electronics

MDPI

Article

Machine Learning Methods for Preterm Birth Prediction: A Review

Tomasz Włodarczyk [1,*], Szymon Płotka [1], Tomasz Szczepański [1], Przemysław Rokita [1], Nicole Sochacki-Wójcicka [2], Jakub Wójcicki [2], Michał Lipa [2] and Tomasz Trzciński [1,3,*]

[1] Institute of Computer Science, Warsaw University of Technology, 00-661 Warsaw, Poland; plotkaszymon@gmail.com (S.P.); tmk.szczepanski@gmail.com (T.S.); pro@ii.pw.edu.pl (P.R.)

[2] 1st Department of Obstetrics and Gynecology, Medical University of Warsaw, 02-091 Warsaw, Poland; nicole.wojcicki@googlemail.com (N.S.-W.); jakub.wojcicki@googlemail.com (J.W.); michallipa1@gmail.com (M.L.)

[3] Tooploox, 53-601 Wroclaw, Poland

* Correspondence: wlodarczyk.tomasz@gmail.com (T.W.); tomasz.trzcinski@pw.edu.pl (T.T.)

Abstract: Preterm births affect around 15 million children a year worldwide. Current medical efforts focus on mitigating the effects of prematurity, not on preventing it. Diagnostic methods are based on parent traits and transvaginal ultrasound, during which the length of the cervix is examined. Approximately 30% of preterm births are not correctly predicted due to the complexity of this process and its subjective assessment. Based on recent research, there is hope that machine learning can be a helpful tool to support the diagnosis of preterm births. The objective of this study is to present various machine learning algorithms applied to preterm birth prediction. The wide spectrum of analysed data sets is the advantage of this survey. They range from electrohysterogram signals through electronic health records to transvaginal ultrasounds. Reviews of works on preterm birth already exist; however, this is the first review that includes works that are based on a transvaginal ultrasound examination. In this work, we present a critical appraisal of popular methods that have employed machine learning methods for preterm birth prediction. Moreover, we summarise the most common challenges incurred and discuss their possible application in the future.

Keywords: artificial intelligence; deep learning; machine learning; preterm birth

Citation: Włodarczyk, T.; Płotka, S.; Szczepański, T.; Rokita, P.; Sochacki-Wójcicka, N.; Wójcicki, J.; Lipa, M.; Trzciński, T. Machine Learning Methods for Preterm Birth Prediction: A Review. *Electronics* **2021**, *10*, 586. https://doi.org/10.3390/electronics10050586

Academic Editor: Juan Manuel Corchado Rodriguez

Received: 30 December 2020
Accepted: 25 February 2021
Published: 3 March 2021

1. Introduction

Preterm births affect around 15 million children a year worldwide [1]. This is the leading cause of infant mortality, developmental delays, and long-term disability. Complications of preterm birth are the single largest direct causes of neonatal deaths, being responsible for 35% of the world's 3.1 million deaths a year [2]. In almost all high- and middle-income countries of the world, preterm birth (PTB), which is also known as premature birth, is the leading cause of child death.

PTB is defined by WHO as all births before 37 completed weeks of gestation or fewer than 259 days since the first day of a woman's last menstrual period [3]. Preterm birth can be further subdivided based on gestational age: extremely preterm (<28 weeks), very preterm (28–<32 weeks), and moderate preterm (32–<37 weeks of gestation). Premature birth can result in long-term complications, with the frequency and severity of unfavourable outcomes increasing with decreasing gestational age and decreasing quality of care [4]. The aforementioned 37-week limit is somewhat arbitrary and, although the risk that is associated with preterm birth is greater the lower the gestational age, babies born at 37 or 38 weeks are still at greater risk than those born at 40 weeks gestation [5].

There are two types of PTB—spontaneous and iatrogenic. In cases affected with spontaneous PTB, the contractions start before the 37th week without any clinical interventions, mostly due to cervical insufficiency or intrauterine infection. On the other hand, iatrogenic

PTB occurs in severe, gestational complications, such as preeclampsia (PE) or fetal growth restriction (FGR). In this group, preterm delivery is recommended due to endangered fetal or maternal wellbeing. In high-income countries with widely available access to healthcare professionals, the iatrogenic preterm birth occurs significantly more frequently than in low-income countries. However, the prevalence of PTB is similar worldwide, regardless of the part of the globe [6].

The natural and desirable date of birth occurs after 37 weeks of pregnancy. The earlier the newborn is delivered, the higher the risk of prematurity complications and the need for a longer stay in the neonatal intensive care unit (NICU). Moreover, a prolonged stay in NICU can cause significant stress among the patient's family and generates costs for the healthcare system. According to the literature, certain screening methods may identify patients with an increased risk of PTB and imply subsequent, prophylactic steps [7]. Ultrasonographic, transvaginal measurement of the cervical length (CL) between 18 + 0 and 22 + 0 weeks of gestation is a recognised, popular screening method for estimating the risk of a PTB that has become a standard in prenatal care worldwide. Nonetheless, a significant rate of PTBs occurs in patients that are identified as low-risk at the mid-trimester scan [8]; thus, we believe that further studies may improve future prediction models.

Ultrasonographic measurement depends on several factors, such as the quality of the ultrasound system, the experience of the sonographer, and the technique of the examination. Because many factors affect the final result, each measurement may be altered due to various conditions of certain ultrasound examinations, and many details may not be visible for the human eye. We hope that ML methods can help in reducing the number of PTBs. These methods, which have already been used in signal and image analysis to generate artificial data, could improve predictions by analysing multiplanar, ultrasonographic images [9,10]. Additionally, they might discover some new features that may be incorporated into current screening strategies. Machine learning can help to analyse the quality of known markers [11] or lead to the discovery of new ones. New algorithms for text analysis allow for using descriptions of medical examinations performed during pregnancy as an input of prediction models, which opens up many new possibilities [12].

In this work, we review articles on preterm birth prediction using ML methods that may potentially be incorporated into perinatal medicine. We have chosen publications that contribute to this area from almost the very beginning of scientists' interest in this topic. Because of the small number of works, we were able to analyse a wide time frame, from 1994 until today. The studies to date have mainly used electronic health records (EHR) statistics as well as electrohysterography (EHG) and uterine electromyography (EMG) records. In recent years, fruitful attempts at using transvaginal (TVS) ultrasound image data have emerged .

The early days of research on this topic were not hopeful. The results that were attained by the created models were worse than the toss of a coin, but over the years this has significantly improved. In our review, we can see a variety of tools being used, from expert systems, through SVM classifiers [13] to deep neural networks [14]. What is important, in all of the reviewed articles, predictions were made on data that were gathered before labour. This is key to developing a predictive system to detect the risk of preterm birth before it happens. The structure of our work is based on the chronology of publications with an additional division into the data types that were used for research.

We hope that our work indicates noteworthy directions of development and also confirms the need to analyse the topic of premature births due to the seriousness of this matter and still a lot of room for improvement. The results presented in this paper are promising despite the difficulty of diagnosis of preterm delivery given the lack of understanding about its causes.

The remainder of this work is organised in the following manner. In Section 2 we present preliminaries: PTB problem description, difficulties, future challenges and data imbalance problem. In Section 3 we present four medical fields that are utilised for preterm birth detection, such as electrohysterography (EHG), electronic health records

(EHR), transvaginal ultrasound (TVS), and uterine electromyography (EMG) (Table 3). In Section 4, we present a discussion on the future of ML applications and noteworthy directions for cooperation between doctors and data scientists. Finally, in Section 5 we conclude the paper.

2. Preliminaries

2.1. Preterm Birth Prediction

The rate of preterm birth has not been significantly reduced throughout the last 30–40 years, although there were many efforts and studies to reduce this number. PTB is considered the main cause of infant mortality, impaired neurodevelopment and long-term disability. This problem is not only limited to low-income countries. The United States and Brazil are both at the forefront in terms of preterm birth rates (at the 6th and 10th position, respectively). On the other hand, Belarus ranks first with the lowest rate of preterm births, corresponding to 4.1 per 100 births [15].

For most of the 20th century, premature births were considered to be an unpredictable and unavoidable fact of life. According to WHO, 1.1 million of born too soon babies dies in the postnatal period due to prematurity complications [16]. It is estimated that three-quarters of them could survive if certain medical management was administered [17]. In recent years, the simultaneous development of advanced medical technologies and machine learning (ML) has allowed for an increase in the quality of healthcare. Many clinical issues remain unsolved despite the above mentioned constant progress in the various fields of science.

For several decades, many researchers have been trying to solve premature births by applying various types of machine learning algorithms. Hence, different methods have been developed to address either sPTB detection or classification. Particularly, the authors analyze ML-based approaches like Support Vector Machine (SVM) [18], K-Nearest Neighbors (KNN) [19], and Convolutional Neural Networks (CNNs) [9,10]. A fundamental motivation for this topic is that with preterm birth prediction, the lives of many children can be saved or spared the many consequences of preterm birth. With the help of early detection of preterm birth, steps can be taken to maintain the pregnancy. Modern medicine has the tools to accomplish this task. All that is missing is a crucial element: early warning [20]. Using machine learning methods to create a prediction model can allow for this foresight.

2.2. Difficulties and Future Challenges

Timely detection of pregnancies at high risk of spontaneous preterm birth (sPTB) is a challenge that can help reduce the number of miscarriages and side effects later in life for premature babies. Nearly half of all sPTBs are found in women with no known clinical risk factors. Current sPTB diagnostic methods, such as obstetric interview, maternal features and transvaginal ultrasound examination of the cervix, did not lower the PTB rate. It is the reason why PTB is a difficult and complex real-world problem. This challenge stems from the nature of pregnancy data, which changes dynamically, is noisy, and often contains missing data for important groups of variables (e.g., genetic data) [20]. Accurate classification and prediction of PTB are challenging tasks considering the large variety of potential factors and the constant lack of reliable data on variables. Another challenge is the waiting time for hardly available data (due to pregnancy duration), the acquisition and processing of which requires the Medical Ethics Committee's approval. The current etiological factors that influence PTB are still mostly unknown. A better understanding of the underlying variables and the use of machine learning methods to develop new methods to predict PTB better may prove crucial.

2.3. Data Imbalance

The class imbalance problem arises when the class of interest is relatively rare compared with other class(es) [21]. Many traditional algorithms to machine learning assume that the target classes share similar prior probabilities. In many real-world applications, this assumption is not valid. Supervised learning methods require labelled training data, and in classification problems, each data sample belongs to a known class. In a binary classification problem with data samples from two groups, class imbalance occurs when one class, the minority group, contains significantly fewer samples than the other class, the majority group [22].

In imbalanced dataset almost all the instances are labelled as one class, while far fewer instances are labelled as the other class, usually the more important class [23]. A well-known class imbalanced machine learning scenario is the medical diagnosis task of detecting disease, where the majority of the patients are healthy and detecting disease is of greater interest. One can find examples in medical field [24]—Grzymala et al. [25] propose how to increase the sensitivity of preterm birth prediction. PTB occurrence is an example of skewed distribution, so solving class imbalance is here an important issue.

Standard machine learning algorithms tend to be overwhelmed by the majority class and ignore the minority class since they classify most of the data into the majority class. Class imbalance causes suboptimal classification performance. Most algorithms do not work correctly when the data sets are highly imbalanced. The minority class has much lower precision and recall than the majority class. Many practitioners have observed that for extremely skewed class distributions, the recall of the minority class is often equal to zero [26].

