Machine learning on cardiotocography data to classify fetal outcomes: A scoping review

Computers in Biology and Medicine 172 (2024) 108220

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Machine learning on cardiotocography data to classify fetal outcomes: A scoping review

Farah Francis [a,*], Saturnino Luz [a], Honghan Wu [b], Sarah J. Stock [a], Rosemary Townsend [a]

[a] Usher Institute, University of Edinburgh, UK
[b] Institute of Health Informatics, University College London, UK

ARTICLE INFO

Keywords:
Machine learning
Cardiotocography
Fetal hypoxia
Signal processing
Obstetrics

ABSTRACT

Introduction: Uterine contractions during labour constrict maternal blood flow and oxygen delivery to the developing baby, causing transient hypoxia. While most babies are physiologically adapted to withstand such intrapartum hypoxia, those exposed to severe hypoxia or with poor physiological reserves may experience neurological injury or death during labour. Cardiotocography (CTG) monitoring was developed to identify babies at risk of hypoxia by detecting changes in fetal heart rate (FHR) patterns. CTG monitoring is in widespread use in intrapartum care for the detection of fetal hypoxia, but the clinical utility is limited by a relatively poor positive predictive value (PPV) of an abnormal CTG and significant inter and intra observer variability in CTG interpretation. Clinical risk and human factors may impact the quality of CTG interpretation. Misclassification of CTG traces may lead to both under-treatment (with the risk of fetal injury or death) or over-treatment (which may include unnecessary operative interventions that put both mother and baby at risk of complications).
Machine learning (ML) has been applied to this problem since early 2000 and has shown potential to predict fetal hypoxia more accurately than visual interpretation of CTG alone. To consider how these tools might be translated for clinical practice, we conducted a review of ML techniques already applied to CTG classification and identified research gaps requiring investigation in order to progress towards clinical implementation.
Materials and method: We used identified keywords to search databases for relevant publications on PubMed, EMBASE and IEEE Xplore. We used Preferred Reporting Items for Systematic Review and Meta-Analysis for Scoping Reviews (PRISMA-ScR). Title, abstract and full text were screened according to the inclusion criteria.
Results: We included 36 studies that used signal processing and ML techniques to classify CTG. Most studies used an open-access CTG database and predominantly used fetal metabolic acidosis as the benchmark for hypoxia with varying pH levels. Various methods were used to process and extract CTG signals and several ML algorithms were used to classify CTG. We identified significant concerns over the practicality of using varying pH levels as the CTG classification benchmark. Furthermore, studies needed to be more generalised as most used the same database with a low number of subjects for an ML study.
Conclusion: ML studies demonstrate potential in predicting fetal hypoxia from CTG. However, more diverse datasets, standardisation of hypoxia benchmarks and enhancement of algorithms and features are needed for future clinical implementation.

1. Introduction

Fetal hypoxia occurs when there is a lack of oxygen supply to the baby during labour. Fetal hypoxic injury can cause a wide range of devastating damage, such as intrapartum stillbirth, asphyxia, neonatal encephalopathy, neonatal death and neurodevelopmental impairment [1–5]. In European hospitals, the overall incidence of fetal hypoxia ranges from 0.06% to 2.8% [6]. Globally, it is estimated that intrapartum fetal hypoxia leads to approximately 1.3 million stillbirths during childbirth, 0.9 million neonatal deaths, and 0.6 to 1 million instances of long-term disability resulting from neonatal hypoxic-ischemic encephalopathy every year [7,8]. Therefore, this issue should be addressed to prevent further cases of hypoxia.

A degree of hypoxic stress can be anticipated during labour when uterine contractions (UC) may impair maternal perfusion of the placenta, thus compromising oxygen delivery to the fetus. The primary

* Corresponding author.NINE. 9 Little France Road, Edinburgh BioQuarter, Edinburgh, EH16 5XP, UK.
E-mail address: farah.francis@ed.ac.uk (F. Francis).

https://doi.org/10.1016/j.compbiomed.2024.108220
Received 9 June 2023; Received in revised form 2 February 2024; Accepted 25 February 2024
Available online 7 March 2024

challenge clinicians face is identifying that small number of babies where the natural physiological protective mechanisms fail to compensate for the hypoxic stress of labour, contributing to significant cerebral injury [9]. Fetal monitoring during labour is crucial to prevent the devastating effects of fetal hypoxia on babies and families. It must also be sufficiently discriminatory to minimise unnecessary iatrogenic interventions in the form of surgical birth (caesarean section) that carry their own risks to both mother and baby [10].

Cardiotocography (CTG) is an electronic fetal monitoring tool commonly used to detect fetal hypoxia in the womb [11]. CTG records two measurements: fetal heart rate (FHR) and UC [12]. In standard clinical use, CTG recordings are visually interpreted by clinicians using these criteria: baseline, variability, accelerations and decelerations. Based on these indicators, clinicians will categorise the recording; in the UK the NICE guidelines use the classifiers reassuring, non-reassuring and pathological [13–15]. Depending on the classification, clinicians may take steps to ameliorate hypoxia (including change in maternal position, intravenous fluids and tocolysis to reduce contractions) or expedite birth via assisted vaginal birth or caesarean section, to reduce the adverse effect on new-borns while ensuring the safety of the mother. CTG is widely used in maternity care in most high-resource settings and is usually limited to women with existing risk factors for fetal hypoxia [16,17].

Since CTG was first introduced in the 1950s without randomised clinical trials, the interobserver variation in visual interpretation of CTG amongst clinicians has been consistently shown to result in a delayed response due to the time taken to achieve an agreement [18–23]. Furthermore, some decision-making can be subjective and with some level of ambiguity which may contribute to discrepancies in CTG interpretation [24].

Multiple studies have reported increased caesarean section rates where CTG is used. At least some of these interventions are presumed to have been unnecessary - there has been a fivefold increase in caesarean sections, while cerebral palsy prevalence remains the same [25]. Conversely, false-negative cases occur when a fetus is misidentified as normal (healthy), resulting in birth injury with damaging results for newborns and families [2,26]. Obstetrics and gynaecology account for 50% of the total value of clinical negligence claims in the UK (£2.3 billion), with CTG interpretation alone or in combination with mismanagement of labour and cerebral palsy contributing to a large proportion of claims [27,28]. This co-existence of over and under-intervention demonstrates that CTG, as currently implemented, is neither sensitive enough to reliably detect fetal hypoxia nor specific enough to avoid unnecessary surgical birth.

One approach to tackle the shortcomings of visual CTG has been the introduction of computerised CTG analysis, theoretically standardising interpretations and allowing a quicker response to compromised fetuses. A randomised controlled trial and retrospective studies have shown that computerised CTG decision support could improve the quality of interpretations while minimising decision-making time [29–31]. A subsequent meta-analysis of six studies showed no significant improvement in fetal well-being between visual and computerised CTG [12]. This highlights the limitations of computerised analysis based on existing visual interpretation parameters and also the contextual barriers to implementing effective novel fetal monitoring strategies [32].

Machine learning (ML) is a discipline of artificial intelligence that has received increasing attention within medical research due to its potential to support decision-making for clinicians. ML can learn and identify patterns from collected data and make predictions [33,34]. Researchers who used ML on CTG data have demonstrated promising results in predicting and classifying FHR. Existing studies have reported a variety of ML models and inconsistent selections of FHR features using a range of clinical endpoints - this is likely to have contributed to variable ML performance.

We aimed to summarise the evidence relating to signal processing, feature extraction and ML techniques used on CTG data to detect fetal hypoxia. We mapped research gaps within ML approaches to CTG interpretation and identified actions that can improve the robustness and practicality of clinical implementation of computerised CTG. We have adopted a scoping review approach, rather than the more traditional systematic review, in order to summarise a high-level overview of this field with the technical and implementation challenges for the application of ML to analyse the CTG [35].

2. Methods

2.1. Search strategy

A full detailed protocol of this study was published in Open Science Framework (https://osf.io/) – Machine Learning for Cardiotocography Data to Classify Fetal Outcome [36]. In brief, we conducted this review based on Arkesy and O'Malley [37], Levac et al. [38] and the Joanna Briggs Institute recommendations [39]. We conducted a comprehensive search on electronic databases such as PubMed, EMBASE and IEEE Xplorefor relevant literature published from January 1, 2000 to September 1, 2023. The search strategy MeSH keywords for this study were support vector machine, random forest, k-nearest neighbours, extreme gradient boosting, decision trees, naive Bayes, artificial neural network, convoluted neural network, adaptive boosting, linear discriminant analysis, machine learning, artificial Intelligence, deep learning, genetic algorithm, cardiotocography and fetal heart rate. No restriction was applied during the search and the keyword combinations are listed in supplementary. The search results were imported to Covidence (https://www.covidence.org/), where duplicate references were removed.

2.2. Inclusion and exclusion criteria

The inclusion criteria for this scoping review were: 1) studies analysed raw intrapartum CTG during labour using ML techniques, 2) studies written in English (reviewers are not familiar with any other languages), 3) studies published from the year 2000 onwards, 4) journal articles and 5) fetus without intrauterine growth restriction (IUGR). The exclusion criteria for this study are: 1) studies that analysed antenatal (prelabour) CTG, 2) animal studies, 3) studies that did not use ML techniques, 4) studies that analysed electrocardiogram data and combinations of data with CTG, 5) studies that did not specify if CTG recordings are during intrapartum or antenatal, 6) studies with less than twenty subjects (low data sample will contribute to model overfitting, leading to a biased model), 7) conference papers, 8) studies with fetal IUGR only and 9) studies without clinical endpoint for classifying CTG, in order to ensure that the results of the included studies are clinically applicable (Table 1).

2.3. Study selection

The first reviewer (FF) identified relevant publications based on the title and abstract according to the inclusion and exclusion criteria. Next, full-text screening was conducted to refine inclusion and exclusion criteria further. A second reviewer (RT) reviewed 10% of the articles to ensure they met eligibility criteria. The discrepancy between reviewers on article eligibility was discussed to reach an agreement.

2.4. Data extraction

Eligible studies that meet the inclusion criteria went through data charting where results from articles were summarised based on the author(s), publication year, authors' background, data set and type of data used and geographical regions of the primary institution. We extracted the following study characteristics: number of participants for both normal and abnormal FHR, maternal risk factors/condition, the standard used for outcome definition, model classifiers, features used,

Table 1
Summarises the inclusion and exclusion criteria of the included studies.

Inclusion Criteria	Exclusion Criteria
1) studies analysed raw intrapartum CTG during labour using ML techniques	1) studies that analysed antenatal (prelabour)
2) studies written in English (reviewers are not familiar with any other languages)	2) animal studies
3) studies published from the year 2000 onwards	3) studies that did not use ML techniques
4) journal articles	4) studies that analysed electrocardiogram data and combinations of data with CTG
5) fetus without intrauterine growth restriction (IUGR)	5) studies that did not specify if CTG recordings are during intrapartum or antenatal
	6) studies with less than twenty subjects (low data sample will contribute to model overfitting, leading to a biased model)
	7) conference papers
	8) studies with fetal IUGR only
	9) studies without clinical endpoint for classifying CTG

number of features, type of features, validation methods, data pre-processing techniques, model interpretability and performance measures. Information regarding the risk of bias was also extracted. Five articles were used as a pilot to test if the form identified all the information relevant to the research questions. One reviewer (FF) extracted all the information and experts (SS, RT, HW, SL) were consulted when there were uncertainties.

2.5. Quality assessment

The quality of the study design for selected studies was evaluated using the Recommendations for Reporting Machine Learning Analyses in Clinical Research [40]. Studies were categorised as 1) sufficient (missing one element or none), 2) medium (2 missing elements) and 3) insufficient (more than two missing elements).

2.6. Collating, summarising and reporting the results

Results from the charting were analysed according to the questions predetermined from the protocol mentioned above. Narrative (descriptive) summaries were provided on a qualitative attribute, such as ML classifiers used on the CTG dataset. Quantifications were carried out on numerical data, such as the number of participants involved in this study and the number of classes for ML model development. The results from the literature search, screening processes and study selections were reported in a flow diagram based on the PRISMA extension for scoping review (Fig. 1) [41].

3. Results

3.1. Studies background

We included 36 studies with 9923 women in this scoping review that met the inclusion and exclusion criteria (Fig. 1). Most of the first authors were from China, as shown in Fig. 2. We found that 22 (61.1%) publications used the same open-access dataset from the Czech Technical University in Prague and the University of Brno (CTU-CHB), four (12.9%) studies used a mixture of CTU-CHB and in-house dataset and 10 (27.8%) use in-house only (Table 2) [42]. Various 'gold standards' were used as surrogate measures for fetal hypoxia, as illustrated in Fig. 3 and Table 2, where the majority used metabolic acidosis taken from the blood cord (n = 26). In addition, none of the studies incorporated clinical factors that could contribute to abnormal CTG readings. Based on Table 2, studies used various numbers of positive cases (hypoxic) where some of them had the same number of both normal and cases, while some had an imbalanced distribution where the cases were much less than normal (Fig. 4).

We assessed the quality of the included studies based on the guidelines mentioned above. None of the studies were good as all at least have two missing reporting elements. 11 studies were insufficient and 25 were medium (Table 2). None of the studies calibrated their model or published their code for reproducibility. We found that studies extracted CTG at different stages of labour - the CTU-CHB database provided the last hour of the second stage of labour. One study used the first stage of labour, one at the end of labour and others did not provide this information in their paper.

3.2. Signal processing

As mentioned above, while CTG comprises FHR and UC, some researchers chose to analyse FHR only (n = 26), while others chose FHR in combination with UC (n = 10).

