor to the higher likelihood of fetal compromise was FHR and Grad-CAM showed the decelerative regions of FHR impacted the model strongly. The slightly higher MAD_{dtrd} and slightly higher β_0 also contributed positively, whereas nulliparous and low maternal age slightly impacted the model output negatively. In the false negative case, since no activating regions of FHR were identified by the model and visualized through Grad-CAM, the SHAP values for FHR were negative, minimizing the likelihood of fetal compromise. Nulliparous and low maternal age further help to minimize the likelihood of fetal compromise. Although slightly higher MAD_{dtrd} and β_0 positively impacted the model, they were not strong enough to counter the effect of FHR. Additional Grad-CAMs and SHAP bar plots for cases with extreme tabular features are shown in the supplementary materials.

IV. DISCUSSION

Electronic health record features that capture personalized clinical context along with human-crafted FHR features and signal quality measures have been shown to be predictive of fetal compromise in previous studies. However, these features have not been comprehensively fused with a deep learning based FHR model to assess their combined predictive potential. This work presents an integrated model that achieves state-of-the-art performance in fetal compromise detection while enhancing interpretability, marking a significant step toward clinically applicable and explainable computer-assisted CTG analysis.

This study first investigated twelve tabular features composed of extracted FHR features, EHR features, and signal quality features, ranked using the mRMR feature selection algorithm, for fetal compromise detection using two SVM models. The top five rank features were β_0, Parity, MAD_{dtrd}, Gestation, and Maternal Age based on their maximum relevance and minimum redundancy. This is further confirmed by the minimum correlation shown among these features in the correlation plot shown in Fig. 4b. These features were next combined with the FHR times series using a deep learning model. This Fusion ResNet model trained on a large dataset of 9,887 FHR records, showed a superior AUC performance of 0.84 compared to the SVM models and the current state-of-the-art MCNN model performance in the benchmark CTU-UHB dataset.

The availability of a large training dataset (n=9,887) enabled us to train the proposed model and evaluate its performance on the external CTU-UHB dataset (n = 552), allowing for a fair comparison with previous studies that used the same benchmark. This approach reduces potential bias associated with training and reporting results on small or locally sourced datasets, while also demonstrating the model's robustness across diverse populations. For example, Spilka et al. [33] employed a Sparse-SVM trained on three features (β_0, MAD_{dtrd}, and H) derived from the Lyon database consisting of 1,288 recordings, achieving an AUC of 0.79 on a CTU-UHB subset with less than 50 percent signal loss. Ogasawara et al. [19] trained a CNN using FHR and UC time series data from a private dataset of 2,116 recordings from Keio University Hospital and reported a mean AUC of 0.68 on a CTU-UHB subset with signal loss below 16 percent. McCoy et al. [34] conducted a comparative study of six deep-learning models trained on a private dataset of 10,182 FHR recordings with less than 30 percent signal loss. Their best-performing model, InceptionTime, achieved AUCs of 0.76 and 0.72 on CTU-UHB with and without transfer learning, respectively. Another recent study evaluated a CNN model using a multicenter retrospective database comprising cases from three teaching hospitals of Assistance Publique des Hôpitaux de Paris (APHP), SPaM (Workshop on Signal Processing and Monitoring in Labor), and CTU-UHB, and found that it outperformed transformer-based models. The CNN achieved an AUC of 0.74 (90% CI: 0.67–0.81) on the CTU-UHB dataset when trained on the other

