MELBA Journal Machine Learning for Biomedical Imaging

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 **Cromb** [1], Mary A. **Rutherford** [1], Lisa **Story** [1,3], Emma C. **Robinson** [2], Jana **Hutter** [1,4],

1 Early Life Imaging department, King's College London, London, UK
2 Biomedical Computing department, King's College London, London, UK
3 Women's Health department, School of Life Course and Population Sciences, King's College London, London, UK
4 Institute for Information Processing, Leibniz University Hannover, Hannover, Germany

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 333 control cases and 93 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. Performance was evaluated using stratified 10-fold cross-validation. The pipeline achieves an R^2 score of 0.13 and a mean absolute error of 2.74 weeks. It also achieves a 0.77 accuracy, 0.59 sensitivity, and 0.82 specificity across folds. The predominant features selected by the pipeline include cervical length and statistics derived from 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. Our code is available at `https://github.com/dfajardorojas/ml-for-preterm-birth-`.

Keywords

Machine Learning, Preterm Birth, Fetal MRI

Article informations

`https://doi.org/10.59275/j.melba.2026-f34b`

Volume 2026, Received: 2026-02-28, Published 2026-06-05

Corresponding author: diego.fajardo_rojas@kcl.ac.uk

1. Introduction

Preterm birth is defined as a live birth before 37 completed weeks of gestation (WHO, 2018). It is estimated that every year 13.4 million babies are born prematurely, corresponding to a global preterm rate of around 9.9% (Ohuma et al., 2023). 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 (Perin et al., 2021). Complications associated with preterm birth are also the leading cause of neonatal mortality, accounting for 36% of these deaths (Perin et al., 2021). 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 (Ancel et al., 2015; Santhakumaran et al., 2017; Bell et al., 2022). Despite advances in perinatal and neonatal care (Ancel et al., 2015; Santhakumaran et al., 2017; Bell et al., 2022; Blencowe et al., 2012; Cheong et al., 2020; Boland et al., 2021) 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 (Cheong et al., 2020; Boland

et al., 2021; Allen et al., 2011). Short-term outcomes of premature birth include infections, bronchopulmonary dysplasia, retinopathy, necrotising enterocolitis, and brain disorders (Costeloe et al., 2012). 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 (Vanes et al., 2021; Allen et al., 2011; Jarjour, 2014). Similar to mortality rates, the incidence and severity of short- and long-term consequences of preterm birth are inversely related to GA at birth (Costeloe et al., 2012; Moore et al., 2012; Moster and Markestad, 2008). GA at birth is also correlated to social aspects later in life such as income and education level (Moster and Markestad, 2008).

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 (Waitzman et al., 2021). 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 (Petrou et al., 2019).

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) (WHO, 2018), with further categorisation by clinical presentation: medically induced (or iatrogenic) and spontaneous (Moutquin, 2003). 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) (Goldenberg et al., 2008; Morken et al., 2006), the aetiologies underlying spontaneous preterm birth are complex, varied, and poorly understood (Frey and Klebanoff, 2016). 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 (Goldenberg et al., 2008; Suff et al., 2022). However, definitive causes are registered for only 50% (Menon, 2008; Muglia and Katz, 2010) of cases. As such, spontaneous preterm birth should more broadly be considered a syndrome resulting from multiple intricate causes (Goldenberg et al., 2008; Romero et al., 2006).

Despite this complexity, several risk factors have been identified (Goldenberg et al., 2008; Frey and Klebanoff, 2016; Cobo et al., 2020) (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 (Kramer et al., 2011).

Table 1: Most common risk factors for preterm birth (Goldenberg et al., 2008; Frey and Klebanoff, 2016; Cobo et al., 2020).

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) (Suff et al., 2018; Goldenberg et al., 2008). The other two biomarkers are mid-trimester cervical length below 25mm (Suff et al., 2018; Romero et al., 2014), 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% (Goldenberg et al., 2008; Suff et al., 2018). These factors have been combined within clinical practice to improve their predictive capabilities (Carter et al., 2019; Watson et al., 2021). In other analyses, the combination of these predictors also reduced the average cost of high-risk pregnancies (Desplanches et al., 2018; Baaren et al., 2013).

Imaging modalities beyond ultrasound are not currently incorporated into routine preterm birth risk assessment. Magnetic Resonance Imaging (MRI) is an imaging modality

with good potential to investigate preterm birth. Fetal MRI can be 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 (Tocchio et al., 2015; Lum and Tsiouris, 2020; Ray et al., 2016). Another key advantage of MRI is that it offers multiple complementary contrasts that can support comprehensive functional evaluation of fetal and maternal tissues (Story et al., 2018). Available contrasts include T2-weighted anatomical imaging, T2* relaxometry (which provides an indirect measure of oxygenation (Sørensen et al., 2020)), diffusion MRI (which can quantify alterations in tissue microstructure (Avena-Zampieri et al., 2022; Lee et al., 2009; Slator et al., 2019; Kristi B et al., 2020)), flow measurements and T2 relaxometry. Past studies have largely focused on individual organs such as investigating changes to lung (Story et al., 2019b), thymus volumes (Story et al., 2020), or assessing placental microstructure by measuring T2* and ADC values (Hutter et al., 2019). One study measured umbilical vein T2 values as a potential marker of intrauterine growth restriction (Zhu et al., 2015).

2. Related Works

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 (Wlodarczyk et al., 2021) reported accuracies of approximately 0.77 (Esty et al., 2018; Wlodarczyk et al., 2019; Prema and Pushpalatha, 2019), while studies based on electrohysterography reported values above 0.94 (Chen and Xu, 2020; Despotović et al., 2018; Sadi-Ahmed et al., 2017), with the latter, however, only including records of women with recorded contractile activity (Sadi-Ahmed et al., 2017; Fele-Zorz et al., 2008).

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 (Kojita et al., 2021; Shen et al., 2022). Namburete et al. (2015) 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.

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 (Zhou, 2012). Stacking was introduced by Wolpert (1992) to improve the predictions and generalisability of individual classification models. Breiman (1996) showed that stacking was also suitable for regression problems, while Ting and Witten (1999) 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 (Dietterich, 2000).

In recent years, stacking has been successful at various tasks such as genomic prediction (Liang et al., 2021), protein interactions prediction (Yi et al., 2020), or prostate cancer detection (Wang et al., 2019). These works take advantage of more recent ML learning models, e.g. Yi et al. (2020) use Support Vector Machines and XGBoost models as part of their base models, while Wang et al. (2019) explore using a Random Forest as their meta-model.

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.

3. 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. An ablation study, which investigates the impact of each component, is also described in detail. Fig 1 illustrates the workflow of the project. The reader is invited to refer to it repeatedly to complement the description that follows.

3.1 Data

The data set used for this work comprises clinical records, MR data, and parameters manually extracted from ultrasound from 489 singleton pregnancies. From the 489 cases originally considered, 47 cases were excluded as they were lacking GA at delivery. We also removed 16 cases scanned

