

Predicting gestational age at birth in the context of preterm birth from multi-modal fetal MRI

Diego Fajardo-Rojas^{1,2 *}, Megan Hall^{1,3}, Daniel Cromb¹, Mary A. Rutherford¹, Lisa Story^{1,3}, Emma Robinson², Jana Hutter^{1,2,4}.

1 Perinatal Imaging department, King's College London, London, UK

2 Biomedical Engineering department, King's College London, London, UK

3 Women's Health department, school of life course sciences, King's College London, London, UK

4 Smart Imaging Lab, Radiological Institute, University Hospital Erlangen, Erlangen, Germany

* diego.fajardo_rojas@kcl.ac.uk

Abstract

Preterm birth is associated with significant mortality and a risk for lifelong morbidity. The complex multifactorial aetiology hampers accurate prediction and thus optimal care. A pipeline consisting of bespoke machine learning methods for data imputation, feature selection, and regression models to predict gestational age (GA) at birth was developed and evaluated from comprehensive multi-modal morphological and functional fetal MRI data from 176 control cases and 67 preterm birth cases. The GA at birth predictions were classified into term and preterm categories and their accuracy, sensitivity, and specificity were reported. An ablation study was performed to further validate the design of the pipeline. The pipeline achieves an R2 score of 0.51 and a mean absolute error of 2.22 weeks. It also achieves a 0.88 accuracy, 0.86 sensitivity, and 0.89 specificity, outperforming previous classification efforts in the literature. The predominant features selected by the pipeline include cervical length and various placental T2* values. The confluence of fast, motion-robust and multi-modal fetal MRI techniques and machine learning prediction allowed the prediction of the gestation at birth. This information is essential for any pregnancy. To the best of our knowledge, preterm birth had only been addressed as a classification problem in the literature. Therefore, this work provides a proof of concept. Future work will increase the cohort size to allow for finer stratification within the preterm birth cohort.

Author summary

Preterm birth is defined as the birth of a baby before the 37th week of pregnancy. It poses a serious risk to the life of a newborn and it is associated with a variety of severe lifelong health problems. Currently, the causes of preterm birth are not completely understood and therefore predicting when a baby will be born prematurely remains a challenging problem. Fetal MRI is an imaging technique that can provide detailed information about the development of the fetus and it is used to support the care of pregnancies at high-risk of preterm birth. Our work combines machine learning techniques with fetal MRI to predict gestational age at birth. The ability to predict this information is crucial for providing adequate care and effective delivery planning. The

main contribution of our study is demonstrating that it is possible to make use of all the information obtained from fetal MRI to estimate the delivery date of a baby. To the best of our knowledge, this is the first study to combine machine learning with such a rich data set to produce these important predictions.

Introduction

Preterm birth is defined as a live birth before 37 completed weeks of gestation [1]. It is estimated that every year 13.4 million babies are born prematurely, corresponding to a global preterm rate of around 9.9% [2]. Prematurity is the leading cause of mortality among children under 5 years accounting for 17.7% of the 5.3 million yearly deaths in this age group [3]. Complications associated with preterm birth are also the leading cause of neonatal mortality, accounting for 36% of these deaths [3]. The chances of survival of preterm babies are directly related to their gestational age (GA) at birth, with survival chances increasing from less than 18% for babies born at 22 weeks to over 95% for babies born at 29 weeks or later [4–6]. Despite advances in perinatal and neonatal care [4–9] survival critically depends on every additional week in-utero.

A continuous rise in survival rate has not translated into a decrease of the short- or long-term morbidity associated with preterm birth [8–10]. Short-term outcomes of premature birth include infections, bronchopulmonary dysplasia, retinopathy, necrotising enterocolitis, and brain disorders [11]. Long-term consequences include an increased risk of neuropsychiatric disorders such as psychosis, neurodevelopmental disabilities such as cerebral palsy and neuromotor dysfunction, adverse sensory outcomes such as hearing and visual impairment, as well as disabilities encompassing learning, cognition, and behaviour [10, 12, 13]. Similar to mortality rates, the incidence and severity of short- and long-term consequences of preterm birth are inversely related to GA at birth [11, 14, 15]. GA at birth is also correlated to social aspects later in life such as income and education level [15].

Reducing the incidence of preterm birth and the impact of its consequences would not only alleviate the burden on individual patients and their families, but also on entire healthcare systems, since the lifetime cost of preterm births in the USA (in 2016) was estimated to be \$25.2 billion [16]. Unsurprisingly, a review of the literature on the economic consequences of preterm birth found a prevailing inverse relation between economic costs and GA at birth, regardless of methodology, date, or country of publication [17].

Preterm birth is classified into three subcategories: extremely preterm (less than 28 weeks), very preterm (28 to 32 weeks), and late preterm (32 to 37 weeks) [1], with further categorisation by clinical presentation: medically induced (or iatrogenic) and spontaneous [18]. While maternal and fetal indicators for iatrogenic preterm birth are well characterised and include conditions such as pre-eclampsia and fetal growth restriction (associated with 30.1% of cases) [19, 20], the aetiologies underlying spontaneous preterm birth are complex, varied, and poorly understood [21]. Causes include -but are not restricted to- infection or inflammation, vascular disease (leading to uterine ischaemia), uterine overdistention, and cervical injury. The latter can be a consequence of LLETZ procedures, cervical cone biopsies for abnormal smear tests, and injuries resulting from emergency C-sections in previous pregnancies [19, 22]. However, definitive causes are registered for only 50% [23, 24] of cases. As such, spontaneous preterm birth should more broadly be considered a syndrome resulting from multiple intricate causes [19, 25].

Despite this complexity, several risk factors have been identified [19, 21, 26] (see Table 1) and are useful, both to provide insights and to help identify at-risk women. The wide variety of factors thereby matches the aetiological complexity of preterm birth.

Even within the same clinical subtype, some factors can have opposite effects. For example, low maternal body mass index (BMI) is a risk factor for fetal growth restriction but protective against preeclampsia, whereas these roles are reversed for maternal obesity [27].

Table 1. Most common risk factors for preterm birth [19, 21, 26].

Risk Factors for Preterm Birth.
African-American ethnicity
Depression
Family history of preterm birth
History of cervical excision
Infections (genitourinary or extragenital)
Low educational attainment
Low socio-economic status
Maternal age (low and high)
Maternal body mass index (low and high)
Multiple gestation (twins, triplets, etc)
Periodontal disease
Prior preterm birth
Stress
Stillbirth or induced abortion history
Tobacco use
Use of assisted reproductive technologies
Uterine anomalies

Currently there are three leading indicators used in clinical practice to identify women at high risk. The strongest predictor is a history of previous preterm birth or cervical surgery or injury (32% chance of recurrent preterm birth) [19, 28]. The other two biomarkers are mid-trimester cervical length below 25mm [28, 29], measured via vaginal ultrasonography; and the presence of more than 50ng/mL fetal fibronectin, a glycoprotein usually absent in cervicovaginal fluid from 18 weeks of gestation and an indicator of choriodecidual disruption. The absence of any of these factors suggests the likelihood of delivering within the following 7 days is only around 1% [19, 28]. These factors have been combined within clinical practice to improve their predictive capabilities [30, 31]. In other analyses, the combination of these predictors also reduced the average cost of high-risk pregnancies. [32, 33].

Another modality with good potential to investigate preterm birth is Magnetic Resonance Imaging (MRI). Fetal MRI is used as a complementary modality to the commonly used ultrasound screening due to its higher resolution, operator independence and suitability for use on women with a higher BMI. It is also non-invasive with no evidence indicating any risk to the fetus or mother [34–36]. Another key advantage of MRI is that it offers multiple complementary contrasts that can support comprehensive functional evaluation of fetal and maternal tissues [37]. Available contrasts include T2-weighted anatomical imaging, T2* relaxometry (which provides an indirect measure of oxygenation [38]), diffusion MRI (which can quantify alterations in tissue microstructure [39–42]), flow measurements and T2 relaxometry. Past studies have largely focused on individual organs such as investigating changes to lung [43], thymus volumes [44], or assessing placental microstructure by measuring T2* and ADC values [45]. One study measured umbilical vein T2 values as a potential marker of intrauterine growth restriction [46]).

While predictive machine learning (ML) models have enjoyed an ever-increasing popularity, preterm birth has only been addressed as a classification problem. Models based on electronic health records, uterine electromyography and transvaginal

ultrasound [47] reported accuracies of approximately 0.77 [48–50], while studies based on electrohysterography reported values above 0.94 [51–53], with the latter, however, only including records of women with recorded contractile activity [53,54].

Machine learning applied to structural MR measurements has been successful at predicting GA at the time of the scan during pregnancy. For example, Convolutional Neural Networks trained on fetal brain MRI have been able to outperform current clinical methods to estimate GA at the time of scan [55,56]. A different study managed to obtain a mean absolute error of 6.1 days by developing bespoke features from 3D ultrasound and using a regression forest for prediction [57].

For this work, a stacking approach was chosen to predict GA at the time of birth. Stacking is an ensembling technique that consists of combining the predictions of individual base models by training a meta-model [58]. Stacking was introduced by Wolpert in 1992 to improve the predictions and generalisability of individual classification models [59]. In 1996 Breiman [60] showed that stacking was also suitable for regression problems, while in 1999 Ting and Witten [61] generalised the technique further by stacking three different types of base models and exploring different meta-models than the ones used in previous work. Ensemble methods such as stacking have the statistical advantage of reducing the risk of overfitting to the training data by taking into account the predictions of all the base models, as well as the representational advantage of expanding the space of available models by combining the base models into meta-models [62].

In recent years, stacking has been successful at various tasks such as genomic prediction [63], protein interactions prediction [64], or prostate cancer detection [65]. These works take advantage of more recent ML learning models, e.g., Yi et. al. [64] use Support Vector Machines and XGBoost models as part of their base models, while Wang et. al. explore using a Random Forest as their meta-model [65].

The present study combines a uniquely rich MR data acquisition including both anatomical and functional scans of multiple fetal organs, and multimodal MRI of the placenta, with a ML pipeline based on stacking. To the best of our knowledge, this is the first work to leverage the advantages of stacking methods together with a comprehensive multi-modal data set to predict GA at birth.

Methods

This section contains a detailed outline of the development of the ML pipeline introduced in this work. The pipeline was designed to address the challenges presented by the data. These include: a large number of derived features relative to the number of training examples, data imbalance, and missing data. These problems were addressed through feature selection, balanced training, and feature imputation. Throughout the development of the pipeline, different design options were investigated including changing data threshold for imputation, and models for feature selection and regression. The end product is a meta-model where predictions are stacked to obtain a final predicted GA at birth. The last subsection describes an ablation study, which investigates the impact of each component. Fig 1 illustrates the workflow of the project. The reader is invited to refer to it repeatedly to complement the description that follows.

Fig 1. Schematic representation of the project pipeline. The boxes in a darker shade denote steps of the pipeline where different design options were explored. The top part of the figure shows the flow of the pipeline with fixed designed options, while the bottom part explicitly indicates the different design choices available for each step.

Data

The data set used for this work comprises clinical records, MR data, and parameters manually extracted from ultrasound from 313 singleton pregnancies, acquired as part of four ethically approved studies: 14/LO/1169 (Placenta Imaging Project, Fulham Research Ethics Committee, approval received September 23, 2016), 19-SS-0032 (Inflammation study in pregnancy, South East Scotland Ethics Committee, approval received March 7, 2019), 21/WA/0075 (Congenital Heart Imaging Programme, Wales Research Ethics Committee, approval received March 8, 2021), and 21/SS/0082 (Individualised Risk prediction of adverse neonatal outcome in pregnancies that deliver preterm using advanced MRI techniques and machine learning, South East Scotland Ethics Committee, approval received March 2022). Informed consent was obtained in all instances.

From the 313 cases originally considered, 59 cases were excluded as they were lacking GA at delivery. We also removed 11 cases scanned after 37 weeks, since - in the context of predicting preterm birth - these would bias training of the model. This resulted in a final data set of 243 cases (see S1 Fig).

Recruitment for all the considered studies was opportunistic, with two studies particularly recruiting women at high risk of preterm birth based on obstetric history, ultrasound and biomarker findings. However, the stated difficulty in accurately predicting preterm birth renders this task difficult, and as a result recruitment and thus the data set available is biased towards term birth.

In all experiments data was split into train, validation, and test sets with an 8:1:1 ratio, keeping an equal proportion of term and preterm birth cases in each set.