3.2.1. Pre-processing

Studies adopted various techniques and methods to pre-process CTG signals which involves removing artefacts and noise that can reduce signal quality. Mother and fetal movements can cause interference during recording. Most researchers removed spikes and interpolated missing data points. Some researchers filtered out frequencies more than 0.5Hz because they were considered noise. A detailed method used for each study is illustrated in Table 3. Various methods of interpolation, mainly linear and cubic spline methods, were used for missing recordings.

3.2.2. Feature extraction

3.2.2.1. Handcrafted features. The summary of features used by studies is in Table 3. Studies predominantly used the International Federation of Gynaecology and Obstetrics guidelines (FIGO) as a guide to extract FHR [14] (acceleration, deceleration, baseline and variability) which are also known as morphological features. Some studies also used features defined by the Royal College of Obstetricians and Gynaecologists (RCOG) [79] and the National Institute for Health and Care Excellence (NICE) [79]. In addition, studies used other signal-based features such as time series, frequency domain, linear, and non-linear. Although most studies primarily used FHR signals for ML modelling, studies like Petrozziello et al [63] and Ben M'Barek et al. (2023) included UC signals as a characteristic for categorising fetal hypoxia. There is an inconsistent number and type of feature used between studies. Some justifications include using what was used in previous studies, while others did not report any reasoning for selecting specific features. However, all studies included morphological features outlined by FIGO.

3.2.2.2. Image-based features. Papers using convoluted neural network (CNN) techniques did not extract handcrafted features. Instead, they used images of the CTG, either the FHR or both FHR and UC. Studies pre-processed raw CTG before feeding the image into CNN to remove noise such as abrupt change and interpolate missing values. Some studies split the whole CTG of each subject into smaller frames or windows to generate multiple samples. For example, studies by Deng et al. [52] and Liang et al. (2022) used the same open-source database and split the CTG into several frames to increase the sample size (Table 2).

3.3. ML model development

3.3.1. Pre-processing and feature engineering

Most of the datasets used by studies are imbalanced, with more normal numbers than abnormal CTG. Nine studies used the Synthetic Minority Oversampling and one used the Adaptive Synthetic Sampling

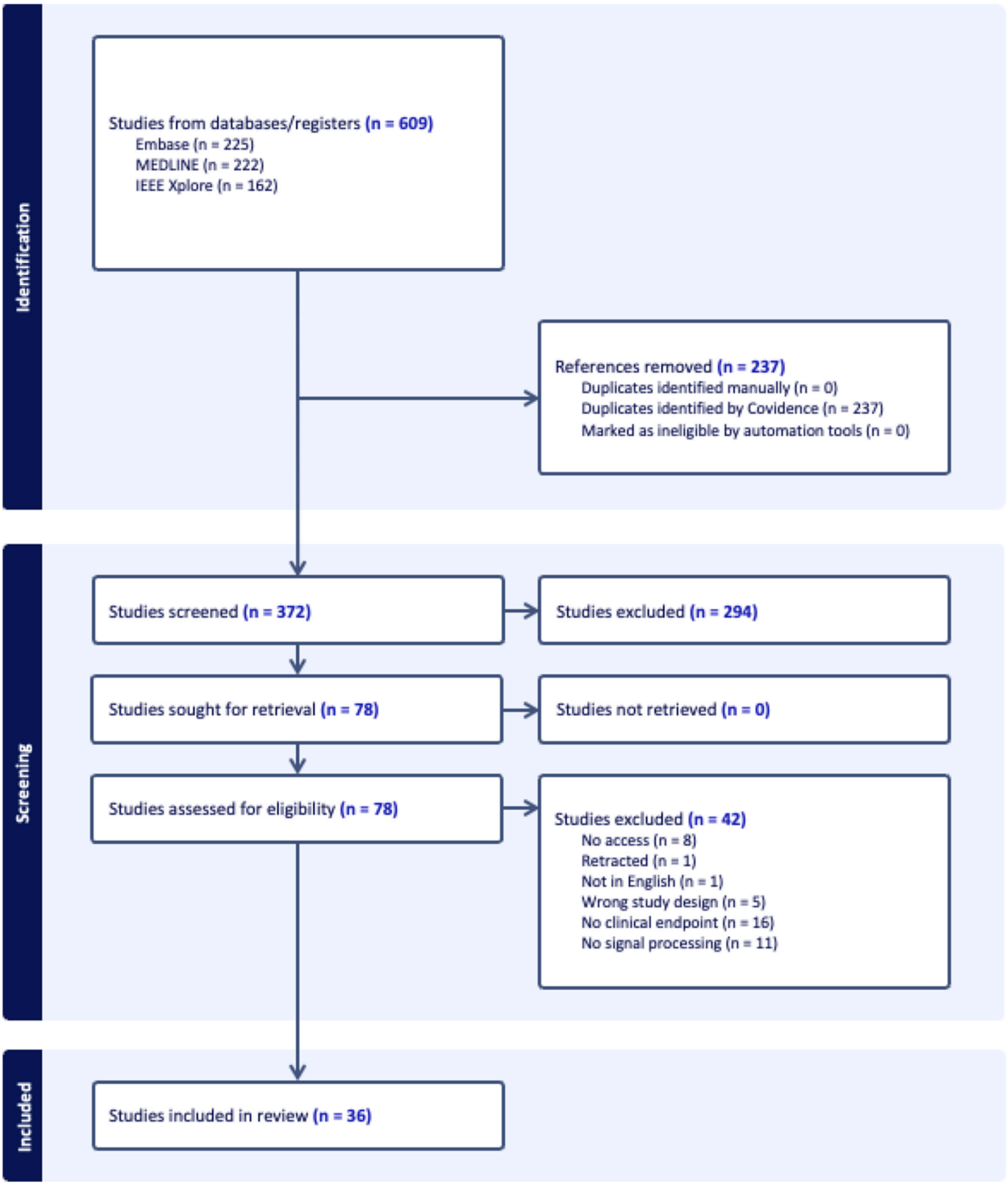

Fig. 1. Illustrates a PRISMA flowchart of the processes of identifying relevant studies in this review.

Method to synthetically increase the number of cases to the same level as the normal (Table 3). The remaining studies also randomly selected a balanced number of normal and cases from resampling, where they split the CTG into several segments (Table 3) [73]. However, some studies did not mention if the dataset was augmented or if modelling was carried out on an imbalanced dataset. This could have an effect on the training model and generalisability. For example, if the training set has more cases, the model will have more training on the cases and might perform poorly when identifying the cases.

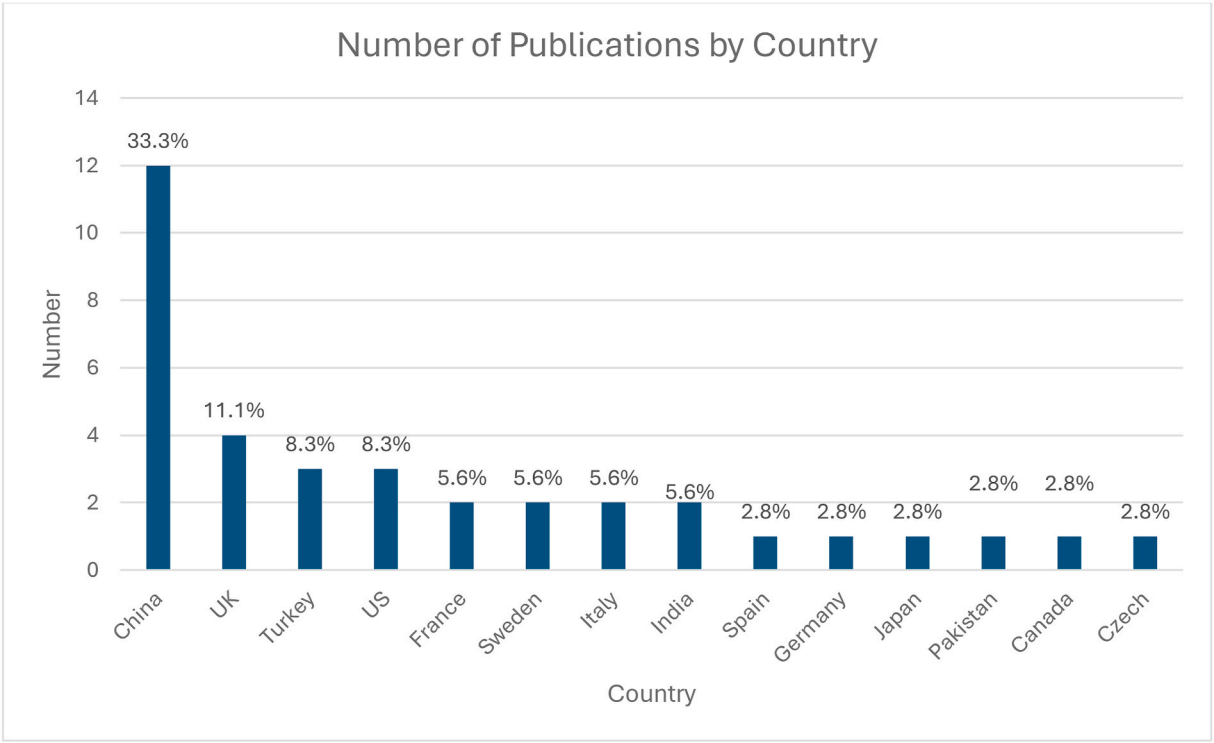

Fig. 2. Shows the number of publications by the first author's country of institution.

3.3.2. Classifiers

All studies used supervised learning to classify abnormal CTG and clustering. Fig. 5 shows the distribution of classifiers used by included studies. Various types of deep learning (neural networks) were also used. We found that most studies used support vector machines (SVM). Other methods used include decision tree (DT), random forest (RT), k-nearest neighbour (KNN), CNN, hierarchical Dirichlet process Gaussian mixture model (HDPGM), naïve Bayes (NB), linear discriminant analysis (LDA), Fisher LDA (FLDA), generative model (NB and first order Markov-chain and maximum *a posteriori* decision-based method), LSTM, artificial neural network (ANN) (also known as multi-level perceptron), ADABOOST and XGBOOST. Some studies used several algorithms and compared classification performance between different algorithms, while others used only one. Interestingly, three studies combined several classifiers in their modelling, such as FLDA with RF and SVM, FLDA with RF only, FLDA with SVM only and RF with SVM only [54]. Ensemble methods used were the Ensemble Cost-sensitive SVM (ECSVM) [71], ADABOOST [64,65], DECORATE and XGBOOST [64]. Details of classifiers used for each study are summarised in Table 3.

3.3.3. Model validation

20 studies split their datasets into training and testing sets (hold-out validation) for internal validation, while others did not specify the methods used for validation. The internal validation method used was cross-validation (CV), including predominantly k-fold, followed by stratified k-fold, single loop and double loop nested CV. 14 studies split and cross-validated their dataset. We found 16 studies that either did not perform internal validation or failed to mention it in their publications. Only four studies used an external dataset to validate their model performance [43,46,63,69]. Details of internal validation used are summarised in Table 3.

3.3.4. Performance measures

Most studies used specificity, sensitivity, and area under the receiving operating characteristics curve (AUROC). Other performance measures include quality metric, accuracy, quality measure, F-measure, weight relative accuracy (WRA), balance error rate (BER), geometric-mean, percentage of correct diagnosis, positive predictive value, negative predicted value, mean square error (MSE), quality metric, quality index, false positive rate (FPR), true positive, false positive and Matthew's correlation coefficient. The number of studies that used each performance metric is quantified and illustrated in a bar chart in Fig. 6 below, and details of the performance measures used are summarised in Table 3. Based on Table 3, we plotted the top three common performance metrics to compare the relationship between the number of positive cases and model performances. Lollipop plots were plotted to illustrate the pooled results achieved by studies using sensitivity, Fig. 8 illustrates specificity, Fig. 9 illustrates AUROC and Fig. 10 illustrates accuracy.

4. Discussion

This descriptive scoping review draws on 36 studies to summarise signal processing and ML techniques applied to CTG data, identify gaps in current studies, and guide future research directions. Compared to a recent review of ML for intrapartum CTG classification by O'Sullivan et al. (2021), our study systematically searched for relevant literature and screened them based on eligibility criteria to ensure an unbiased selection of pooled studies. We also included a detailed summary of CTG signal pre-processing methods, the type of signal features extracted and a more detailed summary of ML techniques and results than another previous review [80]. Our initial search identified a large number of manuscripts reporting the use of machine learning in analysing CTG data. Because we are primarily interested in moving to clinical translation, we included only those studies that linked their machine learning to clinical evidence of fetal hypoxia.

4.1. Study characteristics

CTGs are used globally, and it is essential for CTGs to provide an accurate classification. Most studies (61.1%) used the same open-access database, highlighting an urgent need for further open-access CTG

Table 2
Shows the characteristics of datasets and gold standards used by each study. Studies used variable thresholds when determining metabolic acidosis pH levels, ranging from pH less than 7.00 to pH less than 7.20 defined as low pH. The table also shows the proportion number of subject in cases and normal, country of subjects, source of dataset and the quality of each study.