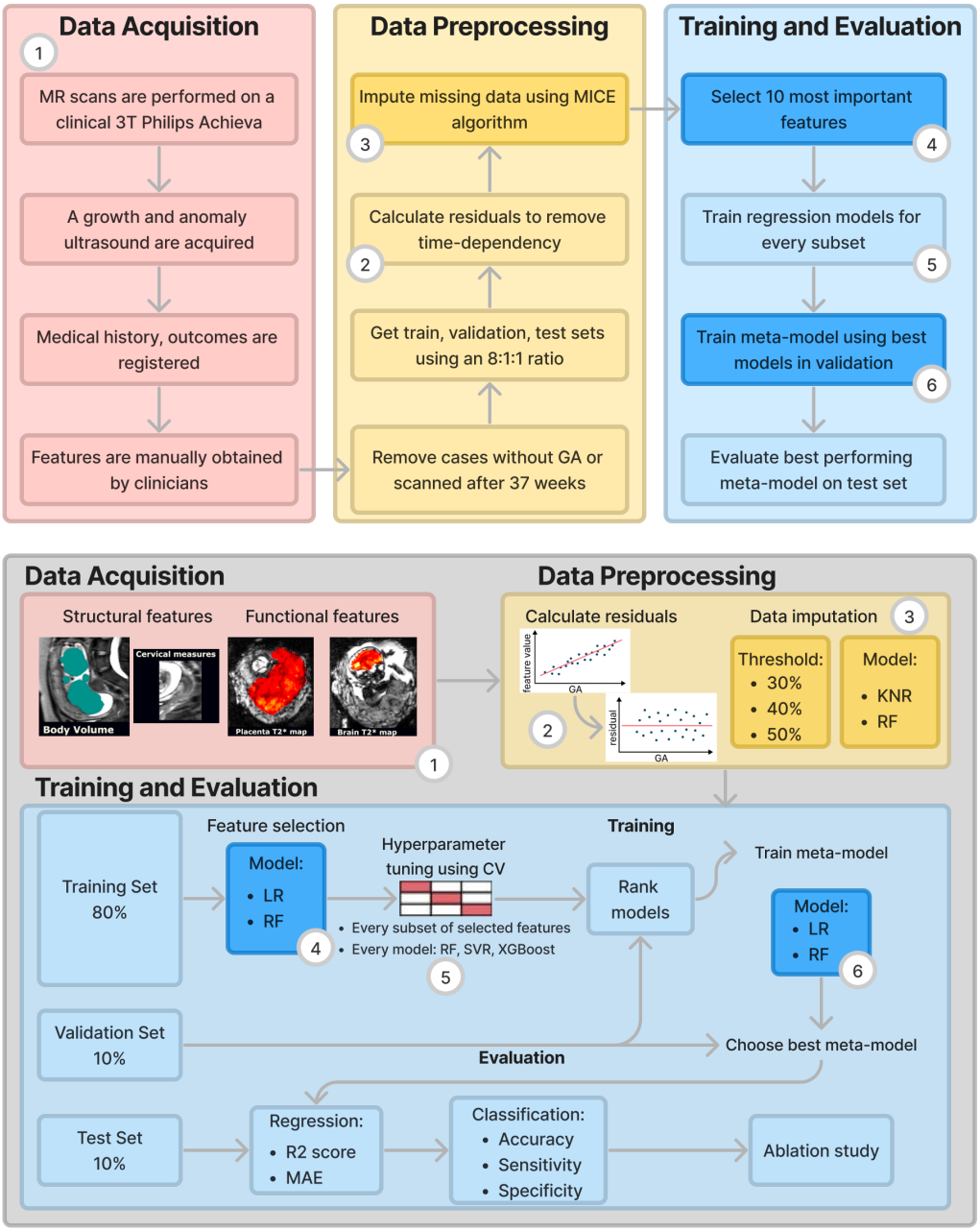


Figure 1: **Schematic representation of the project pipeline.** The boxes in a darker shade denote steps of the pipeline where different design options were explored. The top part of the figure shows the flow of the pipeline with fixed design options, while the bottom part explicitly indicates the different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

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 426 cases (see Appendix D).

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.

Data were split into 10 stratified folds of the same size, i.e. keeping an equal proportion of term and preterm birth cases in each set. These folds were used to obtain train, validation, and test sets with an 8:1:1 ratio across cross-validation iterations.

3.2 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). All mothers were scanned in supine position. 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 (Uus et al., 2020) and learning-based segmentation (Uus et al., 2023) 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 Hutter et al. (2019), 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.

3.3 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 Appendix A 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.

3.4 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.

3.4.1 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 (Story et al., 2019b, 2021, 2020; Papageorghiou et al., 2014), 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 (Cook and Weisberg, 1986). See index 2 in Fig 1.

3.4.2 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 426 cases. For this reason, a regression-based approach to imputation, known as Multivariate imputation by chained equations (MICE) (Azur et al., 2011), was investigated (see Appendix E). Following guidance from the literature, ten iterations of the model were performed (Azur et al., 2011; Jäger et al., 2021), with two different regression models: weighted K-Nearest Neighbours (KNR) (Bicego and Loog, 2016) and Random Forests (RF) (Breiman, 2001). Both models were implemented in the standard way using Sci-kit Learn (Pedregosa et al., 2011). Imputation should not be applied

to features with arbitrarily large amounts of missing data (Bertsimas et al., 2018; Jäger et al., 2021). Thus the impact of discarding features with more than 30%, 40%, or 50% missing values was investigated (see Appendix B 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.

3.5 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 (Wolpert, 1992). Base models consisted of: Random Forests (RF) (Breiman, 2001), Support Vector Regression (SVR) (Smola and Schölkopf, 2004), and XGBoost (Chen and Guestrin, 2016). Each was chosen due to unique strengths: RF are interpretable and robust to overfitting (Gzar et al., 2022); SVR are robust to outliers and well-suited to small data sets (Fernández-Delgado et al., 2019; Kinaneva et al., 2021); XGBoost offers state-of-the-art performance from sparse data sets (Chen and Guestrin, 2016). Importantly they are all capable of capturing non-linear relationships but approach regularisation in different ways (Fernández-Delgado et al., 2019). 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 Appendix A), feature selection was also performed to discourage overfitting. Two simple models were explored: Linear Regression (LR) 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, with hyperparameters (see Appendix C) optimised using 3-fold cross-validated grid search (Krstajic et al., 2014). The metric used for optimisation was the coefficient of determination (R^2) (Casella et al., 2002). 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.

The procedure described in this subsection was performed independently across 10 cross-validation iterations using the splits introduced in subsection 3.1.

3.6 Ablation Study

An ablation study was conducted to validate the design of the proposed pipeline, with results compared against the best performing meta-model. Due to the computational cost of the full pipeline, the study was performed on the cross-validation iteration with the highest performance on the test set. 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.

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

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.

4. Correcting data imbalance: base models were trained with imputation and upsampling preterm cases in the training set, without performing any upsampling or feature selection.

5. Feature selecting and upsampling: This equates to evaluating the best performing base model, obtained without ensembling.

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.

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

3.7 Comparison with Existing Classification Studies

As mentioned in the Introduction, preterm birth prediction has primarily been addressed as a classification problem. Although the primary objective of this work is to predict GA at birth as a regression task, these predictions can be classified into term or preterm, allowing comparison with previous classification-based approaches. A comparative analysis was performed to place the results of this work within the context of existing classification-based studies.

4. Results

4.1 Data Exploration

Table 2 shows key demographics, clinical information, and outcomes, divided into preterm and control cohorts. Specifically, the data set consisted of 333 control cases and 93 preterm cases. The distribution of the data according to the four temporal categories was 78.2% term, 10.8% late preterm, 4.9% very preterm, and 6.1% 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 Appendix B.

4.2 Meta-model

The best performing meta-model was obtained 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. After training RF, SVR, and XGBoost models with every non-empty subset of the selected features, the 18 models with the highest R^2 score on the validation set were used as input for a RF meta-model. In what follows, this meta-model will be referred to by abbreviating its components, i.e. 50-KNR-RF.

Although different features were selected in each of the ten cross-validation iterations, cervical length, mean placental T2*, placental T2* lacunarity, and bi-parietal diameter were always selected. Fig. 3 (a) shows the average mean decrease in impurity for the ten features with the highest average importance across folds. This is the metric used by Random Forests to quantify the importance of each feature (Nembrini et al., 2018).

For every iteration, the metrics used for evaluation 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. Table 3 shows the performance of the meta-model across the ten cross-validation folds. On average, the model achieved an R^2 score of 0.13 and a MAE of 2.74 weeks, as well as 0.77 accuracy, 0.59 sensitivity, and 0.82 specificity. In eight of the ten folds, the MAE was less than 3 weeks. Sensitivity greater than 0.66 was observed in six of the ten folds, although it remained below 0.5 in the remaining four. On the other hand, specificity remained consistently high (≥ 0.67), showing a more reliable identification of term cases. The predictions made by 50-KNR-RF on all folds are depicted in Fig. 3 (b).

To explore whether performance differed across clinically relevant obstetric subgroups, we evaluated the final meta-model stratified by parity and previous preterm birth history. The reported metrics correspond to the average performance across the ten cross-validation folds and are presented in Table 4. Performance in both parity subgroups was comparable to that observed in the full cohort, with similar MAE (2.75 weeks), accuracy (0.77), and sensitivity (0.59). This is also the case for the cohort of women without a previous preterm birth history. In contrast, the model performance worsened within the subgroup of women with a previous history of preterm birth, demonstrating reduced accuracy (0.66) and increased MAE (3.18 weeks), although sensitivity remained consistent with that of the full cohort (0.61).

4.3 Ablation study

In line with the optimisation metric used during training, Fold 9 was selected for the ablation study, as it achieved the highest R^2 score among the cross-validation folds (Table 3). The performance of each of the models in the ablation study is reported in Table 5.

These results are helpful to understand the contributions and limitations of every element of the pipeline. In experiments 1) and 2) it can be seen that the out-of-the-box XGBoost model outperforms XGBoost trained after feature deconfounding, which is consistent with XGBoost's predictive capabilities with minimal feature preprocessing (Chen and Guestrin, 2016).

The sensitivity of all models trained without upsampling is less than 0.6, while the sensitivity of all but one of the models trained with upsampling is greater than 0.66. This suggests that upsampling was effective at addressing class imbalance. Applying feature selection in addition to upsampling results in a better overall performance than using upsampling alone: the RF obtained in experiment 5) achieves better regression and classification metrics than all models from previous experiments. It can also be noted that for individual models (i.e. experiments 3) against 4)) the increase in sensitivity is observed alongside a higher MAE. This trade-off was not observed for the final meta-model (i.e. experiments 6) against 7)) where the complete pipeline

Table 2: **Key demographics, clinical information, and outcomes of participants.** Reported values are the mean ± SD in the case of continuous variables and percentages for discrete variables. The numbers in brackets are ranges.