Image Acquisition and Processing

Imaging protocols were similar for each study: MR scans were performed on a clinical 3T Philips Achieva scanner between 15 and 40 weeks of gestation using a 32-channel cardiac coil (as is standard process for fetal imaging). For maternal comfort, padding was provided, imaging time was limited to under 90 minutes, and there was frequent verbal interaction and monitoring of heart rate and blood pressure. The protocol included both anatomical T2-weighted imaging and functional MR sequences. For the work presented here only T2* relaxometry among the functional sequences was used.

Anatomical information was acquired with a 2D multi-slice Turbo-Spin-Echo sequence in four to ten planes covering the fetal brain and uterus. Next, to allow for image-based shimming on the 3T scanner, a map of the B0 field was obtained; then shimming was performed for the organ of interest. Afterwards, functional MRI of the entire uterus was performed in coronal orientation using free-breathing multi-echo Gradient Echo with Echo Planar read-out.

In addition to the MRI, two ultrasound scans were performed: an anomaly scan (clinically performed between 19 and 21 weeks of gestation) and a second growth scan (including Doppler ultrasound) that was generally performed within one week of the MRI. In both cases, morphological measurements were manually extracted including abdominal and head circumference, bi-parietal diameter (i.e. the cross-sectional diameter of the skull), femur length, and expected weight. From the growth scan blood flow pulsatility indices were also estimated for the umbilical, uterine and mid-cerebral arteries.

The obtained MR images were processed to obtain quantitative values. For the anatomical data, slice-to-volume reconstruction [66] and learning-based segmentation [67] were applied separately for the brain, body and the placenta. Then regional volumes were calculated.

From the functional data, no motion correction was applied, since all echos for each

slice were acquired within 200ms. Using the method described in [45], quantitative T2* values were obtained by fitting the signal of data from subsequent echo times for the entire uterine field of view (FOV). Values over 300ms were clipped to limit partial volume effects following common practise. Segmentation of the placenta, brain and lungs was performed manually. From these segmentations, regional volumes were calculated, as well as the mean, kurtosis, and skewness of their T2* distributions. These data acquisition steps are represented by index 1 in Fig 1.

Summary of Derived Features

In addition to imaging derived features, demographics, obstetric and medical history of the patients (including previous pregnancies, miscarriages and preterm births) were recorded as well as any relevant information from the current pregnancy such as diagnosis of pre-eclampsia, gestational diabetes, fetal growth restriction or any other fetal or maternal pathology. Finally, the outcomes of the pregnancy were obtained, including gestational age at birth, birth weight-centile and any occurrence of major complications. Collectively the full set of features used by our models was summarised as follows (see S1 Table for more details):

1. **Clinical variables:** e.g. number of previous preterm deliveries, and maternal body mass index.
2. **Structural MRI metrics:** describing sizes of structures e.g. volumes of different brain regions or bi-parietal diameter of the foetal head; these were extracted from both the anatomical and functional scans
3. **Functional MRI metrics:** statistics derived from T2* distributions of the placenta, brain and lungs.
4. **Ultrasound metrics:** from both anomaly and growth ultrasounds - manually extracted by a trained sonographer e.g. the fetal head circumference, femur length.

Feature cleaning

Prior to training it was vital to address the confound effect of gestational age at scan, as well as address the impact of missing data.

Deconfounding

While GA at scan is a feature that would normally be available in a clinical setting, its impact on any learning model could lead to data leakage (e.g. by acting as a lower bound for GA at delivery). Moreover, as all features change dramatically with age [43,44,68,69], it is necessary to disentangle the dominant effect of GA from more subtle signatures that might robustly predict preterm birth. For these reasons, GA was linearly regressed from all features using the method of internally studentised residuals [70]. See index 2 in Fig 1.

Data Imputation

There was significant heterogeneity in the availability of features across the data set. Fetal and maternal motion, maternal discomfort, and clerical errors led to loss of data, with different features available for each of the 243 cases. For this reason, a regression-based approach to imputation, known as Multivariate imputation by chained equations (MICE) [71], was investigated (see S1 Algorithm). Following guidance from the literature, ten iterations of the model were performed [71,72], with two different

regression models: weighted K -Nearest Neighbours (KNR) [73] and Random Forests [74]. Both models were implemented in the standard way using Sci-kit Learn [75]. Imputation should not be applied to features with arbitrarily large amounts of missing data [72, 76]. Thus the impact of discarding features with more than 30%, 40%, or 50% missing values was investigated (see S2 Table for the missing percentages of each feature). Features with a greater percentage of missing values than the respective threshold were discarded. All remaining features were normalised (mean 0, std 1) afterwards.

Data imputation corresponds to index 3 in Fig 1. The boxes corresponding to this step are emphasised by a darker shade to represent that different options were investigated as part of the pipeline design process. The top part of the figure shows the flow of the pipeline with fixed design choices (e.g. if the choice is made to investigate a pipeline using a RF within the MICE algorithm to impute features with less than 40% of missing data). Conversely, the bottom part of the figure explicitly indicates the design choices that were investigated for this step.

Training

Training was performed using a stacking approach in which a number of different classes of machine learning model were trained and these were ensembled together through the training of a meta model [59]. Base models consisted of: Random Forests (RF) [74], Support Vector Regression (SVR) [77], and XGBoost [78]. Each was chosen due to unique strengths: RF are interpretable and robust to overfitting [79]; SVR are robust to outliers and well-suited to small data sets [80, 81]; XGBoost offers state-of-the-art performance from sparse data sets [78]. Importantly they are all capable of capturing non-linear relationships but approach regularisation in different ways [80]. This suggests that they will perform differently on boundary cases, to produce diverse predictions that could benefit from ensembling.

Since a key challenge of training models on our data set has been the high number of features relative to examples (see S1 Table), feature selection was also performed to discourage overfitting. Two simple models were explored: linear regression and Random Forests. For each model trained, 10 features were selected. These two different design options are indicated by the boxes with a darker shade with index 4 in Fig 1.

Models were trained using the Sci-kit learn framework [75], with hyperparameters (see S3 Table) optimised using 3-fold cross-validated grid search [82]. The metric used for optimisation was the coefficient of determination (R^2) [83]. Given fixed design choices on the previous steps, training was carried out every non-empty subset of the selected features. Since there are 1023 non-empty subsets of the ten selected features and 3 regression models, 3069 different regression models were trained in total (Index 5 in Fig 1). These were then composed via the training of a meta-model, for which two different methods were explored: Linear Regression and Random Forests (index 6 in Fig 1). Meta-models were trained on the m best performing base models, as validated through their R^2 score on the validation set. The value of m was also optimised using the validation set.

Ablation Study

An ablation study was conducted to validate the design of the proposed pipeline, with results compared against the best performing meta-model. Since XGBoost models may be trained with incomplete data, and without variance normalisation of the features (since the base learners are decision trees) the first two experiments consist of a single XGBoost model trained on unnormalised data. All experiments are described as follows:

1. Out-of-the-box XGBoost: XGBoost without preprocessing.

263
2. XGBoost with deconfounding: one XGboost was trained after linear deconfounding of features.

264

265
3. Imputation: all base predictive models were trained with deconfounding and imputation (using the imputation approach used in the best meta-model), without performing any upsampling or feature selection.

266

267

268
4. Correcting data imbalance: base models were trained with imputation and upsampling preterm cases in the training set.

269

270
5. Combining feature selecting with upsampling: This equates to evaluating the best performing base model, obtained without ensembling.

271

272
6. Meta-model without upsampling: the impact of upsampling in the whole pipeline was explored by turning it off. This is equivalent to the final meta-model without upsampling.

273

274

275
7. Meta-model: Reporting the performance of the proposed meta-model - obtained from the whole pipeline.

276

277

Results

278

Data Exploration

279

Table 2 shows key demographics, clinical information, and outcomes, divided into preterm and control cohorts. Specifically, the data set consisted of 176 control cases and 67 preterm cases. The distribution of the data according to the four temporal categories was 72.4% term, 14.8% late preterm, 7% very preterm, and 5.8% extremely preterm. Fig 2 shows the distribution of the five continuous features and outcomes included in Table 2, namely GA at scan, maternal BMI at scan, maternal age, GA at birth, and birth weight centile. The pairwise relationship between these is also plotted. For a statistical summary of the data set see S2 Table.

280

281

282

283

284

285

286

287

Table 2. Key demographics, clinical information, and outcomes of participants.

	Preterm birth cohort	Control cohort
Current pregnancy		
Gestational age at scan [weeks]	27.62±4.29 [16.86,34.14]	28.39±4.65 [15.00,36.86]
Maternal BMI at scan [kg/m2]	24.26±3.06 [18.31,32.05]	23.59±2.94 [18.21,32.46]
Maternal age at scan [years]	33.30±6.37 [18.82,47.17]	34.50±3.92 [24.43,45.13]
Obstetric history		
Previous preterm birth	16.42%	8.52%
Outcome		
Gestational age at birth	31.67±4.48 [20.14,36.86]	37±42.43 [39.53,1.32]
Birth weight centile	31.07±32.32 [0.00,95.53]	53.23±28.52 [0.82,99.77]
Fetal sex	54.69% female, 45.31% male	55.56% female, 44.44% male

Reported values are the mean ± SD in the case of continuous variables and percentages for discrete variables. The numbers in brackets are ranges.

Fig 2. Data exploration. Distribution and pairwise relationship between GA at scan, maternal BMI at scan, maternal age, GA at birth, and birth weight centile in the data set.

Meta-model

The best performing meta-model achieved an R^2 score of 0.45 and a mean absolute error (MAE) of 2.55 weeks on the validation set. This performance was achieved using the following settings in the pipeline. First, features with more than 50% missing values were discarded. Then, features with 50% or less missing values were imputed using the MICE algorithm with a KNR as its regression model and a RF was used for feature selection. In order of importance, the selected features were:

1. Cervical length measured from the sagittal plane of MR scan.
2. Mean whole placental T2* value measured from the MR scan.
3. End-diastolic flow measured from the growth ultrasound.
4. T2* brain to placenta ratio measured from the MR scan.
5. Bi-parietal diameter measured from the anomaly ultrasound.
6. Placental T2* kurtosis value measured from the MR scan.
7. Fetal head circumference measured from the anomaly ultrasound.
8. Brain T2* kurtosis value measured from the MR scan.
9. Estimated fetal weight at growth ultrasound.
10. Whole brain T2* volume value measured from the MR scan.

Fig 3 a) shows the mean decrease in impurity corresponding to each of these features. This is the metric used by Random Forests to assign importances to each feature [84].

Fig 3. Feature importances and predictions of the meta-model. a) Feature importance score of the 10 most important features used to train the base models in the pipeline. b) Predictions made by the meta-model 50-KNR-RF on the test set, colourised according to their true preterm temporal category.

After training RF, SVR, and XGBoost models with every non-empty subset of these 10 features, the 14 models with the highest R^2 score on the validation set were used as input for a RF meta-model. Specifically, the 14 base models that provided the features used by the meta-model were all SVRs, trained with different subsets of the selected features. In what follows, this meta-model will be referred to by abbreviating its components, i.e. 50-KNR-RF.

The metrics used for evaluation on the test set were the R^2 score and the mean absolute error (MAE) measured in weeks. The cases were labeled as term (≥ 37 weeks) or preterm (< 37 weeks), according to the GA predicted by the meta-model, and accuracy, sensitivity, and specificity were also reported. On the test set 50-KNR-RF achieved an R^2 score of 0.51 and a MAE of 2.22 weeks. After labeling each subject as term or preterm according to their predicted GA at birth, the meta-model achieved a 0.88 accuracy, 0.86 sensitivity, and 0.89 specificity. The predictions made by 50-KNR-RF on the test set are depicted in Fig. 3 b). It is worth noting that the model only misclassified one of the preterm cases and two term ones.

Ablation study

The performance of each of the models resulting from the experiments of the ablation study is reported in Table 3. 50-KNR-RF outperformed every model in the ablation study in every evaluation metric. The best performing model within the experiments that consist of training different versions of the base models (experiments 3), 4), and 5)) was a SVR, which coincides with the type of models used as features for 50-KNR-RF.

Table 3. Evaluation of the models in the ablation study.

Model	R ² score	MAE	Accuracy	Sensitivity	Specificity
1) Out-of-the-box XGBoost	-0.45	3.84	0.60	0.57	0.61
2) XGBoost with deconfounding	-0.34	3.61	0.64	0.43	0.72
3) Imputation, RF	0.01	3.39	0.68	0.43	0.78
3) Imputation, SVR	0.29	2.66	0.80	0.86	0.78
3) Imputation, XGBoost	0.02	3.20	0.68	0.57	0.72
4) Correcting data imbalance, RF	0.10	3.16	0.60	0.57	0.61
4) Correcting data imbalance, SVR	0.34	2.75	0.76	0.86	0.72
4) Correcting data imbalance, XGBoost	-0.28	3.67	0.52	0.43	0.56
5) Combining feature selecting with upsampling, SVR	0.15	3.48	0.72	0.86	0.67
6) Meta-model without upsampling, RF	0.03	2.68	0.72	0.43	0.83
7) Meta-model, 50-KNR-RF	0.51	2.22	0.88	0.86	0.89

Discussion

Comprehensive multi-modal fetal data and ML models combine synergistically to predict GA at delivery. The developed pipeline acknowledges and addresses key challenges such as imbalances and missing features in the data set, both of which are common when investigating preterm birth.