Study ID	Fetal outcome surrogate measure for cases	Data Source	Country of participants	Number of cases	Number of normal	Quality
[43]	Metabolic acidosis (pH < 7.05)	CTU-CHB and in house	France and Czech Republic	In house-56 CTU-CHB -26	France- 1756 Czech-446	Insufficient
[44]	Metabolic acidosis (pH < 7.15)	CTU-CHB	Czech Republic	105	447	Medium
[45]	Metabolic acidosis (pH ≤ 7.05) or low Apgar or resuscitation required	In-house	Spain	17	15	Insufficient
[46]	Metabolic acidosis (pH < 7.05,7.15)	CTU-CHB, external, in house	Czech Republic, France and UK	142	1387	Insufficient
[47]	Metabolic acidosis (pH < 7.20)	CTU-CHB	Czech Republic	177	375	Medium
[48]	Metabolic acidosis (pH ≤ 7.15)	CTU-CHB	Czech Republic	113	439	Medium
[49]	Metabolic acidosis (pH < 7.2)	CTU-CHB	Czech Republic	177	375	Medium
[50]	Metabolic acidosis (pH < 7.15)	CTU-CHB	Czech Republic	Did not specify	Did not specify	Insufficient
[51]	Metabolic acidosis (pH < 7.15)	In-house	US	24	60	Medium
[52]	Metabolic acidosis (pH < 7.05)	CTU-CHB	Czech Republic	44 (106 segments)	508 (106 segments)	Medium
[53]	Type of delivery (vaginal or caesarean)	CTU-CHB	Czech Republic	46	506	Medium
[54]	Type of delivery (vaginal or caesarean)	CTU-CHB	Czech Republic	46	506	Insufficient
[55]	Type of delivery (vaginal or caesarean)	CTU-CHB	Czech Republic	46	506	Medium
[56]	Type of delivery (vaginal or caesarean)	CTU-CHB	Czech Republic	18	354	Insufficient
[57]	Metabolic acidosis (pH < 7.1)	In house	Portugal	20	60	Medium
[58]	Metabolic acidosis (pH < 7.05)	CTU-CHB	Czech Republic	46	508	Medium
[59]	Metabolic acidosis (pH < 7.2)	In-house	US	92	92	Medium
[60]	Metabolic acidosis (pH ≤ 7.15)	CTU-CHB	Czech Republic	105 (2369 segments)	447 (2369 segments)	Medium
[61]	Apgar and FHR deceleration	In-house	China	581	52	Medium
[62]	Other: pH lower than 7.20 or Apgar score lower than 7 at 1 min	In-house	Japan	162	162	Medium
[63]	Metabolic acidosis (pH ≤ 7.05)	CTU-CHB, in-house and Lyon (external)	UK, France and Czech Republic	In-house-1065 External-100	In-house-4249 External-752	Medium
[64]	Type of delivery (vaginal or caesarean)	In-house	Italy	Did not specify	Did not specify	Insufficient
[65]	Type of delivery (vaginal or caesarean)	CTU-CHB	Czech Republic	46	506	Insufficient
[66]	Metabolic acidosis (pH < 7.05)	In-house	France	31	1251	Medium
[67]	Fetal distress: Pathological CTG trace, meconium stain fluid, admission to neonatal intensive care unit and prevalence of fetal distress in labour in the population	In-house	Italy	42	260	Insufficient
[68]	Base deficit: 12 mmol/L and death or evidence of hypoxic-ischemic encephalopathy	In-house	Canada	26	187	Medium
[69]	Metabolic acidosis (pH < 7.15)	CTU-CHB and in -house	Czech Republic	113	439	Medium
[70]	Metabolic acidosis (pH = 7.2, 7.05 and 7.1)	CTU-CHB	Czech Republic	pH 7.05 = 88, pH 7.1 = 122	unclear	Medium
[71]	Type of delivery (vaginal or caesarean)	CTU-CHB	Czech Republic	27	442	Medium
[72]	Metabolic acidosis (pH < 7.05)	CTU-CHB	Czech Republic	40	386	Insufficient
[73]	Metabolic acidosis (pH < 7.05)	CTU-CHB	Czech Republic	40 but oversample to 300	300	Medium
[74]	Metabolic acidosis (pH < 7.15)	CTU-CHB	Czech Republic	105	447	Medium
[75]	Metabolic acidosis (pH = 7.15)	CTU-CHB	Czech Republic	105	477	Medium
[76]	Metabolic acidosis (pH < 7.15)	CTU-CHB	Czech Republic	105	105	Medium
[77]	Metabolic acidosis (pH < 7.15)	CTU-CHB	Czech Republic	Did not specify	Did not specify	Insufficient
[78]	Metabolic acidosis (pH = 7.15)	CTU-CHB	Czech Republic	113	439	Medium

datasets to facilitate external validation and test the generalisability of models developed in different settings. As most studies used the open access dataset without external validation, the model may overfit the population in the Czech Republic, specifically patients that go to the hospitals in Prague and Brno. This can result in bias in performance when applied to other populations. Out of 36 publications, none of the studies adhered to the reporting guidelines was concerning where studies did not fully report their experimental design or adhere to the best practices for predictive modelling, including validating the model on an external dataset based on the Recommendations for Reporting Machine Learning Analyses in Clinical Research to ensure high-quality reporting and reproducibility [40]. As none of those studies shared their codes, researchers should be encouraged to share codes to help increase transparency between studies and facilitate work that builds. This can help future studies replicate pre-existing work, ensuring reproducibility in research and improving existing techniques. Since none of the studies publish their codes, CTG studies are not reproducible and impede clinical implementation in advancing CTG studies. Furthermore, none of the studies calibrated their model. Therefore, they cannot be interested as true probability. The medium to low quality of studies affects the overall generalisability of this field as it undermines the reliability of results achieved and the relevance to real-world scenarios.

Due to the nature of the field, there are fewer cases of hypoxia compared to normal, which causes an imbalance (Fig. 4). Training on imbalanced data has a profound impact on the performance of models by limiting learning experienced by minority classes and causing bias in classification models. A few studies synthetically increased their case

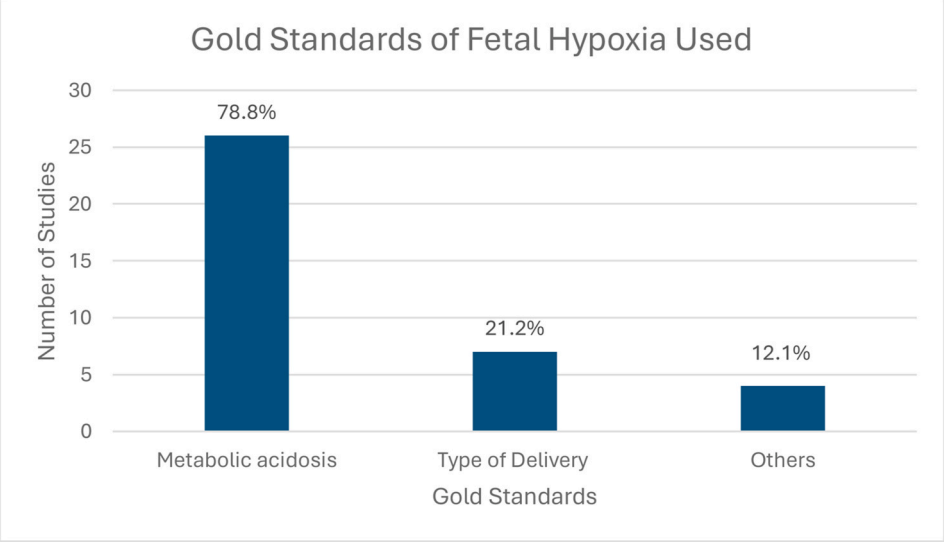

Fig. 3. Shows the number of studies that use different gold standards for fetal hypoxia.

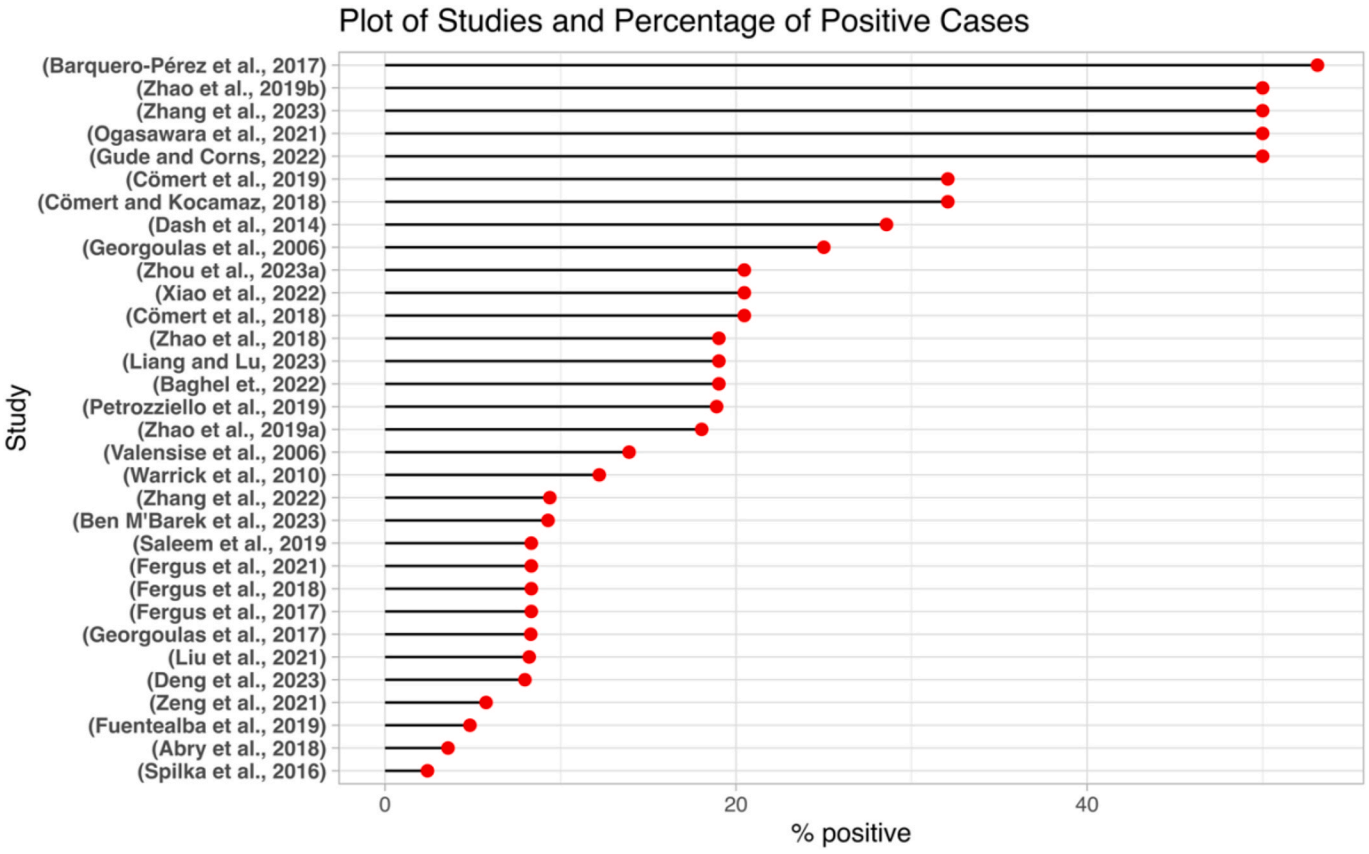

Fig. 4. Shows a dot plot of the percentage of cases in each study.

sample to match the number of normal groups, while others did not. Oversampling produces better performances for prediction models as there are more training data [58,64]. However, synthetically oversampling may not represent real-life events and clinical applications. In addition, a study demonstrated that the minority group that was synthetically oversampled had a high error rate in generating samples where some of the data generated belonged to the majority group [81]. Moreover, synthetically balancing data may cause an overfitting and may create noisy instances [82,83]. Therefore, future studies should use different methods to overcome imbalanced data. Some studies potentially used imbalanced data or failed to report this information. Using imbalanced data may influence poor and biased performance for minority classes due to less training data for cases of hypoxia.

4.2. Surrogate outcomes

We identified several clinical endpoints used to diagnose fetal hypoxia at both (Table 1). 26 studies used metabolic acidosis identified by

Table 3
Describes the pre-processing methods, features extracted by studies, which part of CTG analysed, internal validation method used, classifiers and performance measures used by each study. The performance measures in the table are the highest achieved by classifiers used. Performance is recorded as percentage with (95% Confidence interval (Cl)). Not every study reported 95%Cl. Oversampling columns identify studies that used oversampling technique. LSTM stands for long short-term memory, ADABOOST stands for adaptive boosting, XGBOOST stands for extreme gradient boosting and DECORATE stands for Diverse Ensemble Creation by Oppositional Relabelling of Artificial Training Examples.