	Preterm birth cohort	Control cohort
	Current pregnancy	
Gestational age at scan [weeks]	26.88±4.40 [16.86,34.57]	28.15±5.43 [15.00,36.86]
Maternal BMI at scan [kg/m2]	24.24±3.21 [18.26,32.05]	24.04±2.94 [18.00,32.46]
Maternal age at scan [years]	33.75±6.08 [18.82,48.74]	34.75±3.84 [18.81,45.20]
	Obstetric history	
Previous preterm birth	13.98%	5.71%
	Outcome	
Gestational age at birth	30.94±5.03 [20.14,36.86]	39.67±1.24 [37.00, 42.43]
Birth weight centile	35.34±32.75 [0.00,95.53]	55.65±26.69 [0.00,99.97]
Fetal sex	53.93% female, 46.07% male	53.40% female, 46.60% male

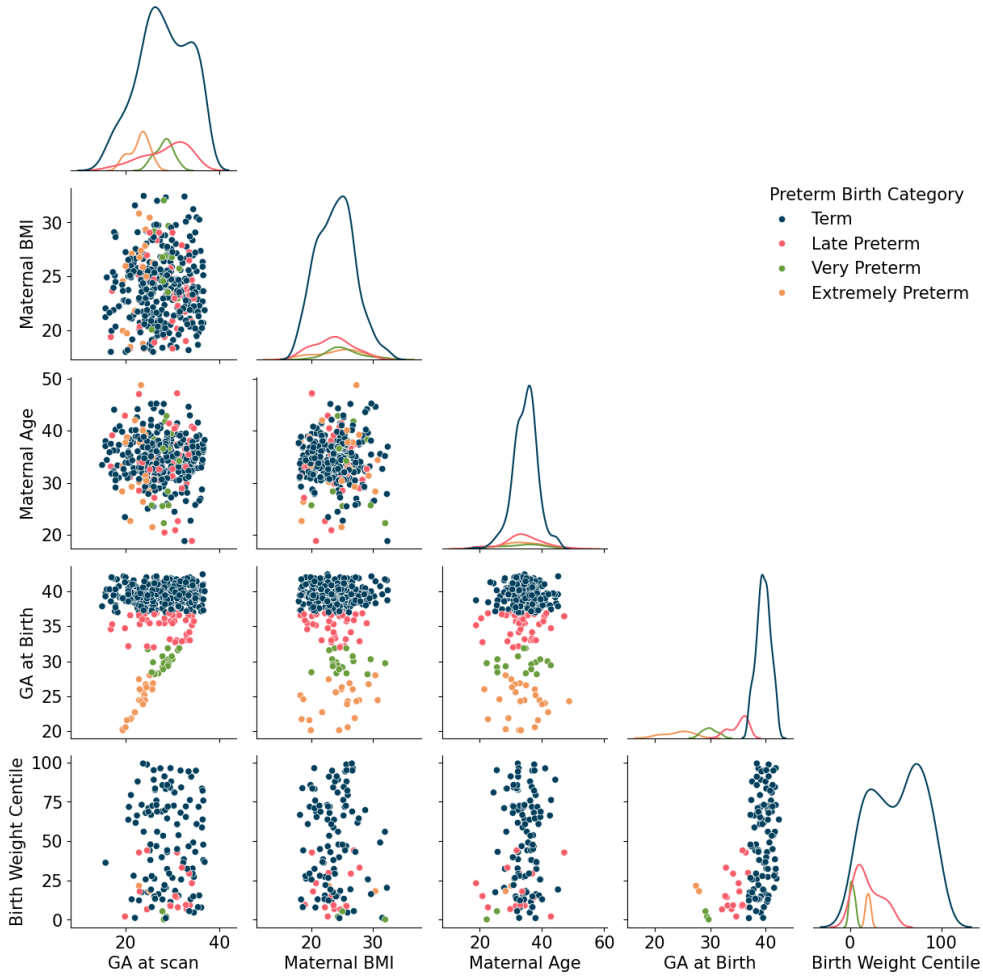


Figure 2: **Data exploration.** Distributions of GA at scan, maternal BMI at scan, maternal age, GA at birth, and birth weight centile (diagonal panels), and their pairwise relationships (off-diagonal panels). Data are colour-coded by preterm birth category.

achieves a better performance across all metrics than its counterpart.

Lastly, the 50-KNR-RF meta-model achieved the best and most balanced performance, exhibiting the highest accuracy (0.83) along with high sensitivity (0.67) and specificity (0.88). While models like the SVR in experiment 3) and the RF in experiment 4) attained higher specificity and sensitivity, it can be seen that these models are highly biased towards predicting term and preterm cases respectively.

4.4 Comparison with Existing Classification Studies

Classifying the predictions of 50-KNR-RF into term or preterm allows for comparison with other models in the

Table 3: **Cross-validation results.**

Fold	R^2 score	MAE	Accuracy	Sensitivity	Specificity
Fold 1	0.05	2.70	0.74	0.22	0.88
Fold 2	0.15	2.62	0.81	0.67	0.85
Fold 3	-0.05	3.50	0.63	0.33	0.71
Fold 4	-0.22	3.80	0.60	0.40	0.67
Fold 5	0.07	2.99	0.77	0.70	0.79
Fold 6	0.42	2.50	0.77	0.70	0.79
Fold 7	0.08	2.64	0.76	0.89	0.73
Fold 8	0.06	2.17	0.79	0.44	0.88
Fold 9	**0.47**	**2.19**	**0.83**	**0.67**	**0.88**
Fold 10	0.29	2.29	0.98	0.89	1.00
Average	**0.13**	**2.74**	**0.77**	**0.59**	**0.82**

Table 4: **Model performance across 10 folds stratified by parity and previous preterm birth history (PBH).**

Group	n	R^2 score	MAE	Accuracy	Sensitivity	Specificity
Parity						
Nulliparous women	302	0.05	2.75	0.77	0.59	0.81
Multiparous women	124	0.24	2.73	0.77	0.59	0.84
Previous preterm birth history						
No previous PBH	394	0.12	2.71	0.78	0.59	0.82
Previous PBH	32	0.01	3.18	0.66	0.61	0.68
Full cohort						
Full cohort	426	0.13	2.74	0.77	0.59	0.82

Table 5: **Evaluation of the models in the ablation study.**

Model	R^2 score	MAE	Accuracy	Sensitivity	Specificity
1) Out-of-the-box XGBoost	0.10	2.75	0.71	0.56	0.76
2) XGBoost with deconfounding	0.03	2.96	0.67	0.56	0.7
3) Imputation, RF	0.18	2.89	0.64	0.33	0.73
3) Imputation, SVR	0.16	2.44	0.83	0.22	1
3) Imputation, XGBoost	0.28	2.66	0.69	0.33	0.79
4) Correcting data imbalance, RF	-0.29	4.09	0.21	0.89	0.03
4) Correcting data imbalance, SVR	0.01	3.21	0.69	0.56	0.73
4) Correcting data imbalance, XGBoost	0.27	2.74	0.64	0.67	0.64
5) Feature selecting and upsampling, RF	0.37	2.48	0.76	0.67	0.79
6) Meta-model without upsampling	0.39	2.38	0.79	0.44	0.88
7) Meta-model, 50-KNR-RF	0.47	2.19	0.83	0.67	0.88

literature. As shown in Table 6, the sensitivity achieved by 50-KNR-RF is lower than that reported by previous studies. However, such comparisons should be interpreted taking into account important methodological differences. Wlodarczyk et al. (2019) developed an automated method for cervical length and anterior cervical angle measurements from transvaginal ultrasound. Subsequently, they make preterm birth predictions based on these features, but explicitly use a balanced data set with precomputed markers for their classification experiments instead of their original data set. (Esty et al., 2018) develop their models on very large population-level registry data sets (over 669,000 subjects) and include features such as obstetrical complications and registers of premature rupture of membranes—not only lead to preterm delivery (Goldenberg et al., 2008) but—under the framework of the present work would be considered a clinical outcome and would generally not be available at the time of imaging studies. (Prema and Pushpalatha, 2019) restrict their analysis to pregnancies complicated by diabetes mellitus or gestational diabetes mellitus. In contrast, 50-KNR-RF is trained and evaluated on an imbalanced cohort representative of real-world prevalence, without the use of pregnancy outcomes as predictors.

To the best of our knowledge, the only other study on predicting GA at delivery using ML is the one by Heinsalu et al. (2021), 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. However, their implementation suffers from data leakage at the imputation stage and was not cross-validated,

Table 6: **Classification performance of the meta-model obtained by this work and other models from recent studies.**

Metric / Model	**Accuracy**	**Sensitivity**	**Specificity**
Esty et al. (2018)	0.72	0.93	0.71
Esty et al. (2018)	0.77	0.84	0.77
Wlodarczyk et al. (2019)	0.78	0.74	0.85
Prema and Pushpalatha (2019)	0.76	0.84	0.73
50-KNR-RF	0.77	0.59	0.82

which makes these results unreliable. Nevertheless, the framework they established is valuable and served as the basis of the present work.

5. 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.

Our 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 (Story et al., 2018). 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 (Story et al., 2019a). At the current level of performance, the MAE of 2.74 weeks indicates that predictions are better suited to inform broader antenatal planning than to guide interventions requiring narrow timing windows, such as corticosteroid administration. The R^2 score of 0.13 reflects that a substantial proportion of the variance in GA at delivery remains unexplained by the model, and predictions should be used as a complement to current clinical management rather than as a stand-alone tool.

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 (Hutter et al., 2019; Purisch and Gyamfi, 2017). The most common clinical indicator, number of previous preterm births, was not a predominant feature.

While our 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 and very preterm subjects available during training help explain the poor performances on these categories. A stratified cross-validation framework was adopted to provide a reliable evaluation of performance and generalisability. Although the MAE remained within the 2-3 weeks range for eight of the ten folds, sensitivity showed substantial variability across folds demonstrating limited ability to detect preterm cases on different data distributions. On the other hand, specificity was higher and more consistent. This is further evidence of the meta-model struggle to overcome strong class imbalance. The variability observed across folds highlights the influence of specific data distribution on performance and underscores the need for further evaluation in larger and more balanced data sets.