Over the last ten years, machine learning and mostly deep learning methods have grown in popularity and were used successfully in many fields. However, very little statistical work has been done which properly evaluates techniques for handling class imbalance using deep learning. In fact, many researchers agree that the subject of deep learning with class imbalanced data is insufficiently researched [27]. However, some methods for coping with the imbalanced class in Convolutional Neural Networks are reviewed by Buda et al. [28]. In 2012 most existing imbalance learning techniques were only designed for and tested in two-class scenarios [29]. However, recent applications extend binary imbalance classifiers to multiclass data using the decomposition methods (e.g., One-vs-All) or adapt the intrinsic process in building the decision trees or adopt ensemble-based approaches [30].

Addressing class imbalance with traditional machine learning techniques has been studied extensively over the last two decades. First publications and surveys in this topic come from the turn of the 2000s [23,26,31–35]. For twenty years, many literature reviews on the topic have been developed—for a more in-depth study of the topic, the reader may familiarise with them. Comparison of results using 35 different benchmark datasets on 7 sampling techniques and 11 commonly-used learning algorithms can be found in the analysis performed by Van Hulse et al. [36].

The majority class bias can be diminished by altering the training data to decrease imbalance or modifying the model's underlying learning or decision process to increase sensitivity towards the minority group. Methods for handling class imbalance are grouped into data-level techniques (e.g., data sampling) and algorithm-level methods (e.g., cost-sensitive and ensemble learning).

Data-level techniques can be subdivided into categories:

- undersampling:
  - non-heuristic random undersampling,
  - one-sided selection [33],
  - wilson's editing [37],
- oversampling:
  - non-heuristic random oversampling,
  - Synthetic Minority Oversampling Technique (SMOTE) [38],
  - borderline-SMOTE [39],
  - safe-level-SMOTE [40],
  - cluster-based over-sampling [41],

– Adaptive Synthetic Sampling (ADASYN) [42].

Random undersampling discards random samples from the majority group, while random oversampling duplicates random samples from the minority group. The downside of oversampling is that it increases training time and can be the cause of overfitting [43]. Random over-sampling does not actually increase any information and fails in solving the fundamental 'lack of data' problem. Overfitting occurs when a model fits too closely to the training data and cannot generalise to new data. Commonly used SMOTE method creates new non-replicated examples by interpolating neighbouring minority class instances. However, their broadened decision regions are still error-prone by synthesising noisy and borderline examples [44]. Therefore undersampling is often preferred to over-sampling [45]. Although undersampling does not introduce false dependencies and features to the data, it also is not without its shortcomings. The elimination of some samples from the training dataset may have two adverse effects:

- information loss—due to the elimination of informative or useful samples, classification effectiveness deteriorates,
- data cleaning—because of eliminating irrelevant, redundant, or even noisy samples, classification effectiveness is falsely improved.

In addition to the appropriate approach to the data, we should also focus on the classifier and the learning process. In learning extremely imbalanced data, the overall classification accuracy is often not an adequate measure of performance. A trivial classifier that predicts every case as the majority class can still achieve very high accuracy. We use metrics, such as true negative rate, true positive rate, weighted accuracy, precision, and recall, in order to evaluate learning algorithms' performance on imbalanced data. We want to provide an intuition of negative influence of imbalanced data on classifiers for the most popular ones (that we also recommend to use in review):

- SVM classifier—for a highly imbalanced classification, the majority class pushes the ideal decision boundary toward the minority class [21],
- random forest—classifier induces each constituent tree from a bootstrap sample of the training data [46]. In learning extremely imbalanced data, there is a significant probability that a bootstrap sample contains few or even none of the minority class, which results in a tree with poor performance for predicting the minority class [47].

Algorithmic methods for handling class imbalance do not alter the training data distribution. Instead, the learning or decision process is adjusted in a way that increases the importance of the positive class. Most commonly, algorithms are modified to take a class penalty or weight into consideration, or the decision threshold is shifted in a way that reduces bias towards the negative class. In cost-sensitive learning, penalties are assigned to each class through a cost matrix. Increasing the cost of the minority group is equivalent to increasing its importance, decreasing the likelihood that the learner will incorrectly classify instances from this group [48]. One of the biggest challenges in cost-sensitive learning is the assignment of an effective cost matrix.

Another type of methods for handling data imbalance are hybrid methods. They combine different sampling-techniques and algorithm-based modifications. Ensemble learning methods such as Boosting and Bagging are the most successful approaches. These methods combine the results of many classifiers. Most commonly met in literature methods are: AdaBoost [49], Rare-Boost [50], SMOTEBoost [51], AsymBoost [52], AdaCost [53], MetaCost [54], EasyEnsemble [35]and BalanceCascade [35].

Contemporary classification methods that are based on deep convolutional neural networks can also follow classic class re-sampling or cost-sensitive training. However, Huang et al. [44] proposed that, given an imagery dataset with imbalanced class distribution, our goal is to learn a Euclidean embedding from an image into a feature space. The embedded features are discriminative without any possible local class imbalance. The presented in work Large Margin Local Embedding (LMLE) approach offers crucial feature representations for the following classification to perform well on imbalanced data. After we have applied LMLE, we can use an appropriate classifier to learn from the obtained features.

State-of-the-art methods for imbalanced class data are reviewed by Bi et al. [30]. They perform analysis on 19 publicly available datasets and measure the performance of different methods, i.e., accuracy, F-score and computation time. The best results are achieved by proposed by them Diversified Error Correcting Output Codes (DECOC) algorithm, though it is very computationally expensive.

The use of described methods for preterm birth data depends mostly from the type of data. SMOTE and ADASYN should help reducing imbalance when using EHR or EHG data. Using SMOTE one should be careful not use to high oversampling because of overfitting risk. For imagery data analysis, we can use LMLE or an oversampling approach specific to neural networks optimised with stochastic gradient descent—the class-aware sampling [55].

3. Methods

This section presents works that are related to different types of medical examination (different data sources) used for preterm birth detection with their advantages and shortcomings. We describe 24 publications, out of which 14 are based on EHR data, six on EHG data, two on TVS data, and two on EMG data. Because of the different character of data, these datasets require various data preprocessing steps and different machine learning methods that we list in each of the descriptions.

For the first time, the use of the TVS data type for preterm birth prediction is reviewed. There are high hopes for improved prediction quality with TVS data, upon which we elaborate in Section 4. Our analysis also covers current preterm birth prediction research challenges and their promising future directions.

In the works, we describe many machine learning algorithms that are used, among others: SVM, k-NN [56], decision tree [57], random forest [58], logistic regression [59], stochastic gradient boosting [60], and neural networks [61]. Within neural networks, we can distinguish convolutional and recurrent networks [62]. The use of each of these algorithms primarily depends on the type of data, but also on the size of the learning set. For small data sets, neural networks are not advisable; however, good results can be expected while using an SVM classifier. The rationale for choosing between traditional machine learning algorithms (SVM, logistic regression) and neural networks must be directly sought in our needs, namely whether we need only classification or whether we also need feature extraction. We would also like to point out that the choice of algorithm among the described works, in addition to the type of data, also depend on the time of publication and the current state-of-the-art algorithms.

3.1. Data Availability and Use

Collecting data on premature births is difficult. In addition to the ethics and data anonymisation that we mention later in the discussion (Section 4), difficulties arise in the nature of the data. After the study is performed, there is a significant waiting time for the annotation of the collected data. Depending on the country in which the study is developed, it is difficult to obtain data from premature babies due to the relatively low rate of premature births as compared to the general population. There is also a significant difference of incidence between countries [16,63]. The rarity of the occurring phenomenon is usually solved by over-sampling a group of premature babies [64] or under-sampling a group born on time [35]. The latter is safer in terms of generalisation, but it significantly reduces the data set. Over-sampling, on the other hand, can lead to cases that do not naturally occur. The difficult issue is a proper choice of preterm birth cases for the dataset that is a decision, if, for example, twin pregnancies should be included or if the classification system should be made only for singleton pregnancies. Twin pregnancy is a strong factor for preterm pregnancy, but it may be followed by singleton pregnancy for the same mother [65,66]. Iatrogenic (illness caused by medical examination or treatment) preterm deliveries are to be excluded from the dataset that was used to predict sPTB. Although performed during pregnancy, some tests are made too late and do not carry the appropriate diagnostic value for extreme preterm birth (before 28th week) detection that accounts for most of the deaths of newborns [12]. As an example of early examination, one can provide cervical length measured at 23 weeks of gestation that provides an accurate prediction of early preterm delivery [67]. The proper assessment of gestational age is another question [68]. It is also difficult to find data from standard procedures, as many countries have different requirements for performing mandatory diagnostics during pregnancy. Many differences in perinatal care are based on government policy, recommendations from a scientific or professional society, or no written policy. Along with the policies, not all women are commissioned to perform certain types of tests [69–72]. Most of the researches come from only one medical centre [12,73]. The model that is created in this way may be difficult to generalise in the case of other hospitals, what is more, it may require the introduction of the same types of tests, which, at a given moment, may not be a standard procedure in another medical centre. Datasets that are gathered from many medical centres should be curated [18,74], because medical documentation standards may differ among them and be inconsistent throughout the whole gestation process. Table 1 presents all of the datasets used in the works presented in this review. They are compared based on the dataset's size, the type of data, the ratio of preterm births to the total, and the source where they were collected. It can be seen that what stage of pregnancy they refer is only clearly stated in the case of a few datasets.

Table 1. Comparison of the preterm birth datasets.

Author	Data Type	Group Size	PTB %	Gestation Age (Week)	Data Source
Grzymała-Busse et al. [75]	EHR	18,890	-	-	St. Luke's Regional Perinatal Center, Healthdyne Perinatal Services, Tokos Corporation
Woolery et al. [76]	EHR	18,899	-	-	St. Luke's Regional Perinatal Center, Healthdyne Perinatal Services, Tokos Corporation
Mercer et al. [77]	EHR	2929	10.55	23–24	Maternal-Fetal Medicine Units Network
Goodwin et al. [78]	EHR	63,167	22	-	Duke University's Medical Center
Frize et al. [79]	EHR	113,000	17	-	The Pregnancy Risk Assessment Monitoring System (PRAMS) database
Vovsha et al. [18]	EHR	2929	10.55	-	Maternal-Fetal Medicine Units Network
Tran et al. [11]	EHR	18,836	6.81	-	Royal North Shore (RNS) hospital
Weber et al. [80]	EHR	336,214	1.02	-	-
Esty et al. [81]	EHR	782,000	7.09	-	BORN (Better Outcomes Registry Network) Information System, PRAMS (Pregnancy Risk Monitoring Assessment)
Gao et al. [12]	EHR	25,689	8.09	-	Vanderbilt University Medical Center
Prema et al. [82]	EHR	124	14.52	-	Local hospitals of Mysuru, Karnataka state, India
Lee et al. [83]	EHR	596	7.21	18–24	Anam Hospital in Seoul, Korea
Rawashdeh et al. [84]	EHR	274	9.49	-	Fetal medicine unit in a tertiary hospital in NSW, Australia
Koivu et al. [85]	EHR	15,883,784	9.65	-	CDC - National Center of Health Statistics
Fergus et al. [19]	EHG	300	12.67	-	TPEHG
Hussain et al. [64]	EHG	300	12.67	-	TPEHG
Sadi-Ahmed et al. [86]	EHG	-	-	-	TPEHG
Despotovic et al. [87]	EHG	160	11.73	22–25	TPEHG
Chen et al. [88]	EHG	31	41.94	-	TPEHGT
Degbedzui et al. [89]	EHG	300	12.67	22–32	TPEHG
Włodarczyk et al. [9]	TVS	354	10.97	-	King's College London, Medical University of Warsaw
Włodarczyk et al. [10]	TVS	359	11.98	-	King's College London, Medical University of Warsaw
Maner et al. [90]	EMG	185	27.57	-	University of Texas Medical Branch
Most et al. [91]	EMG	87	100	-	-