Study ID	Pre-processing method	Type of features extracted	Part of CTG used	Clinicians as author (s)	Hold-out-validation	Cross validation method	ML classifier(s) used	Model interpretability	Performance measure(s)	Oversampling
[43]	Sliding median filter and linear spline interpolation	FIGO, spectral and scale-free	FHR	Both	Yes	Single loop and double loop CV	Sparse-SVM	Partially	SLCV: Sensitivity = 72.0% Specificity = 67.0% Balance error rate = 70.0%	No
[44]	Gaussian Butterworth filtering high cut 20Hz and low 200Hz	Image	FHR	Non-clinician	Yes	k-fold	1D-CNN	No	Accuracy = 99.99% AUROC = 91.90% F1 = 98.68%	No
[45]	Dividing signal into a set of short sliding windows (5-min segments), denoise, and linear interpolation	Time, moments and frequency domain	FHR	Non-clinician	No	Leave-one-out	KNN and SVM	Partially	KNN: Accuracy for frequency domain = 77% Accuracy for moments = 88% Accuracy for time domain = 70%	No
[46]	Linear interpolation	FHR FIGO features and UC	FHR and UC	Non-clinician	Yes	k-fold	LG	Yes	LG: AUROC = 74%	No
[47]	Cubic Hermite interpolation	morphological, time, frequency, wavelet transform, statistical analysis, nonlinear	FHR	Non-clinician	No	k-fold	ANN, SVM and kNN	No	SVM: Accuracy = 77.81% Sensitivity = 76.83% Specificity = 78.27% Geometric mean = 77.29% F measure = 68.48% AUROC = 84%	Yes
[48]	Outlier detection, cubic spine interpolation and segmenting	morphological, time, frequency	FHR	Non-clinician	No	None	SVM	Partially	Accuracy = 65.41%, Sensitivity = 63.45% Specificity = 65.88% Quality Index (QI) = 64.04% F measure = 42.17%	No
[49]	Segment selection, outlier detection and interpolation	morphological, linear, nonlinear and frequency domain	FHR	Non-clinician	No	k-fold	SVM, ANN, KNN, DT	Partially	Weighted SVM Accuracy = 88.58% Sensitivity = 77.40% Specificity = 93.8% QI = 85.23% F measure = 81.30%	No
[50]	Did not specify	morphological	FHR and UC	Non-clinician	No	k-fold	MLP, bagging, RF ands SVM	Partially	RF Sensitivity = 96.4% Specificity = 98.4% Accuracy = 96.7% Precision = 96.8%	Yes
[51]	Segmentation feature discretization	NICHD features with heart rate variability and	FHR and UC	Non-clinician	Yes	stratified k-fold	Naive bayes GM and first order markov chain	Partially	Naive bayes GM: Specificity = 82.0%	No

(*continued on next page*)

Table 3 (*continued*)

Study ID	Pre-processing method	Type of features extracted	Part of CTG used	Clinicians as author (s)	Hold-out-validation	Cross validation method	ML classifier(s) used	Model interpretability	Performance measure(s)	Oversampling
		non-linear features					GM compared to SVM		Sensitivity = 61.0%	
[52]	Remove spikes, interpolate, and segment into 20 min	Wavelet packet decomposition image	FHR	Non-clinician	No	k-fold	2DCNN	No	CNN: Accuracy = 95.24% Sensitivity = 90.4% Specificity = 100%	No
[53]	Signal recordings were filtered using a 6th-order low-pass Butterworth filter. Noise and missing values were removed using cubic Hermite spline interpolation	FIGO and NICE, morphological, time series, frequency domain, non-linear	FHR	Non-clinician	No	K-fold	RF, deep learning and fishers linear discriminant analysis	No	Deep learning Sensitivity = 94.0% Specificity = 91.0% AUROC = 99.0% F measure = 100.0% mean squared error (MSE) = 1.0%	Yes
[54]	Filtered using a finite impulse response 6th order high pass filter and cubic Hermite spline interpolation	FIGO, Accelerations, decelerations and non-linear features	FHR	Non-clinician	Yes	k-fold	Fishers Linear Discriminant Analysis, RF, SVM and combinations of classifiers: FLDA_FR_SVM, FLDA_RF, FLDA_SVM and RF_SVM	Partially	Ensemble classifier: FLDA_RF_SVM-sensitivity = 87.0% (95% Cl: 86.0%, 88.0%), Specificity = 90.0% (95% CI: 89.0%, 91.0%), AUROC = 96.0% (95% CI: 96.0%, 97.0%) and MSE = 9.0% (95% CI: 9.0%, 10.0%)	Yes
[55]	Cubic Hermite spline interpolation	Image	FHR	Non-clinician	Yes	Did not specify	1DCNN, RF and SVM	No	Window 200: Sensitivity = 80.0% Specificity = 79.0% AUROC = 86.0%	Yes
[56]	Denoising, baseline estimation, floating line computation, signal detrending and signal decomposition	Time domain, frequency domain, non-linear and time-variant	FHR	Non-clinician	No	K-fold	SVM, LDA and KNN	Partially	SVM: Quality metric Including modal spectral = 81.7%	Yes
[57]	Removing noise	Time domain, frequency domain and morphological features	FHR	Non-clinician	Yes	None	SVM and KNN	Partially	SVM: Geometric mean = 77.1% Accuracy = 78.8% AUROC = 78.0%	No
[58]	Artifact rejection and Hermite spline interpolation	FIGO based, time domain, frequency domain and non-linear domain	FHR	Non-clinician	Yes	Stratified k-fold	least square SVM	Partially	1-Balance Error Rate (BER) = 73.1% Geometric mean = 72.9% F measure = 25.2% Matthew's correlation coefficient (MCC) = 28.5%	Yes
[59]	Smoothing	morphological and statistical	FHR and UC	Non-clinician	No	k-fold	NN, RF, clustering and SVM	Partially	Ensemble combination-NN, RF, k-means and SVM: Accuracy = 92.30%	No

(*continued on next page*)

Table 3 (*continued*)

Study ID	Pre-processing method	Type of features extracted	Part of CTG used	Clinicians as author (s)	Hold-out-validation	Cross validation method	ML classifier(s) used	Model interpretability	Performance measure(s)	Oversampling
[60]	Processing outliers and removing spike using moving average	Image	FHR and UC	Non-clinician	Yes	Did not specify	1D-CNN and bidirectional Gate Recurrent Unit (BiGRU)	No	Accuracy = 95.15% Sensitivity = 96.20% Specificity = 94.09%, Precision = 94.21% F measure = 95.20% AUROC = 99.29%	No
[61]	Did not specify	Image	FHR	Non-clinician	Yes	Did not specify	CNN	No	AUROC = 72.3% Sensitivity = 52.8%	No
[62]	Denoising, smoothing, Hilbert transform and peak detection steps	Number of accelerations and the total area of the decelerations in each case were calculated	FHR and UC	Both	No	k-fold	CNN with 3 convolution layers & LSTM	No	CNN: F measure = 67%	No
[63]	Abrupt increases and decreases were removed and missing values were linearly interpolated. The signals were then averaged down to 0.25Hz	Image	FHR and UC	Both	Yes	k-fold	Multimodal CNN and stacked MCNN	No	Tested on external dataset MCNN: 14% FPR = 58.0% (95% CI: 53.0%–60.0%) 35%FPR = 80.0% (95% CI: 75.0%–85.0%) Stacked MCNN: 14%FPR = 55.0% (95% CI: 53.0%–60.0%) and 35%FPR = 83.0% (95% CI: 75.0%–88.0%)	No
[64]	Cubic spline interpolation	FIGO, time domain, frequency domain and non-linear features	FHR	Non-clinician	No	k-fold	DT, ADABOOST, RF, Gradient boosted tree and DECORATE	Yes	RF: Accuracy = 91.1%Precision = 90.0% Sensitivity = 92.2% and AUROC = 96.7%	Yes
[65]	Sliding mean, 6th order Butterworth filter with 0.5Hz as cut-off	Non-linear and non-stationary (time variant)	FHR	Non-clinician	Yes	Did not specify	DT, SVM with Gaussian kernel and AdaBoost with 20 weak learner (assemble classifiers)	Partially	ADABOOST: Sensitivity = 91.8%, Specificity = 95.5%, AUROC = 98.0% MSE = 5.0%	Yes
[66]	Sliding median and cubic spline interpolation	FIGO, frequency domain and time domain and frequency (multiscale multifractal analysis)	FHR and UC	Both	No	Did not specify	SVM and sparse SVM	No	Sparse SVM with DLCV: Sensitivity = 73.0%, Specificity = 75.0%, AUROC = 77.0%, TP=N:26 and FP=N:305	No
[67]	Did not specify	RCOG with adaptations	FHR		Yes	Did not specify	ANN	No	AUROC = 62.0%, Sensitivity = 56.0% Specificity = 91.0% PPV = 53.0% NPV = 92.0% Accuracy = 86.0%	No

(*continued on next page*)

Table 3 (*continued*)

Study ID	Pre-processing method	Type of features extracted	Part of CTG used	Clinicians as author (s)	Hold-out-validation	Cross validation method	ML classifier(s) used	Model interpretability	Performance measure(s)	Oversampling
[68]	Interpolate interruptions that is less than 15s, remove segments containing longer interruptions and segmenting signals	Frequency domain	FHR and UC	Both	Yes	k-fold	SVM with gaussian kernel	No	AUROC = 13.1% FPR = 7%	No
[69]	Outlier detection and linear interpolation	linear and nonlinear, extract feature using CNN & LSTM	FHR	Non-clinician	No	k-fold	SVM and CNN-BiLSTM	Partially	SVM: Sensitivity = 56.97% Specificity = 73.35% QI = 63.91%	No
[70]	Piecewise cubic Hermite polynomial interpolation	time frequency	FHR	Non-clinician	Yes	k-fold	hierarchical Dirichlet process gaussian model	Partially	HDPGMs sensitivity = 65.0%, specificity = 86.7%, WRA = 51.7%	No
[71]	Signal quality interpolation	Time frequency and linear features	FHR and UC	Non-clinician	No	Did not specify	Ensemble Cost-sensitive SVM (ECSVM), DT, NB and SVM	Partially	Ensemble Cost-sensitive SVM: AUROC = 77%	No
[72]	Did not specify	Image based and text	FHR	Non-clinician	Yes	Stratified k-fold	CNN	No	MMIF-1 (ViT-B/16): Accuracy = 96.3% F measure = 96.3% AUROC = 96.2%	No
[73]	Did not specify	Image	FHR	Non-clinician	Yes	k-fold	KNN, NB, SVM, DT, RF, ADABOOST, XGBOOST	No	XGBOOST: Accuracy = 96.3% Precision = 95.4% Recall = 97.3% F measure = 96.4% AUROC = 95.9%	No
[74]	Denoising, remove spike and spline interpolation	Morphological, time domain, frequency domain and non-linear features	FHR	Non-clinician	Yes	k-fold	SVM, DT and ADABOOST	Partially	ADABOOST: Accuracy = 92.0%, Sensitivity = 92.0%, Specificity = 92.0%, AUROC = 91.0%	No
[75]	Cubic spline interpolation	Image	FHR		No	k-fold	CNN	No	Accuracy = 98.34%, Sensitivity = 98.22%, Specificity = 94.87%, Quality index = 96.53%, AUROC = 97.82%	No
[76]	Gap detection, interpolation, outlier detection and detrending.	Image	FHR	Non-clinician	Yes	k-fold	CNN	No	Accuracy = 98.36%, Sensitivity = 99.05% Specificity = 97.67% AUROC = 98.36%	No
[77]	Lagrange interpolation	Image	FHR	Non-clinician	No	Did not specify	Double Trend Accumulation Former CNN	No	Accuracy = 90.6%	No
[78]	Lagrange interpolation	curve classification	FHR	Non-clinician	Yes	k-fold	Trend-Guided Long CNN	No	Accuracy = 89.80%	No

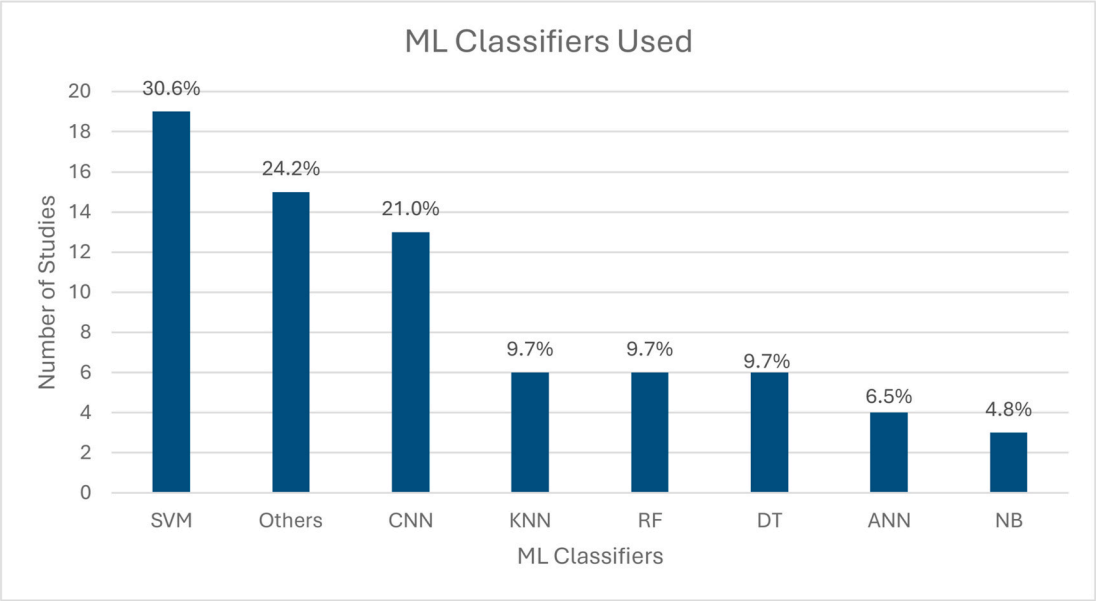

Fig. 5. Shows a bar chart of studies using different ML classifiers to classify CTG. The bar for others is grouped classifiers with only 1 study using them. Ensemble classifiers are considered as others.