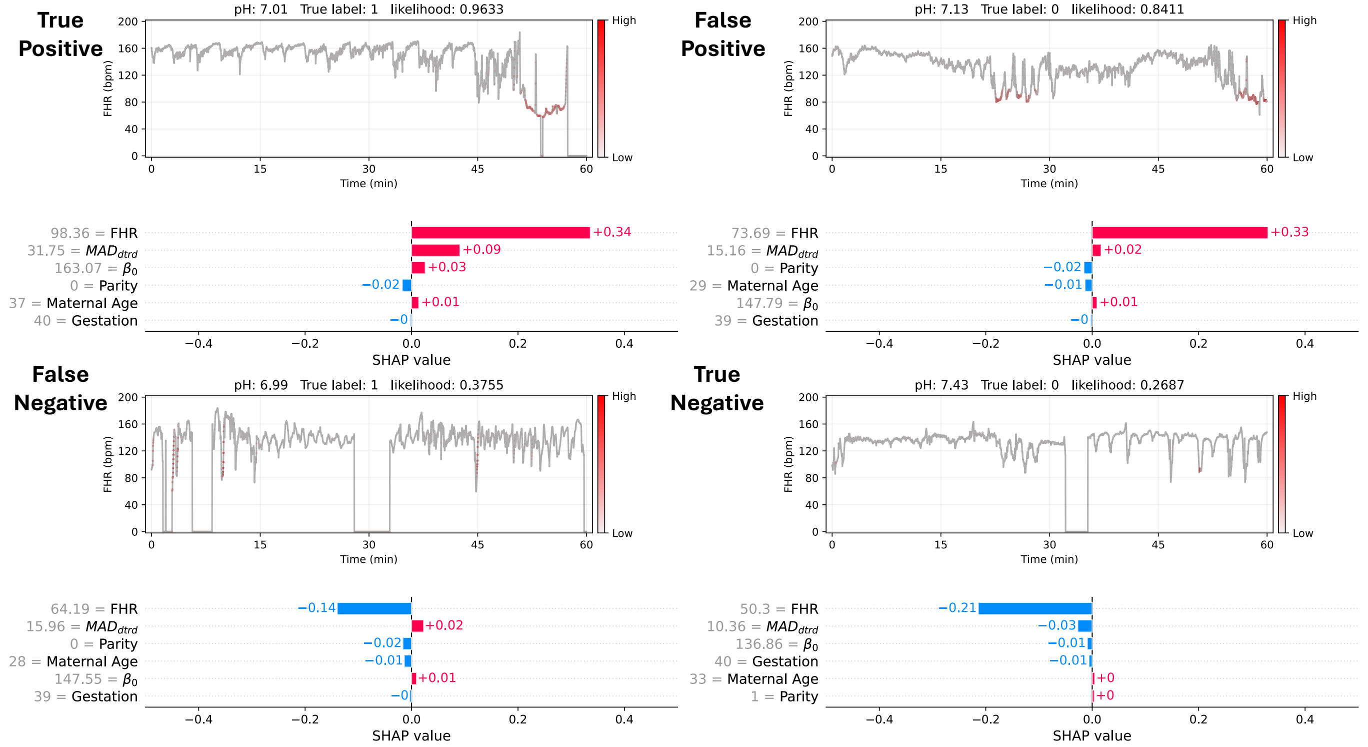

Fig. 8. The Grad-CAM and SHAP bar plots for true positive, false positive, false negative, and true negative cases. The Grad-CAM highlights FHR regions that strongly impact the model output. The SHAP bar plots show the tabular feature contributions to the model output.

four cohorts. Petrozziello et al. [7] proposed two deep learning architectures, the MCNN and a stacked MCNN, which achieved state-of-the-art AUCs of 0.81 and 0.82 respectively on the CTU-UHB. However, the authors noted that MCNN performed better than the stacked MCNN in their internal testing. These models were trained on a dataset of over 35,000 CTG recordings using FHR, UC, and a quality feature vector based on signal loss percentage.

The main limitation of these existing studies is that they do not integrate clinical parameters that are known to adversely affect fetal outcomes as fetal compromise is rare, heterogeneous, and challenging to detect with each fetal injury resulting from a complex, multifactorial interaction [7], [35]. For instance, a study analyzing 2,241 term singleton pregnancies observed that nulliparity was significantly frequent in fetal acidosis cases whereas no relationship was observed with gestational diabetes and hypertension [36]. A prospective database of over 35,000 women from a multicenter investigation of singletons showed a significant association with miscarriage, chromosomal abnormalities, congenital anomalies, gestational diabetes, placenta previa, and cesarean delivery with increasing maternal age, and an increased risk of low birth weight and perinatal mortality in women aged $\geq$ 40 years [37]. Similarly, other retrospective studies have also shown that thick meconium, gestational diabetes, and hypertensive disorders are associated with adverse fetal outcomes [38], [39]. Petrozziello et al. [7] in their study, suggested that combining deep learning with domain-specific knowledge in future studies may improve the detection of fetal compromise.