3.2. Electrohysterography

Electrohysterography (EHG) is a non-invasive measurement of the electrical activity underlying uterine contractions. EHG is a recording of the electrical currents through contact electrodes at the maternal abdomen. The first EHG signal ever reported in the literature was measured in 1931 as the deflection of a galvanometer's needle that is caused by a uterine contraction [92]. The need for a non-invasive and reliable method for testing uterine activity, predicting delivery, and understanding the processes underlying the onset of labour results in the steadily increasing interest in EHG. Since the first application of EHG, the recording techniques have significantly advanced, and computer technology has allowed for new methods of signal analysis, including ML methods [93]. One of the most important factors in using the EHG data, which emerges from the reviewed articles, is an extraction of features from the recorded signals. In EHG signal analysis for the preterm birth prediction task, many researchers approached this challenge thanks to Gašper Fele-Zorz et al., who first worked on the Term-Preterm EHG Database (TPEHG) records dataset [73,94]. In the following years, many types of research applied ML methods on this data in the hope of improving the results. It is noticeable that only 38 EHG samples are collected from patients whose gestation ended in preterm delivery, while the other 262 EHG samples are from patients with normal term delivery (Table 2). The week of the pregnancy in which examination of the patient is performed is the value that can hardly be overestimated for future diagnostics. Records from the normal term deliveries consist of 143 records that were recorded early (before the 26th week of gestation) and 119 records later. The 38 PTB records consist of 19 records that were recorded early, before the 26th week of gestation, and 19 records recorded later. The results of the developed models can be assessed based on such quality measures as accuracy, specificity, sensitivity, and area under the curve (AUC) of the receiver characteristic operator curve (ROC) [95,96].

Table 2. Details of the Term-Preterm EHG Database (TPEHG) recordings dataset [73,94].

Recording	Recording Week Median	Term Delivery	Preterm Delivery
Early Term	23	143	19
Late term	30	119	19

Accuracy (1) is the ratio of how many correct predictions the model made out of the total number of predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Sensitivity (True Positive Ratio) (2) tells us what proportion of the positive class were correctly classified and specificity (True Negative Rate) (3) tells us what proportion of the negative class became correctly classified.

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

The ROC curve is an evaluation metric for binary classification problems. It is a probability curve that plots the True Positive Ratio against False Positive Ratio at various threshold values and separates the 'signal' from the 'noise'. The AUC is the measure of the ability of a classifier to distinguish between classes and it is used as a summary of the ROC curve. Figure 1 depicts an example of an EHG workflow.

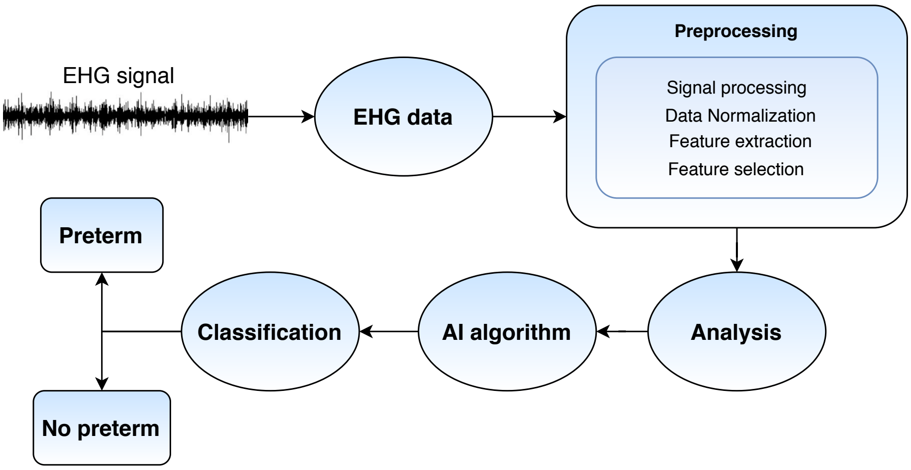

Figure 1. An example of the diagram of preterm birth classification workflow. From top-left: raw EHG signal as an entry to quantifier block, preprocessing step serves as a feature extractor, and then ML algorithm being employed to classify preterm labor.

Fergus et al. [19] focus their attention on the possibility of preterm birth prediction using EHG signals—the type of data that was used for the first time by Manner et al. Fergus' work is based on the TPEHG dataset. In this work, models are developed using various classifiers, namely: density-based, linear, and polynomial-based and nonlinear-based classifiers. The features used include: root mean squares, peak frequency, median frequency, and sample entropy. The best performing classifier oversampled the original TPEHG dataset using the synthetic minority oversampling technique (SMOTE) [38], which is a tree classifier that scores 90% sensitivity, 83% specificity. and an 89% AUC value. The use of new features (clinical data), which was added later to original TPEHG dataset, allows for the generation of a new dataset that consists of cases equally split between preterm and term births. As a result, this significantly improves the accuracy as well as the specificity of SMOTE. The polynomial classifier has a sensitivity of 97%, specificity of 90%, and AUC of 95%.

Unlike Fergus' attempt using traditional ML methods, Hussain et al. [64] improve the result on the original TPEHG dataset, scoring 89% sensitivity, 91% specificity, and a 93% AUC value with a self-organised neural network inspired by the immune algorithm (SONIA) [97]. Such a model has improved generalisation capability over a plain neural network. For balancing the dataset, Hussain oversamples the minority class. The authors also propose a dynamic self-organised network (DSIA), the results of which lie just below SONIA's results.

Sadi-Ahmed et al.'s [86] approach differs from the two previous works by using the Huang–Hilbert transform (HHT) [98]. HHT is a combination of two procedures: empirical mode decomposition (EMD) and Hilbert transform (HT). EMD adaptively decomposes any signal, with no prior knowledge, into a sum of oscillating single components, called intrinsic mode functions (IMF). Basis functions of EMD are specific to the signal, which makes it suitable for the analysis of non-stationary and non-linear signals. HT is used to calculate instantaneous frequency (IF) and amplitude (IA) of each of these IMFs. For the best combination of features, the linear SVM classifier achieves sensitivity 98%, specificity 93%, and AUC 95%. The work's advantage is a high diagnostic performance that is due to the very high sensitivity and low computational cost.

The attempts described so far use 30 min long recordings, while Despotovic et al. [87] achieve better results by splitting the 30-min records into two 15 min ones. What is worth mentioning is that the authors use records that were made between the 22nd and 25th week of gestation, that is very important for the early prediction. They propose new features that exploit the signal's nonstationarity and empirical mode decomposition. The problem of imbalance is solved using an adaptive synthetic sampling (ADASYN) [42] approach for imbalanced learning. A random forest [46] classifier combined with artificial sampling using 10-fold cross-validation on 322 samples, out of which 38 are preterm, achieves 99% accuracy, 98% sensitivity and AUC of 99%. The obtained results are suspiciously good. Perhaps the model has over-fitted the data and cross-validation may not be able to detect it in the case of a small number of data [99]. It is very likely that the model will not generalise well to new data.

Degbedzui et al. [89] propose the classification of EHG signals using the following ML algorithms: K-NN and SVM. The best result obtained is 96.16% accuracy for K-NN and 99.74% for the SVM algorithm. The work is distinguished by the high accuracy that was achieved in classification.

Chen et al. [88] propose a solution that is based on deep neural networks on the TPEHGT dataset [94,100]. This is different from previous solutions that used classic ML algorithms. It obtains the results of 98.2% sensitivity, 97.74% specificity, and accuracy of 97.9%. According to the authors, adding more EHG signals can improve the accuracy of predicting preterm birth.

3.3. Electronic Health Records

Electronic health records (EHR) contain data regarding the course of the pregnancy as well as information about patients medical history and necessary personal data collected through a medical interview. They allow us to assess the risk of premature birth before conception and during pregnancy [12,80]. However, this type of data is very difficult to analyse due to the lack of standardisation among hospitals in the world, which does not allow for an easy automation of predictions and the fast adaptation of the created models for various research centres. Despite these difficulties, many researchers have attempted to analyse the available data to detect risk factors favouring premature birth [11]. It is difficult to compare the results of the following works, because all of the models are trained on different datasets. The only differences that we can highlight are related to data imputation or methods of balancing the dataset. Figure 2 depicts an example of an EHR workflow.

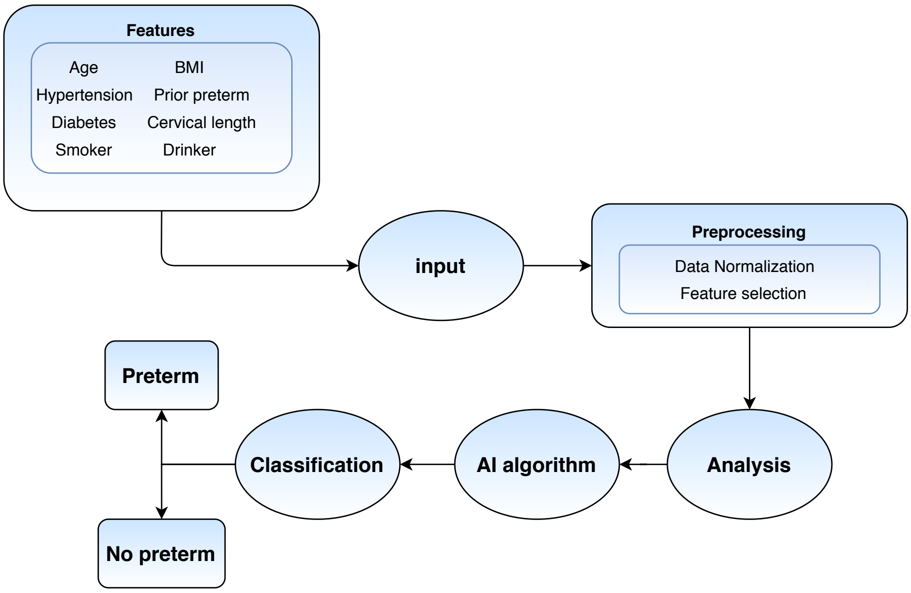

Figure 2. An example of the diagram of preterm birth classification workflow. From top-left: raw electronic health records (EHR) features as an entry to quantifier block, preprocessing step serves as a feature extractor, and then ML algorithm employed to classify preterm labor.

Woolery et al. [76] developed a prototype system for predicting preterm labour risk based on rough set theory, a method for managing uncertainty in knowledge acquisition. Applying that method to the dataset, the authors create 520 usable rules and then validate them by entering them into a dedicated expert system. The proposed system is 53–88% accurate in predicting preterm birth for 9419 patients.

Grzymala-Busse et al.’s [75] research showed that the performance of computer-based methods for the prediction of preterm birth is significantly better than the performance of manual methods. The authors try to identify regularities that are hidden in datasets by utilising the ML program LERS (Learning from Examples based on Rough Sets) [101]. Classification rules for preterm prediction are induced by applying genetic algorithms to the dataset to improve the accuracy. The experimental results show that the prediction rate of rule sets with the appropriate classification scheme is 68–90% accurate.