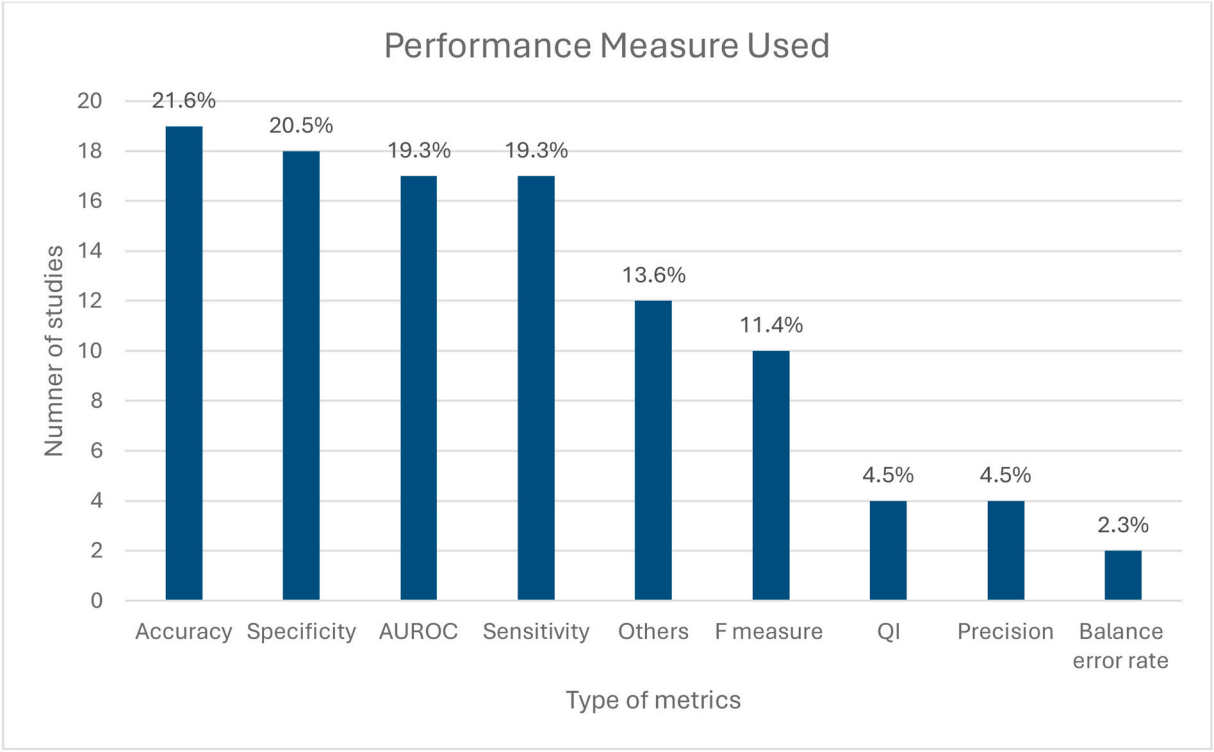

Fig. 6. Shows the performance measures used by studies in this review. The bar for others is grouped performance measures with only one study using them.

the umbilical cord pH after birth. A wide range of cut-off values were used, limiting our ability to compare model performance directly. In addition, several studies used a much higher pH cut-off point than is likely to be clinically relevant - for example, pH < 7.2. Poor neonatal and childhood outcomes are most strongly associated with pH values < 7.0 [84–86]. Models developed using inappropriate surrogate biomarkers to classify the CTG would be likely to result in a high false positive rate in clinical practice. This may result in unnecessary intervention, exposing mothers and newborns to the risk of unnecessary caesarean section.

Other surrogate outcomes for intrapartum fetal hypoxia identified in the included studies were types of delivery, a mixture of Apgar scores and pH and multiple definitions of fetal distress (Table 2). When comparing studies that used similar open-access datasets, the number of subjects in each class (normal and cases) differed in almost every study. The variations in pH values used as benchmarks contribute to differences in the proportion of hypoxic cases (see Table 2), thereby influencing the performance of ML models. Consequently, direct comparisons of results between studies cannot be undertaken.

4.3. Signal processing

The pre-processing stage is of utmost importance in the extraction and interpretation of signals. It is anticipated that a range of approaches would have been employed, as these depend on the inherent characteristics of the raw CTG signal and are subject to the researcher's subjective choices. As evidenced in Table 3, studies that utilised the same open-access database varied in their pre-processing techniques. For instance, studies employed different interpolation methods to fill in missing data, such as linear or cubic spline interpolation, demonstrating the diversity of techniques employed. While these interpolation methods are commonly used, each has its own set of advantages and disadvantages, necessitating researchers to carefully select appropriate methods based on the quality of the CTG signal. For example, linear interpolation can be clearly visualised, but the curve is not smooth and inaccurate for non-linear trend data. On the other hand, the cubic spline produces a smooth curve, but it may introduce oscillations between points for uneven data. Research suggests that data that has been improperly pre-processed may perform worse than the original data [87]. Consequently, signals suffer some degree of loss of information that could be useful for distinguishing hypoxia, which may lead to variability and misinterpretation in the results achieved [88].

4.4. Feature extraction

Our review demonstrated that studies used different parts of CTG to build detection models. Some studies used FHR alone, whereas others used a combination of FHR and UC. Using FHR only is of particular concern in the intrapartum (in labour) context because the relationship between uterine contractions and FHR patterns is the key to fetal physiological resilience to labour. The FHR will often decelerate in response to UC; this is a normal physiological response and typically recovers rapidly [13,90]. During visual interpretation, clinicians use UC as a reference and assess the speed of recovery and depth of the deceleration to determine the probability of fetal hypoxia. As most studies only used FHR for modelling, it challenges the clinical validity of those studies. It is likely that the features extracted in ML models relate to these visually evident features, but ML may be able to identify early and subtle changes more reliably than clinician assessors. Clinicians also consider other factors that can cause FHR changes that do not necessarily indicate that the fetus is experiencing hypoxia, such as maternal movement, umbilical cord compression, and environmental factors [89,90]. Hence, when extracting hand-crafted features, studies should consider other factors that can cause FHR pattern changes, particularly UC. One included study produced an open-access software called CTC-OAS for processing CTG signals. Still, this software excludes UC data and therefore, any CTG research using the software is likely to be clinically irrelevant in the intrapartum setting [91].

Most studies extracted hand-crafted features according to the medical guidelines, also known as morphological features [14]. Additional features commonly extracted were from the time series domain, frequency domain, linear, non-linear and statistical features. These types of features are the generic features relating to the signal processing technique. Our results revealed that a majority of the studies utilised features that were consistent with the clinical guidelines, specifically the morphological changes of CTG. However, these studies did not adequately explain why they incorporated additional features from the time, frequency, and non-linear domains into their model. Without employing feature selection techniques, it is impossible to determine the significance of each feature in classifying CTG. Those CTG characteristics that are not visible by the naked eye might help improve the prediction of fetal hypoxia beyond visual interpretation. Most studies extracted some similar but varied feature types, which may explain the inconsistency in ML modelling performances even using the same open-access dataset. Only one investigated the effects on ML performances of different feature sets using several feature selection methods, where the important feature set varies based on the algorithm used. Only one study combined raw CTG data with other modalities, combining images of CTG with images of the text of the CTG interpretation [72].

4.5. Classifiers

Studies used a variety of classifiers, SVM is the most used due to its versatility, separating linear and non-linear classification (Fig. 3). SVM also works well with high-dimensionality data and it has different ways to interpret SVM outcomes [92–95]. The next most frequently used classifiers are the CNN, either by itself or combined with other classifiers such as LSTM [62,69]. Recently, more CTG-ML studies have used CNN, particularly studies from China, suggesting a growing interest in modelling fetal hypoxia using images, indicating that researchers are exploring different techniques to advance CTG research. Other commonly used classifiers are KNN, DT and RF. The advantages of those classifiers are that they are more explainable and instinctive, making it straightforward for non-AI professionals to understand how decisions were made, which is highly desirable in healthcare settings [96]. Studies mainly reported several classifiers used to build CTG models. This is a good practice to compare how different classifiers distinguish decision boundaries of the same dataset, where some classifiers perform better than the rest. However, this can be limited by recourse and cost of using multiple classifiers. The drawback of only using one classifier is that there might be other methods that work better for the CTG dataset. Studies should also choose appropriate classifiers for modelling CTG, as there are algorithms that can handle class imbalances better, such as ensemble learning, if authors did not perform any data augmentations [97,98].

Clinical interpretability is becoming essential and future studies should consider this when developing prediction models. Based on Table 3, most studies used non-interpretable classifiers when building their model. We regard models such as decision trees, logistic regression, linear regression and naïve Bayes as inherently interpretable, random forest and support vector machines are partially interpretable and deep learning models are not interpretable. This categorisation is partly based on Molnar's [99]. Only one study by Zhang et al [73] attempted to explain the feature outputs using the Shapely Additive exPlanations (widely known as SHAP) algorithm to explain features used in their models to increase the interpretability. SHAP uses the game theoretic approach that aims to elucidate the output of any machine learning model. By employing optimal credit allocation and local explanations, it incorporates the classical Shapley values from game theory and their associated extensions [100,101]. Researchers ought to consider the trade-off between the complexity and interpretability of models when aiming to improve fetal hypoxia detection. Although complex models may perform better, employing interpretability techniques to elucidate clinical decision-making could be advantageous. Nonetheless, using complex models for intricate data may prove beneficial in achieving high performance.

4.6. Performances

Specificity was used most, followed by AUROC and sensitivity to evaluate model performances. Sensitivity, specificity, negative predictive values and positive predictive values are the most clinically relevant measures of model performance. AUROC evaluates performance across all thresholds, including clinically relevant and clinically irrelevant ones [102]. Diagnostic and prognostic tests are generally conceptualised when describing gains and losses to specific patients where AUROC lacks, which makes it clinically challenging to interpret results where many healthcare practitioners have limited ML or statistical knowledge [102–105]. Several studies used accuracy to measure their classification model, which is unsuitable for an imbalanced model where it generates misleading high results caused by systematically predicting the majority class [106]. Trevethan [102] suggested that high positive predictive

values (minimum false-positive outcomes) and high negative predictive values (minimum false-negative outcomes) are preferable in healthcare settings. In addition [106], recommends using balanced accuracy to reduce classification error. The most outstanding performance was attained by a CNN algorithm employing image-based features, as reported by Zhou et al. (2023): an accuracy of 98%, a sensitivity of 99.05%, a specificity of 97.67%, and AUROC of 98.36%. As for non-image features, the highest performance was demonstrated in a study by Fergus et al. [53] using deep learning, using the CTU-UHB database to classify CTG based on delivery type: a sensitivity of 94.0%, a specificity of 91.0%, an AUROC of 99.0%, an F measure of 100.0%, and an MSE of 1.0%. Notably, this study exhibited a lower quantity of cases compared to others, yet it achieved commendable results. However, to enhance the classification of CTG, this study employed synthetic oversampling to augment the number of cases. Based on the plots in Figs. 4 and 7–9, the results achieved are expected for study with an equal number of cases and normal where the model has matched training numbers. A very high model performance produced by studies could indicate overfitting if the studies are not externally validated. Another factor contributing to overfitting is balancing the distribution using oversampling techniques, which can lead to overfitting because the model learns from the same example as other drawbacks previously mentioned. For the imbalanced model, Although these model performances are encouraging, it is important to note that the medical guidelines for visual interpretation features were integrated with signal-processing features during model building. Therefore, we cannot directly compare the performances of existing studies with features proposed by clinical guidelines for interpreting CTG. Moreover, previous research has employed non-interpretable classifiers for modelling, which may impede clinical implementation due to the difficulty clinicians face in interpreting the decisions made by the model.

4.7. Strengths and limitations

The biggest strength of this review is that we summarised the various techniques for automatically predicting fetal hypoxia using CTG during labour. We provided a concise summary of the dataset, methods used to pre-process raw CTG, feature extraction, surrogate outcome selection, ML modelling and performance measures. We emphasised how different techniques used by studies resulted in various model performances, even when using the same dataset.

This review is limited to English publications. Since ML in healthcare is expanding and CTG data are publicly available, we would aim to include CTG-ML research reported in other languages in future work. The SVM classifier may be over-represented in our review as it was included in our keyword search strategy, but we have mitigated this by careful hand-searching and a comprehensive search strategy. We observed that keywords such as 'machine learning' and 'artificial intelligence' did not capture all ML-CTG publications and therefore included specific ML classifiers as keywords in our search strategy to maximise our sensitivity.