Only a few studies have explored the integration of clinical features into prediction models for fetal compromise. O'Sullivan et al. [13] incorporated parity, gestation, and hypertension into a logistic regression model alongside CTG features, which improved AUC performance from 0.79 to 0.82 in a 5-fold cross-validation on a subset of CTU-UHB data (n = 333), where at-risk cases were defined by pH $\leq$ 7.0 and Apgar score at 5 minutes $\leq$ 6. Including the durations of Stage 1 and Stage 2 labor further increased the AUC to 0.83 and 0.86, respectively. However, these durations are only available during labor or retrospectively in the case of Stage 2 duration and thus are not suitable for real-time prediction. Similarly, Houze de l'Aulnoit et al. [36] used a logistic regression model combining FHR and clinical features (parity, gestational age, fetal sex, and time from recording to delivery) to predict acidosis (pH $\leq$ 7.15), achieving an AUC of 0.79 in a retrospective analysis of 2,241 singleton deliveries. A more recent study introduced DeepCTG® 1.5 [14], which applied logistic regression to integrate CTG features with clinical data (maternal age, fetal growth restriction, diabetes, hypertension, and meconium), showing an AUC of 0.71 compared to 0.70 when using CTG features alone. This model was trained on a private dataset of 1,264 cases and evaluated on a subset of CTU-UHB (n = 550) using pH $<$ 7.05 as the outcome criterion.

In the present study, the Sparse SVM and Classical SVM performance in multivariate feature analysis reached maximum performance with only the top three features (β_0, Parity, and MAD_{dtrd}) in 5-fold CV on the MHW-pH dataset with both models achieving a mean AUC of 0.69. When more features were added, the Classical SVM performance degraded while the Sparse SVM maintained a relatively similar performance. This

TABLE V

PERFORMANCE COMPARISON OF DEEP LEARNING MODELS IN PRIOR STUDIES EXTERNALLY TESTED ON THE CTU-UHB.

Study	DL Model	Input	Outcome	Training Dataset	Testing Dataset	AUC
Petrozziello et al., 2019 [7]	Stacked MCNN	FHR+UC+Quality	pH<7.05	Last 60 min of Stage 1 and the last 30 min of Stage 2 Oxford at 0.25 Hz	CTU-UHB	0.82
Petrozziello et al., 2019 [7]	MCNN	FHR+UC+Quality	pH<7.05	Last 60 min Oxford at 0.25 Hz	CTU-UHB	0.81
Ogasawara et al., 2021 [19]	CNN	FHR+UC	pH<7.05 or Apgar1<7	Last 30 min Keio data at 1 Hz	CTU-UHB (subset of 78 cases)	0.68 ± 0.03
McCoy et al., 2024 [34]	Inception Time	FHR	pH<7.05	Last 60 min Pennsylvania at 0.25 Hz	CTU-UHB	0.72
Ben M'Barek et al., 2025 [24]	CNN	FHR+UC	pH<7.05	Last 60 min Multicenter APHP + SPaM at 1 Hz	CTU-UHB	0.74 (90% CI: 0.67–0.81)
Mendis et al., 2025 [10]	ResNet	FHR	pH<7.05	Last 60 min MHW-pH at 1 Hz	CTU-UHB	0.81 ± 0.01
This Work	**Fusion ResNet**	**FHR + Top 4 Tabular Features**	**pH<7.05**	**Last 60 min MHW-pH at 1 Hz**	**CTU-UHB**	**0.84 ± 0.02**

can be attributed to the capability of sparse selection of weights and features by the Sparse SVM model, effectively not selecting underperforming features [9]. The trend of the first three features achieving the optimum performance was consistent with the external CTU-UHB evaluation in both models where Sparse SVM achieved an AUC of 0.75 while Classical SVM achieved an AUC of 0.74. Similar findings were reported by Spilka et al. [9], where they found β_0, MAD_{dtrd}, and H as the top three features among 20 features for fetal compromise detection using Sparse-SVM, achieving an AUC of 0.79. The performance differences compared to our SVM models may be attributed to the use of pH $\leq$ 7.05 as the outcome criterion, the focus on the last 20 minutes of the first stage of labor, and the selection of a CTU-UHB subset that excluded recordings with more than 50% signal loss. Additionally, the H feature may not have ranked among the top features in the MHW-pH dataset due to its positive correlation with MAD_{dtrd} and DC_{PRSA}, as all three features capture aspects of FHR variability.