Mercer et al. [77] aim to develop a risk assessment system for sPTB prediction at 23 to 24 weeks’ gestation. They evaluate more than 100 parameters, like age, medical history, complications of current gestation, body mass, or Bishop score that may be considered to be clinical risk factors for subsequent spontaneous preterm delivery. In this work, the authors apply univariate analysis and multivariate logistic regression on a random selection. Despite many input factors, the risk assessment system has relatively low sensitivity (24.2% and 18.2%) and positive predictive value (28.6% and 33.3%), for multiparous and nulliparous women, respectively. Nulliparous is a term that describes a woman who has not given birth to a child, but it does not mean that she has never been pregnant. Someone who has had a miscarriage, stillbirth, or elective abortion, but has never given birth to a live baby, is referred to as nulliparous.

Goodwin et al. [78] focused on developing tools and techniques to help understand the causes of preterm birth. The authors perform five different modelling techniques that use neural networks, logistic regression, Classification And Regression Trees (CART) [102], and software, called PVRuleMiner and FactMiner. However, the results do not perform as well as previous studies using smaller datasets and inductive ML methods and found only small differences in all proposed methods.

Frize et al. [79] compared two preterm births classification methods: the first is an artificial neural network with weight-elimination (ANN-we); the second is a combined decision-tree (DT) for the elimination of input features that have little impact on the results and artificial neural network with weight-elimination (ANN-we). At first, both methods are evaluated on a relatively small dataset of adults in an intensive care unit (ICU). The performance of both methods is reported as a mean and standard deviation for three output measures: specificity, sensitivity, and ROC. Next, the better performing classifier, the DT-ANN, is applied to a large database collected in the United States before 23 weeks of gestation in order to predict premature births. The dataset contains over 113,000 cases. The superior DT-ANN classifier is also shown to be effective in predicting PTB. For the parous cases, a sensitivity of 66% and specificity of 84% are achieved. The classifier for nulliparous cases achieves a sensitivity of 65% and specificity of 71%.

Vovsha et al. [18] used the “Preterm Prediction Study”, a clinical trial dataset that was collected between 1992 to 1994, which means that current PTB treatments were not then in use, that is why dataset depicts natural incidence of PTB. A valuable feature of the dataset is the presence of screenings performed at an early stage of the pregnancy. In this article, particular emphasis is placed on both predicting preterm birth in nulliparous mothers and understanding its complex etiologies. Vovsha’s approach makes use of SVM with linear and non-linear kernels and logistic regression. The best performing classifier is an SVM with radial basis function (RBF) kernel scores, on average, for all populations: 0.57 sensitivity and 0.69 specificity. The positive aspect of this work is the fact that the best model is tested on previously unseen data. However, Vovsha’s analysis lacks a more elaborate explanation of the approach to the missing data and imputation. What is more, the results may be distorted due to the training of the SVM classifier on an unbalanced set.

Tran et al. [11] paid considerable attention to the proper preparation of data. The research team takes measures to prevent a “leakage” problem—records that implicitly indicate the outcome to be predicted, e.g., they contain procedures and tests that are performed late in the course of the pregnancy, and are not considered as features. Features that occurred before the 25th week of gestation are explicitly extracted. Data are balanced through under-sampling the majority class. This work utilises, as previously proposed by the authors, reducing instability under correlated data, which is a stabilised sparse logistic regression (SSLR). To estimate the upper-bound of model accuracy, Tran uses Randomised Gradient Boosting (RGB)—a hybrid of Random Forests and Stochastic Gradient Boosting. Before building the model, the group visualises dataset by embedding data points into 2D space using t-SNE [103]—that leads them to the assumption that there are no simple linear hyperplanes that can separate the preterm births from the rest. Results are very well presented and documented. The top three risk factors found are multiple fetuses, cervix incompetence and prior preterm births. The highest AUC using RGB is in the range of 0.80–0.81 and the proposed SSLR method is only slightly worse with the AUC in the range of 0.79. The authors also propose a simplified prediction model with only 10 features, achieving a similar overall quality of AUC 0.77, which, in turn, allows for better transparency and interpretability.

Weber et al. [80] used data that were collected on a large number (2+ million) of patients to predict preterm birth. Weber's work focuses on nulliparous women preterm birth. The final analytic sample does not include records with missing critical data, limits the scope to early sPTB (20–32 weeks) and nulliparous women of the desired ethnicity. The authors treat missing data with Multiple Imputation by Chained Equations (MICE) [104]. The following ML methods are used: logistic regression, random forest, k-nearest neighbours, generalised additive models, lasso regression, ridge regression, and elastic net with mixing parameter. All of the algorithms have similar performance (five-fold cross-validation—AUC). The results for logistic regression of cross-validated AUCs are 0.62 and 0.63 for non-Hispanic blacks and whites, respectively. Having combined racial-ethnic groups prediction improves the AUC to 0.67—that is a similar value to others who use bio-markers. This article concludes that the resolution of administrative data is inadequate for the precise prediction of risk for early sPTB, despite the use of advanced statistical methods. The results may be limited by the data, rather than the statistical tools. Perhaps a better feature engineering could improve the results.

Esty et al. [81] analysed two datasets containing data on births. One of the objectives of this work is to create a model that uses data that are collected during maternal prenatal medical visits to surpass the prediction quality of fibronectin marker predictions, for which screening is expensive and highly invasive. The authors note that there is a large number of missing variables in both of the datasets, although they do not provide exactly to what extent. The missing features in both datasets are mostly missing at random, which allows them to be imputed based on other related features in the dataset. However, features with more than 50% of missing data are removed. Imbalance in datasets is cured with down-sampling the majority class. For classification, the authors use a C5.0 Decision Tree [105]. Unfortunately, they do not provide features that the model considers to be the most decisive. The results of the proposed model are sensitivity: 90.9%, specificity: 71.8%, and ROC: 80.9%. According to the authors, their model presents a trade-off between increasing sensitivity and decreasing specificity for the previous work done. An increase in sensitivity is desirable in the case of preterm prediction.

Gao et al. [12] tried to predict extreme preterm birth (EPB)—infants that were born before the 28th week of gestational age. The data include patient demographics, diagnoses and procedures, prescribed medications, and laboratory test results. To solve bias problems that are caused by the unbalanced dataset, the controls in the development are under-sampled. The evaluation set is not balanced. The model preparation workflow consists of four elements: word embedding, cohort construction (30 cohorts), which uses bootstrapping to undersample controls, ML, and medical concept ranking based on statistical models. Gao et al. use both bag of words (BOW) and word embedding to represent each medical concept. For BOW, term frequency-inverse document frequency (TF-IDF) is used to normalise the importance of each medical concept. For word embedding, a skip-gram is used to find the most related words for a given word (representation that can predict the surrounding medical concepts for a particular concept of interest). The created models use linear regression (LR), SVM, and gradient boosting (GB). A long short-term memory (LSTM) [106] model is used to characterise sequential information. For non-neural networks with word embedding, LR has the highest AUC of 0.769, and with a combination of BOW and word embedding, LR also achieves the best performance, with an AUC of 0.780. LSTM using word2vec [107] scores an AUC of 0.811. The authors create an ensemble of the LSTM with word2vec models, which achieves an AUC of 0.827, 96.5% sensitivity, and 69.8% specificity. The positive predictive value (PPV) of the best ensemble model is extremely low, at 3.3%, which indicates that most of the predicted EPB are non-EPB. PPV is equal to the ratio of the number of true positive results to the sum of true and false positives.

Lee et al. [83] proposed the application and comparison of six ML algorithms for the prediction of preterm birth: the artificial neural network (ANN), logistic regression, decision tree, Naive Bayes, random forest, and SVM. The following features are used for the analysis: prior preterm birth, diabetes mellitus, drinker, smoker, hypertension, in vitro fertilisation, age, BMI, parity, and cervical length. The model achieves a classification accuracy of 91.14% using ANN and 91.80% multinominal logistic regression. The study shows that the most important features used for classification are: hypertension, BMI, cervical length, and age. The study was conducted on a small statistical group of 596 women.

Rawashdeh et al. [84] proposed the preterm birth prediction for women with cervical cerclage based on EHR data, which contains such features as age, prior preterm, cervical length before and after cerclage, or uterine anomaly. The authors compare several classifiers, such as random forest, K-NN, and neural networks. In the best variant, they achieve 98% accuracy and specificity at the level of 94%.

Prema et al. [82] proposed a solution that is based on two classifiers: logistic regression and the SVM algorithm. The main risk factors for spontaneous preterm birth, such as age, number of times pregnant, obesity, diabetes mellitus (DM and GDM), and hypertension of the pregnant women are featured as entry into the classification. It shows that the identified risk factors are helpful in sPTB prediction. The results show that both GDM and DM are major preterm birth risks.

Koivu et al. [85] proposed experimentation to discover novel risk models that could be utilised in clinical settings. They use a large dataset of almost 16 million observations. They use three state-of-the-art ML algorithms, such as logistic regression, artificial neural network, and gradient boosting decision tree. The best performing ML algorithms achieved 0.76 AUC for early stillbirth, 0.63 for late stillbirth, and 0.64 for preterm birth.

3.4. Transvaginal Ultrasound

Transvaginal ultrasound is a non-invasive transvaginal imaging examination that is based on ultrasound waves, allowing for visualising and diagnosing many pathologies. It consists of inserting the probe into the vagina, thanks to which the doctor can better visualise and more accurately assess the reproductive organ. Each time, the probe is protected with a disposable latex cover, thanks to which the test is safe for the woman—there is no possibility of transmission of infection. The first work was based on transvaginal ultrasound images that used ML algorithms was the work of Włodarczyk et al. [9]. Włodarczyk proposed the segmentation of the cervix using the convolutional neural network—U-Net [108]. On the segmentation results, he estimates two biomarkers—CL (cervical length) and ACA (anterior cervical angle). Both of the biomarkers are used to manually assess the risk of spontaneous preterm birth (sPTB) by doctors [109]. Subsequently, he classifies the obtained results using classic ML algorithms, such as SVM or Naive Bayes. The best results are obtained by Naive Bayes algorithm: accuracy of 7.5%, precision of 85%, recall of 74%, and AUC of 78.13%. The sPTB prediction results are better than those that were presented in [109] by gynaecologists. In the second study, Włodarczyk et al. [10] proposed an end-to-end solution. It is a simultaneous segmentation and classification of the cervix on transvaginal ultrasound images. The solution is based on a convolutional neural network that was inspired by the Y-Net network [110]. The paper has state-of-the-art results in IoU, recall, and precision on transvaginal ultrasound images. Figure 3 depicts an example of workflow utilising transvaginal ultrasound.

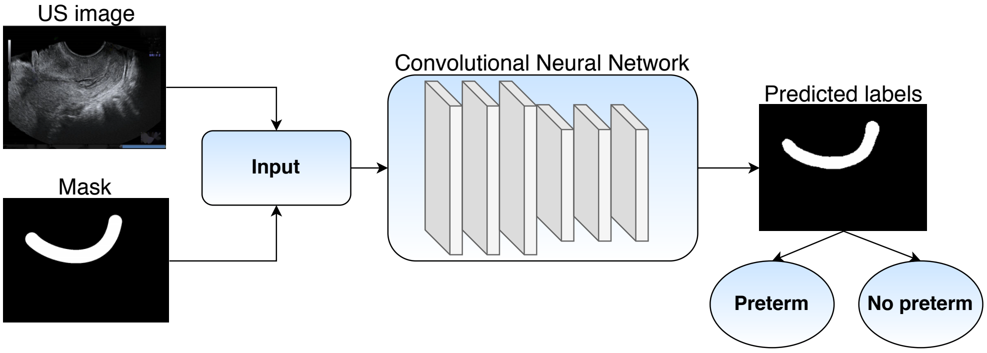

Figure 3. An example of a diagram of the preterm birth classification. From left: transvaginal ultrasound image with a mask showing cervix as entry into convolutional neural network, simplified diagram of convolutional neural network and segmentation results with *preterm/control* class as a outcome.