4.8. Future research

We demonstrated a lack of the gold standard for identifying fetal hypoxia as many studies used surrogate markers that were clinically irrelevant. An agreed common outcome would make more CTG-ML studies more consistent. From the results, there is also a need to establish guidelines for processing and interpreting CTG-ML for research purposes. No studies took clinical factors into account, but clinicians are well aware that events in labour, uterine activity and maternal pre-existing risk factors are key to accurately identifying fetuses experiencing intrapartum hypoxia because these factors modify the fetal reserves and response to hypoxic insult. Researchers must clearly understand the relationship between FHR and UC to develop clinically

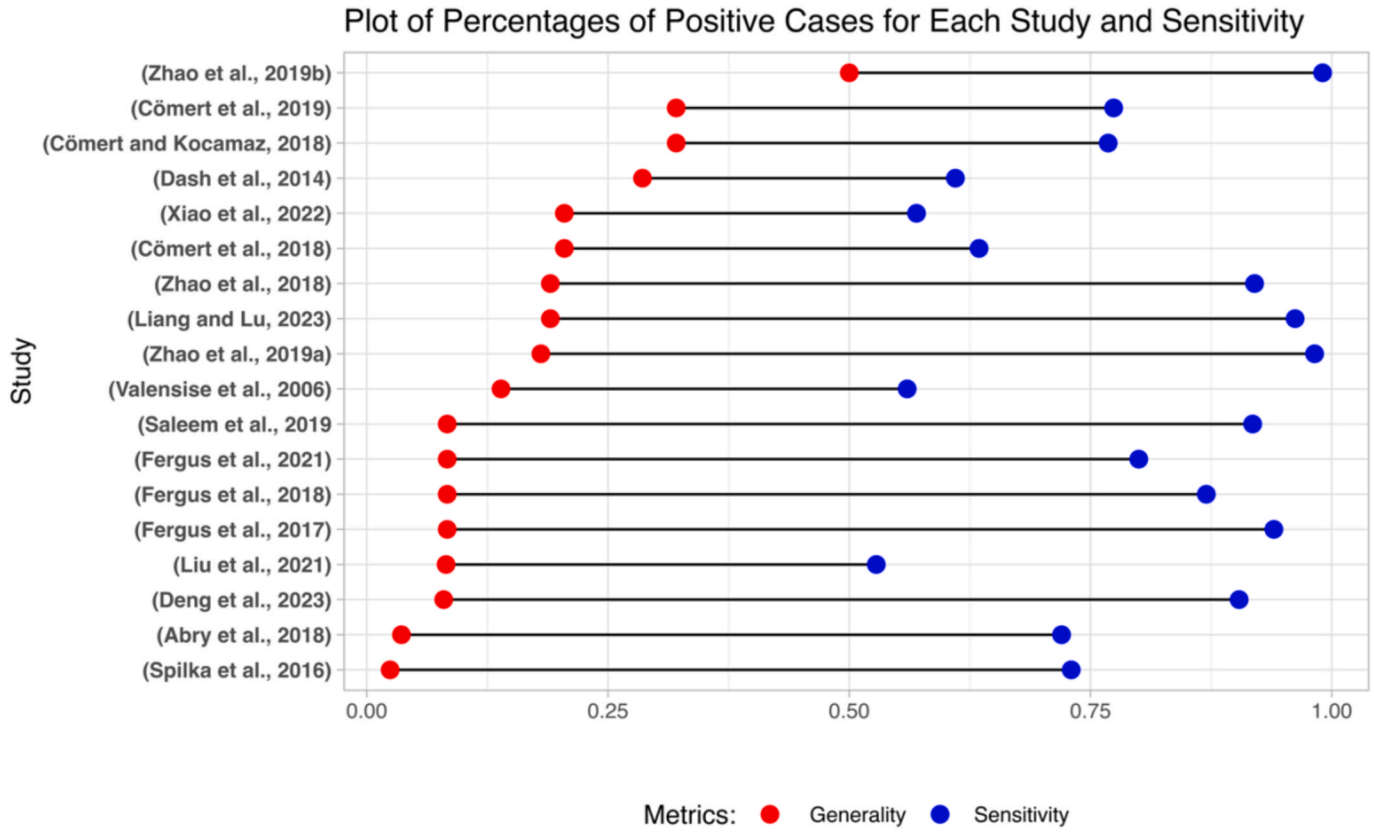

Fig. 7. Lollipop graph showing the relationship between the percentage of hypoxic cases in a dataset and the sensitivity of models.

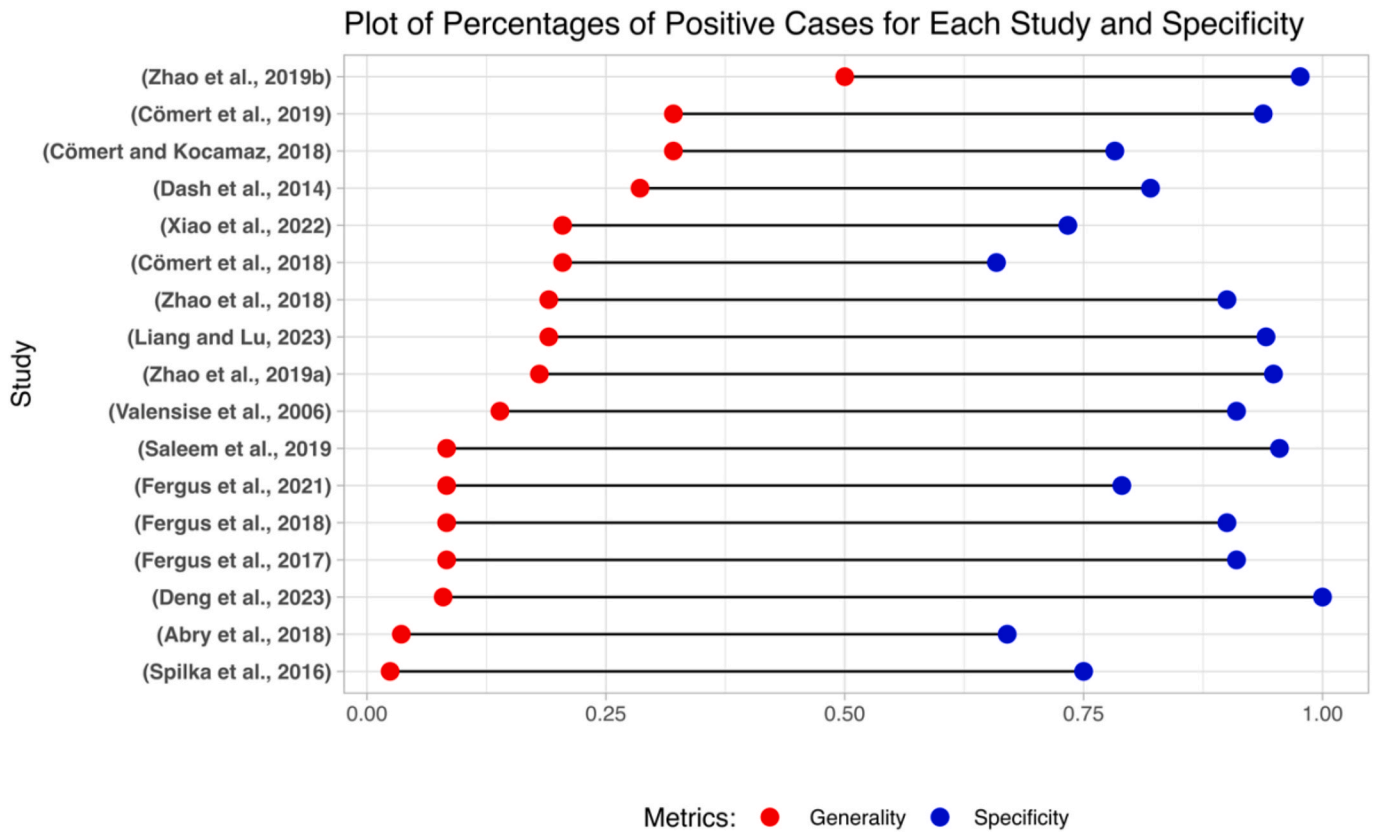

Fig. 8. Lollipop graph showing the relationship between the percentage of hypoxic cases in a dataset and the specificity of models.

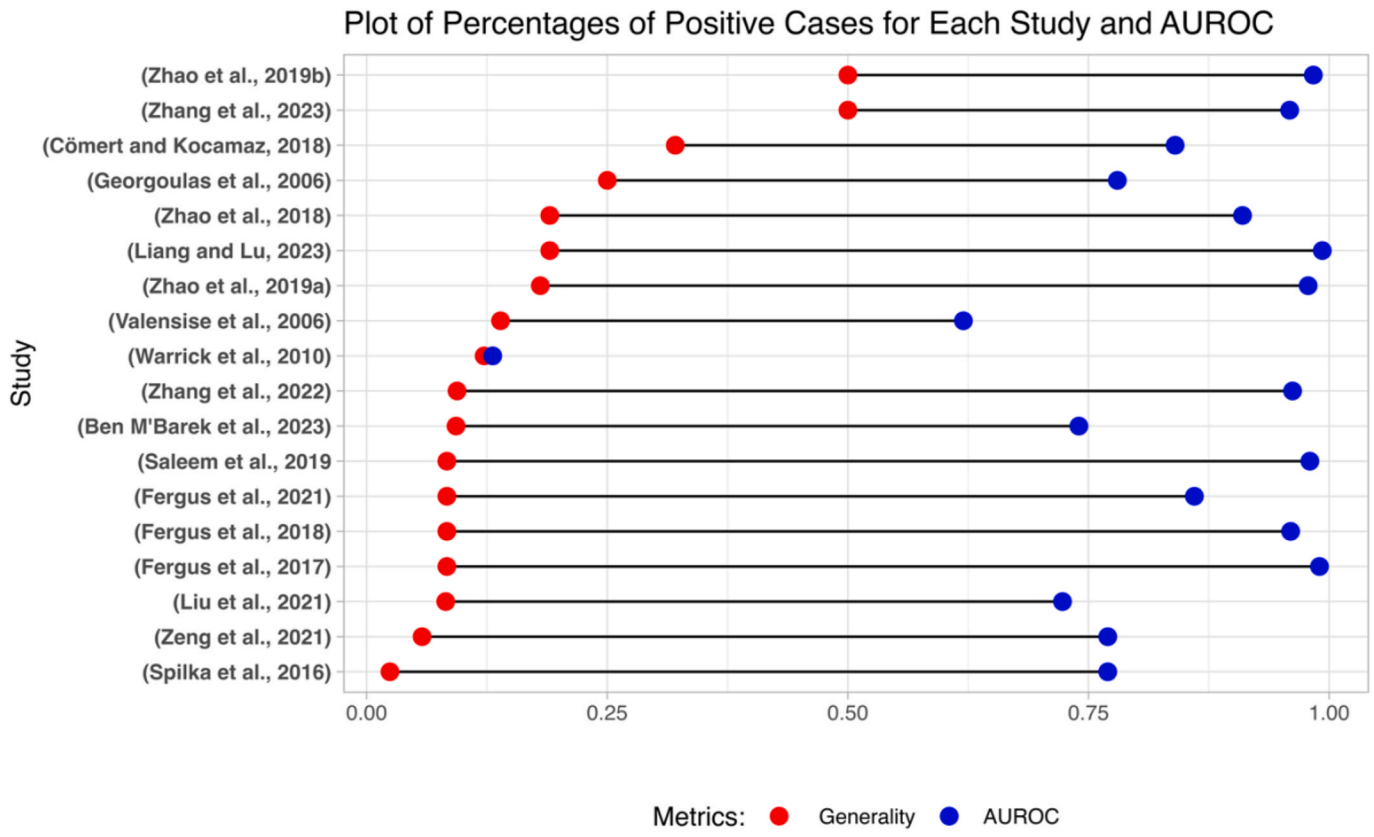

Fig. 9. Lollipop graph showing the relationship between the percentage of hypoxic cases in a dataset and the AUROC of models.

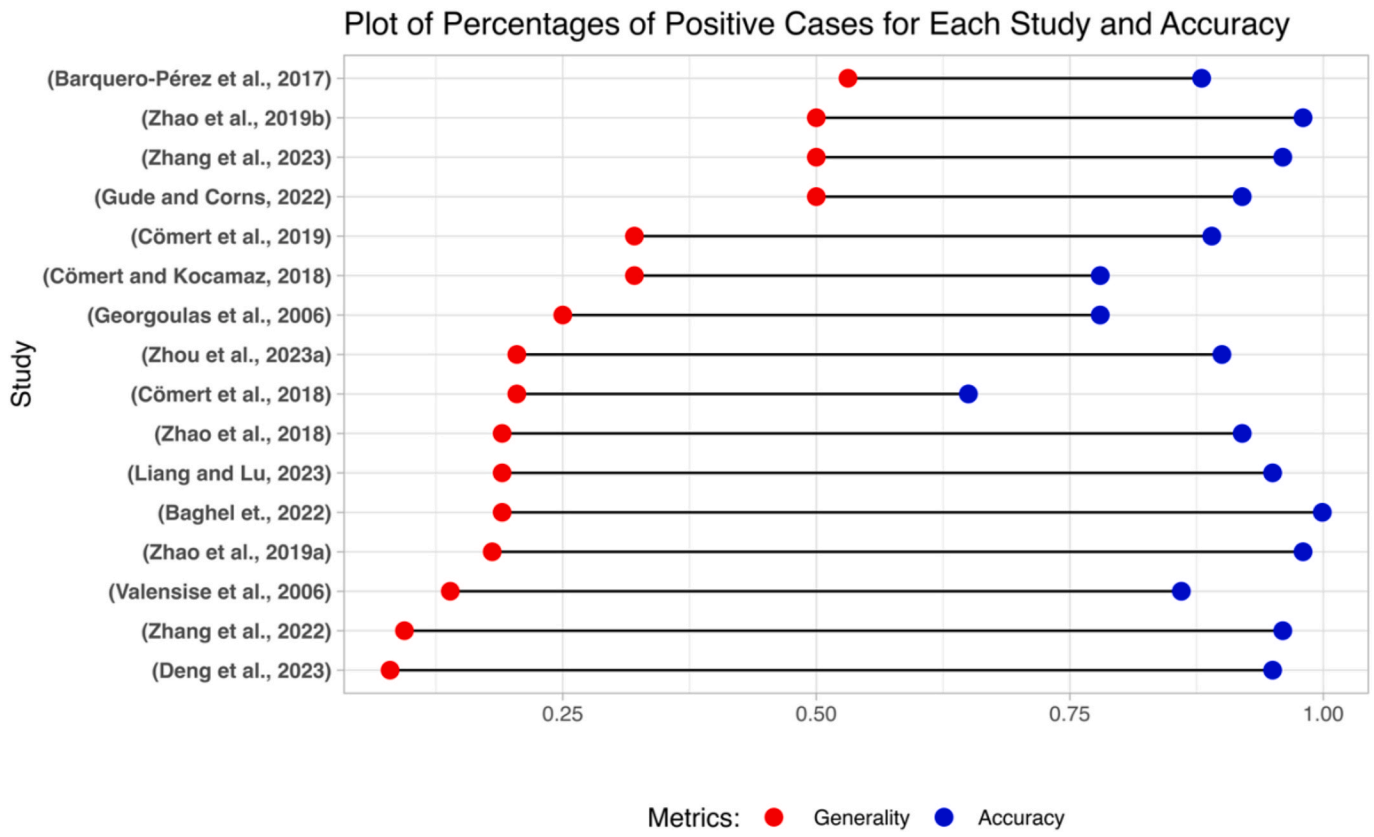

Fig. 10. Lollipop graph showing the relationship between the percentage of hypoxic cases in a dataset and the accuracy of models.

relevant ML algorithms. We highly recommend experts in this field come to a consensus on the best way to analyse and interpret CTG using ML to encourage clinical application to reduce the adverse effects of fetal hypoxia. We propose that future studies employ consistent performance metrics to facilitate the comparison of results and advance the field.