In the deep learning approach of the present study, the top five tabular features were combined with the FHR time series using a Fusion ResNet model for improving fetal compromise detection. The internal evaluation using a 5-fold CV of the proposed Fusion ResNet model on MHW-pH showed the highest mean AUC of 0.77 with only the top four tabular features (β_0, Parity, MAD_{dtrd}, and Gestation) fused with the time series FHR. This highest-performing model, when externally tested with CTU-UHB, achieved a state-of-the-art AUC of 0.84 (74% TPR at 20% FPR, 67% TPR at 15% FPR), outperforming existing prediction models including the MCNN (AUC=0.81) and stacked MCNN (AUC=0.82). Its performance was also superior to typical clinical practice which generally lies between TPR of 31% to 48% at 16% and 21% FPR [7], [8]. A comparison of the performance of existing deep learning approaches and current work is given in Table V. The mean area under the precision-recall curve (AUPRC), mean F1 scores at 5, 10, 15, and 20% FPR of our selected model on CTU-UHB were 0.39, 0.43, 0.40, 0.37, and 0.33, respectively. We refer the reader to the supplementary materials for additional results of AUPRC and F1 scores.

The superior performance of our Fusion ResNet model can be attributed to its ability to integrate domain and clinical knowledge captured in tabular features with complex, non-linear patterns learned from the FHR time series. In clinical practice, these tabular features including EHR features such as Parity and Gestation are readily accessible to support informed decision-making, while FHR features such as β_0 and MAD_{dtrd} representing baseline and variability metrics are consistent with FIGO guidelines for fetal monitoring and clinical assessment. Furthermore, the Fusion ResNet model performance was superior to both SVM models, demonstrating that deep learning outperforms classical machine learning methods for fetal compromise detection as identified by existing works [17], [19]. Another advantage of the Fusion ResNet model is the ability of its FHR time-series branch to process variable input lengths, which enhances its suitability for real-world deployment. This flexibility is achieved through the use of a ResNet architecture followed by a global average pooling layer in the FHR branch [10]. This design also contributes to the model's low computational overhead, primarily achieved through the use of 1D convolutions, a global average pooling layer that eliminates the need for large dense layers, and a lightweight tabular branch consisting of two fully connected layers. The selected model contained 505,587 trainable parameters and required 3.6 GFLOPs per 60-minute FHR segment. The mean inference time per 60-minute segment was 15.33 ms on an NVIDIA A100 GPU and 76.44 ms on a standard Intel Core i7-1185G7 CPU. These architectural choices ensure that the proposed model remains computationally efficient and suitable for real-time deployment in intrapartum fetal monitoring systems.

A further limitation of previous ML and DL studies for FHR evaluation was the limited explainability of the model predictions. Clinicians often express reluctance to adopt predictive ML and DL models due to limited transparency regarding how input features contribute to the model's decision-making process [40], [41]. This lack of interpretability hinders clinical trust and restricts integration into routine practice, particularly in high-stakes settings such as intrapartum care. The present study addressed this limitation by investigating the interpretability of the deep learning model fused with the top five tabular features using two explainable artificial intelligence techniques: Grad-CAM and SHAP. A previous study by our research group [10] used CAM to explain the model predictions, but to the best of our knowledge, this is the first study to use both methods to explain a model used for fetal compromise detection. The global SHAP values shown by the beeswarm plot in Fig. 7 demonstrate that higher feature values of all features except for Gestation contribute highly to the likelihood of fetal compromise. Particularly, this aligns with domain knowledge where high maternal age, high FHR baseline (β_0), and longer depth of decelerations (MAD_{dtrd}) are associated with fetal compromise [4], [33], [37]. Furthermore, the ability of Grad-CAM to identify regions of high importance and SHAP values to show feature importance in explaining individual case predictions of true positive, false positive, true negative, and false negative were illustrated. These visual representations, if made available during the clinical decision-making process, could offer valuable insights that support more informed, reliable, and accurate assessment by clinicians.

Interestingly, both the SVM and deep learning models showed reduced performance when Parity was added to the β_0 feature when tested on the CTU-UHB dataset. This pattern was not observed during the 5-fold cross-validation on the MHW-pH dataset. A possible explanation is the higher proportion of nulliparous compromised cases in the CTU-UHB dataset compared to the training distribution in the MHW-pH dataset, as shown in Fig. 4a. Previous studies have reported a consistent association between nulliparity and fetal compromise [36]. Similarly, significant differences between Normal and Compromised cases were also observed in both MHW-pH and CTU-UHB datasets. However, the low prevalence of compromised cases (1.1%) and the small number of nulliparous cases within the compromised group in the MHW-pH dataset may have limited the model's ability to learn from this feature during training. This is also reflected in the low SHAP values observed for low Parity cases. The negative impact was offset when MAD_{dtrd} was included, demonstrating its strong contribution to model predictions. This is further supported by the higher SHAP values of MAD_{dtrd} compared to other tabular features, as shown in Fig. 7.