3.5. Uterine Electromyography

Uterine electromyography (EMG) is electrical activity of the myometrium, or uterine muscle, which is responsible for myometrial contractions.

Maner et al. [90] addressed the problem of classification preterm labour pregnant patients in gestational ages that ranged from 24 to 41 weeks, by using an artificial neural network (ANN) on uterine electromyography (EMG) data. The authors use spectral analysis to quantify EMG signals by finding the means and standard deviations of the peak frequency and then measure parameters, like burst duration, number of bursts per unit time, and total activity. All of those extracted variables serve as inputs to the Kohonen ANN, which groups the output data into four categories: term labour (TL), term non-labour (TN), preterm labour (PTL), and preterm non-labour (PTN). The obtained accuracy of correctly identified patients is 79%, 92%, 86%, and 71%, for TL, PTL, TN, and PTN, respectively.

Most et al. [91] also focussed on analysing the electrical activity of the uterine muscle. The authors study non-invasive transabdominal electrical uterine myographic monitoring (EUM) and utilise it as a preterm birth predictor with a comparison of fetal fibronectin (fFN) and cervical length measurement. To evaluate this hypothesis, they perform logistic regression and obtain the following results: 41% sensitivity and 92% specificity. It proves that a combination of EUM and the history of preterm delivery is significantly correlated to the risk of PTB. Furthermore, a combination of the following three risk factors: EUM, fFN, and cervical length (CL) can increase the negative predictive value from 79% for the only fFN to 92% for all inputs.

4. Discussion

Despite global efforts to prevent PTB, the worldwide rate of newborns delivered prematurely has remained stable throughout the last 30–40 years. Many studies and screening models were prepared to reduce this rate. However, millions of neonates continue to be born before 37 weeks of gestation every year. Preterm delivery is associated with an increased risk of respiratory distress syndrome (RDS), intracranial haemorrhage (ICH), necrotising enterocolitis (NEC), or cerebral palsy. Complications that are associated with prematurity may lead to neonatal death, impaired neurodevelopment, and long-term disability. Because we may prevent some of the PTB due to population screening, further steps need to be taken to improve current strategies.

It is noteworthy that prenatal care standards differ worldwide. In some countries, every pregnant patient is subjected to for the three ultrasound scans (e.g., Poland), in others two scans (e.g., the United Kingdom), and in some countries only one scan in the second trimester (e.g., Norway). Regardless of the country, every patient shall be offered a mid-trimester scan to measure the cervical length (CL). The length of the uterine cervix strictly corresponds to the risk of PTB—the shorter cervix, the higher risk of PTB. Several studies show that the cut off value for the increased risk of PTB is 25 mm. When the CL at the mid-trimester scan (between 18 + 0 and 22 + 0 weeks) is less than 25 mm, the patient is at higher risk of PTB, and shall remain under strict prenatal surveillance. In high-risk group patients, we may administer means that have proven efficacy in prolonging the gestation, such as progesterone or reducing adverse neonatal outcomes, such as the administration of corticosteroids.

In spite of proper screening based on CL measurement, current screening strategies cannot identify a significant percentage of patients who will deliver prematurely. This fact points out that we shall continue extensive study in this field and look for alternative solutions. One of the most popular theories associated with preterm delivery is the one that was proposed by Romero et al. [111]. According to this theory, we should analyse the preterm birth from a broader perspective, not only by shortening the uterine cervix, but also by including other factors, such as immunological factors, uterine distention, or inflammation. Unfortunately, we cannot predict whether one of these additional risk factors will occur, and it seems that cervical measurement is the best tool for predicting preterm birth. There were many studies that analysed the practical use of biochemical markers of preterm birth detected in the vagina (e.g., insulin-like growth factor-1 (IGF-1), insulin-like growth factor-binding protein IGFBP-1, or fetal fibronectin) [112,113]. Although tje results pointed to statistically significant differences in these markers' expression among patients who delivered prematurely, they have not become a part of routine screening. We may apply many methods to assess the risk of preterm birth. However, it seems that cervical measurement is the only method applied worldwide, and we shall focus on increasing its accuracy.

The length of the cervical canal is the key element of the cervical measurement. The risk of preterm birth is increased in patients with a cervical length canal shorter than 25 mm at the mid-trimester scan. However, the cervical length does not identify every patient who will deliver soon. From our perspective, ultrasound images that are taken during transvaginal measurement of the uterine cervix provide much more data than cervical length alone. According to this point of view, more studies were evaluating additional aspects of these ultrasound images. Sochacki et al. [109] proposed measuring the anterior cervical angle to identify patients with an increased risk of preterm birth. Volpe et al. [114] described a so-called cervical sliding sign in patients endangered with preterm birth. Banos et al. [115] reported that the mid-trimester cervical consistency index performs better than CL alone in the population at high risk of preterm birth. The last two research studies point out that looking for new markers of preterm birth provided in the ultrasound images makes sense, and we should continue our efforts.

ML methods allow for us to analyse data on the binary level, which is impossible for a human eye. We believe that, when we incorporate various uterine cervix assessment methods into the deep learning networks, we may improve the detection rates and observe new phenomena based on the computer analysis of certain areas in the uterine cervix. Cervical length measurement is based on the subjective assessment of the sonographer. For inexperienced hands, the intraobserver measurement differences can be as high as 4 mm. However, when we consider how many factors affect the final result, a new standardisation method of the cervix's objective evaluation may improve perinatal care quality. ML may help clinicians in further management of the patient's treatment. After a primary assessment, deep learning networks could assist medical professionals in evaluating the obtained images.

Moreover, a dynamic evaluation of the uterine cervix could be possible due to the deep learning networks' application. Ultrasound images only provide a single sample for each patient, while short cine loops with hundreds of frames may be analysed successively with dedicated algorithms. As we mentioned before, we believe that the place of ML in this field begins where the abilities of the human eye end. Thanks to the frame-by-frame investigation, which would be very time-consuming, we may obtain additional data on how the cervical tissue responds to applied pressure with the ultrasound probe and investigate changes among the cervix automatically.

From our perspective, machines are meant not to replace humans in medicine, but to support them. Deep learning methods may help to evaluate ultrasound images in a novel way, especially where the human eye is not enough. We shall develop further studies investigating these ideas' clinical application and evaluate whether they improve perinatal results. Unfortunately, ML methods have weaknesses and limitations in predicting spontaneous preterm birth. The main limitation is obtaining an extensive and good-quality database that would allow new correlations between features to be tested on a larger statistical sample. An additional limitation is the collection time of the dataset—researchers have to wait until the pregnancy is completed to find out about the outcome, health, and other characteristics of the baby that may be useful in the analysis, such as weight and week at birth. The research paths in which we can see the potential may be ultrasound imaging during the abdominal examination and transvaginal data. Ultrasound images in solving the problem of predicting spontaneous preterm births are not well researched yet. There are many research opportunities and discoveries for new biomarkers in ultrasound images. An example of the possibility of further work on ultrasound images may be the examination of tissue density around the cervical canal or fetal biometry.

In this work, we present a table that compares the datasets of analysed articles in Table 1. Among reviewed works, we distinguish dataset size—number of patients or pregnancies—the percentage of spontaneous preterm birth in this group and gestation week of pregnant women during an examination or data recording, depending on the data type. It can be seen that datasets significantly vary in size. Electronic health records are the easiest to obtain in large amounts and, very often, they do not need annotations. However, they may contain a lot of noise and information that itself suggest the result of pregnancy—so the most effort must be put into data preparation. A common feature of almost all datasets is connected with preterm birth specificity, which is unbalanced data because of an average 10% frequency of sPTB. The researchers, predominantly to counter this difficulty, reach out for the oversampling techniques, but they rarely point out that it creates artificial data that may not have much in common with real observations—instead, they focus on better accuracy results.

Difficulties of collecting preterm data arise not only from the specificity of the phenomenon of the premature birth itself. Gathering medical data, in general, is connected with important issues that are associated with ethical aspects, data privacy, and accountability. ML's use in medicine is accompanied by relevant challenges beyond the simple use of the algorithms. Medical and human health work touches sensitive topics that require high standards and accuracy. One issue that requires special attention is that of ethics. In many countries, the Medical Ethics Committee's approval is required to conduct preterm birth research [116]. Moreover, the use of the collected data requires additional preparatory steps, e.g., to implement ML projects, all medical and personal data should be anonymised according to GDPR standards [117]. Finally, attention should also be paid to the purpose and use of the tools created. ML systems should only support medical diagnosis and not replace the doctor to whom the final decision belongs.

In our work, we also provide a summary table of all the publications that we are analysing (Table 3). It makes a comparison of the obtained results and used data types easier. We can see that 75% of publications come from the 2010s, which may represent an increase in interest in the topic, but this phenomenon is also correlated with the ever-increasing possibilities of using machine learning to predict preterm birth. The most represented data types are EHR and EHG with the latter primarily due to the publicly available TPEHG dataset [73,94]. What might be found interesting is access to other datasets. Some of them can be directly downloaded, while others require written enquiry. To our best knowledge, the currently available datasets sources are:

- National Institute of Child Health and Human Development (NICHD)—Maternal-Fetal Medicine Units Network (MFMU) (https://mfmunetwork.bsc.gwu.edu/PublicBSC/MFMU/MFMUPublic/datasets/ (accessed on 28 December 2020)),
- Better Outcomes Registry Network (BORN) Information System (https://www.bornontario.ca/en/data/data-dictionary-and-library.aspx (accessed on 28 December 2020)),
- Pregnancy Risk Monitoring Assessment (PRAMS) (https://www.cdc.gov/prams/state-success-stories/data-to-action-success.html (accessed on 28 December 2020)),
- Centers for Disease Control and Prevention (CDC)—National Center of Health Statistics (NCHS) (https://www.cdc.gov/nchs/data_access/ftp_data.htm (accessed on 28 December 2020)),
- Term-Preterm EHG Database (TPEHG) (https://physionet.org/content/tpehgdb/1.0.1/ (accessed on 28 December 2020)), and
- Term-Preterm EHG Dataset with Tocogram (TPEHGT) (https://physionet.org/content/tpehgt/1.0.0/ (accessed on 28 December 2020)) [94,100].

To summarise the recent results obtained for each type of data, one can distinguish:

- EHG—Degbedzui et al. [89] who achieve accuracy of 0.997, recall 0.995, and specificity 1.0—using SVM classifier,
- EHR—Rawashdeh et al. [84] who achieve accuracy of 0.95, recall 1.0, and specificity 0.94—using random forest,
- TVS—Włodarczyk et al. [10] who achieve a recall of 0.68, and specificity 0.97—using convolutional neural networks, and
- EMG—Most et al. [91] who achieve a recall of 0.41 and specificity 0.92—using logistic regression.

We must remember that these results should be interpreted in the context of a data set. From these results, we also conclude which tools to use, depending on the type of data. Generally, for classification, the best choice is to use SVM or random forest. If we care about automatic feature extraction, it is advisable to use neural networks. For image data, the best results are obtained while using convolutional neural networks.