5. Conclusion

In conclusion, we have summarised 36 international studies attempting to improve the classification of CTG by using ML. These have shown the ability of ML to detect subtle changes in the intrapartum FHR and, therefore, potential clinical utility in aiding decision-making in maternity units. The steps for CTG modelling are: 1) pre-processing of raw signals, 2) extracting features, 3) feature engineering, 4) model building, and 5) performance evaluation. We found various methods to process and extract CTG signals. Implementation is limited by the fact that ML algorithms used need to be interpretable. Our work also demonstrates how studies using blood pH as clinical endpoints and the same data source have different distributions of the number of hypoxic and normal fetuses, indicating that the degree of data imbalance is highly dependent on the range of pH benchmarks. We can also see a similar pattern in the model performances, highlighting the complexity of this field.

We have highlighted the gap in this field where there is a need for more open-source CTG datasets, transparency of code and modelling strategies, consensus-derived meaningful clinical endpoints and consideration of baseline risk when implementing new fetal monitoring strategies. We also identified the gaps in CTG processing, including inconsistent use of FHR and UC for morphological analysis, features, and classifiers. Our research emphasised the lack of consistency of CTG-ML research from choosing the gold standard of hypoxia to evaluating model performances and future research should address these shortcomings for clinical application.

Funding

This work was supported by the Medical Research Council [MR/R01566X/1].

CRediT authorship contribution statement

Farah Francis: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Saturnino Luz:** Writing – review & editing, Visualization, Supervision, Formal analysis, Conceptualization. **Honghan Wu:** Writing – review & editing, Supervision. **Sarah J. Stock:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Rosemary Townsend:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2024.108220.

References

[1] B. Petterson, J. Bourke, H. Leonard, P. Jacoby, C. Bower, Co-occurrence of birth defects and intellectual disability, Paediatr. Perinat. Epidemiol. 21 (2007) 65–75.
[2] G. Bogdanovic, A. Babovic, M. Rizvanovic, D. Ljuca, G. Grgic, J. Djuranovic-Milicic, Cardiotocography in the prognosis of perinatal outcome, Med. Arch. 68 (2014) 102–105.

[3] J. Klumper, J.J. Kaandorp, E. Schuit, F. Groenendaal, C. Koopman-Esseboom, E. J. Mulder, F. Van Bel, M.J. Benders, B.W. Mol, R.M. Van Elburg, Behavioral and neurodevelopmental outcome of children after maternal allopurinol administration during suspected fetal hypoxia: 5-year follow up of the ALLO-trial, PLoS One 13 (2018) e0201063.

[4] I. Aliyu, T. Lawal, B. Onankpa, Hypoxic-ischemic encephalopathy and the Apgar scoring system: the experience in a resource-limited setting, Journal of Clinical Sciences 15 (2018) 18–21.

[5] C.E. Wood, M. Keller-Wood, Current paradigms and new perspectives on fetal hypoxia: implications for fetal brain development in late gestation, Am. J. Physiol. Regul. Integr. Comp. Physiol. 317 (2019) R1–R13.

[6] D. Ayres-De-Campos, Obstetric Emergencies, Springer, 2016.

[7] J.E. Lawn, H. Blencowe, P. Waiswa, A. Amouzou, C. Mathers, D. Hogan, V. Flenady, J.F. Frøen, Z.U. Qureshi, C. Calderwood, S. Shiekh, F.B. Jassir, D. You, E.M. Mcclure, M. Mathai, S. Cousens, Stillbirths: rates, risk factors, and acceleration towards 2030, Lancet 387 (2016) 587–603.

[8] S. Ariff, A.C. Lee, J. Lawn, Z.A. Bhutta, Global Burden, epidemiologic trends, and prevention of intrapartum-related deaths in low-resource settings, Clin. Perinatol. 43 (2016) 593–608.

[9] L.P. Thompson, S. Crimmins, B.P. Telugu, S. Turan, Intrauterine hypoxia: clinical consequences and therapeutic perspectives, Res. Rep. Neonatol. 5 (2015) 79–89.

[10] J. Sandall, R.M. Tribe, L. Avery, G. Mola, G.H.A. Visser, C.S.E. Homer, D. Gibbons, N.M. Kelly, H.P. Kennedy, H. Kidanto, P. Taylor, M. Temmerman, Short-term and long-term effects of caesarean section on the health of women and children, Lancet 392 (2018) 1349–1357.

[11] Z. Alfirevic, G.M.L. Gyte, A. Cuthbert, D. Devane, Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour, Cochrane Database Syst. Rev. (2) (2017).

[12] R.M. Grivell, Z. Alfirevic, G.M.L. Gyte, D. Devane, Antenatal Cardiotocography for Fetal Assessment. The Cochrane Database of Systematic Reviews, 2015, 2015, p. CD007863. CD007863.

[13] A. Sweha, T.W. Hacker, J. Nuovo, Interpretation of the electronic fetal heart rate during labor, Am. Fam. Physician 59 (1999) 2487–2500.

[14] D. Ayres-De-Campos, C.Y. Spong, E. Chandraharan, FIGO intrapartum fetal monitoring expert consensus panel, FIGO consensus guidelines on intrapartum fetal monitoring: cardiotocography, Int. J. Gynecol. Obstet. 131 (2015) 13–24.

[15] S. Pereira, E. Chandraharan, Recognition of chronic hypoxia and pre-existing foetal injury on the cardiotocograph (Ctg): urgent need to think beyond the guidelines, Porto Biomedical Journal 2 (2017) 124–129.

[16] P. Holmes, L.W. Oppenheimer, S.W. Wen, The relationship between cervical dilatation at initial presentation in labour and subsequent intervention, Br. J. Obstet. Gynaecol. 108 (2001) 1120–1124.

[17] S.K. Tracy, E. Sullivan, Y.A. Wang, D. Black, M. Tracy, Birth outcomes associated with interventions in labour amongst low risk women: a population-based study, Women Birth 20 (2007) 41–48.

[18] E.H. Hon, Apparatus for continuous monitoring of the fetal heart rate, Yale J. Biol. Med. 32 (1960) 397–399.

[19] K. Hammacher, [New method for the selective registration of the fetal heart beat], Geburtshilfe Frauenheilkd 22 (1962) 1542–1543.

[20] H. Alvarez, R. Caldeyro-Barcia, [The normal and abnormal contractile waves of the uterus during labour], Gynaecologia 138 (1954) 190–212.

[21] F.K. Lotgering, H.C. Wallenburg, H.J. Schouten, Interobserver and intraobserver variation in the assessment of antepartum cardiotocograms, Am. J. Obstet. Gynecol. 144 (1982) 701–705.

[22] S.C. Blackwell, W.A. Grobman, L. Antoniewicz, M. Hutchinson, C. Gyamfi Bannerman, Interobserver and intraobserver reliability of the NICHD 3-tier fetal heart rate interpretation system, Am. J. Obstet. Gynecol. 205 (2011) 378.e1–378.e5.

[23] E. Gyllencreutz, I. Hulthén Varli, P.G. Lindqvist, M. Holzmann, Reliability in cardiotocography interpretation - impact of extended on-site education in addition to web-based learning: an observational study, Acta Obstet. Gynecol. Scand. 96 (2017) 496–502.

[24] S. Das, H. Mukherjee, K. Roy, C.K. Saha, Shortcoming of visual interpretation of cardiotocography: a comparative study with automated method and established guideline using statistical analysis, SN Computer Science 1 (2020) 179.

[25] A.H. Maclennan, S.C. Thompson, J. Gecz, Cerebral palsy: causes, pathways, and the role of genetic variants, Am. J. Obstet. Gynecol. 213 (2015) 779–788.

[26] D.A. Grimes, J.F. Peipert, Electronic fetal monitoring as a public health screening program: the arithmetic of failure, Obstet. Gynecol. 116 (2010) 1397–1400.

[27] Nhs, *Annual Report and Accounts 2018/19* [Online]. NHS Digital, 2019. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/824345/NHS_Resolution_Annual_Report_and_accounts_print.pdf, 8 April 2021.

[28] C.W.H. Yau, B. Leigh, E. Liberati, D. Punch, M. Dixon-Woods, T. Draycott, Clinical negligence costs: taking action to safeguard NHS sustainability, BMJ 368 (2020) m552.

[29] T. Todros, C.U. Preve, C. Plazzotta, M. Biolcati, P. Lombardo, Fetal heart rate tracings: observers versus computer assessment, Eur. J. Obstet. Gynecol. Reprod. Biol. 68 (1996) 83–86.

[30] S. Schiermeier, G. Westhof, A. Leven, H. Hatzmann, J. Reinhard, Intra- and interobserver variability of intrapartum cardiotocography: a multicenter study comparing the FIGO classification with computer analysis software, Gynecol. Obstet. Invest. 72 (2011) 169–173.

[31] G.S. Dawes, M. Lobb, M. Moulden, C.W. Redman, T. Wheeler, Antenatal cardiotocogram quality and interpretation using computers, BJOG An Int. J. Obstet. Gynaecol. 121 (Suppl 7) (2014) 2–8.

[32] R. Keith, The INFANT study - a flawed design foreseen, Lancet 389 (2017) 1697–1698.

[33] B.M. Lake, T.D. Ullman, J.B. Tenenbaum, S.J. Gershman, Building Machines that Learn and Think like People, vol. 40, Behavioral and brain sciences, 2017.

[34] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, WIREs Data Mining and Knowledge Discovery 9 (2019) e1312.

[35] Z. Munn, M.D.J. Peters, C. Stern, C. Tufanaru, A. Mcarthur, E. Aromataris, Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach, BMC Med. Res. Methodol. 18 (2018) 143.

[36] F. Francis, S. Townsend, H. Wu, S. Luz, S. Stock, Machine Learning for Cardiotocography Data to Classify Fetal Outcomes, 2021 [Online], https://osf.io/kh9af/:Opensciencefframework, 01/10/2021 2021.

[37] H. Arksey, L. O'malley, Scoping studies: towards a methodological framework, Int. J. Soc. Res. Methodol. 8 (2005) 19–32.

[38] D. Levac, H. Colquhoun, K.K. O'brien, Scoping studies: advancing the methodology, Implement. Sci. 5 (2010) 69.

[39] M. Peters, C. Godfrey, P. Mcinerney, C. Soares, H. Khalil, D. Parker, The Joanna Briggs Institute Reviewers' Manual 2015: Methodology for JBI Scoping Reviews, 2015.

[40] L.M. Stevens, B.J. Mortazavi, R.C. Deo, L. Curtis, D.P. Kao, Recommendations for reporting machine learning analyses in clinical research, Circulation: Cardiovascular Quality and Outcomes 13 (2020) e006556.

[41] A.C. Tricco, E. Lillie, W. Zarin, K.K. O'brien, H. Colquhoun, D. Levac, D. Moher, M.D.J. Peters, T. Horsley, L. Weeks, S. Hempel, E.A. Akl, C. Chang, J. Mcgowan, L. Stewart, L. Hartling, A. Aldcroft, M.G. Wilson, C. Garritty, S. Lewin, C. M. Godfrey, M.T. Macdonald, E.V. Langlois, K. Soares-Weiser, J. Moriarty, T. Clifford, Ö. Tunçalp, S.E. Straus, PRISMA extension for scoping reviews (PRISMA-Scr): checklist and explanation, Ann. Intern. Med. 169 (2018) 467–473.

[42] V. Chudáček, J. Spilka, M. Burša, P. Janků, L. Hruban, M. Huptych, L. Lhotská, Open access intrapartum Ctg database, BMC Pregnancy Childbirth 14 (2014) 16.

[43] P. Abry, J. Spilka, R. Leonarduzzi, V. Chudáček, N. Pustelnik, M. Doret, Sparse learning for Intrapartum fetal heart rate analysis, Biomedical Physics & Engineering Express 4 (2018) 034002.

[44] N. Baghel, R. Burget, M.K. Dutta, 1d-Fhrnet: automatic diagnosis of fetal acidosis from fetal heart rate signals, Biomed. Signal Process Control 71 (2022) 102794.

[45] Ó. Barquero-Pérez, R. Santiago-Mozos, J.M. Lillo-Castellano, B. García-Viruete, R. Goya-Esteban, A.J. Caamaño, J.L. Rojo-Álvarez, C. Martín-Caballero, Fetal heart rate analysis for automatic detection of perinatal hypoxia using normalized compression distance and machine learning, Front. Physiol. 8 (2017) 113.

[46] I. Ben M'barek, G. Jauvion, J. Vitrou, E. Holmström, M. Koskas, P.F. Ceccaldi, Deepctg® 1.0: an interpretable model to detect fetal hypoxia from cardiotocography data during labor and delivery, Front Pediatr 11 (2023) 1190441.

[47] Z. Cömert, A.F. Kocamaz, Open-access software for analysis of fetal heart rate signals, Biomed. Signal Process Control 45 (2018) 98–108.

[48] Z. Cömert, A.F. Kocamaz, V. Subha, Prognostic model based on image-based time-frequency features and genetic algorithm for fetal hypoxia assessment, Comput. Biol. Med. 99 (2018) 85–97.

[49] Z. Cömert, A. Şengür, Ü. Budak, A.F. Kocamaz, Prediction of intrapartum fetal hypoxia considering feature selection algorithms and machine learning models, Health Inf. Sci. Syst. 7 (2019) 17.