The lower performance observed in the subgroup with a previous history of preterm birth (Table 4) further illustrates the challenges imposed by data imbalance. The small number of cases in this group ($n = 32$) likely reduced the model's ability to generalise within this clinically important population. This may reflect both sample size constraints and variability in the underlying mechanisms of recurrent preterm birth.

All MR features were acquired at a single GA for each patient. The biological pathways that lead to preterm birth change progressively across pregnancy (Goldenberg et al., 2008). While deconfounding via internally studentised residuals addresses temporal dependency with respect to GA, the data offer no longitudinal description of the trajectory of the feature values.

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

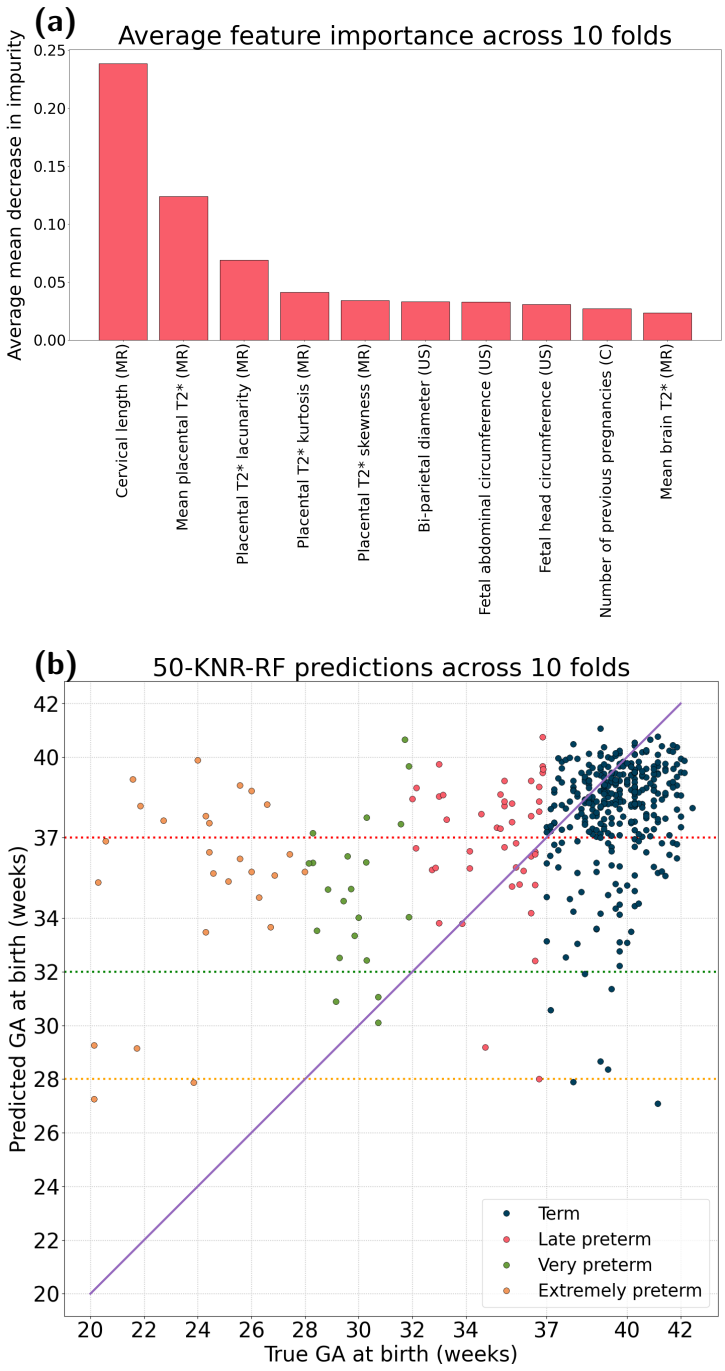


Figure 3: **Feature importances and predictions of the meta-model.** (a) Average mean decrease in impurity for the ten highest-ranked features based on their average importance across cross-validation folds. (b) Predictions made by the meta-model 50-KNR-RF across cross-validation folds, colourised according to their true preterm temporal category.

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 (Garcia-Eulate et al., 2011). 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. Similarly, as the data used for this paper comprised only singleton pregnancies, the generalisability of the model to multiple gestations remains to be established. Evaluation in twin pregnancies would provide further validation of the robustness of the pipeline across different cohorts.

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 (Wlodarczyk et al., 2019) and to estimate GA at scan from fetal brain MRI (Kojita et al., 2021; Shen et al., 2022).

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

This work was supported by funding from the EPSRC Centre for Doctoral Training in Smart Medical Imaging (EP/S022104/1) to Diego Fajardo-Rojas, from the Wellcome/EPSRC Centre for Medical Engineering (WT203148/Z/16/Z a UKRI FLF (MR/T018119/1), and DFG Heisenberg funding through the High Tech Agenda Bavaria (502024488) to Jana Hutter, and from the NIHR Advanced Fellowship (NIHR3016640) and the MRC grant (MR/W019469/1) to Lisa Story.

The authors acknowledge the invaluable help of the

radiographers and midwives while acquiring the data presented here.

Ethical Standards

The data used for this work were 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.

Conflicts of Interest

We declare we do not have conflicts of interest.

Data availability

All data and code used in this paper are available online at `https://github.com/dfajardorojas/ml-for-preterm-birth`.

References

Marilee Allen, Elizabeth Cristofalo, and Christina Kim. Outcomes of preterm infants: Morbidity replaces mortality. *Clinics in perinatology*, 38:441–54, 09 2011. .

Pierre-Yves Ancel, François Goffinet, Pierre Kuhn, Bruno Langer, Jacqueline Matis, Xavier Hernandorena, Pierre Chabanier, Laurence Joly-Pedespan, Bénédicte Lecomte, Françoise Vendittelli, Michel Dreyfus, Bernard Guillois, Antoine Burguet, Pierre Sagot, Jacques Sizun, Alain Beuchée, Florence Rouget, Amélie Favreau, Elie Saliba, and Monique Kaminski. 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*, 169, 01 2015. .

Carla L. Avena-Zampieri, Jana Hutter, Mary Rutherford, Anna Milan, Megan Hall, Alexia Egloff, David F.A. Lloyd, Surabhi Nanda, Anne Greenough, and Lisa Story. Assessment of the fetal lungs in utero. *American Journal of Obstetrics & Gynecology MFM*, 4(5):100693, 2022. ISSN 2589-9333. . URL `https://www.sciencedirect.com/science/article/pii/S2589933322001252`.

Melissa Azur, Elizabeth Stuart, Constantine Frangakis, and Philip Leaf. Multiple imputation by chained equations: What is it and how does it work? *International journal of methods in psychiatric research*, 20:40–9, 03 2011. .

Gert-Jan Baaren, Jolande Vis, William Grobman, Patrick Bossuyt, Brent Opmeer, and Ben W Mol. Cost-effectiveness anlysis of cervical length measurement and fibronectin testing in women with threatened preterm labor. *American journal of obstetrics and gynecology*, 209, 06 2013. .

Edward Bell, Susan Hintz, Nellie Hansen, Carla Bann, Myra Wyckoff, Sara Demauro, Michele Walsh, Betty Vohr, Barbara Stoll, Waldemar Carlo, Krisa Meurs, Matthew Rysavy, Ravi Patel, Stephanie Merhar, Pablo Sánchez, Abbot Laptook, Anna Maria Hibbs, Charles Cotten, Carl D'Angio, and Abhik Das. Mortality, in-hospital morbidity, care practices, and 2-year outcomes for extremely preterm infants in the us, 2013-2018. *JAMA*, 327:248–263, 01 2022. .

Dimitris Bertsimas, Colin Pawlowski, and Ying Zhuo. From predictive methods to missing data imputation: An optimization approach. *Journal of Machine Learning Research*, 18:1–39, 04 2018.

Manuele Bicego and Marco Loog. Weighted k-nearest neighbor revisited. pages 1642–1647, 12 2016. .

Hannah Blencowe, Simon Cousens, Mikkel Oestergaard, Doris Chou, Ann-Beth Moller, Rajesh Narwal, Alma Adler, Claudia Garcia, Sarah Rohde, Lale Say, and Joy Lawn. 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*, 379:2162–72, 06 2012. .

Rosemarie Boland, Jeanie Cheong, and Lex Doyle. Changes in long-term survival and neurodevelopmental disability in infants born extremely preterm in the post-surfactant era. *Seminars in Perinatology*, 45:151479, 08 2021. .

Leo Breiman. Stacked regressions. *Mach. Learn.*, 24(1): 49–64, 1996. URL `http://dblp.uni-trier.de/db/journals/ml/ml24.html#Breiman96a`.

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 10 2001. .

Jenny Carter, Paul Seed, Helena Watson, Anna David, Jane Sandall, Andrew Shennan, and Rachel Tribe. 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*, 55, 08 2019. .

George Casella, Roger L. Berger, and Brooks/Cole Publishing Company. *Statistical Inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning, 2002. ISBN 9780534243128. URL https://books.google.co.uk/books?id=0x_vAAAAMAAJ.

Lili Chen and Huoyao Xu. Deep neural network for semi-automatic classification of term and preterm uterine recordings. *Artificial Intelligence in Medicine*, 105:101861, 04 2020. .

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754, 2016. URL http://arxiv.org/abs/1603.02754.

Jeanie LY Cheong, Alicia J Spittle, Alice C Burnett, Peter J Anderson, and Lex W Doyle. Have outcomes following extremely preterm birth improved over time? In *Seminars in Fetal and Neonatal Medicine*, volume 25, page 101114. Elsevier, 2020.