As mentioned in the Introduction, preterm birth had so far only been addressed as a classification problem. The pipeline constructed in this work attempts to make more precise predictions by developing a regression model to predict GA at birth. In turn, these predictions can be classified as term or preterm to compare the performance of 50-KNR-RF to other classification models in the literature. Table 4 shows a comparison between 50-KNR-RF and the best performing models obtained by other recent studies. Out of the 5 models, 50-KNR-RF has the highest accuracy and specificity. The only model that achieves a higher sensitivity than 50-KNR-RF is one of the models developed by by Esty et al. [48]. However, the 0.93 sensitivity achieved by that model comes with the trade-off of having an accuracy and specificity lower than 0.73. In contrast, 50-KNR-RF achieved a score greater than 0.85 in every metric.

Table 4. Classification performance of the meta-model obtained by this work and other models from recent studies.

Metric Model	Accuracy	Sensitivity	Specificity
Esty et al. [48]	0.72	0.93	0.71
Esty et al. [48]	0.77	0.84	0.77
Włodarczyk et al. [49]	0.78	0.74	0.85
Prema et. al [50]	0.76	0.84	0.73
50-KNR-RF	0.88	0.86	0.89

To the best of our knowledge, the only other study on predicting GA at delivery using ML is the one by Heinsalu et al. [85], where they investigated models using a simpler version of the pipeline displayed in this work. Their best performing model achieves an R^2 score of 0.66 and a MAE of 1.60, but their implementation suffers from data leakage at the imputation stage which makes these results unreliable. Nevertheless, the framework they established is valuable and served as the basis of the present work.

This study provides a proof of concept, but the clinical implementation of a reliable model that could predict the timing of delivery would have important benefits. These include ensuring women are transferred to appropriate neonatal care facilities. A timely transfer helps reduce neonatal mortality and decrease costs [37]. Another crucial example is the targeting of therapies to mitigate the effects of prematurity. Specifically, corticosteroids administration can help reduce intra-ventricular haemorrhage and promote lung maturity. The timing of this therapy is highly important, since it works best when administered within a week before delivery, and repeated doses increase the risk of adverse effects such as reduction in birthweight [86].

The results of the ablation study in Table 3 are helpful to understand the importance of every element of the pipeline. In experiments 1) and 2) it can be seen that even though XGBoost has a built-in mechanism to be trained using data sets with missing values, the results achieved with minimal preprocessing steps are among the worst in the study. In experiments 3) and 4) RF, SVR, and XGBoost models were trained using the data set obtained after imputation, without a feature selection selection step, and using upsampling in the case of experiment 4). On average, results in these experiments were better than in experiments 1) and 2) but upsampling did not translate in a significant difference between experiments 3) and 4). Contrastingly, the impact of upsampling in the whole pipeline can be appreciated by comparing the meta-model obtained in experiment 5) and 50-KNR-RF. The only difference between these two experiments is the use of upsampling during training. The sensitivity score of 0.43 obtained by the meta-model in experiment 5) evidences its inadequacy to address data imbalance, while the 0.86 score achieved by 50-KNR-RF shows that upsampling is effective in tackling this problem. Lastly, it can also be observed that 50-KNR-RF generalises well, i.e. it is not overfitting. 50-KNR-RF achieves an R^2 score of 0.4530 and a MAE of 2.5489 weeks on the validation set and an R^2 score of 0.5143 and a MAE of 2.2202 weeks on the test set. This is line with the literature, that suggests that ensemble models tend to generalise well [59]. The reason why the SVR model in experiment 6) has lower scores than its counterparts in experiments 3) and 4) in spite of being the best model prior to the ensembling of 50-KNR-RF is because it is not generalising well. It is the best performing model on the validation set, where it achieves an R^2 score of 0.4878 and a MAE of 2.6418 weeks, but its performance on the test set is significantly worse. Combining the predictions of the 14 best models via 50-KNR-RF helps to solve this problem.

The features selected as the most important are in line with the literature. The importance of cervical length as a predictor in clinical practice is reflected by its consistent use in the models. Placental features obtained from MRI scans were other prominent features, which is in line with the current understanding of the mechanisms leading to iatrogenic preterm birth [45,87]. The most common clinical indicator, number of previous preterm births, was not a predominant feature. This could be explained by the decision of approaching the problem as a regression instead of a classification one.

While this data set could be considered large given the comprehensive data acquisition, including a fetal MR scan in a cohort of women requiring a high level of medical care, its size is an important limitation for ML methods. The few examples of extremely preterm subjects available during training help explain the poor performance on the test set for this category. Taking into account the performance of 50-KNR-RF in

the validation and test sets, the meta-model seems to generalise well. However, the size of the data set makes it hard to predict if the performance of the meta-model would generalise to new data.

Another limitation is the lack of information on the clinical presentation of preterm birth for every patient in the data set. Iatrogenic and spontaneous preterm births have different aetiologies and training separate models for each case could not only yield better predictions, but also help improve the understanding of each clinical presentation by differentiating their most predictive features. Future work will focus on such subgroups and on extracting relevant phenotypes associated with the different types of preterm birth.

Data obtained on a 1.5T and a 0.55T scanner were available for this study. However, these were not included as there is not a straightforward way to extrapolate the signals acquired by scanners with different magnetic field strengths [88]. Future experiments that include these types of data could test the adequacy of the elements of the pipeline, such as the method of internally studentised residuals, to make accurate predictions regardless of field strength.

There are other directions future research can take to expand or improve the methodology presented in this work. The implementation of the models is ready to benefit from larger or more complete data sets. Adding features known for their predictive power, such as quantitative fibronectin measurements, could improve the results. The performance of the meta-model demonstrates that structural and functional information obtained from MRI can be used to predict GA at delivery. An interesting direction is to make predictions directly from the images making use of deep learning techniques, bypassing the problem of missing data and the need of time-consuming measurements made by experienced clinicians. These techniques have been explored to classify preterm and term patients by automatic measurements of cervical length from transvaginal ultrasound [49] and to estimate GA at scan from fetal brain MRI [55, 56].

MRI remains an expensive modality, however, with an increasing use of fetal MRI, the pipeline presented in this study helps to address a question essential for any pregnancy, and can find an application regardless of the indication of the scan. One of the fundamental contributions of this work is that it shows that fetal MR data acquired as part of diagnostic care or research can be used to obtain useful predictions on the GA at delivery, which in turn can inform the care provided to all pregnancies.

Acknowledgments

The authors acknowledge the invaluable help of the radiographers and midwives while acquiring the data presented here.

Author Contributions

Conceptualization: Diego Fajardo-Rojas, Emma Robinson, Jana Hutter.
Data Curation: Megan Hall, Daniel Cromb, Lisa Story.
Formal Analysis: Diego Fajardo-Rojas.
Investigation: Diego Fajardo-Rojas.
Methodology: Diego Fajardo-Rojas, Emma Robinson, Jana Hutter.
Software: Diego Fajardo-Rojas.
Supervision: Mary A. Rutherford, Lisa Story, Emma Robinson, Jana Hutter.
Validation: Diego Fajardo-Rojas.
Visualization: Diego Fajardo-Rojas.
Writing – Original Draft Preparation: Diego Fajardo-Rojas.

Writing – Review & Editing: Emma Robinson, Jana Hutter.

Supporting information

S1 Table. Description of the features. Features and outcomes available in the original data set. The first column is the name of the feature, the second column its type (C = continuous, Cat = categorical, D = discrete), the third column provides a short description, and the last column registers how each feature was acquired: clinical background (CB), clinical outcome (CO), structural MRI (sMRI), functional MRI (fMRI), growth ultrasound (GUS), and anomaly ultrasound (AUS).

Feature	Type	Description	Origin
GA scan (<i>tag_ga</i>)	C	GA of fetus at time of MR scan	-
GA ROM (<i>tag_garom</i>)	C	GA of fetus at time of rupture of the membranes	CO
Cohort type (<i>tag_typ</i>)	Cat	Cohort type	CO
Scanner type (<i>tag_scanner</i>)	D	1.5 Tesla, 3 Tesla	-
Age (<i>tag_age</i>)	C	Maternal age	CB
BMI (<i>tag_bmi</i>)	C	Maternal body mass index	CB
LOC (<i>tag_loc</i>)	D	Placental location.	GU
GA delivery (<i>tag_gadel</i>)	C	Gestational age of fetus at delivery	CO
MOD (<i>tag_mod</i>)	Cat	Modality of the delivery i.e. C-section.	CO
Sex (<i>tag_sex</i>)	Cat	Sex of the fetus (Male, Female)	CO
BWG (<i>tag_bwg</i>)	C	Birth weight (grams)	CO
BWC (<i>tag_bwc</i>)	C	Birth weight centile	CO
Parity (<i>tag_parity</i>)	D	Number of previous pregnancies.	CB
Parity (<i>tag_prev_ptb</i>)	D	Number of previous preterm births.	CB
BPD (<i>tag_bpd</i>)	C	Biparietal diameter of the fetal brain	sMRI
BPD Cent (<i>tag_bpd.cent</i>)	C	Biparietal diameter centile of the fetus	sMRI
TCD (<i>tag_ce.tcd</i>)	C	Transcerebral diameter of the fetus	sMRI
TCD Cent (<i>tag_ce.tcd.cent</i>)	C	Transcerebral diameter centile of the fetus	sMRI
Post Hor Diam (<i>tag_post_hor_diam</i>)	C	Diameter of the posterior horn of the fetus	sMRI
VOL Body (<i>tag_vol_body</i>)	C	Volume of the fetus	sMRI
Diabetes (<i>tag_diabetes</i>)	Cat	Diagnosis of diabetes	CB
HC (<i>tag_hc</i>)	C	Fetal head circumference at birth	CO
HC Cent (<i>tag_hcc</i>)	C	Fetal head circumference centile at birth	CO
CPTR (<i>tag_cpnr</i>)	C	T2* brain to placenta ratio	fMRI
Anom Pi Left (<i>tag_anom_pi_left</i>)	C	Pulsatility index of the left uterine artery	AUS
Anom Pi Right (<i>tag_anom_pi_right</i>)	C	Pulsatility index of the right uterine artery	AUS
Anom GA (<i>tag_anom_ga</i>)	C	Gestational age at anomaly US	AUS
Anom LOC (<i>tag_anom_loc</i>)	D	Placental location at anomaly US	AUS
Anom Cord (<i>tag_anom_cord</i>)	D	Umbilical cord type at anomaly US	AUS
Anom Cord Ins (<i>tag_anom_cord.ins</i>)	D	Umbilical cord insertion at anomaly US	AUS
Anom HC (<i>tag_anom_hc</i>)	C	Fetal head circumference at anomaly US	AUS
Anom AC (<i>tag_anom_ac</i>)	C	Abdominal circumference at anomaly US	AUS
Anom BPD (<i>tag_anom_bpd</i>)	C	Bi-parietal diameter at anomaly US	AUS
Anom FL (<i>tag_anom_fl</i>)	C	Femur length at anomaly US	GUS
GU GA (<i>tag_gu_ga</i>)	C	Gestational age at growth US	GUS
GU HC (<i>tag_gu_hc</i>)	C	Head circumference at growth US	GUS
GU AC (<i>tag_gu_ac</i>)	C	Abdominal circumference at growth US	GUS
GU BPD (<i>tag_gu_bpd</i>)	C	Bi-parietal diameter at growth US	GUS
GU FL (<i>tag_gu_fl</i>)	C	Femur length at growth US	GUS
GU Pi (<i>tag_gu_pi</i>)	C	Pulsatility index of the umbilical vein	GUS
GU EFW (<i>tag_gu_ewf</i>)	C	Estimated foetal weight at growth US	GUS
GU MCA PI (<i>tag_gu_mca_pi</i>)	C	Pulsatility index of the mid-cerebral artery	GUS
GU MCA PSV (<i>tag_gu_mca_psc</i>)	C	Peak systolic velocity of the mid-cerebral artery	GUS
GU MCA CPR (<i>tag_gu_mca_cpr</i>)	C	Cerebroplacental ratio	GUS
GU Notch (<i>tag_gu_notch</i>)	C	Notching present in the uterine artery	GUS
GU Pi Left (<i>tag_gu_pi_left</i>)	C	Pulsatility index of the left uterine artery	GUS
GU Pi Right (<i>tag_gu_pi_right</i>)	C	Pulsatility index of the right uterine artery	GUS
GU EDF (<i>tag_gu_edf</i>)	C	End-diastolic flow	GUS