Another important observation was that both the SVM models and the Fusion ResNet models performed better on the external testing set CTU-UHB than on the internal validation set MHW-pH. Similar trends have been reported in previous studies that externally evaluated model performance on CTU-UHB [7], [10], [33]. One possible explanation is that the CTU-UHB dataset is smaller and less heterogeneous, with data collected over a period of less than three years and a high prevalence of fetal compromise (7.2%). In contrast, the MHW-pH dataset is approximately 18 times larger, spanning 12 years of clinical practice, and reflects a low prevalence rate of 1.1%. These differences highlight the importance of validating predictive models on large, diverse, and representative datasets. Overreliance on smaller, less heterogeneous cohorts may lead to an overestimation of model performance and limit generalizability in real-world clinical settings.

V. Limitations and Future Work

This study has several limitations. First, the tabular feature set was limited to twelve tabular features, including six EHR features selected based on their availability in both datasets and four human-crafted FHR features identified from previous studies. Because the recordings were not pre-filtered for signal loss, the handcrafted FHR features extracted using classical methods may have been affected. Second, feature selection was limited to the mRMR method, and future work could extend this to compare other feature selection methods, including strategies embedded within neural networks. Third, several of the features used in this study have previously been identified as predictive within the CTU-UHB dataset, as demonstrated in the work by Spilka et al. [33]. Since the proposed model was developed and validated using only the MHW-pH and CTU-UHB datasets, representing cohorts from Australia and the Czech Republic, its generalizability to broader populations needs to be confirmed. Fourth, although the model achieved a TPR of 67-72% at a 15-20% FPR on the external CTU-UHB dataset, surpassing typical clinician performance of 31–48% TPR at 16–21% FPR [7], [8], this trade-off between sensitivity and specificity must be carefully considered for real-world deployment. While higher sensitivity is desirable for timely intervention, further improvement in specificity will reduce unnecessary cesarean sections and alarm fatigue in practice. Fifth, while the SHAP and Grad-CAM explanations demonstrate strong alignment with known physiological patterns, their clinical plausibility warrants validation through expert consultation. Future work will involve collaboration with obstetricians to review interpretability maps and conduct case-based clinical evaluations. Therefore, further validation in more diverse retrospective datasets and prospective clinical trials is needed to assess the generalizability of the proposed model and to determine whether its integration into clinical practice can lead to improved fetal outcomes.

VI. Conclusion

In conclusion, this cross-center study involving over 10,000 FHR recordings demonstrates that integrating domain-specific and clinical features with FHR time series data using deep learning significantly enhances the detection of fetal compromise. The proposed Fusion ResNet model achieved state-of-the-art performance, with an AUC of 0.84 on the open-access CTU-UHB dataset. For the first time, model predictions were interpreted using two explainable artificial intelligence techniques, providing visual insights into the contribution of both tabular features and FHR signal regions to the model's output. These approaches enhance clinical interpretability and may help address concerns among clinicians regarding the perceived lack of transparency in deep learning models, thereby improving their clinical acceptability. The improved predictive performance and interpretability of the proposed method represent a step toward the development of a robust computerized CTG analysis system with the potential to reduce adverse perinatal outcomes.

Acknowledgment

L. Mendis is supported by the Melbourne Research Scholarship and a Top-Up Scholarship from the Graeme Clark Institute for Biomedical Engineering at the University of Melbourne. E. Keenan is supported by an Early Career Research Fellowship from the Department of Obstetrics and Gynaecology at the University of Melbourne. F. Brownfoot is supported by an NHMRC Early Career Fellowship (NHMRC #1142636) and a Norman Beischer Clinical Research Fellowship. D. Karmakar is supported by an NHMRC postgraduate scholarship. This research was supported by The University of Melbourne's Research Computing Services and the Petascale Campus Initiative.

Conflict of Interest

E. Keenan, M. Palaniswami, and F. Brownfoot are co-founders and shareholders of Kali Healthcare Pty Ltd. L. Mendis and D. Karmakar have no competing interests to declare.

References

[1] United Nations Inter-agency Group for Child Mortality Estimation (UN IGME), "Never Forgotten: The situation of stillbirth around the globe," United Nations Children's Fund, New York, Report, 2023.