Teresa Cobo, Marian Kacerovsky, and Bo Jacobsson. Risk factors for spontaneous preterm delivery. *International Journal of Gynecology & Obstetrics*, 150:17–23, 07 2020. .

R.Dennis Cook and Sanford Weisberg. *Residuals and Influence in Regression*. Monographs on statistics and applied probability. Chapman and Hall, 1986. URL https://books.google.co.uk/books?id=aMDpswEACAAJ.

Kate L Costeloe, Enid M. Hennessy, Sadia Haider, Fiona Stacey, Neil Marlow, and Elizabeth S. Draper. Short term outcomes after extreme preterm birth in england: comparison of two birth cohorts in 1995 and 2006 (the epicure studies). *The BMJ*, 345, 2012.

Thomas Desplanches, Catherine Lejeune, Cottenet Jonathan, Paul Sagot, and Catherine Quantin. Cost-effectiveness of diagnostic tests for threatened preterm labor in singleton pregnancy in france. *Cost Effectiveness and Resource Allocation*, 16, 06 2018. .

Danica Despotović, Aleksandra Zec, Katarina Mladenovic, Nevena Radin, and Tatjana Loncar-Turukalo. A machine learning approach for an early prediction of preterm delivery. pages 000265–000270, 09 2018. .

Thomas G. Dietterich. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, 2000. URL https://api.semanticscholar.org/CorpusID:56776745.

Alana Esty, Monique Frize, Jeff Gilchrist, and Erika Bariciak. Applying data preprocessing methods to predict premature birth. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6096–6099, 2018. .

Gašper Fele-Zorz, Gorazd Kavsek, Živa Novak-Antolic, and Franc Jager. 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*, 46:911–22, 04 2008. .

Manuel Fernández-Delgado, Manisha Sanjay Sirsat, Eva Cernadas, Sadi Alawadi, Senén Barro, and Manuel Febrero-Bande. An extensive experimental survey of regression methods. *Neural Networks*, 111:11–34, 2019. ISSN 0893-6080. . URL https://www.sciencedirect.com/science/article/pii/S0893608018303411.

Heather Frey and Mark Klebanoff. The epidemiology, etiology, and costs of preterm birth. *Seminars in Fetal and Neonatal Medicine*, 21, 01 2016. .

Reyes Garcia-Eulate, David Garcia-Garcia, Pablo Domínguez, Jose Noguera, Esther Luis, Maria Rodriguez-Oroz, and Jose Zubieta. Functional bold mri: Advantages of the 3 t vs. the 1.5 t. *Clinical imaging*, 35:236–41, 05 2011. .

Robert Goldenberg, Jennifer Culhane, Jay Iams, and Roberto Romero. Epidemiology and causes of preterm birth. *Lancet*, 371:75–84, 02 2008. .

Dunia Abas Gzar, Ali Majeed Mahmood, and Maythem K. Abbas. A comparative study of regression machine learning algorithms: Tradeoff between accuracy and computational complexity. *Mathematical Modelling of Engineering Problems*, 9(5):1217–1224, December 2022. ISSN 2369-0747. . URL http://dx.doi.org/10.18280/mmep.090508.

Riine Heinsalu, Logan Williams, Aditi Ranjan, Carla Avena Zampieri, Alena Uus, Emma Robinson, Mary Rutherford, Lisa Story, and Jana Hutter. *Predicting preterm birth using multimodal fetal imaging*. Springer, July 2021.

Jana Hutter, Paddy J. Slator, Laurence Jackson, Ana Dos Santos Gomes, Alison Ho, Lisa Story, Jonathan O'Muircheartaigh, Rui P. A. G. Teixeira, Lucy C. Chappell, Daniel C. Alexander, Mary A. Rutherford, and Joseph V. Hajnal. Multi-modal functional mri to explore placental function over gestation. *Magnetic Resonance in Medicine*, 81(2):1191–1204, 2019. . URL https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.27447.

Imad Jarjour. Neurodevelopmental outcome after extreme prematurity: A review of the literature. *Pediatric Neurology*, 52, 11 2014. .

Sebastian Jäger, Arndt Allhorn, and Felix Biessmann. A benchmark for data imputation methods. *Frontiers in Big Data*, 4, 07 2021. .

Diyana Kinaneva, Georgi Hristov, Petko Kyuchukov, Georgi Georgiev, Plamen Zahariev, and Rosen Daskalov. 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)*, pages 1–6, 2021. .

Yasuyuki Kojita, Hidetoshi Matsuo, Tomonori Kanda, Mizuho Nishio, Keitaro Sofue, Munenobu Nogami, Atsushi Kono, Masatoshi Hori, and Takamichi Murakami. Deep learning model for predicting gestational age after the first trimester using fetal mri. *European Radiology*, 31, 04 2021. .

Michael Kramer, Aris T. Papageorghiou, Jennifer Culhane, Zulfiqar Bhutta, Robert Goldenberg, Michael Gravett, Jay Iams, Agustin Conde-Agudelo, Sarah Waller, Fernando Barros, Hannah Knight, and José Villar. Challenges in defining and classifying the preterm birth syndrome. *American journal of obstetrics and gynecology*, 206:108–12, 10 2011. .

Anderson Kristi B, Hansen Ditte N, Haals Caroline, Sinding Marianne, Petersen Astrid, Frøkjær Jens B, Peters David A, and Sørensen Anne. Placental diffusion-weighted mri in normal pregnancies and those complicated by placental dysfunction due to vascular malperfusion. *Placenta*, 91:52–58, 2020. ISSN 0143-4004. . URL `https://www.sciencedirect.com/science/article/pii/S0143400420300199`.

Damjan Krstajic, Ljubomir Buturovic, David Leahy, and Simon Thomas. Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics*, 6:10, 03 2014. .

Wonyeoul Lee, Ashlee Krisko, Anil Shetty, Lami Yeo, Sonia Hassan, Francesca Gotsch, Swati Mody, Luis GONCALVES, and Roberto Romero. 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*, 34:673–7, 12 2009. .

Mang Liang, Tianpeng Chang, Bingxing An, Xinghai Duan, Lili Du, Xiaoqiao Wang, Jian Miao, Lingyang Xu, Xue Gao, Lupei Zhang, Junya Li, and Huijiang Gao. A stacking ensemble learning framework for genomic prediction. *Frontiers in Genetics*, 12, 2021. ISSN 1664-8021. . URL `https://www.frontiersin.org/articles/10.3389/fgene.2021.600040`.

Mark Lum and Apostolos Tsiouris. Mri safety considerations during pregnancy. *Clinical Imaging*, 62, 02 2020. .

Ramkumar Menon. Spontaneous preterm birth, a clinical dilemma: Etiologic, pathophysiologic and genetic heterogeneities and racial disparity. *Acta obstetricia et gynecologica Scandinavica*, 87:590–600, 02 2008. .

Tamanna Moore, Enid Hennessy, Jonathan Myles, Samantha Johnson, Elizabeth Draper, Kate Costeloe, and Neil Marlow. Neurological and developmental outcome in extremely preterm children born in england in 1995 and 2006: The epicure studies. *BMJ (Clinical research ed.)*, 345:e7961, 12 2012. .

Nils-Halvdan Morken, Karin Kallen, and Bo Jacobsson. Fetal growth and onset of delivery: A nationwide population-based study of preterm infants. *American journal of obstetrics and gynecology*, 195:154–61, 08 2006. .

Dag Moster and Trond Markestad. Long-term medical and social consequences of preterm birth. *The New England journal of medicine*, 359:262–73, 07 2008. .

Jean-Marie Moutquin. Classification and heterogeneity of preterm birth. *BJOG : an international journal of obstetrics and gynaecology*, 110 Suppl 20:30–3, 04 2003. .

Louis J. Muglia and Michael Katz. The enigma of spontaneous preterm birth. *The New England journal of medicine*, 362 6:529–35, 2010.

Ana Namburete, Richard Stebbing, Bryn Kemp, Mohammad Yaqub, Aris T. Papageorghiou, and Julia Noble. Learning-based prediction of gestational age from ultrasound images of the fetal brain. *Medical Image Analysis*, 30, 01 2015. .

Stefano Nembrini, Inke König, and Marvin Wright. The revival of the gini importance? *Bioinformatics (Oxford, England)*, 34, 05 2018. .

Eric Ohuma, Ann-Beth Moller, Ellen Bradley, Samuel Chakwera, Laith Hussain-Alkhateeb, Alexandra Lewin, Yemisrach Okwaraji, Wahyu Mahanani, Emily Johansson, Tina Lavin, Diana Fernandez, Giovanna Domínguez, Ayesha Costa, Jenny Cresswell, Julia Krasevec, Joy Lawn, Hannah Blencowe, Jennifer Requejo, and Allisyn Moran. National, regional, and global estimates of preterm birth in 2020, with trends from 2010: a systematic analysis. *The Lancet*, 402:1261–1271, 10 2023. .