GU LOC (<i>tag_gu_loc</i>)	D	Placental location at growth US	GUS
GU Cord (<i>tag_gu_cord</i>)	D	Umbilical cord type at growth US	GUS
GU Cord Ins (<i>tag_gu_cord_ins</i>)	C	Umbilical cord insertion at growth US	GUS
APGAR5 (<i>tag_apgar5</i>)	C	APGAR score at 5 minutes	CO
Histo MVM (<i>tag_histo_mvm</i>)	C	Histopathology, maternal villi malperfusion	CO
Histo FVM (<i>tag_histo_fvm</i>)	C	Histopathology, foetal villi malperfusion	CO
Histo weight (<i>tag_histo_weight</i>)	C	Histopathology placental weight	CO
Histo chorio (<i>tag_histo_chorio</i>)	C	Histopathology chorioamnionitis	CO
SMOK (<i>tag_smok</i>)	D	Maternal smoking status	CB
IVF (<i>tag_ivf</i>)	D	In vitro fertilisation (IVF) status	CB
BWC delivery (<i>tag_del_bwc</i>)	C	Birth weight centile (Intergrowth21) at delivery	CO
Plac T2* mean (<i>plac_t2s_mean</i>)	C	Mean whole placental T2* value	fMRI
Plac T2* vol (<i>plac_t2s_vol</i>)	C	Whole placental T2* volume value	fMRI
Plac T2* lacu (<i>plac_t2s_lacu</i>)	C	Placental T2* lacunarity	fMRI
Plac T2* skew (<i>plac_t2s_skew</i>)	C	Placental T2* skewness	fMRI
Plac T2* kurt (<i>plac_t2s_kurt</i>)	C	Placental T2* kurtosis	fMRI
GU EFW Cent (<i>tag_gu_ewf_cen</i>)	C	Expected foetal weight centile (Intergrowth21)	GUS
Brain T2* mean (<i>brain_t2s_mean</i>)	C	Mean brain T2* value	fMRI
Brain T2* vol (<i>brain_t2s_vol</i>)	C	Brain T2* volume value	fMRI
Brain T2* lacu (<i>brain_t2s_lacu</i>)	C	Brain T2* lacunarity	fMRI
Brain T2* skew (<i>brain_t2s_skew</i>)	C	Brain T2* skewness	fMRI
Brain T2* kurt (<i>brain_t2s_kurt</i>)	C	Brain T2* kurtosis	fMRI
T1 (<i>t1_1</i>)	C	Mean T1 from MRI	fMRI
Cervical length (<i>tag_cervix_length</i>)	C	Cervical length from sagittal plane	sMRI
Cross Study ID (<i>tag_complete_id</i>)	C	Anonymous Cross Study Identifier	-
Volume eCSF left (<i>eCSF_L</i>)	C	Volume eCSF left side	sMRI
Volume eCSF right (<i>eCSF_R</i>)	C	Volume eCSF right side	sMRI
Left cortex (<i>Cortex_L</i>)	C	Volume cortex left	sMRI
Right cortex (<i>Cortex_R</i>)	C	Volume cortex right	sMRI
Left white matter (<i>WM_L</i>)	C	White matter volume left	sMRI
Right white matter (<i>WM_R</i>)	C	White matter volume right	sMRI
Left lateral ventricles (<i>Lat_ventricle_L</i>)	C	Lateral ventricles volume left	sMRI
Right lateral ventricles (<i>Lat_ventricle_R</i>)	C	Lateral ventricles volume right	sMRI
Csp volume (<i>tag_cervix_length</i>)	C	Cavum septi pellucidi volume	sMRI
Brainstem volume (<i>Brainstem</i>)	C	Brainstem volume	sMRI
Left cerebellum volume (<i>Cerebellum_L</i>)	C	Cerebellum volume left	sMRI
Right cerebellum volume (<i>Cerebellum_R</i>)	C	Cerebellum volume right	sMRI
Vermis volume (<i>Vermis</i>)	C	Vermis volume	sMRI
Left lentiform volume (<i>Lentiform_L</i>)	C	Lentiform volume left	sMRI
Right lentiform volume (<i>Lentiform_R</i>)	C	Lentiform volume right	sMRI
Left thalamus volume (<i>Thalamus_L</i>)	C	Thalamus volume left	sMRI
Right thalamus volume (<i>Thalamus_R</i>)	C	Thalamus volume right	sMRI
Third ventricle volume (<i>Third_ventricle</i>)	C	Third ventricle volume	sMRI
Category (<i>tag_cat_norm</i>)	D	Assigned category	CO
Brain volume T2 (<i>tag_cervix_length</i>)	C	Complete T2-weighted brain volume	fMRI
Blood pressure systole (<i>tag_bp_sys</i>)	C	Average systole blood pressure	CB
Blood pressure diastole (<i>tag_bp_dias</i>)	C	Average diastole blood pressure	CB
Heart rate (<i>tag_bp_hr</i>)	C	Average heart rate	CB
Fetal body volume (<i>tag_volume_wb</i>)	C	Whole fetal body volume	sMRI
Amniotic fluid volume (<i>tag_volume_amniotic</i>)	C	Amniotic fluid volume	sMRI
Cohort at scan (<i>tag_control_at_scan</i>)	C	Cohort assessment at scan	CO
Mid cerebellar artery ratio (<i>tag_gu_mca_cpr1</i>)	C	Mid cerebellar artery ratio	US
Adc (<i>diff_1</i>)	C	Average adc	fMRI
T2* (<i>diff_2</i>)	C	Average T2*	fMRI
T2* perfusion compartment (<i>diff_3</i>)	C	T2* perfusion compartment	fMRI
T2* diffusing compartment (<i>diff_4</i>)	C	T2* diffusing compartment	fMRI
Adc perfusion compartment (<i>diff_5</i>)	C	Adc perfusion compartment	fMRI
Adc diffusing compartment (<i>diff_6</i>)	C	Adc diffusing compartment	fMRI
T2* perfusing compartment weighted (<i>diff_7</i>)	C	T2* perfusing compartment weighted	fMRI
T2* diffusing compartment weighted (<i>diff_8</i>)	C	T2* diffusing compartment weighted	fMRI
Adc perfusion compartment weighted (<i>diff_9</i>)	C	Adc perfusion compartment weighted	fMRI
Adc diffusing compartment weighted (<i>diff_10</i>)	C	Adc diffusing compartment weighted	fMRI

Perfusion fraction ivim (<i>diff_11</i>)	C	Perfusion fraction ivim	fMRI
Left lung T2* mean (<i>lung_t2s_left_mean</i>)	C	Mean left lung T2* value	fMRI
Left lung T2* vol (<i>lung_t2s_left_vol</i>)	C	Left lung T2* volume value	fMRI
Left lung T2* lacu (<i>lung_t2s_left_lacu</i>)	C	Left lung T2* lacunarity	fMRI
Left lung T2* skew (<i>lung_t2s_left_skew</i>)	C	Left lung T2* skewness	fMRI
Left lung T2* kurt (<i>lung_t2s_left_kurt</i>)	C	Left lung T2* kurtosis	fMRI
Right lung T2* mean (<i>lung_t2s_right_mean</i>)	C	Mean right lung T2* value	fMRI
Right lung T2* vol (<i>lung_t2s_right_vol</i>)	C	Right lung T2* volume value	fMRI
Right lung T2* lacu (<i>lung_t2s_right_lacu</i>)	C	Right lung T2* lacunarity	fMRI
Right lung T2* skew (<i>lung_t2s_right_skew</i>)	C	Right lung T2* skewness	fMRI
Right lung T2* kurt (<i>lung_t2s_right_kurt</i>)	C	Right lung T2* kurtosis	fMRI
Both lungs T2* mean (<i>lung_t2s_both_mean</i>)	C	Mean T2* value of both lungs	fMRI
Both lungs T2* vol (<i>lung_t2s_both_vol</i>)	C	Both lungs T2* volume value	fMRI

S2 Table. Statistical summary of the features. Statistics of the data set after the first preprocessing steps and before imputation.

Feature name	Missing	Mean	Median	STD	Skewness
<i>tag_ga</i>	0 (0.0%)	28.18	28.14	4.56	-0.11
<i>tag_typ</i>	23 (9.47%)	158.21	99.0	300.93	2.6
<i>tag_scanner</i>	0 (0.0%)	1.0	1.0	0.0	0.0
<i>tag_age</i>	10 (4.12%)	34.18	34.3	4.7	-0.27
<i>tag_bmi</i>	21 (8.64%)	23.77	23.65	2.98	0.48
<i>tag_loc</i>	35 (14.4%)	3.53	2.0	2.09	1.76
<i>tag_gadel</i>	0 (0.0%)	37.37	38.86	4.37	-1.73
<i>tag_mod</i>	0 (0.0%)	4.34	4.0	2.48	0.14
<i>tag_sex</i>	8 (3.29%)	1.57	2.0	0.52	0.19
<i>tag_bwg</i>	7 (2.88%)	2832.48	3062.5	917.24	-0.91
<i>tag_bwc</i>	107 (44.03%)	45.75	43.06	29.52	0.17
<i>tag_parity</i>	0 (0.0%)	0.48	0.0	0.78	2.14
<i>tag_prev_ptb</i>	0 (0.0%)	0.11	0.0	0.33	2.84
<i>tag_bpd</i>	122 (50.21%)	73.51	74.0	12.86	-0.29
<i>tag_bpd.cent</i>	124 (51.03%)	49.3	47.0	28.02	0.05
<i>tag_ce.tcd</i>	123 (50.62%)	33.88	32.55	8.42	0.08
<i>tag_ce.tcd.cent</i>	129 (53.09%)	51.94	52.5	23.59	-0.08
<i>tag_post_hor.diam</i>	125 (51.44%)	6.42	6.35	1.61	0.22
<i>eCSF_L</i>	100 (41.15%)	31941.45	33049.6	11969.29	0.17
<i>eCSF_R</i>	100 (41.15%)	30944.74	31214.5	10941.05	0.35
<i>Cortex_L</i>	100 (41.15%)	21351.56	16914.0	12237.17	0.74
<i>Cortex_R</i>	100 (41.15%)	21422.58	17092.2	11960.04	0.73
<i>WM_L</i>	100 (41.15%)	44681.85	41335.1	20056.3	0.35
<i>WM_R</i>	100 (41.15%)	44771.93	41127.5	20324.06	0.37
<i>Lat_ventricle_L</i>	100 (41.15%)	2154.67	2018.12	963.81	1.01
<i>Lat_ventricle_R</i>	100 (41.15%)	2008.0	1891.96	893.35	0.89
<i>CSP</i>	100 (41.15%)	497.36	471.24	224.35	0.85
<i>Brainstem</i>	100 (41.15%)	3514.29	3428.88	1474.82	-0.16
<i>Cerebellum_L</i>	100 (41.15%)	2999.35	2493.24	1914.44	0.5
<i>Cerebellum_R</i>	100 (41.15%)	2977.61	2348.0	1992.55	0.47
<i>Vermis</i>	100 (41.15%)	936.0	831.04	535.0	0.25
<i>Lentiform_L</i>	100 (41.15%)	2120.58	1953.5	1059.44	0.23
<i>Lentiform_R</i>	100 (41.15%)	2041.67	1871.62	1033.36	0.15
<i>Thalamus_L</i>	100 (41.15%)	1642.67	1483.75	819.65	0.29
<i>Thalamus_R</i>	100 (41.15%)	1590.87	1420.96	801.94	0.26
<i>Third.ventricle</i>	100 (41.15%)	140.56	143.68	70.51	-0.08
<i>tag_diabetes</i>	7 (2.88%)	0.5	0.0	2.13	4.3
<i>tag_cat_norm</i>	75 (30.86%)	1.45	1.0	0.52	0.45
<i>tag_hc</i>	64 (26.34%)	33.3	34.0	3.03	-2.23
<i>tag_hcc</i>	121 (49.79%)	53.64	56.6	33.57	-0.12
<i>tag_garom</i>	0 (0.0%)	37.0	38.86	5.07	-1.78
<i>tag_gu_efw_cen</i>	103 (42.39%)	54.11	61.19	32.35	-0.41
<i>tag_vol_t2w_complete</i>	100 (41.15%)	143758.08	134099.98	60465.11	0.26