Many reviewed works use TPEHG dataset in which researchers compare their results with previous works and notice the improvement in accuracy of a few per cent or less and such trend repeats for many following years with new publications that are based on TPEHG. One should realise that the dataset consists of only 38 preterm births (Table 2). One preterm sample out of 38 available is 2.5% of the whole preterm group. One of the important things that can be deduced from this is that one should not focus on accuracy only, but rather more on sensitivity. What is more, one should treat these works more as ML methods comparison, rather than a new model proposal that is encouragingly more accurate and will make a better decision assisting system. It would be beneficial if researchers could get bigger EHG dataset, which would verify whether models created on TPEHG dataset generalise well, because the accuracy of almost 99% may raise suspicion of data overfitting. Another thing to consider is that researchers attempting preterm birth topic should point out which age of gestation patient's examination or recordings were performed. Because of the high mortality of infants born prematurely before the 28th week of gestation, one should focus the most on examinations taken early. Only such ones allow for us to create models that will help doctors to take precautions. The mentioned TPEHG dataset has only 19 preterm deliveries in the group examined early around 23rd week of gestation. It is known that such annotated data are challenging to obtain. However, it seems of little benefit to science to exploit the same TPEHG dataset more in the future.

Table 3. Comparison of the papers based on the obtained results and used data type, sorted by year.

Author	Methods	Results	Data Type	Year
Woolery et al. [76]	LERS, ID3 Tree	Accuracy: 0.53–0.88	EHR	1994
Grzymała-Busse et al. [75]	LERS, genetic algorithm	Accuracy: 0.68–0.90	EHR	1994
Mercer et al. [77]	Univariate analysis and multivariate logistic regression	Recall: 0.18-0.24, Precision: 0.29–0.33	EHR	1996
Goodwin et al. [78]	Neural networks, CART, logistic regression	AUC = 0.76	EHR	2000
Maner et al. [90]	FFT, Kohonen Network	Accuracy = 0.82	EHG	2007

Most et al. [91]	Bivariate analysis, CHI square logistic regression, Fisher's test	Recall = 0.41,
		Specificity = 0.92
			EMG	2008

Frize et al. [79]	Neural network, decision tree	Recall: 0.65–0.66, Specificity: 0.71–0.84	EHR	2011

Fergus et al. [19]	K-NN, decision trees, SVM	AUC = 0.95,
		Recall = 0.97,
		Specificity = 0.90
			EHG	2013

Vovsha et al. [18]	Logistic regression, SVM	Recall = 0.57,
		Specificity = 0.69
			EHR	2014

Hussain et al. [64]	DSIA (Dynamic Self-Organised Network), benchmark: SONIA, MLP, Fuzzy-SONIA, K-NN	AUC = 0.93,
		Recall = 0.89,
		Specificity = 0.91
			EHG	2015

Tran et al. [11]	Logistic regression, randomised gradient boosting, stochastic gradient boosting, random forest	AUC = 0.81	EHR	2016

Sadi-Ahmed et al. [86]	Huang-Hilbert transform (HHT), IMF, SVM	AUC = 0.95,
		Recall = 0.99,
		Specificity = 0.98
			EHG	2017

Weber et al. [80]	Super learning (SL), K-NN, random forest, lasso regression, ridge regression, elastic net, Generalised Additive Models (GAM)	AUC = 0.67	EHR	2018

Despotovic et al. [87]	Random forest, K-NN, SVM	Accuracy = 0.99,
		AUC = 0.99,
		Recall = 0.98
			EHG	2018

Esty et al. [81]	Decision trees, neural networks	AUC = 0.81,
		Recall = 0.91,
		Specificity = 0.72
			EHR	2018

Gao et al. [12]	BOW and word embedding (NLP), recurrent neural network (RNN), regularised logistic regression	AUC = 0.83,
		Recall = 0.966,
		Specificity = 0.70
			EHR	2019

Włodarczyk et al. [9]	Convolutional neural network (CNN), SVM, K-NN, Naive Bayes, Decision trees	Accuracy = 0.78,
		AUC = 0.78,
		Recall = 0.74,
		Precision = 0.85
			TVS	2019

Prema et al. [82]	SVM, logistic regression	Accuracy = 0.76,
		Recall = 0.84,
		Specificity = 0.73,
		Precision = 0.84
			EHR	2019

Table 3. *Cont.*

Author	Methods	Results	Data Type	Year
Lee et al. [83]	Naive Bayes, neural networks, SVM, logistic regression, decision trees, random forest	Accuracy = 0.92	EHR	2019

Chen et al. [88]	Wavelet entropy, Stacked Sparse Autoencoder (SSAE)	Accuracy = 0.98,
		Recall = 0.98,
		Specificity = 0.98
			EHG	2020

Degbedzui et al. [89]	SVM	Accuracy = 0.997,
		Recall = 0.995,
		Specificity = 1.0
			EHG	2020

Rawashdeh et al. [84]	Naive Bayes, decision trees, K-NN, random forest, neural networks	Accuracy = 0.95,
		AUC = 0.98,
		Recall = 1.0,
		Specificity = 0.94
			EHR	2020

Włodarczyk et al. [10]	CNN - FCN, DeepLab, U-Net	Recall = 0.68,
		Specificity = 0.97
			TVS	2020

Koivu et al. [85]	Logistic regression, neural networks, gradient boosting	AUC = 0.64	EHR	2020

In summary, an individual interested in developing a preterm birth prediction system should take the following steps. Firstly, the method of collecting data should be consulted between physicians and data scientists. It is crucial to collect universal data so that as many medical centres as possible can use it. An important aspect is to try to create the most balanced data. When working with imbalanced preterm birth data, that are naturally skewed because of PTB occurrence, we need to use the methods described in Section 2.3. Solving class imbalance is here an important issue, which will largely determine the quality of our model. Next, it is necessary to prepare it properly, i.e., perform data anonymisation, carry out the dataset's filtration, which consists of removing outliers and imputing missing values. The last stage is the proper selection of the machine learning algorithm for both classification and feature extraction. The learning process should be carried out with the division into training, validation, and test sets. The test set must not be used to adjust any parameters of the model. In the case of preterm births, we need to primarily focus on sensitivity and precision, but not solely on accuracy. We should be aware that a good result on the test set does not guarantee success in applying the model in real life.

5. Conclusions

In this paper, we summarise the current application of ML methods that may be incorporated into perinatal medicine. Afterwards, we outline the main four methods used in preterm birth prediction: electrohysterography, electronic health records, transvaginal ultrasound, and uterine electromyography. We present various works concerning each method and compare them based on the achieved results. The reviewed studies suggest that ML methods can improve preterm birth detection rates and contribute additional information to identify women with true sPTB. Finally, they can also produce a powerful, objective tool for assessing labour and earlier intervention.

Author Contributions: Conceptualization, T.W., S.P. and T.S.; methodology, T.W., S.P. and T.S.; writing—original draft preparation, T.W., S.P. and T.S.; writing—review and editing, T.T.; supervision, T.T., P.R., N.S.-W., J.W. and M.L.; funding acquisition, T.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by Warsaw University of Technology Dean's grant number II/2019/GD/2 and Grant of Scientific Discipline of Computer Science and Telecommunications at Warsaw University of Technology.

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

References

1. Blencowe, H.; Cousens, S.; Oestergaard, M.Z.; Chou, D.; Moller, A.B.; Narwal, R.; Adler, A.; Garcia, C.V.; Rohde, S.; Say, L.; et al. National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: A systematic analysis and implications. *Lancet* **2012**, *379*, 2162–2172. [CrossRef]
2. Liu, L.; Johnson, H.L.; Cousens, S.; Perin, J.; Scott, S.; Lawn, J.E.; Rudan, I.; Campbell, H.; Cibulskis, R.; Li, M.; et al. Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. *Lancet* **2012**, *379*, 2151–2161. [CrossRef]
3. Dbstet, A. WHO: Recommended definitions, terminology and format for statistical tables related to the perinatal period and use of a new certificate for cause of perinatal deaths. *Acta Obstet Gynecol Scand* **1977**, *56*, 247–253.
4. Blencowe, H.; Cousens, S.; Chou, D. Born Too Soon: The global epidemiology of 15 million preterm births. *Reprod. Health 10* **2013**, *10*, 1–14. [CrossRef] [PubMed]
5. Marlow, N. Full term; an artificial concept. *Arch. Dis. Childhood Fetal Neonatal* **2012**, F158–F159. [CrossRef] [PubMed]
6. Goldenberg, R.L.; Gravett, M.G.; Iams, J.; Papageorghiou, A.T.; Waller, S.A.; Kramer, M.; Culhane, J.; Barros, F.; Conde-Agudelo, A.; Bhutta, Z.A.; et al. The preterm birth syndrome: Issues to consider in creating a classification system. *Am. J. Obstet. Gynecol.* **2012**, *206*, 113–118. [CrossRef] [PubMed]
7. Ward, R.M.; Beachy, J.C. Neonatal complications following preterm birth. *BJOG Int. J. Obstet. Gynaecol.* **2003**, *110*, 8–16. [CrossRef]
8. Okitsu, O.; Mimura, T.; Nakayama, T.; Aono, T. Early prediction of preterm delivery by transvaginal ultrasonography. *Ultrasound Obstet. Gynecol.* **1992**, *2*, 402–409. [CrossRef] [PubMed]
9. Włodarczyk, T.; Płotka, S.; Trzciński, T.; Rokita, P.; Sochacki-Wójcicka, N.; Lipa, M.; Wójcicki, J. Estimation of Preterm Birth Markers with U-Net Segmentation Network. In *Smart Ultrasound Imaging and Perinatal, Preterm and Paediatric Image Analysis*; Wang, Q., Gomez, A., Hutter, J., McLeod, K., Zimmer, V., Zettinig, O., Licandro, R., Robinson, E., Christiaens, D., Turk, E.A., et al., Eds.; Springer: Cham, Switzerland, 2019; pp. 95–103. [CrossRef]
10. Włodarczyk, T.; Płotka, S.; Rokita, P.; Sochacki-Wójcicka, N.; Wójcicki, J.; Lipa, M.; Trzciński, T. Spontaneous Preterm Birth Prediction Using Convolutional Neural Networks. In *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*; Hu, Y., Licandro, R., Noble, J.A., Hutter, J., Aylward, S., Melbourne, A., Turk, E.A., Barrena, J.T., Eds.; Springer: Cham, Switzerland, 2020; pp. 274–283. [CrossRef]
11. Tran, T.; Luo, W.; Phung, D.; Morris, J.; Rickard, K.; Venkatesh, S. Preterm birth prediction: stable selection of interpretable rules from high dimensional data. In Proceedings of the 1st Machine Learning for Healthcare Conference, Los Angeles, CA, USA, 19–20 August 2016; pp. 164–177.
12. Gao, C.; Osmundson, S.; Edwards, D.R.V.; Jackson, G.P.; Malin, B.A.; Chen, Y. Deep learning predicts extreme preterm birth from electronic health records. *J. Biomed. Inf.* **2019**, *100*, 103334. [CrossRef] [PubMed]
13. Boser, B.E.; Guyon, I.M.; Vapnik, V.N. A training algorithm for optimal margin classifiers. In Proceedings of the COLT '92: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, Pittsburgh, PA, USA, 27–29 July 1992; pp. 144–152. [CrossRef]
14. Bengio, Y. *Learning Deep Architectures for AI*; Now Publishers Inc.: Norwell, MA, USA, 2009; pp. 1–127. [CrossRef]
15. Walani, S.R. Global burden of preterm birth. *Int. J. Gynecol. Obstet.* **2020**, *150*, 31–33. [CrossRef] [PubMed]
16. Beck, S.; Wojdyla, D.; Say, L.; Bertran, A.P.; Meraldi, M.; Requejo, J.H.; Rubens, C.; Menon, R.; Look, P.V. The worldwide incidence of preterm birth: A systematic review of maternal mortality and morbidity. *Bull. World Health Organ.* **2010**, *88*, 31–38. [CrossRef]
17. Institute of Medicine. *Preterm Birth: Causes, Consequences, and Prevention*; National Academies Press: Washington, DC, USA, 2007. [CrossRef]
18. Vovsha, I.; Rajan, A.; Salleb-Aouissi, A.; Raja, A.; Radeva, A.; Diab, H.; Tomar, A.; Wapner, R. Predicting preterm birth is not elusive: Machine learning paves the way to individual wellness. In Proceedings of the 2014 AAAI Spring Symposium Series, Palo Alto, CA, USA, 24–26 March 2014; pp. 82–89.
19. Fergus, P.; Cheung, P.; Hussain, A.; Al-Jumeily, D.; Dobbins, C.; Iram, S. Prediction of Preterm Deliveries from EHG Signals Using Machine Learning. *PLoS ONE* **2013**, *8*, e77154. [CrossRef]
20. Glover, A.V.; Manuck, T.A. Screening for spontaneous preterm birth and resultant therapies to reduce neonatal morbidity and mortality: A review. *Semin. Fetal Neonatal Med.* **2018**, *23*, 126–132. [CrossRef] [PubMed]
21. Tang, Y.; Zhang, Y.Q.; Chawla, N.; Krasser, S. SVMs Modeling for Highly Imbalanced Classification. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **2008**, *39*, 281–288. [CrossRef] [PubMed]
22. Johnson, J.M.; Khoshgoftaar, T.M. Survey on deep learning with class imbalance. *J. Big Data* **2019**, *6*, 1–54. [CrossRef]
23. Guo, X.; Yin, Y.; Dong, C.; Yang, G.; Zhou, G. On the class imbalance problem. In Proceedings of the 2008 Fourth International Conference on Natural Computation, Jinan, China, 18–20 October 2008; pp. 192–201.
24. Mac Namee, B.; Cunningham, P.; Byrne, S.; Corrigan, O.I. The problem of bias in training data in regression problems in medical decision support. *Artif. Intell. Med.* **2002**, *24*, 51–70. [CrossRef]
25. Grzymala-Busse, J.W.; Goodwin, L.K.; Zhang, X. Increasing sensitivity of preterm birth by changing rule strengths. *Pattern Recognit. Lett.* **2003**, *24*, 903–910. [CrossRef]