Aris T. Papageorghiou, Eric O. Ohuma, Douglas G. Altman, Tullia Todros, Leila Cheikh Ismail, Ann Lambert, Yasmin A. Jaffer, Enrico Bertino, Michael G. Gravett,

Manorama B Purwar, Julia Alison Noble, Ruyan Pang, Cesar Gomes Victora, Fernando C Barros, Maria Carvalho, Laurent J Salomon, Zulfiqar Ahmed Bhutta, Stephen H Kennedy, and José Villar. International standards for fetal growth based on serial ultrasound measurements: the fetal growth longitudinal study of the intergrowth-21st project. *The Lancet*, 384:869–879, 2014. URL `https://api.semanticscholar.org/CorpusID:28633594`.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Jamie Perin, Amy Mulick, Diana Yeung, Francisco Villavicencio, Gerard Lopez, Kathleen Strong, David Prieto-Merino, Simon Cousens, and Robert Black. 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*, 6, 11 2021. .

Stavros Petrou, Hei Hang Yiu, and Joseph Kwon. Economic consequences of preterm birth: a systematic review of the recent literature (2009–2017). *Archives of Disease in Childhood*, 104(5):456–465, 2019. ISSN 0003-9888. . URL `https://adc.bmj.com/content/104/5/456`.

Nisana Siddegowda Prema and Mullur Puttabuddi Pushpalatha. *Machine Learning Approach for Preterm Birth Prediction Based on Maternal Chronic Conditions*, pages 581–588. 01 2019. ISBN 978-3-662-53832-6. .

Stephanie Purisch and Cynthia Gyamfi. Epidemiology of preterm birth. *Seminars in Perinatology*, 41, 09 2017. .

Joel Ray, Marian Vermeulen, Aditya Bharatha, Walter Montanera, and Alison Park. Association between mri exposure during pregnancy and fetal and childhood outcomes. *JAMA*, 316:952–961, 09 2016. .

Roberto Romero, Sudhansu Dey, and Susan Fisher. Preterm labor: One syndrome, many causes. *Science (New York, N.Y.)*, 345:760–765, 08 2014. .

Roberto J. Romero, Jimmy Espinoza, Juan Pedro Kusanovic, Francesca Gotsch, Sonia S. Hassan, Offer Erez, Tinnakorn Chaiworapongsa, and Moshe Mazor. The preterm parturition syndrome. *BJOG: An International Journal of Obstetrics & Gynaecology*, 113, 2006.

Nafissa Sadi-Ahmed, Baya Kacha, Hamza Taleb, and Malika Kedir-Talha. Relevant features selection for automatic prediction of preterm deliveries from pregnancy electrohysterograhic (ehg) records. *Journal of Medical Systems*, 41, 11 2017. .

Shalini Santhakumaran, Eugene Statnikov, Daniel Gray, Cheryl Battersby, Deborah Ashby, and Neena Modi. 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*, 103, 09 2017. .

Liyue Shen, Jimmy Zheng, Edward Lee, Katie Shpanskaya, Emily McKenna, Mahesh Atluri, Dinko Plasto, Courtney Mitchell, Lillian Lai, Carolina Guimaraes, Hisham Dahmoush, Jane Chueh, Safwan Halabi, John Pauly, Lei Xing, Quin Lu, Ozgur Oztekin, Beth Kline-Fath, and Kristen Yeom. Attention-guided deep learning for gestational age prediction using fetal brain mri. *Scientific Reports*, 12, 01 2022. .

Paddy J. Slator, Jana Hutter, Marco Palombo, Laurence H. Jackson, Alison Ho, Eleftheria Panagiotaki, Lucy C. Chappell, Mary A. Rutherford, Joseph V. Hajnal, and Daniel C. Alexander. Combined diffusion-relaxometry mri to identify dysfunction in the human placenta. *Magnetic Resonance in Medicine*, 82(1):95–106, 2019. . URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.27733`.

Alex Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14:199–222, 08 2004.

Lisa Story, Jana Hutter, Tong Zhang, Andrew Shennan, and Mary Rutherford. 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*, 222, 01 2018. .

Lisa Story, Nigel Simpson, Anna David, Zarko Z, Phillip Bennett, Matthew Jolly, and Andrew Shennan. Reducing the impact of preterm birth: Preterm birth commissioning in the united kingdom. *European Journal of Obstetrics & Gynecology and Reproductive Biology: X*, 3, 04 2019a. .

Lisa Story, Tong Zhang, Johannes Steinweg, Jana Hutter, Jacqueline Matthew, Theodore Dassios, Paul Seed, Dharmintra Pasupathy, Joanna Allsop, Joseph Hajnal, Anne Greenough, Andrew Shennan, and Mary Rutherford. Foetal lung volumes in pregnant women who deliver very preterm: a pilot study. *Pediatric Research*, 87, 12 2019b. .

Lisa Story, Tong Zhang, Alena Uus, Jana Hutter, Alexia Egloff, Deena Gibbons, Ho Alison, Mudher Al-Adnani, Caroline Knight, Iakovos Theodoulou, Maria Deprez, Paul Seed, Rachel Tribe, Andrew Shennan, and Mary Rutherford. Antenatal thymus volumes in fetuses that delivered ¡32 weeks gestation: An mri pilot study. *Acta Obstetricia et Gynecologica Scandinavica*, 100, 08 2020. .

Lisa Story, Alice Davidson, Prachi Patkee, Bobbi Fleiss, Vanessa Kyriakopoulou, Kathleen Colford, Srividhya Sankaran, Paul Seed, Alice Jones, Jana Hutter, Andrew Shennan, and Mary Rutherford. Brain volumetry in fetuses that deliver very preterm: An mri pilot study. *NeuroImage: Clinical*, 30:102650, 2021. ISSN 2213-1582. . URL https://www.sciencedirect.com/science/article/pii/S2213158221000942.

Natalie Suff, Lisa Story, and Andrew Shennan. The prediction of preterm delivery: What is new? *Seminars in Fetal and Neonatal Medicine*, 24, 09 2018. .

Natalie Suff, Vicky X. Xu, Agnieszka Glazewska-Hallin, Jenny Carter, Shaun Brennecke, and Andrew Shennan. 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*, 271:108–111, 2022. ISSN 0301-2115. . URL https://www.sciencedirect.com/science/article/pii/S0301211522000586.

Anne Sørensen, Jana Hutter, Mike Seed, P. Ellen Grant, and Penny A. Gowland. T2*-weighted placental mri: basic research tool or emerging clinical test for placental dysfunction? *Ultrasound in Obstetrics & Gynecology*, 55(3):293–302, 2020. . URL https://obgyn.onlinelibrary.wiley.com/doi/abs/10.1002/uog.20855.

Kai Ming Ting and Ian H. Witten. Issues in stacked generalization. *Journal of Artificial Intelligence Research*, 10:271–289, May 1999. . URL https://doi.org/10.1613%2Fjair.594.

Shannon Tocchio, Beth Kline-Fath, Emanuel Kanal, Vincent Schmithorst, and Ashok Panigrahy. Mri evaluation and safety in the developing brain. *Seminars in perinatology*, 39, 03 2015. .

Alena Uus, Tong Zhang, Laurence H Jackson, Thomas A Roberts, Mary A Rutherford, Joseph V Hajnal, and Maria Deprez. Deformable slice-to-volume registration for motion correction of fetal body and placenta mri. *IEEE transactions on medical imaging*, 39(9):2750–2759, 2020.

Alena U Uus, Vanessa Kyriakopoulou, Antonios Makropoulos, Abi Fukami-Gartner, Daniel Cromb, Alice Davidson, Lucilio Cordero-Grande, Anthony N Price, Irina Grigorescu, Logan ZJ Williams, et al. Bounti: Brain volumetry and automated parcellation for 3d fetal mri. *bioRxiv*, pages 2023–04, 2023.

Lucy Vanes, Robin Murray, and Chiara Nosarti. Adult outcome of preterm birth: Implications for neurodevelopmental theories of psychosis. *Schizophrenia Research*, 247, 05 2021. .

Norman Waitzman, Ali Jalali, and Scott Grosse. Preterm birth lifetime costs in the united states in 2016: An update. *Seminars in Perinatology*, 45:151390, 01 2021. .

Yuyan Wang, Dujuan Wang, Na Geng, Yanzhang Wang, Yunqiang Yin, and Yaochu Jin. Stacking-based ensemble learning of decision trees for interpretable prostate cancer detection. *Applied Soft Computing*, 77:188–204, 2019. ISSN 1568-4946. . URL https://www.sciencedirect.com/science/article/pii/S1568494619300195.

Helena A. Watson, Naomi Carlisle, Paul T. Seed, Jenny Carter, Katy Kuhrt, Rachel M. Tribe, and Andrew H. Shennan. Evaluating the use of the quipp app and its impact on the management of threatened preterm labour: A cluster randomised trial. *PLoS Medicine*, 18, 2021.

WHO. Preterm birth, Feb 2018. URL https://www.who.int/news-room/fact-sheets/detail/preterm-birth.

Tomasz Wlodarczyk, Szymon Płotka, Tomasz Trzcinski, Przemyslaw Rokita, Nicole Sochacki-Wójcicka, Michal Lipa, and Jakub Wójcicki. Estimation of preterm birth markers with u-net segmentation network. 10 2019.

Tomasz Wlodarczyk, Szymon Płotka, Tomasz Szczepański, Przemyslaw Rokita, Nicole Sochacki-Wójcicka, Jakub Wójcicki, Michal Lipa, and Tomasz Trzcinski. Machine learning methods for preterm birth prediction: A review. *Electronics*, 10:586, 03 2021. .

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. ISSN 0893-6080. . URL https://www.sciencedirect.com/science/article/pii/S0893608005800231.