[2] R. L. Goldenberg, M. S. Harrison, and E. M. McClure, "Stillbirths: The Hidden Birth Asphyxia - US and Global Perspectives," *Clinics in Perinatology*, vol. 43, no. 3, pp. 439–453, Sep. 2016.

[3] A. Pinas and E. Chandraharan, "Continuous cardiotocography during labour: Analysis, classification and management," *Best Practice & Research Clinical Obstetrics & Gynaecology*, vol. 30, pp. 33–47, Jan. 2016.

[4] D. Ayres-de Campos *et al.*, "FIGO consensus guidelines on intrapartum fetal monitoring: Cardiotocography," *International Journal of Gynecology & Obstetrics*, vol. 131, no. 1, pp. 13–24, 2015.

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

[6] Z. Alfirevic *et al.*, "Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour," *Cochrane Database of Systematic Reviews*, no. 2, 2017.

[7] A. Petrozziello *et al.*, "Multimodal Convolutional Neural Networks to Detect Fetal Compromise During Labor and Delivery," *IEEE Access*, vol. 7, pp. 112 026–112 036, 2019.

[8] P. Abry *et al.*, "Sparse learning for Intrapartum fetal heart rate analysis," *Biomedical Physics & Engineering Express*, vol. 4, no. 3, p. 034002, Apr. 2018.

[9] J. Spilka *et al.*, "Sparse Support Vector Machine for Intrapartum Fetal Heart Rate Classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 3, pp. 664–671, May 2017.

[10] L. Mendis *et al.*, "Cross-Database Evaluation of Deep Learning Methods for Intrapartum Cardiotocography Classification," *IEEE Journal of Translational Engineering in Health and Medicine*, pp. 1–1, 2025.

[11] A. Georgieva, C. W. G. Redman, and A. T. Papageorghiou, "Computerized data-driven interpretation of the intrapartum cardiotocogram: a cohort study," *Acta Obstetricia Et Gynecologica Scandinavica*, vol. 96, no. 7, pp. 883–891, Jul. 2017.

[12] L. Mendis *et al.*, "Computerised Cardiotocography Analysis for the Automated Detection of Fetal Compromise during Labour: A Review," *Bioengineering*, vol. 10, no. 9, p. 1007, Sep. 2023.

[13] M. O'Sullivan *et al.*, "Classification of fetal compromise during labour: signal processing and feature engineering of the cardiotocograph," in *2021 29th European Signal Processing Conference (EUSIPCO)*, Dublin, Ireland, Aug. 2021, pp. 1331–1335.

[14] E. Menzhulina *et al.*, "Integration of clinical features in a computerized cardiotocography system to predict severe newborn acidemia," *European Journal of Obstetrics & Gynecology and Reproductive Biology*, vol. 307, pp. 78–83, Apr. 2025.

[15] A. Georgieva *et al.*, "Phase-rectified signal averaging for intrapartum electronic fetal heart rate monitoring is related to acidaemia at birth," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 121, no. 7, pp. 889–894, 2014.

[16] A. Lovers *et al.*, "Advancements in Fetal Heart Rate Monitoring: A Report on Opportunities and Strategic Initiatives for Better Intrapartum Care," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. n/a, no. n/a, 2025.

[17] L. Mendis *et al.*, "Rapid detection of fetal compromise using input length invariant deep learning on fetal heart rate signals," *Scientific Reports*, vol. 14, no. 1, p. 12615, Jun. 2024.

[18] A. Petrozziello *et al.*, "Deep Learning for Continuous Electronic Fetal Monitoring in Labor," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Honolulu, HI, USA, Jul. 2018, pp. 5866–5869.

[19] J. Ogasawara *et al.*, "Deep neural network-based classification of cardiotocograms outperformed conventional algorithms," *Scientific Reports*, vol. 11, no. 1, p. 13367, Jun. 2021.

[20] D. Karmakar *et al.*, "Impact of missing electronic fetal monitoring signals on perinatal asphyxia: a multicohort analysis," *npj Digital Medicine*, vol. 8, no. 1, pp. 1–9, May 2025.

[21] R. Pardasani *et al.*, "Development of a novel artificial intelligence algorithm for interpreting fetal heart rate and uterine activity data in cardiotocography," *Frontiers in Digital Health*, vol. 7, Sep. 2025.