[50] S. Das, H. Mukherjee, K. Roy, C.K. Saha, Fetal health classification from cardiotocograph for both stages of labor-A soft-computing-based approach, Diagnostics 13 (2023).

[51] S. Dash, J.G. Quirk, P.M. Djuri, Fetal heart rate classification using generative models, IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 61 (2014) 2796–2805.

[52] Y. Deng, Y. Zhang, Z. Zhou, X. Zhang, P. Jiao, Z. Zhao, A lightweight fetal distress-assisted diagnosis model based on a cross-channel interactive attention mechanism, Front. Physiol. 14 (2023) 1090937.

[53] P. Fergus, A. Hussain, D. Al-Jumeily, D.-S. Huang, N. Bouguila, Classification of caesarean section and normal vaginal deliveries using foetal heart rate signals and advanced machine learning algorithms, Biomed. Eng. Online 16 (2017) 89.

[54] P. Fergus, M. Selvaraj, C. Chalmers, Machine learning ensemble modelling to classify caesarean section and vaginal delivery types using Cardiotocography traces, Comput. Biol. Med. 93 (2018) 7–16.

[55] P. Fergus, C. Chalmers, C.C. Montanez, D. Reilly, P. Lisboa, B. Pineles, Modelling segmented cardiotocography time-series signals using one-dimensional convolutional neural networks for the early detection of abnormal birth outcomes, IEEE Transactions on Emerging Topics in Computational Intelligence 5 (2021) 882–892.

[56] P. Fuentealba, A. Illanes, F. Ortmeier, Cardiotocographic signal feature extraction through CEEMDAN and time-varying autoregressive spectral-based analysis for fetal welfare assessment, IEEE Access 7 (2019) 159754–159772.

[57] G. Georgoulas, C.D. Stylios, P.P. Groumpos, Predicting the risk of metabolic acidosis for newborns based on fetal heart rate signal classification using support vector machines, IEEE Trans. Biomed. Eng. 53 (2006) 875–884.

[58] G. Georgoulas, P. Karvelis, J. Spilka, V. Chudáček, C.D. Stylios, L. Lhotská, Investigating ph based evaluation of fetal heart rate (Fhr) recordings, Health Technol. 7 (2017) 241–254.

[59] V. Gude, S. Corns, Integrated deep learning and supervised machine learning model for predictive fetal monitoring, Diagnostics 12 (2022).

[60] H. Liang, Y. Lu, A Cnn-Rnn unified framework for intrapartum cardiotocograph classification, Comput. Methods Progr. Biomed. 229 (2023) 107300.

[61] L.C. Liu, Y.H. Tsai, Y.C. Chou, Y.C. Jheng, C.K. Lin, N.Y. Lyu, Y. Chien, Y.P. Yang, K.J. Chang, K.H. Chang, Y.L. Lee, P.H. Wang, T.W. Chu, C.C. Chang, Concordance analysis of intrapartum cardiotocography between physicians and artificial intelligence-based technique using modified one-dimensional fully convolutional networks, J. Chin. Med. Assoc. 84 (2021) 158–164.
[62] J. Ogasawara, S. Ikenoue, H. Yamamoto, M. Sato, Y. Kasuga, Y. Mitsukura, Y. Ikegaya, M. Yasui, M. Tanaka, D. Ochiai, Deep neural network-based classification of cardiotocograms outperformed conventional algorithms, Sci. Rep. 11 (2021) 13367.
[63] A. Petrozziello, C.W.G. Redman, A.T. Papageorghiou, I. Jordanov, A. Georgieva, Multimodal convolutional neural networks to detect fetal compromise during labor and delivery, IEEE Access 7 (2019) 112026–112036.
[64] C. Ricciardi, G. Improta, F. Amato, G. Cesarelli, M. Romano, Classifying the type of delivery from cardiotocographic signals: a machine learning approach, Comput. Methods Progr. Biomed. 196 (2020) 105712.
[65] S. Saleem, S.S. Naqvi, T. Manzoor, A. Saeed, N. Ur Rehman, J. Mirza, A strategy for classification of "vaginal vs. Cesarean section" delivery: bivariate empirical mode decomposition of cardiotocographic recordings, Front. Physiol. 10 (2019).
[66] J. Spilka, J. Frecon, R. Leonarduzzi, N. Pustelnik, P. Abry, M. Doret, Sparse support vector machine for intrapartum fetal heart rate classification, IEEE Journal of Biomedical and Health Informatics 21 (2016) 1, 1.
[67] H. Valensise, F. Facchinetti, B. Vasapollo, F. Giannini, I.D. Monte, D. Arduini, The computerized fetal heart rate analysis in post-term pregnancy identifies patients at risk for fetal distress in labour, Eur. J. Obstet. Gynecol. Reprod. Biol. 125 (2006) 185–192.
[68] P.A. Warrick, E.F. Hamilton, D. Precup, R.E. Kearney, Classification of normal and hypoxic fetuses from systems modeling of intrapartum cardiotocography, IEEE Trans. Biomed. Eng. 57 (2010) 771–779.
[69] Y. Xiao, Y. Lu, M. Liu, R. Zeng, J. Bai, A deep feature fusion network for fetal state assessment, Front. Physiol. 13 (2022) 969052.
[70] K. Yu, J.G. Quirk, P.M. Djurić, Dynamic classification of fetal heart rates by hierarchical Dirichlet process mixture models, PLoS One 12 (2017) e0185417.
[71] R. Zeng, Y. Lu, S. Long, C. Wang, J. Bai, Cardiotocography signal abnormality classification using time-frequency features and Ensemble Cost-sensitive Svm classifier, Comput. Biol. Med. 130 (2021) 104218.
[72] Y. Zhang, Y. Deng, Z. Zhou, X. Zhang, P. Jiao, Z. Zhao, Multimodal learning for fetal distress diagnosis using a multimodal medical information fusion framework, Front. Physiol. 13 (2022) 1021400.
[73] Y. Zhang, Y. Deng, X. Zhang, P. Jiao, X. Zhang, Z. Zhao, Dt-Ctnet: a clinically interpretable diagnosis model for fetal distress, Biomed. Signal Process Control 86 (2023) 105190.
[74] Z. Zhao, Y. Zhang, Y. Deng, A comprehensive feature analysis of the fetal heart rate signal for the intelligent assessment of fetal state, J. Clin. Med. 7 (2018).
[75] Z. Zhao, Y. Deng, Y. Zhang, Y. Zhang, X. Zhang, L. Shao, Deepfhr: intelligent prediction of fetal Acidemia using fetal heart rate signals based on convolutional neural network, BMC Med. Inf. Decis. Making 19 (2019) 286.
[76] Z. Zhao, Y. Zhang, Z. Comert, Y. Deng, Computer-aided diagnosis system of fetal hypoxia incorporating recurrence plot with convolutional neural network, Front. Physiol. 10 (2019) 255, 255.
[77] Z. Zhou, Z. Zhao, X. Zhang, X. Zhang, P. Jiao, Improvement of accuracy and resilience in Fhr classification via double trend accumulation encoding and attention mechanism, Biomed. Signal Process Control 85 (2023) 104929.
[78] Z. Zhou, Z. Zhao, X. Zhang, X. Zhang, P. Jiao, X. Ye, Identifying fetal status with fetal heart rate: deep learning approach based on long convolution, Comput. Biol. Med. 159 (2023) 106970.
[79] Nice, *Intrapartum Care for Healthy Women and Babies* [Online]. Clinical Guideline (Cg190), National Institute for Health and Care Excellence, 2017. Available: https://www.nice.org.uk/guidance/cg190/chapter/recommendations#care-of-the-newborn-baby, 12 June 2021.
[80] J.L. Aeberhard, A.-P. Radan, R. Delgado-Gonzalo, K.M. Strahm, H. B. Sigurthorsdottir, S. Schneider, D. Surbek, Artificial intelligence and machine learning in cardiotocography: a scoping review, Eur. J. Obstet. Gynecol. Reprod. Biol. 281 (2023) 54–62.
[81] A.S. Tarawneh, A.B. Hassanat, G.A. Altarawneh, A. Almuhaimeed, Stop oversampling for class imbalance learning: a review, IEEE Access 10 (2022) 47643–47660.
[82] A. Stando, M. Cavus, P. Biecek, The Effect of Balancing Methods on Model Behavior in Imbalanced Classification Problems, 2023 *arxiv preprint arxiv: 2307.00157*.
[83] F. Grina, Z. Elouedi, E. Lefevre, A preprocessing approach for class-imbalanced data using SMOTE and belief function theory. Intelligent Data Engineering and Automated Learning–IDEAL 2020: 21st International Conference, Guimaraes, Portugal, November 4–6, 2020, Proceedings, Part Ii 21, Springer, 2020, pp. 3–11.
[84] F.P. Vandenbussche, I.L. Van Kamp, D. Oepkes, J. Hermans, J. Bennebroek Gravenhorst, H.H. Kanhai, Blood gas and ph in the human fetus with severe anemia, Fetal Diagn. Ther. 13 (1998) 115–122.
[85] R. Victory, D. Penava, O. Da Silva, R. Natale, B. Richardson, Umbilical cord ph and base excess values in relation to adverse outcome events for infants delivering at term, Am. J. Obstet. Gynecol. 191 (2004) 2021–2028.
[86] P.P. Van Den Berg, W.L. Nelen, H.W. Jongsma, R. Nijland, L.A. Kollée, J. G. Nijhuis, T.K. Eskes, Neonatal complications in newborns with an umbilical artery ph < 7.00, Am. J. Obstet. Gynecol. 175 (1996) 1152–1157.
[87] J. Engel, J. Gerretzen, E. Szymańska, J.J. Jansen, G. Downey, L. Blanchet, L. M. Buydens, Breaking with trends in pre-processing? TrAC, Trends Anal. Chem. 50 (2013) 96–106.
[88] P. Oliveri, C. Malegori, R. Simonetti, M. Casale, The impact of signal pre-processing on the final interpretation of analytical outcomes - a tutorial, Anal. Chim. Acta 1058 (2019) 9–17.
[89] German Society Of Gynecology And Obstetrics, Maternal Fetal Medicine Study Group, German Society Of Prenatal Medicine And Obstetrics & Medicine, G. S. O. P, S1-Guideline on the use of ctg during pregnancy and labor: long version - AWMF registry No. 015/036, Geburtshilfe Frauenheilkd 74 (2014) 721–732.
[90] National Collaborating Centre For, W. S. & Children's, H., National Institute for health and clinical excellence: guidance, in: Intrapartum Care: Care of Healthy Women and Their Babies during Childbirth, RCOG Presscopyright © 2007, National Collaborating Centre for Women's and Children's Health, London, 2007.
[91] Z. Cömert, A.F. Kocamaz, A novel software for comprehensive analysis of cardiotocography signals "CTG-OAS", in: 2017 International Artificial Intelligence and Data Processing Symposium (IDAP), 16-17 Sept. 2017, 2017, pp. 1–6.
[92] S. Karamizadeh, S.M. Abdullah, M. Halimi, J. Shayan, M.J. Rajabi, Advantage and drawback of support vector machine functionality, in: 2014 International Conference on Computer, Communications, and Control Technology (I4CT), 2-4 Sept. 2014, 2014, pp. 63–65.
[93] M. Somvanshi, P. Chavan, S. Tambade, S.V. Shinde, A review of machine learning techniques using decision tree and support vector machine, in: 2016 International Conference on Computing Communication Control and Automation (ICCUBEA), 12-13 Aug. 2016, 2016, pp. 1–7.
[94] A.E. Mohamed, Comparative study of four supervised machine learning techniques for classification, International Journal of Applied 7 (2017).
[95] D.A. Pisner, D.M. Schnyer, Chapter 6 - support vector machine, in: A. MECHELLI, S. Vieira (Eds.), Machine Learning, Academic Press, 2020.
[96] Y. Zhao, Y. Zhang, Comparison of decision tree methods for finding active objects, Adv. Space Res. 41 (2008) 1955–1959.
[97] B. Krawczyk, Learning from imbalanced data: open challenges and future directions, Progress in Artificial Intelligence 5 (2016) 221–232.
[98] J. Błaszczyński, J. Stefanowski, Neighbourhood sampling in bagging for imbalanced data, Neurocomputing 150 (2015) 529–542.
[99] C. Molnar, Interpretable Machine Learning, Leanpublishing, 2019.
[100] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: *Proceedings Of the 31st International Conference On Neural Information Processing Systems*. Long Beach, Curran Associates Inc, California, USA, 2017.
[101] W.J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, B. Yu, Definitions, methods, and applications in interpretable machine learning, Proc. Natl. Acad. Sci. USA 116 (2019) 22071–22080.
[102] R. Trevethan, Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice, Front. Public Health 5 (2017).
[103] S. Halligan, D.G. Altman, S. Mallett, Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach, Eur. Radiol. 25 (2015) 932–939.
[104] E.J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nat. Med. 25 (2019) 44–56.
[105] C.J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, D. King, Key challenges for delivering clinical impact with artificial intelligence, BMC Med. 17 (2019) 195.
[106] P. Thölke, Y.-J. Mantilla-Ramos, H. Abdelhedi, C. Maschke, A. Dehgan, Y. Harel, A. Kemtur, L. Mekki Berrada, M. Sahraoui, T. Young, A. Bellemare Pépin, C. El Khantour, M. Landry, A. Pascarella, V. Hadid, E. Combrisson, J. O'byrne, K. Jerbi, Class imbalance should not throw you off balance: choosing the right classifiers and performance metrics for brain decoding with imbalanced data, Neuroimage 277 (2023) 120253.