Hai-Cheng Yi, Zhu-Hong You, Mei-Neng Wang, Zhen-Hao Guo, Yan-Bin Wang, and Ji-Ren Zhou. Rpi-se: a stacking ensemble learning framework for ncrna-protein interactions prediction using sequence information. *BMC bioinformatics*, 21(1):60, 2020. ISSN 1471-2105. URL https://search.ebscohost.com/login.aspx?direct=true&db=cmedm&AN=32070279&site=ehost-live.

Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms*. CRC press, 2012.

Meng Yuan Zhu, Natasha Milligan, Sarah Keating, Rory Windrim, Johannes Keunen, Varsha Thakur, Annika Ohman, Sharon Portnoy, John Sled, Edmond Kelly, Shi-Joon Yoo, Lars Gross-Wortmann, Edgar Jaeggi, Christopher Macgowan, John Kingdom, and Mike Seed. The hemodynamics of late onset intrauterine growth restriction by mri. *American journal of obstetrics and gynecology*, 214, 10 2015. .

Appendix A. 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 (tag_ga)	C	GA of fetus at time of MR scan	-
GA ROM (tag_garom)	C	GA of fetus at time of rupture of the membranes	CO
Cohort type (tag_typ)	Cat	Cohort type	CO
Scanner type ($tag\_scanner$)	D	1.5 Tesla, 3 Tesla	-
Age (tag_age)	C	Maternal age	CB
BMI (tag_bmi)	C	Maternal body mass index	CB
LOC (tag_loc)	D	Placental location.	GU
GA delivery (tag_gadel)	C	Gestational age of fetus at delivery	CO
MOD (tag_mod)	Cat	Modality of the delivery i.e. C-section.	CO
Sex (tag_sex)	Cat	Sex of the fetus (Male, Female)	CO
BWG (tag_bwg)	C	Birth weight (grams)	CO
BWC (tag_bwc)	C	Birth weight centile	CO
Parity (tag_parity)	D	Number of previous pregnancies.	CB
Parity (tag_prev_ptb)	D	Number of previous preterm births.	CB
BPD (tag_bpd)	C	Biparietal diameter of the fetal brain	sMRI
BPD Cent (tag_bpd_cent)	C	Biparietal diameter centile of the fetus	sMRI
TCD (tag_ce_tcd)	C	Transcerebral diameter of the fetus	sMRI
TCD Cent ($tag\_ce\_tcd\_cent$)	C	Transcerebral diameter centile of the fetus	sMRI
Post Hor Diam ($tag\_post\_hor\_diam$)	C	Diameter of the posterior horn of the fetus	sMRI
VOL Body (tag_vol_body)	C	Volume of the fetus	sMRI
Diabetes ($tag\_diabetes$)	Cat	Diagnosis of diabetes	CB
HC (tag_hc)	C	Fetal head circumference at birth	CO
HC Cent (tag_hcc)	C	Fetal head circumference centile at birth	CO
CPTR (tag_cptr)	C	T2* brain to placenta ratio	fMRI
Anom Pi Left ($tag\_anom\_pi\_left$)	C	Pulsatility index of the left uterine artery	AUS
Anom Pi Right ($tag\_anom\_pi\_right$)	C	Pulsatility index of the right uterine artery	AUS
Anom GA (tag_anom_ga)	C	Gestational age at anomaly US	AUS
Anom LOC (tag_anom_loc)	D	Placental location at anomaly US	AUS
Anom Cord (tag_anom_cord)	D	Umbilical cord type at anomaly US	AUS
Anom Cord Ins ($tag\_anom\_cord\_ins$)	D	Umbilical cord insertion at anomaly US	AUS
Anom HC (tag_anom_hc)	C	Fetal head circumference at anomaly US	AUS
Anom AC (tag_anom_ac)	C	Abdominal circumference at anomaly US	AUS
Anom BPD (tag_anom_bpd)	C	Bi-parietal diameter at anomaly US	AUS
Anom FL (tag_anom_fl)	C	Femur length at anomaly US	GUS
GU GA (tag_gu_ga)	C	Gestational age at growth US	GUS
GU HC (tag_gu_hc)	C	Head circumference at growth US	GUS
GU AC (tag_gu_ac)	C	Abdominal circumference at growth US	GUS
GU BPD (tag_gu_bpd)	C	Bi-parietal diameter at growth US	GUS
GU FL (tag_gu_fl)	C	Femur length at growth US	GUS
GU Pi (tag_gu_pi)	C	Pulsatility index of the umbilical vein	GUS
GU EFW (tag_gu_efw)	C	Estimated foetal weight at growth US	GUS
GU MCA PI ($tag\_gu\_mca\_pi$)	C	Pulsatility index of the mid-cerebral artery	GUS
GU MCA PSV ($tag\_gu\_mca\_psc$)	C	Peak systolic velocity of the mid-cerebral artery	GUS
GU MCA CPR ($tag\_gu\_mca\_cpr$)	C	Cerebroplacental ratio	GUS
GU Notch (tag_gu_notch)	C	Notching present in the uterine artery	GUS
GU Pi Left ($tag\_gu\_pi\_left$)	C	Pulsatility index of the left uterine artery	GUS
GU Pi Right ($tag\_gu\_pi\_right$)	C	Pulsatility index of the right uterine artery	GUS
GU EDF (tag_gu_edf)	C	End-diastolic flow	GUS
GU LOC (tag_gu_loc)	D	Placental location at growth US	GUS
GU Cord (tag_gu_cord)	D	Umbilical cord type at growth US	GUS
GU Cord Ins ($tag\_gu\_cord\_ins$)	C	Umbilical cord insertion at growth US	GUS
APGAR5 (tag_apgar5)	C	APGAR score at 5 minutes	CO
Histo MVM (tag_histo_mvm)	C	Histopathology, maternal villi malperfusion	CO
Histo FVM (tag_histo_fvm)	C	Histopathology, foetal villi malperfusion	CO
Histo weight (tag_histo_weight)	C	Histopathology placental weight	CO
Histo chorio (tag_histo_chorio)	C	Histopathology chorioamnionitis	CO
SMOK (tag_smok)	D	Maternal smoking status	CB
IVF (tag_ivf)	D	In vitro fertilisation (IVF) status	CB
BWC delivery (tag_del_bwc)	C	Birth weight centile (Intergrowth21) at delivery	CO
Plac T2* mean ($plac\_t2s\_mean$)	C	Mean whole placental T2* value	fMRI
Plac T2* vol ($plac\_t2s\_vol$)	C	Whole placental T2* volume value	fMRI
Plac T2* lacu ($plac\_t2s\_lacu$)	C	Placental T2* lacunarity	fMRI
Plac T2* skew ($plac\_t2s\_skew$)	C	Placental T2* skewness	fMRI
Plac T2* kurt ($plac\_t2s\_kurt$)	C	Placental T2* kurtosis	fMRI
GU EFW Cent ($tag\_gu\_efw\_cen$)	C	Expected foetal weigth centile (Intergrowth21)	GUS
Brain T2* mean ($brain\_t2s\_mean$)	C	Mean brain T2* value	fMRI
Brain T2* vol ($brain\_t2s\_vol$)	C	Brain T2* volume value	fMRI
Brain T2* lacu ($brain\_t2s\_lacu$)	C	Brain T2* lacunarity	fMRI
Brain T2* skew ($brain\_t2s\_skew$)	C	Brain T2* skewness	fMRI
Brain T2* kurt ($brain\_t2s\_kurt$)	C	Brain T2* kurtosis	fMRI
T1 ($t1\_1$)	C	Mean T1 from MRI	fMRI
Cervical length (tag_cervix_length)	C	Cervical length from sagittal plane	sMRI
Cross Study ID ($tag\_complete\_id$)	C	Anonymous Cross Study Identifier	-
Volume eCSF left ($eCSF\_L$)	C	Volume eCSF left side	sMRI
Volume eCSF right ($eCSF\_L$)	C	Volume eCSF right side	sMRI
Left cortex ($Cortex\_L$)	C	Volume cortex left	sMRI
Right cortex ($Cortex\_L$)	C	Volume cortex right	sMRI
Left white matter (WM_L)	C	White matter volume left	sMRI
Right white matter (WM_R)	C	White matter volume right	sMRI
Left lateral ventricles ($Lat\_ventricle\_L$)	C	Lateral ventricles volume left	sMRI
Right lateral ventricles ($Lat\_ventricle\_R$)	C	Lateral ventricles volume right	sMRI
Csp volume (tag_cervix_length)	C	Cavum septi pellucidi volume	sMRI
Brainstem volume ($Brainstem$)	C	Brainstem volume	sMRI
Left cerebellum volume ($Cerebellum\_L$)	C	Cerebellum volume left	sMRI