<i>tag_cpnr</i>	102 (41.98%)	0.38	0.36	0.12	0.93
<i>tag_histo_weight</i>	131 (53.91%)	417.62	453.0	130.33	-0.55
<i>tag_histo_mvm</i>	132 (54.32%)	0.22	0.0	0.41	1.4
<i>tag_histo_fvm</i>	132 (54.32%)	0.02	0.0	0.13	7.35
<i>tag_histo_chorio</i>	133 (54.73%)	0.38	0.0	0.49	0.49
<i>tag_anom_pi_left</i>	191 (78.6%)	1.19	1.07	0.52	0.5
<i>tag_anom_pi_right</i>	191 (78.6%)	1.22	1.12	0.56	1.19
<i>tag_anom_ga</i>	53 (21.81%)	20.02	20.0	0.9	-3.34
<i>tag_anom_loc</i>	35 (14.4%)	3.53	2.0	2.09	1.76
<i>tag_anom_cord</i>	69 (28.4%)	1.11	1.0	0.56	5.32
<i>tag_anom_cord_ins</i>	182 (74.9%)	5.7	8.0	2.86	-0.66
<i>tag_anom_hc</i>	58 (23.87%)	172.3	171.8	9.81	1.51
<i>tag_anom_ac</i>	58 (23.87%)	150.4	150.2	11.03	0.83
<i>tag_anom_bpd</i>	71 (29.22%)	47.56	47.65	3.15	-2.64
<i>tag_anom_fl</i>	58 (23.87%)	31.72	31.7	2.53	0.23
<i>tag_gu_ga</i>	94 (38.68%)	28.65	28.43	4.63	-0.37
<i>tag_gu_hc</i>	98 (40.33%)	259.63	265.3	44.46	-0.89
<i>tag_gu_ac</i>	98 (40.33%)	239.14	239.3	49.19	-0.41
<i>tag_gu_bpd</i>	100 (41.15%)	73.0	74.8	13.2	-0.72
<i>tag_gu_fl</i>	97 (39.92%)	52.53	53.25	10.87	-0.68
<i>tag_gu_pi</i>	112 (46.09%)	1.08	1.03	0.22	1.65
<i>tag_gu_efw</i>	103 (42.39%)	1378.66	1229.5	727.21	0.53
<i>tag_gu_mca_pi</i>	133 (54.73%)	1.85	1.84	0.36	-0.12
<i>tag_gu_mca_psc</i>	138 (56.79%)	43.83	42.7	13.47	0.71
<i>tag_gu_mca_cpr</i>	158 (65.02%)	1.82	1.83	0.45	0.0
<i>tag_gu_notch</i>	153 (62.96%)	1.1	1.0	0.5	5.23
<i>tag_gu_pi_left</i>	153 (62.96%)	0.87	0.85	0.26	0.71
<i>tag_gu_pi_right</i>	153 (62.96%)	0.87	0.86	0.26	0.91
<i>tag_gu_edf</i>	112 (46.09%)	1.11	1.0	0.62	6.12
<i>tag_gu_loc</i>	110 (45.27%)	3.61	2.0	2.22	1.53
<i>tag_gu_cord</i>	140 (57.61%)	1.06	1.0	0.44	8.1
<i>tag_gu_cord_ins</i>	161 (66.26%)	3.28	4.0	1.28	-1.24
<i>tag_apgar5</i>	26 (10.7%)	9.25	10.0	1.5	-3.74
<i>tag_smok</i>	33 (13.58%)	1.41	1.0	1.08	2.35
<i>tag_ivf</i>	25 (10.29%)	1.87	2.0	0.46	-2.15
<i>tag_bp_sys</i>	105 (43.21%)	106.22	104.22	12.78	0.54
<i>tag_bp_dias</i>	105 (43.21%)	65.36	64.06	10.72	0.53
<i>tag_bp_hr</i>	118 (48.56%)	77.5	76.4	10.1	0.27
<i>tag_volume_wb</i>	220 (90.53%)	763487.03	683058.59	313741.75	0.59
<i>tag_volume_amniotic</i>	220 (90.53%)	460971.62	526734.02	242231.28	-0.32
<i>tag_control_at_scan</i>	0 (0.0%)	0.62	1.0	0.49	-0.49
<i>tag_gu_mca_cpr1</i>	138 (56.79%)	1.79	1.83	0.46	-0.22
<i>tag_del_bwc</i>	7 (2.88%)	47.22	48.04	31.14	-0.0
<i>plac.t2s_mean</i>	0 (0.0%)	57.92	59.69	19.11	-0.01
<i>plac.t2s_vol</i>	0 (0.0%)	366.89	340.31	173.32	0.77
<i>plac.t2s_lacu</i>	0 (0.0%)	19.47	19.08	6.16	2.53
<i>plac.t2s_skew</i>	0 (0.0%)	13.33	2.36	30.53	3.49
<i>plac.t2s_kurt</i>	0 (0.0%)	1.34	0.68	1.89	2.37
<i>lung.t2s_left_mean</i>	159 (65.43%)	65.26	62.16	18.45	1.21
<i>lung.t2s_left_vol</i>	159 (65.43%)	13.88	11.08	8.71	1.4
<i>lung.t2s_left_lacu</i>	159 (65.43%)	20.06	18.43	8.72	1.58
<i>lung.t2s_left_skew</i>	159 (65.43%)	7.27	1.78	13.54	2.99
<i>lung.t2s_left_kurt</i>	159 (65.43%)	0.99	0.58	1.16	1.81
<i>lung.t2s_both_mean</i>	155 (63.79%)	67.06	62.62	19.26	1.17
<i>lung.t2s_both_vol</i>	155 (63.79%)	31.95	26.15	20.13	1.2
<i>tag_vol_body</i>	100 (41.15%)	1009.95	971.36	456.13	8.94
<i>tag_cervix_length</i>	26 (10.7%)	30.7	31.48	10.39	-0.71
<i>brain.t2s_mean</i>	101 (41.56%)	165.0	174.29	42.69	-0.56
<i>brain.t2s_vol</i>	101 (41.56%)	128.08	109.09	72.89	1.32
<i>brain.t2s_lacu</i>	101 (41.56%)	68.79	73.92	19.64	-0.77
<i>brain.t2s_skew</i>	101 (41.56%)	2.27	0.17	6.07	4.43
<i>brain.t2s_kurt</i>	101 (41.56%)	0.91	0.73	0.67	1.56

<i>lung_t2s_right_mean</i>	158 (65.02%)	67.97	63.24	19.56	1.07
<i>lung_t2s_right_vol</i>	158 (65.02%)	19.64	15.85	13.1	1.15
<i>lung_t2s_right_lacu</i>	158 (65.02%)	22.13	19.95	9.71	1.22
<i>lung_t2s_right_skew</i>	158 (65.02%)	7.66	2.42	14.4	4.1
<i>lung_t2s_right_kurt</i>	158 (65.02%)	1.14	0.82	1.19	2.08
<i>t1_1</i>	199 (81.89%)	1003.34	1016.46	243.97	-0.8
<i>diff_1</i>	140 (57.61%)	55.75	57.47	19.34	-0.22
<i>diff_2</i>	140 (57.61%)	0.0	0.0	0.0	1.14
<i>diff_3</i>	140 (57.61%)	0.0	0.0	0.0	2.41
<i>diff_4</i>	140 (57.61%)	0.0	0.0	0.0	0.0
<i>diff_5</i>	140 (57.61%)	0.01	0.0	0.01	3.3
<i>diff_6</i>	140 (57.61%)	0.0	0.0	0.0	0.0
<i>diff_7</i>	140 (57.61%)	52.51	52.21	32.07	0.11
<i>diff_8</i>	140 (57.61%)	0.03	0.03	0.01	1.02
<i>diff_9</i>	140 (57.61%)	76.01	75.25	22.84	0.99
<i>diff_10</i>	140 (57.61%)	0.0	0.0	0.0	0.62
<i>diff_11</i>	140 (57.61%)	0.33	0.3	0.17	0.68
<i>tag_complete_id</i>	0 (0.0%)	2400682.76	1000187.0	2864221.51	1.98

S3 Table. Hyperparameters of the base models. Hyperparameters investigated by the grid search for each of the base models

Random Forests		Support Vector Regression		XGBoost	
Hyperparameter	Values	Hyperparameter	Values	Hyperparameter	Values
<i>max_depth</i>	[3, 5, 10, 20, 50, 100]	<i>C</i>	[0.001, 0.01, 0.1, 1, 10, 100, 1000]	<i>max_depth</i>	[3, 5, 7, 10, 20, 50]
<i>max_features</i>	['auto', 'sqrt', 'log2']	<i>gamma</i>	['scale', 'auto']	<i>learning_rate</i>	[0.01, 0.05, 0.1, 0.3, 0.5]
<i>n_estimators</i>	[5, 10, 20, 50, 100, 250]	<i>kernel</i>	['rbf', 'poly', 'sigmoid', 'linear']	<i>min_child_weight</i>	[1, 3, 5, 7]
		<i>epsilon</i>	[0.001, 0.01, 0.1, 0.5, 1]	<i>gamma</i>	[0.1, 0.5, 0.8, 2, 5, 10]
		<i>degree</i>	[2, 3]	<i>colsample_bytree</i>	[0.3, 0.5, 0.7]

S1 Fig. Details of First Preprocessing Steps. Number of subjects kept after each initial preprocessing step.

S1 Algorithm. Pseudocode of the MICE algorithm.

Algorithm 1 Pseudocode of the MICE algorithm	
Input: Data matrix $D_{ij} = (d_{ij})$, Number of iterations MaxIter.	
Output: Imputed matrix $\hat{D}_{ij}$.	
1:	Impute D_{ij} with column means: $\hat{D}_{ij} = (\hat{d}_{ij}) \leftarrow \text{FillWithMeans}(D_{ij})$
2:	$k \leftarrow 1$
3:	while $k \leq \text{MaxIter}$ do
4:	for each column D_j in D_{ij} do
5:	Set imputations for D_j to missing in $\hat{D}_{ij}$
6:	$f_j \leftarrow \text{FitRegressionModel}(\hat{D}_{ij}, \hat{D}_j)$
7:	for each row D_i in D_{ij} do
8:	if $\hat{d}_{ij}$ is missing in $\hat{D}_{ij}$ then
9:	Impute missing value: $\hat{d}_{ij} \leftarrow f_j(\hat{D}_i \setminus \{\hat{d}_{ij}\})$
10:	end if
11:	end for
12:	end for
13:	$k \leftarrow k + 1$
14:	end while
15:	return $\hat{D}_{ij}$

References

1. Organization WH. Preterm birth; 2018. Available from: <https://www.who.int/news-room/fact-sheets/detail/preterm-birth>.
2. Ohuma E, Moller AB, Bradley E, Chakwera S, Hussain-Alkhateeb L, Lewin A, et al. National, regional, and global estimates of preterm birth in 2020, with trends from 2010: a systematic analysis. *The Lancet*. 2023;402:1261–1271. doi:10.1016/S0140-6736(23)00878-4.
3. Perin J, Mulick A, Yeung D, Villavicencio F, Lopez G, Strong K, et al. Global, regional, and national causes of under-5 mortality in 2000–19: an updated systematic analysis with implications for the Sustainable Development Goals. *The Lancet Child Adolescent Health*. 2021;6. doi:10.1016/S2352-4642(21)00311-4.
4. Ancel PY, Goffinet F, Kuhn P, Langer B, Matis J, Hernandorena X, et al. Survival and Morbidity of Preterm Children Born at 22 Through 34 Weeks' Gestation in France in 2011: Results of the EPIPAGE-2 Cohort Study. *JAMA pediatrics*. 2015;169. doi:10.1001/jamapediatrics.2014.3351.
5. Santhakumaran S, Statnikov E, Gray D, Battersby C, Ashby D, Modi N. Survival of very preterm infants admitted to neonatal care in England 2008-2014: time trends and regional variation. *Archives of disease in childhood Fetal and neonatal edition*. 2017;103. doi:10.1136/archdischild-2017-312748.
6. Bell E, Hintz S, Hansen N, Bann C, Wyckoff M, Dem Mauro S, et al. Mortality, In-Hospital Morbidity, Care Practices, and 2-Year Outcomes for Extremely Preterm Infants in the US, 2013-2018. *JAMA*. 2022;327:248–263. doi:10.1001/jama.2021.23580.
7. Blencowe H, Cousens S, Oestergaard M, Chou D, Moller AB, Narwal R, 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;379:2162–72. doi:10.1016/S0140-6736(12)60820-4.
8. Cheong JL, Spittle AJ, Burnett AC, Anderson PJ, Doyle LW. Have outcomes following extremely preterm birth improved over time? In: *Seminars in Fetal and Neonatal Medicine*. vol. 25. Elsevier; 2020. p. 101114.
9. Boland R, Cheong J, Doyle L. Changes in long-term survival and neurodevelopmental disability in infants born extremely preterm in the post-surfactant era. *Seminars in Perinatology*. 2021;45:151479. doi:10.1016/j.semperi.2021.151479.
10. Allen M, Cristofalo E, Kim C. Outcomes of Preterm Infants: Morbidity Replaces Mortality. *Clinics in perinatology*. 2011;38:441–54. doi:10.1016/j.clp.2011.06.011.
11. Costeloe KL, Hennessy EM, Haider S, Stacey F, Marlow N, Draper ES. Short term outcomes after extreme preterm birth in England: comparison of two birth cohorts in 1995 and 2006 (the EPICure studies). *The BMJ*. 2012;345.
12. Vanes L, Murray R, Nosarti C. Adult outcome of preterm birth: Implications for neurodevelopmental theories of psychosis. *Schizophrenia Research*. 2021;247. doi:10.1016/j.schres.2021.04.007.
13. Jarjour I. Neurodevelopmental Outcome After Extreme Prematurity: A Review of The Literature. *Pediatric Neurology*. 2014;52. doi:10.1016/j.pediatrneurol.2014.10.027.