26. Kotsiantis, S.; Kanellopoulos, D.; Pintelas, P. Handling imbalanced datasets: A review. *GESTS Int. Trans. Comput. Sci. Eng.* **2006**, *30*, 25–36.
27. Wang, S.; Liu, W.; Wu, J.; Cao, L.; Meng, Q.; Kennedy, P.J. Training deep neural networks on imbalanced data sets. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 4368–4374.
28. Buda, M.; Maki, A.; Mazurowski, M.A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw.* **2018**, *106*, 249–259. [CrossRef]
29. Wang, S.; Yao, X. Multiclass imbalance problems: Analysis and potential solutions. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **2012**, *42*, 1119–1130. [CrossRef] [PubMed]
30. Bi, J.; Zhang, C. An empirical comparison on state-of-the-art multi-class imbalance learning algorithms and a new diversified ensemble learning scheme. *Knowl. Based Syst.* **2018**, *158*, 81–93. [CrossRef]
31. Japkowicz, N. Concept-learning in the presence of between-class and within-class imbalances. In *Conference of the Canadian Society for Computational Studies of Intelligence*; Stroulia E., Matwin S., Eds.; Springer: Berlin, Germany, 2001; pp. 67–77.
32. Japkowicz, N.; Stephen, S. The class imbalance problem: A systematic study. *Intell. Data Anal.* **2002**, *6*, 429–449. [CrossRef]
33. Kubat, M.; Matwin, S. Addressing the curse of imbalanced training sets: one-sided selection. In Proceedings of the Fourteenth International Conference on Machine Learning (ICML 1997), Nashville, TN, USA, 8–12 July 1997; pp. 179–186.
34. Holte, R.C.; Acker, L.; Porter, B.W. Concept Learning and the Problem of Small Disjuncts. In Proceedings of International Joint Conference on Artificial Intelligence (IJCAI-89), Detroit, MI, USA, 20–25 August 1989; pp. 813–818.
35. Liu, X.Y.; Wu, J.; Zhou, Z.H. Exploratory Undersampling for Class-Imbalance Learning. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **2008**, *39*, 539–550. [CrossRef]
36. Van Hulse, J.; Khoshgoftaar, T.M.; Napolitano, A. Experimental perspectives on learning from imbalanced data. In Proceedings of the 24th international conference on Machine learning, Corvallis, OR, USA, 20–24 June 2007; pp. 935–942.
37. Barandela, R.; Valdovinos, R.M.; Sánchez, J.S.; Ferri, F.J. The imbalanced training sample problem: Under or over sampling? In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*; Fred A., Caelli T.M., Duin R.P.W., Campilho A.C., de Ridder D., Eds.; Springer: Berlin, Germany, 2004; pp. 806–814.
38. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
39. Chang, F.; Ma, L.; Qiao, Y. Target Tracking Under Occlusion by Combining Integral-Intensity-Matching with Multi-block-voting. In *Lecture Notes in Computer Science*; Huang, D.S., Zhang, X.P., Huang, G.B., Eds.; Springer: Berlin, Germany, 2005; pp. 77–86. [CrossRef]
40. Bunkhumpornpat, C.; Sinapiromsaran, K.; Lursinsap, C. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*; Theeramunkong, T., Kijsirikul, B., Cercone, N., Ho, T.B., Eds.; Springer: Berlin, Germany, 2009; pp. 475–482.
41. Jo, T.; Japkowicz, N. Class imbalances versus small disjuncts. *ACM SIGKDD Explor. Newsl.* **2004**, *6*, 40–49. [CrossRef]
42. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–8 June 2008; pp. 1322–1328.
43. Chawla, N.V.; Japkowicz, N.; Kotcz, A. Special issue on learning from imbalanced data sets. *ACM SIGKDD Explor. Newsl.* **2004**, *6*, 1–6. [CrossRef]
44. Huang, C.; Li, Y.; Loy, C.C.; Tang, X. Learning Deep Representation for Imbalanced Classification. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, Las Vegas, NV, USA, 27–30 June 2016; pp. 5375–5384.
45. Drummond, C.; Holte, R.C. C4. 5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. In Proceedings of the Workshop on Learning from Imbalanced Datasets II, ICML, Washington, DC, USA, 21 August 2003; pp. 1–8.
46. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
47. Chen, C.; Liaw, A.; Breiman, L. Using Random Forest to Learn Imbalanced Data. Available online: https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf (accessed on 28 December 2020).
48. Krawczyk, B. Learning from imbalanced data: open challenges and future directions. *Prog. Artif. Intell.* **2016**, *5*, 221–232. [CrossRef]
49. Freund, Y.; Schapire, R.E. Experiments with a new boosting algorithm. In Proceedings of the Thirteenth International Conference (ICML '96), Bari, Italy, 3–6 July 1996; pp. 148–156.
50. Joshi, M.V.; Kumar, V.; Agarwal, R.C. Evaluating boosting algorithms to classify rare classes: Comparison and improvements. In Proceedings 2001 IEEE International Conference on Data Mining, San Jose, CA, USA, 29 November–2 December 2001; pp. 257–264.
51. Chawla, N.V.; Lazarevic, A.; Hall, L.O.; Bowyer, K.W. SMOTEBoost: Improving Prediction of the Minority Class in Boosting. In *Knowledge Discovery in Databases: PKDD 2003*; Lavrac, N., Gamberger, D., Todorovski, L., Blockeel, H., Eds.; Springer: Berlin, Germany, 2003; pp. 107–119. [CrossRef]
52. Viola, P.; Jones, M. Fast and robust classification using asymmetric adaboost and a detector cascade. *Adv. Neural Inf. Process. Syst.* **2001**, *14*. Available online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.332.9301&rep=rep1&type=pdf (accessed on 3 March 2021).

53. Fan, W.; Stolfo, S.J.; Zhang, J.; Chan, P.K. AdaCost: Misclassification cost-sensitive boosting. In Proceedings of the Sixteenth International Conference on Machine Learning (ICML 1999), Bled, Slovenia, 27–30 June 1999; pp. 97–105.
54. Domingos, P. Metacost: A general method for making classifiers cost-sensitive. In Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 15–18 August 1999; pp. 155–164.
55. Shen, L.; Lin, Z.; Huang, Q. Relay backpropagation for effective learning of deep convolutional neural networks. In *European Conference on Computer Vision*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer: Cham, Switzerland, 2016, pp. 467–482.
56. Altman, N.S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **1992**, *46*, 175–185. [CrossRef]
57. Rokach, L.; Maimon, O. Decision trees. In *Data Mining and Knowledge Discovery Handbook*; Maimon, O., Rokach, L., Eds; Springer: Boston, MA, USA, 2005; pp. 165–192.
58. Ho, T.K. Random decision forests. In Proceedings of 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 14–16 August 1995; pp. 278–282.
59. Cramer, J.S. The origins of logistic regression. *SSRN* **2002**. [CrossRef]
60. Friedman, J.H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367–378. [CrossRef]
61. Anthony, M.; Bartlett, P.L. *Neural Network Learning: Theoretical Foundations*; Cambridge University Press: Cambridge, UK, 2009.
62. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, UK, 2016.
63. Zeitlin, J.; Szamotulska, K.; Drewniak, N.; Mohangoo, A.; Chalmers, J.; Sakkeus, L.; Irgens, L.; Gatt, M.; Gissler, M.; et al. Preterm birth time trends in Europe: A study of 19 countries. *BJOG Int. J. Obstet. Gynaecol.* **2013**, *120*, 1356–1365. [CrossRef] [PubMed]
64. Hussain, A.; Fergus, P.; Al-Askar, H.; Al-Jumeily, D.; Jager, F. Dynamic neural network architecture inspired by the immune algorithm to predict preterm deliveries in pregnant women. *Neurocomputing* **2015**, *151*, 963–974. [CrossRef]
65. Menard, M.; Newman, R.B.; Keenan, A.; Ebelingc, M. Prognostic significance of prior preterm twin delivery on subsequent singleton pregnancy. *Am. J. Obstet. Gynecol.* **1996**, *174*, 1429–1432. [CrossRef]
66. Facco, F.L.; Nash, K.; Grobman, W.A. Are women who have had a preterm twin delivery at greater risk of preterm birth in a subsequent singleton pregnancy? *Am. J. Obstet. Gynecol.* **2007**, *197*, 253.e1–253.e3. [CrossRef] [PubMed]
67. Heath, V.C.F.; Southall, T.R.; Souka, A.P.; Elisseou, A.; Nicolaides, K.H. Cervical length at 23 weeks of gestation: prediction of spontaneous preterm delivery. *Ultrasound Obstet. Gynecol.* **1998**, *12*, 312–317. [CrossRef] [PubMed]
68. Quinn, J.A.; Munoz, F.M.; Gonik, B.; Frau, L.; Cutland, C.; Mallett-Moore, T.; Kissou, A.; Wittke, F.; Das, M.; Nunes, T.; et al. Preterm birth: Case definition & guidelines for data collection, analysis, and presentation of immunisation safety data. *Vaccine* **2016**, *34*, 6047–6056. [CrossRef]
69. Renzo, G.D.; O Herlihy, C.; van Geijn, H.; Copray, F. Organization of perinatal care within the European community. *Eur. J. Obstet. Gynecol. Reprod. Biol.* **1992**, *45*, 81–87. [CrossRef]
70. Zeitlin, J.; Papiernik, E.; Bréart, G. Regionalization of perinatal care in Europe. *Semin. Neonatol.* **2004**, *9*, 99–110. [CrossRef]
71. Iams, J.D.; Romero, R.; Culhane, J.F.; Goldenberg, R.L. Primary, secondary, and tertiary interventions to reduce the morbidity and mortality of preterm birth. *Lancet* **2008**, *371*, 164–175. [CrossRef]
72. Skirton, H.; Goldsmith, L.; Jackson, L.; Lewis, C.; Chitty, L. Offering prenatal diagnostic tests: European guidelines for clinical practice. *Eur. J. Hum. Genet.* **2013**, *22*, 580–586. [CrossRef] [PubMed]
73. Fele-Žorž, G.; Kavšek, G.; Novak-Antolič, Ž.; Jager, F. A comparison of various linear and non-linear signal processing techniques to separate uterine EMG records of term and pre-term delivery groups. *Med. Biol. Eng. Comput.* **2008**, *46*, 911–922. [CrossRef]
74. Bedathur, S.; Srivastava, D.; Valluri, S.R. (Eds.) Big Data Curation. In Proceedings of the 20th International Conference on Management of Data, Hyderabad, India, 17–19 December 2014.
75. Grzymala-Busse, J.W.; Woolery, L.K. Improving prediction of preterm birth using a new classification scheme and rule induction. In Proceedings of the AMIA Annual Symposium on Computer Application in Medical Care, Washington, DC, USA, 5–9 November 1994; p. 730.
76. Woolery, L.K.; Grzymala-Busse, J. Machine Learning for an Expert System to Predict Preterm Birth Risk. *J. Am. Med. Inf. Assoc.* **1994**, *1*, 439–446. [CrossRef]
77. Mercer, B.; Goldenberg, R.; Das, A.; Moawad, A.; Iams, J.; Meis, P.; Copper, R.; Johnson, F.; Thom, E.; McNellis, D.; et al. The preterm prediction study: A clinical risk assessment system. *Am. J. Obstet. Gynecol.* **1996**, *174*, 1885–1895. [CrossRef]
78. Goodwin, L.; Maher, S. Data mining for preterm birth prediction. In Proceedings of the 2000 ACM Symposium on Applied Computing—Volume 1, Como, Italy, 19–21 March 2000; pp. 46–51. [CrossRef]
79. Frize, M.; Yu, N.; Weyand, S. Effectiveness of a hybrid pattern classifier for medical applications. *Int. J. Hybrid Intell. Syst.* **2011**, *8*, 71–79. [CrossRef]
80. Weber, A.; Darmstadt, G.L.; Gruber, S.; Foeller, M.E.; Carmichael, S.L.; Stevenson, D.K.; Shaw, G.M. Application of machine-learning to predict early spontaneous preterm birth among nulliparous non-Hispanic black and white women. *Ann. Epidemiol.* **2018**, *28*, 783–789.e1. [CrossRef]
81. Esty, A.; Frize, M.; Gilchrist, J.; Bariciak, E. Applying Data Preprocessing Methods to Predict Premature Birth. In Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 18–21 July 2018; pp. 6096–6099. [CrossRef]
82. Prema, N.S.; Pushpalatha, M.P. Machine Learning Approach for Preterm Birth Prediction Based on Maternal Chronic Conditions. In *Lecture Notes in Electrical Engineering*; Sridhar, V., Padma, M., Rao, K., Eds; Springer: Singapore, 2019; pp. 581–588. [CrossRef]