[22] M. J. Khan, M. Vatish, and G. Davis Jones, "PatchCTG: A Patch Cardiotocography Transformer for Antepartum Fetal Health Monitoring," *Sensors*, vol. 25, no. 9, p. 2650, Jan. 2025.

[23] K. E. Gumilar *et al.*, "Artificial intelligence-large language models (AI-LLMs) for reliable and accurate cardiotocography (CTG) interpretation in obstetric practice," *Computational and Structural Biotechnology Journal*, vol. 27, pp. 1140–1147, Mar. 2025.

[24] I. Ben M'Barek *et al.*, "DeepCTG® 2.0: Development and validation of a deep learning model to detect neonatal acidemia from cardiotocography during labor," *Computers in Biology and Medicine*, vol. 184, p. 109448, 2025.

[25] V. Chudáček *et al.*, "Open access intrapartum CTG database," *BMC Pregnancy and Childbirth*, vol. 14, no. 1, p. 16, Jan. 2014.

[26] J. Spilka *et al.*, "Automatic Evaluation of FHR Recordings from CTU-UHB CTG Database," in *Proceedings of the 4th International Conference, ITBAM 2013*, M. Bursa, S. Khuri, and M. E. Renda, Eds. Prague, Czech Republic: Springer, Berlin, Heidelberg, 2013, pp. 47–61.

[27] J. Spilka, "Complex approach to fetal heart rate analysis: A hierarchical classification model," Ph.D dissertation, Czech Technical University in Prague, Faculty of Electrical Engineering, Department of Cybernetics, 2013.

[28] Z. Zhao *et al.*, "DeepFHR: intelligent prediction of fetal Acidemia using fetal heart rate signals based on convolutional neural network," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, p. 286, Dec. 2019.

[29] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," in *Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003*, Aug. 2003, pp. 523–528.

[30] P. A. Warrick *et al.*, "Classification of Normal and Hypoxic Fetuses From Systems Modeling of Intrapartum Cardiotocography," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 771–779, Apr. 2010.

[31] R. R. Selvaraju *et al.*, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, Feb. 2020.

[32] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., Dec. 2017, pp. 4768–4777.

[33] J. Spilka *et al.*, "Intrapartum Fetal Heart Rate Classification: Cross-Database Evaluation," in *Proceedings of the XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016*, ser. IFMBE Proceedings, E. Kyriacou, S. Christofides, and C. S. Pattichis, Eds. Paphos, Cyprus: Springer, Cham, 2016, pp. 1199–1204.

[34] J. A. Mccoy *et al.*, "Intrapartum electronic fetal heart rate monitoring to predict acidemia at birth with the use of deep learning," *American Journal of Obstetrics & Gynecology*, vol. 0, no. 0, Apr. 2024.

[35] J. M. Turner, M. D. Mitchell, and S. S. Kumar, "The physiology of intrapartum fetal compromise at term," *American Journal of Obstetrics & Gynecology*, vol. 222, no. 1, pp. 17–26, Jan. 2020.

[36] A. Houzé de l'Aulnoit *et al.*, "Use of automated fetal heart rate analysis to identify risk factors for umbilical cord acidosis at birth," *Computers in Biology and Medicine*, vol. 115, p. 103525, Dec. 2019.

[37] J. Cleary-Goldman *et al.*, "Impact of Maternal Age on Obstetric Outcome," *Obstetrics & Gynecology*, vol. 105, no. 5 Part 1, p. 983, May 2005.

[38] O. Gluck *et al.*, "The effect of meconium thickness level on neonatal outcome," *Early Human Development*, vol. 142, p. 104953, Mar. 2020.

[39] Y.-W. Lin *et al.*, "Population-based study on birth outcomes among women with hypertensive disorders of pregnancy and gestational diabetes mellitus," *Scientific Reports*, vol. 11, no. 1, p. 17391, Aug. 2021.

[40] R. Dlugatch, A. Georgieva, and A. Kerasidou, "AI-driven decision support systems and epistemic reliance: a qualitative study on obstetricians' and midwives' perspectives on integrating AI-driven CTG into clinical decision making," *BMC Medical Ethics*, vol. 25, no. 1, p. 6, Jan. 2024.

[41] A. Vellido, "The importance of interpretability and visualization in machine learning for applications in medicine and health care," *Neural Computing and Applications*, vol. 32, no. 24, pp. 18 069–18 083, Dec. 2020.