14. Moore T, Hennessy E, Myles J, Johnson S, Draper E, Costeloe K, et al. Neurological and Developmental Outcome in Extremely Preterm Children Born in England in 1995 and 2006: The EPICure Studies. *BMJ (Clinical research ed)*. 2012;345:e7961. doi:10.1136/bmj.e7961.
15. Moster D, Markestad T. Long-Term Medical and Social Consequences of Preterm Birth. *The New England journal of medicine*. 2008;359:262–73. doi:10.1056/NEJMoA0706475.
16. Waitzman N, Jalali A, Grosse S. Preterm Birth Lifetime Costs in the United States in 2016: An Update. *Seminars in Perinatology*. 2021;45:151390. doi:10.1016/j.semperi.2021.151390.
17. Petrou S, Yiu HH, Kwon J. Economic consequences of preterm birth: a systematic review of the recent literature (2009–2017). *Archives of Disease in Childhood*. 2019;104(5):456–465. doi:10.1136/archdischild-2018-315778.
18. Moutquin JM. Classification and heterogeneity of preterm birth. *BJOG : an international journal of obstetrics and gynaecology*. 2003;110 Suppl 20:30–3. doi:10.1016/S1470-0328(03)00021-1.
19. Goldenberg R, Culhane J, Iams J, Romero R. Epidemiology and Causes of Preterm Birth. *Lancet*. 2008;371:75–84. doi:10.1016/S0140-6736(08)60074-4.
20. Morken NH, Kallen K, Jacobsson B. Fetal growth and onset of delivery: A nationwide population-based study of preterm infants. *American journal of obstetrics and gynecology*. 2006;195:154–61. doi:10.1016/j.ajog.2006.01.019.
21. Frey H, Klebanoff M. The epidemiology, etiology, and costs of preterm birth. *Seminars in Fetal and Neonatal Medicine*. 2016;21. doi:10.1016/j.siny.2015.12.011.
22. Suff N, Xu VX, Glazewska-Hallin A, Carter J, Brennecke S, Shennan A. Previous term emergency caesarean section is a risk factor for recurrent spontaneous preterm birth; a retrospective cohort study. *European Journal of Obstetrics Gynecology and Reproductive Biology*. 2022;271:108–111. doi:https://doi.org/10.1016/j.ejogrb.2022.02.008.
23. Menon R. Spontaneous preterm birth, a clinical dilemma: Etiologic, pathophysiologic and genetic heterogeneities and racial disparity. *Acta obstetrica et gynecologica Scandinavica*. 2008;87:590–600. doi:10.1080/00016340802005126.
24. Muglia LJ, Katz M. The enigma of spontaneous preterm birth. *The New England journal of medicine*. 2010;362 6:529–35.
25. Romero RJ, Espinoza J, Kusanovic JP, Gotsch F, Hassan SS, Erez O, et al. The preterm parturition syndrome. *BJOG: An International Journal of Obstetrics & Gynaecology*. 2006;113.
26. Cobo T, Kacerovsky M, Jacobsson B. Risk factors for spontaneous preterm delivery. *International Journal of Gynecology Obstetrics*. 2020;150:17–23. doi:10.1002/ijgo.13184.
27. Kramer M, Papageorgiou AT, Culhane J, Bhutta Z, Goldenberg R, Gravett M, et al. Challenges in defining and classifying the preterm birth syndrome. *American journal of obstetrics and gynecology*. 2011;206:108–12. doi:10.1016/j.ajog.2011.10.864.

28. Suff N, Story L, Shennan A. The prediction of preterm delivery: What is new? *Seminars in Fetal and Neonatal Medicine*. 2018;24. doi:10.1016/j.siny.2018.09.006.
29. Romero R, Dey S, Fisher S. Preterm Labor: One Syndrome, Many Causes. *Science (New York, NY)*. 2014;345:760–765. doi:10.1126/science.1251816.
30. Carter J, Seed P, Watson H, David A, Sandall J, Shennan A, et al. Development and validation of prediction models for the QUiPP App v.2: a tool for predicting preterm birth in women with symptoms of threatened preterm labor. *Ultrasound in Obstetrics Gynecology*. 2019;55. doi:10.1002/uog.20422.
31. Watson HA, Carlisle N, Seed PT, Carter J, Kuhrt K, Tribe RM, et al. Evaluating the use of the QUiPP app and its impact on the management of threatened preterm labour: A cluster randomised trial. *PLoS Medicine*. 2021;18.
32. Desplanches T, Lejeune C, Jonathan C, Sagot P, Quantin C. Cost-effectiveness of diagnostic tests for threatened preterm labor in singleton pregnancy in France. *Cost Effectiveness and Resource Allocation*. 2018;16. doi:10.1186/s12962-018-0106-y.
33. Baaren GJ, Vis J, Grobman W, Bossuyt P, Opmeer B, Mol BW. Cost-effectiveness analysis of cervical length measurement and fibronectin testing in women with threatened preterm labor. *American journal of obstetrics and gynecology*. 2013;209. doi:10.1016/j.ajog.2013.06.029.
34. Tocchio S, Kline-Fath B, Kanal E, Schmithorst V, Panigrahy A. MRI Evaluation and Safety in the Developing Brain. *Seminars in perinatology*. 2015;39. doi:10.1053/j.semperi.2015.01.002.
35. Lum M, Tsiouris A. MRI safety considerations during pregnancy. *Clinical Imaging*. 2020;62. doi:10.1016/j.clinimag.2020.02.007.
36. Ray J, Vermeulen M, Bharatha A, Montanera W, Park A. Association Between MRI Exposure During Pregnancy and Fetal and Childhood Outcomes. *JAMA*. 2016;316:952–961. doi:10.1001/jama.2016.12126.
37. Story L, Hutter J, Zhang T, Shennan A, Rutherford M. The use of antenatal fetal Magnetic Resonance Imaging in the assessment of patients at high risk of preterm birth. *European Journal of Obstetrics Gynecology and Reproductive Biology*. 2018;222. doi:10.1016/j.ejogrb.2018.01.014.
38. Sørensen A, Hutter J, Seed M, Grant PE, Gowland P. T2*-weighted placental MRI: basic research tool or emerging clinical test for placental dysfunction? *Ultrasound in Obstetrics & Gynecology*. 2020;55(3):293–302. doi:https://doi.org/10.1002/uog.20855.
39. Avena-Zampieri CL, Hutter J, Rutherford M, Milan A, Hall M, Egloff A, et al. Assessment of the fetal lungs in utero. *American Journal of Obstetrics Gynecology MFM*. 2022;4(5):100693. doi:https://doi.org/10.1016/j.ajogmf.2022.100693.
40. Lee W, Krisko A, Shetty A, Yeo L, Hassan S, Gotsch F, et al. Noninvasive Fetal Lung Assessment Using Diffusion Weighted Imaging. *Ultrasound in obstetrics gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology*. 2009;34:673–7. doi:10.1002/uog.7446.

41. Slator PJ, Hutter J, Palombo M, Jackson LH, Ho A, Panagiotaki E, et al. Combined diffusion-relaxometry MRI to identify dysfunction in the human placenta. *Magnetic Resonance in Medicine*. 2019;82(1):95–106. doi:<https://doi.org/10.1002/mrm.27733>.
42. Kristi B A, Ditte N H, Caroline H, Marianne S, Astrid P, Jens B F, et al. Placental diffusion-weighted MRI in normal pregnancies and those complicated by placental dysfunction due to vascular malperfusion. *Placenta*. 2020;91:52–58. doi:<https://doi.org/10.1016/j.placenta.2020.01.009>.
43. Story L, Zhang T, Steinweg J, Hutter J, Matthew J, Dassios T, et al. Foetal lung volumes in pregnant women who deliver very preterm: a pilot study. *Pediatric Research*. 2019;87. doi:10.1038/s41390-019-0717-9.
44. Story L, Zhang T, Uus A, Hutter J, Egloff A, Gibbons D, et al. Antenatal thymus volumes in fetuses that delivered ≥32 weeks gestation: An MRI pilot study. *Acta Obstetrica et Gynecologica Scandinavica*. 2020;100. doi:10.1111/aogs.13983.
45. Hutter J, Slator PJ, Jackson L, Gomes ADS, Ho A, Story L, et al. Multi-modal functional MRI to explore placental function over gestation. *Magnetic Resonance in Medicine*. 2019;81(2):1191–1204. doi:<https://doi.org/10.1002/mrm.27447>.
46. Zhu MY, Milligan N, Keating S, Windrim R, Keunen J, Thakur V, et al. The hemodynamics of late onset intrauterine growth restriction by MRI. *American journal of obstetrics and gynecology*. 2015;214. doi:10.1016/j.ajog.2015.10.004.
47. Włodarczyk T, Płotka S, Szczepański T, Rokita P, Sochacki-Wójcicka N, Wójcicki J, et al. Machine Learning Methods for Preterm Birth Prediction: A Review. *Electronics*. 2021;10:586. doi:10.3390/electronics10050586.
48. Esty A, Frize M, Gilchrist J, Bariciak E. Applying Data Preprocessing Methods to Predict Premature Birth. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2018. p. 6096–6099.
49. Włodarczyk T, Płotka S, Trzcinski T, Rokita P, Sochacki-Wójcicka N, Lipa M, et al. Estimation of Preterm Birth Markers with U-Net Segmentation Network. 2019;.
50. N S P, Pushpalatha M. In: Machine Learning Approach for Preterm Birth Prediction Based on Maternal Chronic Conditions; 2019. p. 581–588.
51. Chen L, Xu H. Deep neural network for semi-automatic classification of term and preterm uterine recordings. *Artificial Intelligence in Medicine*. 2020;105:101861. doi:10.1016/j.artmed.2020.101861.
52. Despotović D, Zec A, Mladenovic K, Radin N, Loncar-Turukalo T. A Machine Learning Approach for an Early Prediction of Preterm Delivery; 2018. p. 000265–000270.
53. Sadi-Ahmed N, Kacha B, Taleb H, Kedir-Talha M. Relevant Features Selection for Automatic Prediction of Preterm Deliveries from Pregnancy ElectroHysterographic (EHG) records. *Journal of Medical Systems*. 2017;41. doi:10.1007/s10916-017-0847-8.
54. Fele-Zorz G, Kavsek G, Novak-Antolic Z, 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. *Medical biological engineering computing*. 2008;46:911–22. doi:10.1007/s11517-008-0350-y.