83. Lee, K.S.; Ahn, K.H. Artificial Neural Network Analysis of Spontaneous Preterm Labor and Birth and Its Major Determinants. *J. Korean Med. Sci.* **2019**, *34*. [CrossRef]
84. Rawashdeh, H.; Awawdeh, S.; Shannag, F.; Henawi, E.; Faris, H.; Obeid, N.; Hyett, J. Intelligent system based on data mining techniques for prediction of preterm birth for women with cervical cerclage. *Comput. Biol. Chem.* **2020**, *85*, 107233. [CrossRef]
85. Koivu, A.; Sairanen, M. Predicting risk of stillbirth and preterm pregnancies with machine learning. *Health Inf. Sci. Syst.* **2020**, *8*. [CrossRef]
86. Sadi-Ahmed, N.; Kacha, B.; Taleb, H.; Kedir-Talha, M. Relevant Features Selection for Automatic Prediction of Preterm Deliveries from Pregnancy ElectroHysterograhic (EHG) records. *J. Med. Syst.* **2017**, *41*. [CrossRef] [PubMed]
87. Despotovic, D.; Zec, A.; Mladenovic, K.; Radin, N.; Turukalo, T.L. A Machine Learning Approach for an Early Prediction of Preterm Delivery. In Proceedings of the 2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY), Subotica, Serbia, 13–15 September 2018; pp. 265–270. [CrossRef]
88. Chen, L.; Xu, H. Deep neural network for semi-automatic classification of term and preterm uterine recordings. *Artif. Intell. Med.* **2020**, *105*, 101861. [CrossRef] [PubMed]
89. Degbedzui, D.K.; Yüksel, M.E. Accurate diagnosis of term–preterm births by spectral analysis of electrohysterography signals. *Comput. Biol. Med.* **2020**, *119*, 103677. [CrossRef]
90. Maner, W.L.; Garfield, R.E. Identification of Human Term and Preterm Labor using Artificial Neural Networks on Uterine Electromyography Data. *Ann. Biomed. Eng.* **2007**, *35*, 465–473. [CrossRef]
91. Most, O.; Langer, O.; Kerner, R.; David, G.B.; Calderon, I. Can myometrial electrical activity identify patients in preterm labor? *Am. J. Obstet. Gynecol.* **2008**, *199*, 378.e1–378.e6. [CrossRef]
92. Bode, O. Das elektrohysterogramm. *Archiv Gynäkologie* **1931**, *146*, 123–128. [CrossRef]
93. Rabotti, C. *Characterization of Uterine Activity by Electrohysterography*; Eindhoven University of Technology: Eindhoven, The Netherlands, 2010. [CrossRef]
94. Goldberger, A.L.; Amaral, L.A.N.; Glass, L.; Hausdorff, J.M.; Ivanov, P.C.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.K.; Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* **2000**, *101*. [CrossRef]
95. Sammut, C.; Webb, G.I. (Eds.) *Encyclopedia of Machine Learning*; Springer: Cham, Switzerland, 2010. [CrossRef]
96. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
97. Widyanto, M.R.; Nobuhara, H.; Kawamoto, K.; Hirota, K.; Kusumoputro, B. Improving recognition and generalization capability of back-propagation NN using a self-organized network inspired by immune algorithm (SONIA). *Appl. Soft Comput.* **2005**, *6*, 72–84. [CrossRef]
98. Huang, N.E.; Shen, Z.; Long, S.R.; Wu, M.C.; Shih, H.H.; Zheng, Q.; Yen, N.C.; Tung, C.C.; Liu, H.H. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. Lond. A.* **1998**, *454*, 903–995. [CrossRef]
99. Cawley, G.C.; Talbot, N.L. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* **2010**, *11*, 2079–2107.
100. Jager, F.; Libenšek, S.; Geršak, K. Characterization and automatic classification of preterm and term uterine records. *PLoS ONE* **2018**, *13*, e0202125. [CrossRef]
101. Grzymala-Busse, J.W. LERS-A System for Learning from Examples Based on Rough Sets. In *Intelligent Decision Support*; Słowiński, R., Eds.; Springer: Dordrecht, The Netherlands, 1992; pp. 3–18. [CrossRef]
102. Vega, F.; Matías, J.; Andrade, M.; Reigosa, M.; Covelo, E. Classification and regression trees (CARTs) for modelling the sorption and retention of heavy metals by soil. *J. Hazard. Mater.* **2009**, *167*, 615–624. [CrossRef]
103. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*. Available online: https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf?fbclid=IwAR0Bgg1eA5TFmqOZeCQXsIoL6PKrVXUFaskUKtg6yBhVXAFFvZA6yQiYx-M (accessed on 3 March 2021).
104. Azur, M.J.; Stuart, E.A.; Frangakis, C.; Leaf, P.J. Multiple imputation by chained equations: What is it and how does it work? *Int. J. Methods Psychiatr. Res.* **2011**, *20*, 40–49. [CrossRef] [PubMed]
105. Salzberg, S.L. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Mach. Learn.* **1994**, *16*, 235–240. [CrossRef]
106. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
107. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed representations of words and phrases and their compositionality. *arXiv* **2013**, arXiv:1310.4546.
108. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Lecture Notes in Computer Science*; Navab, N., Hornegger, J., Wells, W., Frangi, A., Eds.; Springe: Cham, Switzerland, 2015; pp. 234–241. [CrossRef]
109. Sochacki-Wójcicka, N.; Wojcicki, J.; Bomba-Opon, D.; Wielgos, M. Anterior cervical angle as a new biophysical ultrasound marker for prediction of spontaneous preterm birth. *Ultrasound Obstet. Gynecol.* **2015**, *46*, 377–378. [CrossRef]
110. Mehta, S.; Mercan, E.; Bartlett, J.; Weaver, D.; Elmore, J.G.; Shapiro, L. Y-Net: Joint segmentation and classification for diagnosis of breast biopsy images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Frangi, A., Schnabel, J., Davatzikos, C., Alberola-López, C., Fichtinger, G., Eds.; Springer: Cham, Switzerland, 2018; pp. 893–901.
111. Romero, R.; Dey, S.K.; Fisher, S.J. Preterm labor: One syndrome, many causes. *Science* **2014**, *345*, 760–765. [CrossRef]

112. Ravi, M.; Beljorie, M.; Masry, K.E. Evaluation of the quantitative fetal fibronectin test and PAMG-1 test for the prediction of spontaneous preterm birth in patients with signs and symptoms suggestive of preterm labor. *J. Matern. Fetal Neonatal Med.* **2018**, *32*, 3909–3914. [CrossRef] [PubMed]
113. Nikolova, T.; Uotila, J.; Nikolova, N.; Bolotskikh, V.M.; Borisova, V.Y.; Renzo, G.C.D. Prediction of spontaneous preterm delivery in women presenting with premature labor: A comparison of placenta alpha microglobulin-1, phosphorylated insulin-like growth factor binding protein-1, and cervical length. *Am. J. Obstet. Gynecol.* **2018**, *219*, 610.e1–610.e9. [CrossRef]
114. Volpe, N.; Schera, G.B.L.; Dall Asta, A.; Pasquo, E.D.; Ghi, T.; Frusca, T. Cervical sliding sign: New sonographic marker to predict impending preterm delivery in women with uterine contractions. *Ultrasound Obstet. Gynecol.* **2019**, *54*, 557–558. [CrossRef]
115. Baños, N.; Julià, C.; Lorente, N.; Ferrero, S.; Cobo, T.; Gratacos, E.; Palacio, M. Mid-Trimester Cervical Consistency Index and Cervical Length to Predict Spontaneous Preterm Birth in a High-Risk Population. *Am. J. Perinatol. Rep.* **2018**, *08*, e43–e50. [CrossRef] [PubMed]
116. Baer, G.R.; Nelson, R.M. Preterm Birth: Causes, Consequences, and Prevention. C: A Review of Ethical Issues Involved in Premature Birth. Available online: https://www.ncbi.nlm.nih.gov/books/NBK11389/ (accessed on 28 December 2020)
117. Phillips, M. International data-sharing norms: From the OECD to the General Data Protection Regulation (GDPR). *Hum. Genet.* **2018**, *137*, 575–582. [CrossRef]