55. Kojita Y, Matsuo H, Kanda T, Nishio M, Sofue K, Nogami M, et al. Deep learning model for predicting gestational age after the first trimester using fetal MRI. *European Radiology*. 2021;31. doi:10.1007/s00330-021-07915-9.
56. Shen L, Zheng J, Lee E, Shpanskaya K, McKenna E, Atluri M, et al. Attention-guided deep learning for gestational age prediction using fetal brain MRI. *Scientific Reports*. 2022;12. doi:10.1038/s41598-022-05468-5.
57. Namburete A, Stebbing R, Kemp B, Yaqub M, Papageorgiou AT, Noble J. Learning-Based Prediction of Gestational Age from Ultrasound Images of the Fetal Brain. *Medical Image Analysis*. 2015;30. doi:10.1016/j.media.2014.12.006.
58. Zhou ZH. Ensemble methods: foundations and algorithms. CRC press; 2012.
59. Wolpert DH. Stacked generalization. *Neural Networks*. 1992;5(2):241–259. doi:[https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1).
60. Breiman L. Stacked Regressions. *Mach Learn*. 1996;24(1):49–64.
61. Ting KM, Witten IH. Issues in Stacked Generalization. *Journal of Artificial Intelligence Research*. 1999;10:271–289. doi:10.1613/jair.594.
62. Dietterich TG. Ensemble Methods in Machine Learning. In: *International Workshop on Multiple Classifier Systems*; 2000. Available from: <https://api.semanticscholar.org/CorpusID:56776745>.
63. Liang M, Chang T, An B, Duan X, Du L, Wang X, et al. A Stacking Ensemble Learning Framework for Genomic Prediction. *Frontiers in Genetics*. 2021;12. doi:10.3389/fgene.2021.600040.
64. Yi HC, You ZH, Wang MN, Guo ZH, Wang YB, Zhou JR. RPI-SE: a stacking ensemble learning framework for ncRNA-protein interactions prediction using sequence information. *BMC bioinformatics*. 2020;21(1):60.
65. Wang Y, Wang D, Geng N, Wang Y, Yin Y, Jin Y. Stacking-based ensemble learning of decision trees for interpretable prostate cancer detection. *Applied Soft Computing*. 2019;77:188–204. doi:<https://doi.org/10.1016/j.asoc.2019.01.015>.
66. Uus A, Zhang T, Jackson LH, Roberts TA, Rutherford MA, Hajnal JV, et al. Deformable slice-to-volume registration for motion correction of fetal body and placenta MRI. *IEEE transactions on medical imaging*. 2020;39(9):2750–2759.
67. Uus AU, Kyriakopoulou V, Makropoulos A, Fukami-Gartner A, Cromb D, Davidson A, et al. BOUNTI: Brain vOlumetry and aUtomated parcellatiON for 3D feTal MRI. *bioRxiv*. 2023; p. 2023–04.
68. Story L, Davidson A, Patkee P, Fleiss B, Kyriakopoulou V, Colford K, et al. Brain volumetry in fetuses that deliver very preterm: An MRI pilot study. *NeuroImage: Clinical*. 2021;30:102650. doi:<https://doi.org/10.1016/j.nicl.2021.102650>.
69. Papageorgiou AT, Ohuma EO, Altman DG, Todros T, Ismail LC, Lambert A, et al. International standards for fetal growth based on serial ultrasound measurements: the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project. *The Lancet*. 2014;384:869–879.
70. Cook RD, Weisberg S. Residuals and Influence in Regression. *Monographs on statistics and applied probability*. Chapman and Hall; 1986. Available from: <https://books.google.co.uk/books?id=aMDpswEACAAJ>.

71. Azur M, Stuart E, Frangakis C, Leaf P. Multiple Imputation by Chained Equations: What is it and how does it work? *International journal of methods in psychiatric research*. 2011;20:40–9. doi:10.1002/mpr.329.
72. Jäger S, Allhorn A, Biessmann F. A Benchmark for Data Imputation Methods. *Frontiers in Big Data*. 2021;4. doi:10.3389/fdata.2021.693674.
73. Bicego M, Loog M. Weighted K-Nearest Neighbor revisited; 2016. p. 1642–1647.
74. Breiman L. Random Forests. *Machine Learning*. 2001;45:5–32. doi:10.1023/A:1010950718922.
75. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011;12:2825–2830.
76. Bertsimas D, Pawlowski C, Zhuo Y. From predictive methods to missing data imputation: An optimization approach. *Journal of Machine Learning Research*. 2018;18:1–39.
77. Smola A, Schölkopf B. A tutorial on support vector regression. *Statistics and Computing*. 2004;14:199–222. doi:10.1023/B
78. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *CoRR*. 2016;abs/1603.02754.
79. Gzar DA, Mahmood AM, Abbas MK. A Comparative Study of Regression Machine Learning Algorithms: Tradeoff Between Accuracy and Computational Complexity. *Mathematical Modelling of Engineering Problems*. 2022;9(5):1217–1224. doi:10.18280/mmep.090508.
80. Fernández-Delgado M, Sirsat MS, Cernadas E, Alawadi S, Barro S, Febrero-Bande M. An extensive experimental survey of regression methods. *Neural Networks*. 2019;111:11–34. doi:https://doi.org/10.1016/j.neunet.2018.12.010.
81. Kinaneva D, Hristov G, Kyuchukov P, Georgiev G, Zahariev P, Daskalov R. Machine Learning Algorithms for Regression Analysis and Predictions of Numerical Data. In: 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA); 2021. p. 1–6.
82. Krstajic D, Buturovic L, Leahy D, Thomas S. Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics*. 2014;6:10. doi:10.1186/1758-2946-6-10.
83. Casella G, Berger RL, Company BP. Statistical Inference. Duxbury advanced series in statistics and decision sciences. Thomson Learning; 2002. Available from: https://books.google.co.uk/books?id=0x_vAAAAAAAJ.
84. Nembrini S, König I, Wright M. The revival of the Gini Importance? *Bioinformatics (Oxford, England)*. 2018;34. doi:10.1093/bioinformatics/bty373.
85. Heinsalu R, Williams L, Ranjan A, Avena Zampieri C, Uus A, Robinson E, et al. In: Predicting preterm birth using multimodal fetal imaging. Springer; 2021.
86. Story L, Simpson N, David A, Z Z, Bennett P, Jolly M, et al. Reducing the Impact of Preterm Birth: Preterm Birth Commissioning in the United Kingdom. *European Journal of Obstetrics Gynecology and Reproductive Biology: X*. 2019;3. doi:10.1016/j.eurox.2019.100018.

87. Purisch S, Gyamfi C. Epidemiology of preterm birth. *Seminars in Perinatology*. 2017;41. doi:10.1053/j.semperi.2017.07.009.
88. Garcia-Eulate R, Garcia-Garcia D, Domínguez P, Noguera J, Luis E, Rodriguez-Oroz M, et al. Functional bold MRI: Advantages of the 3 T vs. the 1.5 T. *Clinical imaging*. 2011;35:236–41. doi:10.1016/j.clinimag.2010.07.003.

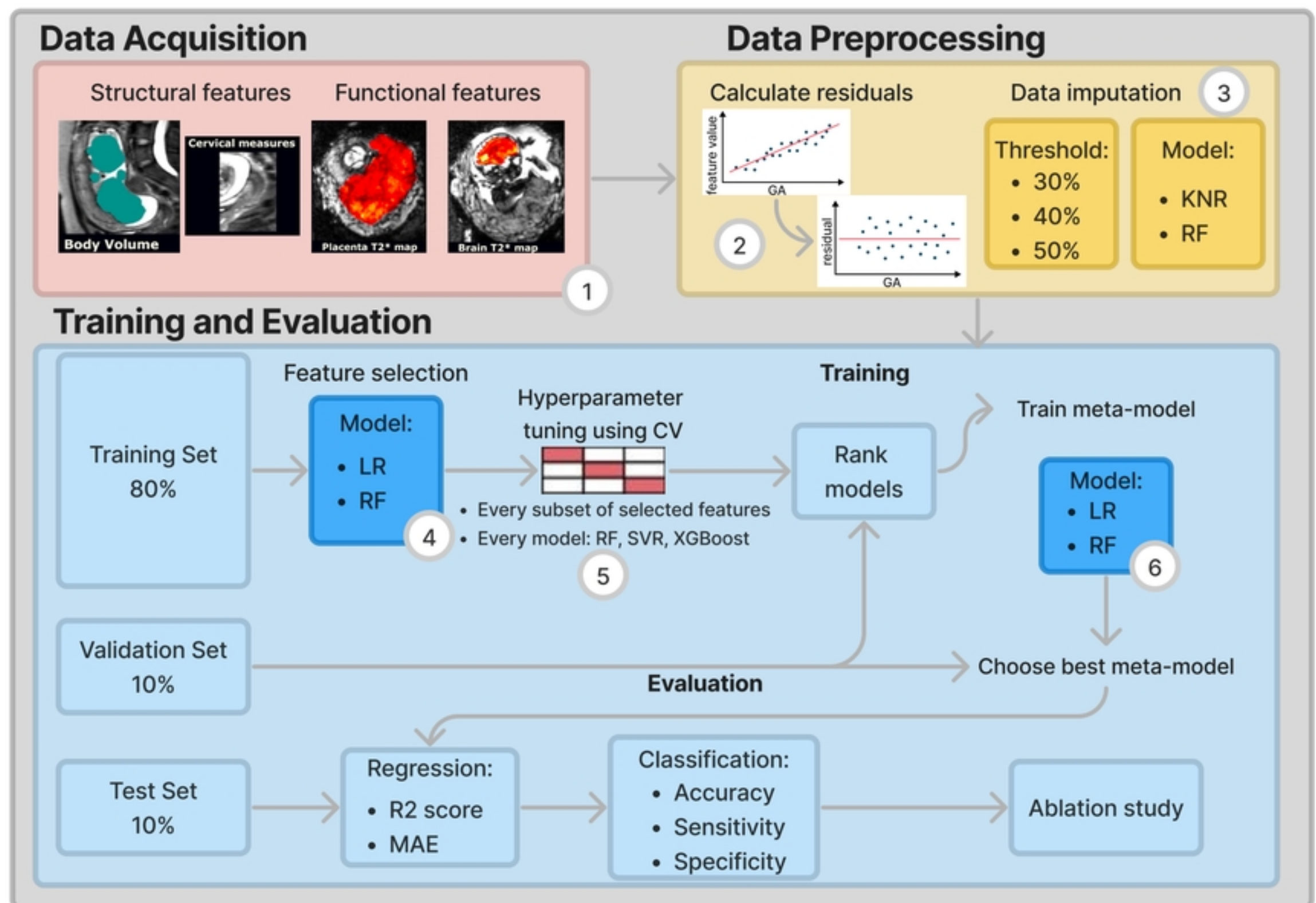
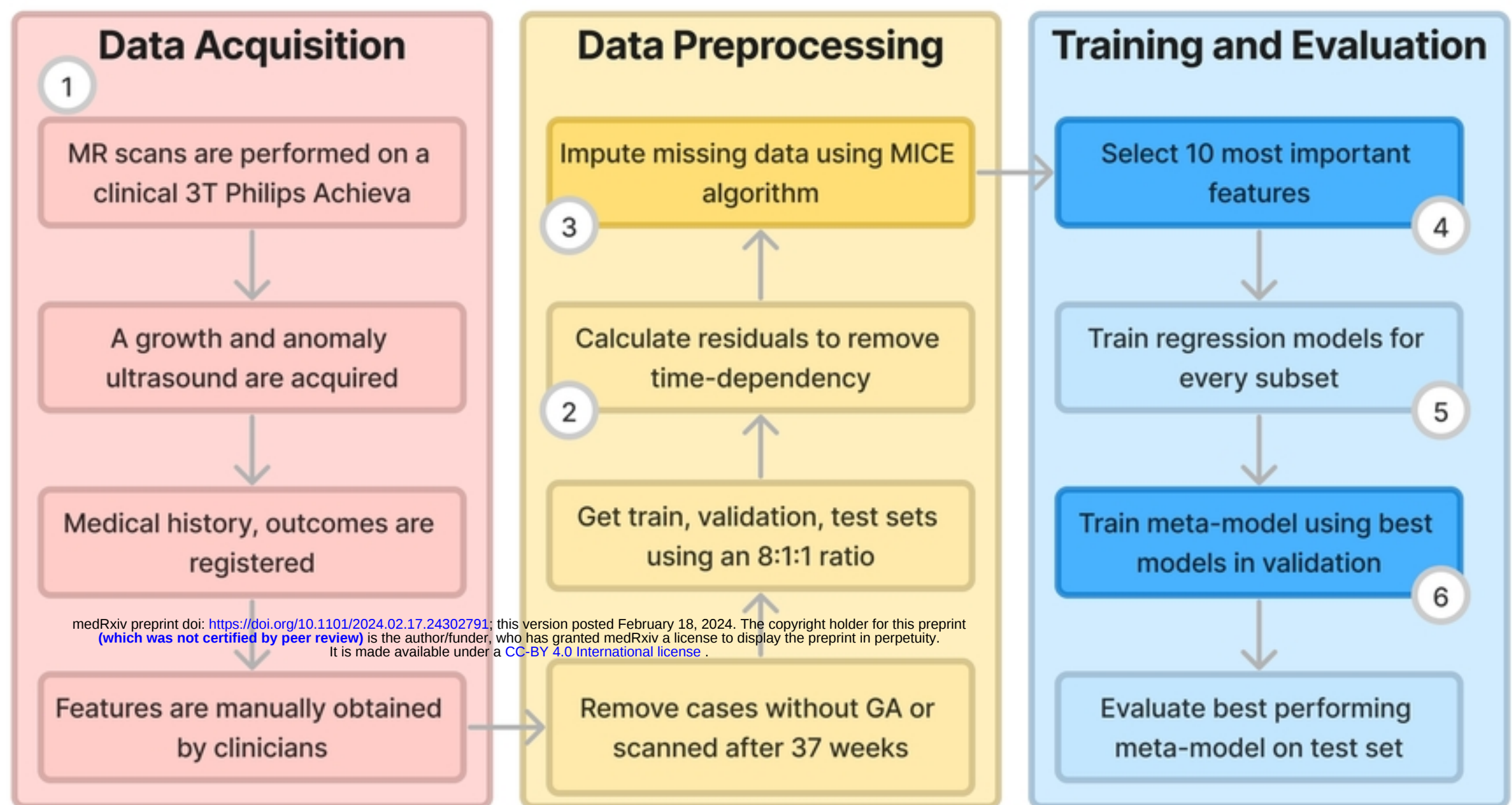


Figure 1

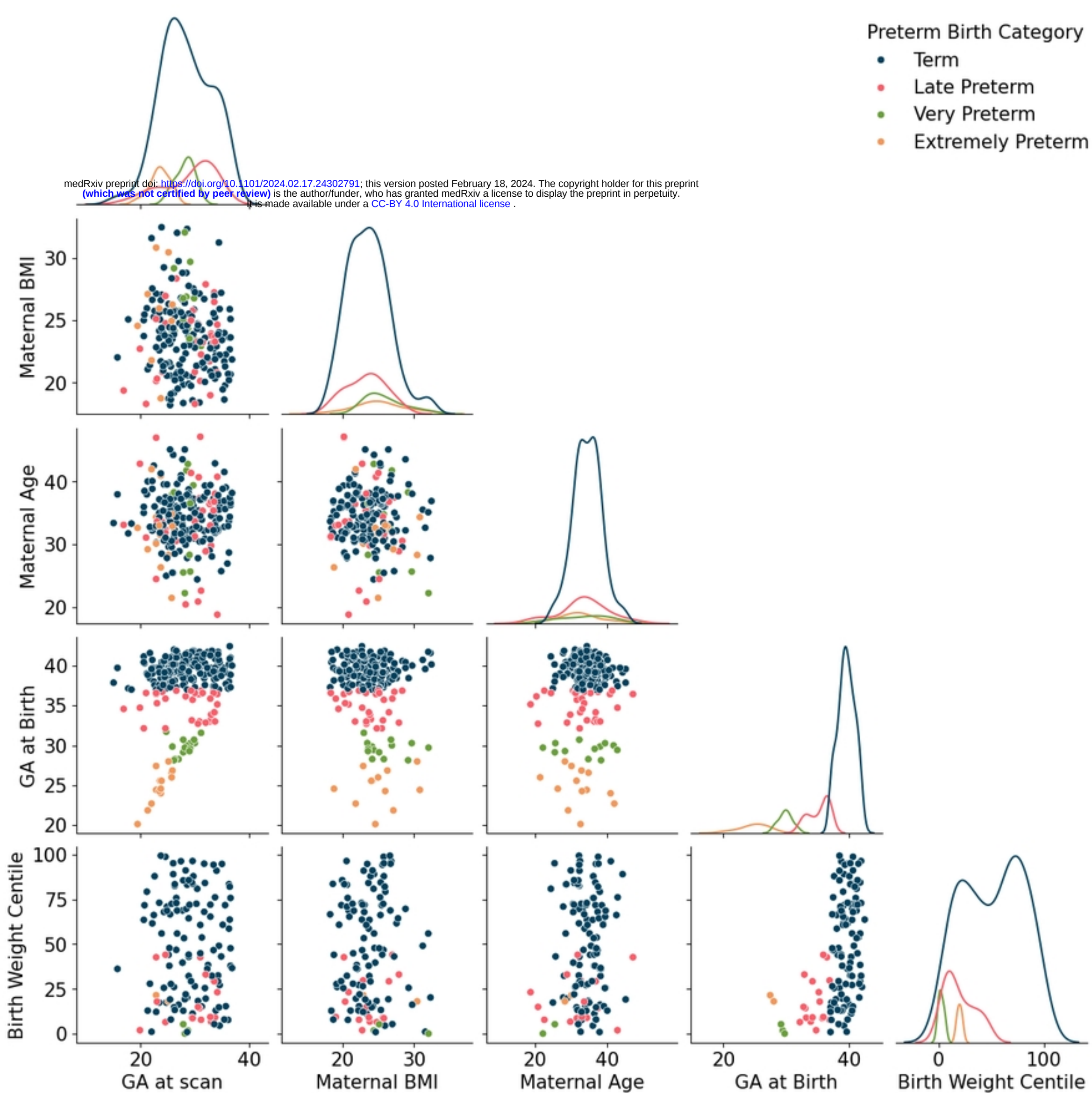


Figure 2

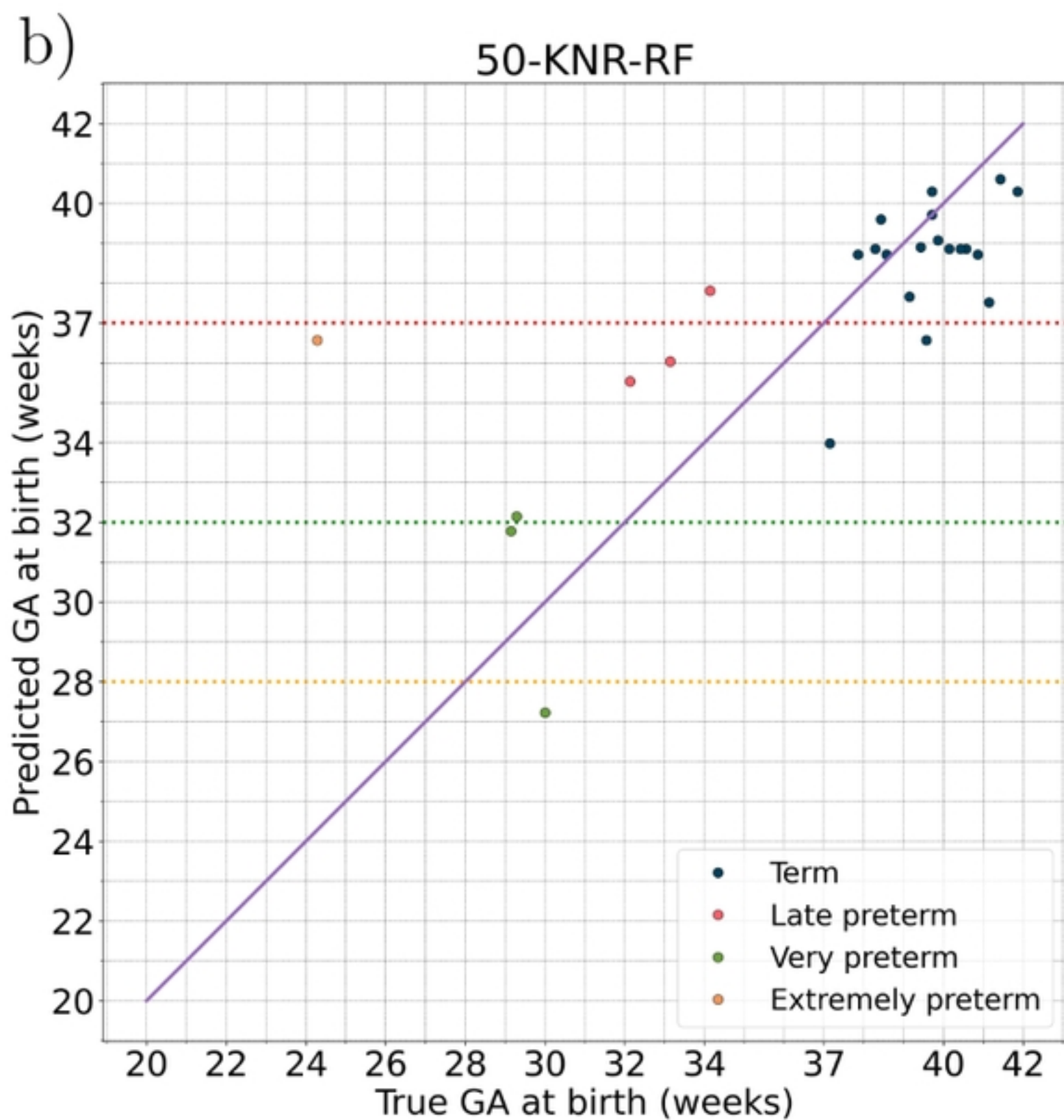
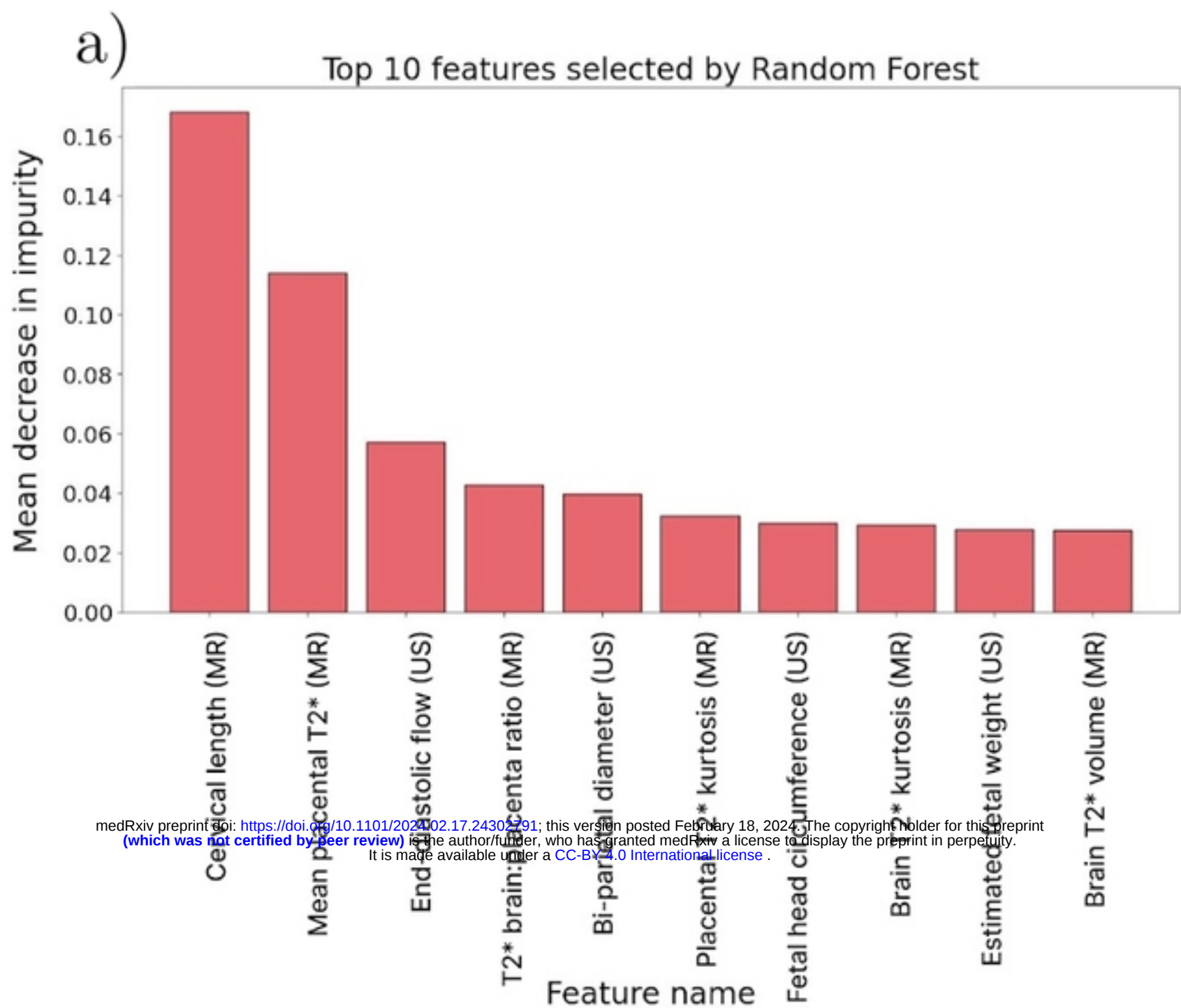


Figure 3