Right cerebellum volume ($Cerebellum\_R$)	C	Cerebellum volume right	sMRI
Vermis volume ($Vermis$)	C	Vermis volume	sMRI
Left lentiform volume ($Lentiform\_L$)	C	Lentiform volume left	sMRI
Right lentiform volume ($Lentiform\_R$)	C	Lentiform volume right	sMRI
Left thalamus volume ($Thalamus\_L$)	C	Thalamus volume left	sMRI
Right thalamus volume ($Thalamus\_R$)	C	Thalamus volume right	sMRI
Third ventricle volume ($Third\_ventricle$)	C	Third ventricle volume	sMRI
Category (tag_cat_norm)	D	Assigned category	CO
Brain volume T2 (tag_cervix_length)	C	Complete T2-weighted brain volume	fMRI
Blood pressure systole (tag_bp_sys)	C	Average systole blood pressure	CB
Blood pressure diastole (tag_bp_dias)	C	Average diastole blood pressure	CB
Heart rate (tag_bp_hr)	C	Average heart rate	CB
Fetal body volume (tag_volume_wb)	C	Whole fetal body volume	sMRI
Amniotic fluid volume ($tag\_volume\_amniotic$)	C	Amniotic fluid volume	sMRI
Cohort at scan ($tag\_control\_at\_scan$)	C	Cohort assessment at scan	CO
Mid cerebellar artery ratio ($tag\_gu\_mca\_cpr1$)	C	Mid cerebellar artery ratio	US
Adc ($diff\_1$)	C	Average adc	fMRI
T2* ($diff\_2$)	C	Average T2*	fMRI
T2* perfusion compartment ($diff\_3$)	C	T2* perfusion compartment	fMRI
T2* diffusing compartment ($diff\_4$)	C	T2* diffusing compartment	fMRI
Adc perfusion compartment ($diff\_5$)	C	Adc perfusion compartment	fMRI
Adc diffusing compartment ($diff\_6$)	C	Adc diffusing compartment	fMRI
T2* perfusing compartment weighted ($diff\_7$)	C	T2* perfusing compartment weighted	fMRI
T2* diffusing compartment weighted ($diff\_8$)	C	T2* diffusing compartment weighted	fMRI
Adc perfusion compartment weighted ($diff\_9$)	C	Adc perfusion compartment weighted	fMRI
Adc diffusing compartment weighted ($diff\_10$)	C	Adc diffusing compartment weighted	fMRI
Perfusion fraction ivim ($diff\_11$)	C	Perfusion fraction ivim	fMRI
Left lung T2* mean ($lung\_t2s\_left\_mean$)	C	Mean left lung T2* value	fMRI
Left lung T2* vol ($lung\_t2s\_left\_vol$)	C	Left lung T2* volume value	fMRI
Left lung T2* lacu ($lung\_t2s\_left\_lacu$)	C	Left lung T2* lacunarity	fMRI
Left lung T2* skew ($lung\_t2s\_left\_skew$)	C	Left lung T2* skewness	fMRI
Left lung T2* kurt ($lung\_t2s\_left\_kurt$)	C	Left lung T2* kurtosis	fMRI
Right lung T2* mean ($lung\_t2s\_right\_mean$)	C	Mean right lung T2* value	fMRI
Right lung T2* vol ($lung\_t2s\_right\_vol$)	C	Right lung T2* volume value	fMRI
Right lung T2* lacu ($lung\_t2s\_right\_lacu$)	C	Right lung T2* lacunarity	fMRI
Right lung T2* skew ($lung\_t2s\_right\_skew$)	C	Right lung T2* skewness	fMRI
Right lung T2* kurt ($lung\_t2s\_right\_kurt$)	C	Right lung T2* kurtosis	fMRI
Both lungs T2* mean ($lung\_t2s\_both\_mean$)	C	Mean T2* value of both lungs	fMRI
Both lungs T2* vol ($lung\_t2s\_both\_vol$)	C	Both lungs T2* volume value	fMRI

Appendix B. 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
tag_ga	0 (0.0%)	28.18	28.14	4.56	-0.11
tag_typ	23 (9.47%)	158.21	99.0	300.93	2.6
$tag\_scanner$	0 (0.0%)	1.0	1.0	0.0	0.0
tag_age	10 (4.12%)	34.18	34.3	4.7	-0.27
tag_bmi	21 (8.64%)	23.77	23.65	2.98	0.48
tag_loc	35 (14.4%)	3.53	2.0	2.09	1.76
tag_gadel	0 (0.0%)	37.37	38.86	4.37	-1.73
tag_mod	0 (0.0%)	4.34	4.0	2.48	0.14
tag_sex	8 (3.29%)	1.57	2.0	0.52	0.19
tag_bwg	7 (2.88%)	2832.48	3062.5	917.24	-0.91
tag_bwc	107 (44.03%)	45.75	43.06	29.52	0.17
tag_parity	0 (0.0%)	0.48	0.0	0.78	2.14
tag_prev_ptb	0 (0.0%)	0.11	0.0	0.33	2.84
tag_bpd	122 (50.21%)	73.51	74.0	12.86	-0.29
tag_bpd_cent	124 (51.03%)	49.3	47.0	28.02	0.05
tag_ce_tcd	123 (50.62%)	33.88	32.55	8.42	0.08
$tag\_ce\_tcd\_cent$	129 (53.09%)	51.94	52.5	23.59	-0.08
$tag\_post\_hor\_diam$	125 (51.44%)	6.42	6.35	1.61	0.22
$eCSF\_L$	100 (41.15%)	31941.45	33049.6	11969.29	0.17
$eCSF\_R$	100 (41.15%)	30944.74	31214.5	10941.05	0.35
$Cortex\_L$	100 (41.15%)	21351.56	16914.0	12237.17	0.74
$Cortex\_R$	100 (41.15%)	21422.58	17092.2	11960.04	0.73
WM_L	100 (41.15%)	44681.85	41335.1	20056.3	0.35
WM_R	100 (41.15%)	44771.93	41127.5	20324.06	0.37
$Lat\_ventricle\_L$	100 (41.15%)	2154.67	2018.12	963.81	1.01
$Lat\_ventricle\_R$	100 (41.15%)	2008.0	1891.96	893.35	0.89
CSP	100 (41.15%)	497.36	471.24	224.35	0.85
$Brainstem$	100 (41.15%)	3514.29	3428.88	1474.82	-0.16
$Cerebellum\_L$	100 (41.15%)	2999.35	2493.24	1914.44	0.5
$Cerebellum\_R$	100 (41.15%)	2977.61	2348.0	1992.55	0.47

$Vermis$	100 (41.15%)	936.0	831.04	535.0	0.25
$Lentiform\_L$	100 (41.15%)	2120.58	1953.5	1059.44	0.23
$Lentiform\_R$	100 (41.15%)	2041.67	1871.62	1033.36	0.15
$Thalamus\_L$	100 (41.15%)	1642.67	1483.75	819.65	0.29
$Thalamus\_R$	100 (41.15%)	1590.87	1420.96	801.94	0.26
$Third\_ventricle$	100 (41.15%)	140.56	143.68	70.51	-0.08
$tag\_diabetes$	7 (2.88%)	0.5	0.0	2.13	4.3
tag_cat_norm	75 (30.86%)	1.45	1.0	0.52	0.45
tag_hc	64 (26.34%)	33.3	34.0	3.03	-2.23
tag_hcc	121 (49.79%)	53.64	56.6	33.57	-0.12
tag_garom	0 (0.0%)	37.0	38.86	5.07	-1.78
$tag\_gu\_efw\_cen$	103 (42.39%)	54.11	61.19	32.35	-0.41
$tag\_vol\_t2w\_complete$	100 (41.15%)	143758.08	134099.98	60465.11	0.26
tag_cptr	102 (41.98%)	0.38	0.36	0.12	0.93
tag_histo_weight	131 (53.91%)	417.62	453.0	130.33	-0.55
tag_histo_mvm	132 (54.32%)	0.22	0.0	0.41	1.4
tag_histo_fvm	132 (54.32%)	0.02	0.0	0.13	7.35
tag_histo_chorio	133 (54.73%)	0.38	0.0	0.49	0.49
$tag\_anom\_pi\_left$	191 (78.6%)	1.19	1.07	0.52	0.5
$tag\_anom\_pi\_right$	191 (78.6%)	1.22	1.12	0.56	1.19
tag_anom_ga	53 (21.81%)	20.02	20.0	0.9	-3.34
tag_anom_loc	35 (14.4%)	3.53	2.0	2.09	1.76
tag_anom_cord	69 (28.4%)	1.11	1.0	0.56	5.32
$tag\_anom\_cord\_ins$	182 (74.9%)	5.7	8.0	2.86	-0.66
tag_anom_hc	58 (23.87%)	172.3	171.8	9.81	1.51
tag_anom_ac	58 (23.87%)	150.4	150.2	11.03	0.83
tag_anom_bpd	71 (29.22%)	47.56	47.65	3.15	-2.64
tag_anom_fl	58 (23.87%)	31.72	31.7	2.53	0.23
tag_gu_ga	94 (38.68%)	28.65	28.43	4.63	-0.37
tag_gu_hc	98 (40.33%)	259.63	265.3	44.46	-0.89
tag_gu_ac	98 (40.33%)	239.14	239.3	49.19	-0.41
tag_gu_bpd	100 (41.15%)	73.0	74.8	13.2	-0.72
tag_gu_fl	97 (39.92%)	52.53	53.25	10.87	-0.68
tag_gu_pi	112 (46.09%)	1.08	1.03	0.22	1.65
tag_gu_efw	103 (42.39%)	1378.66	1229.5	727.21	0.53
$tag\_gu\_mca\_pi$	133 (54.73%)	1.85	1.84	0.36	-0.12
$tag\_gu\_mca\_psc$	138 (56.79%)	43.83	42.7	13.47	0.71
$tag\_gu\_mca\_cpr$	158 (65.02%)	1.82	1.83	0.45	0.0
tag_gu_notch	153 (62.96%)	1.1	1.0	0.5	5.23
$tag\_gu\_pi\_left$	153 (62.96%)	0.87	0.85	0.26	0.71
$tag\_gu\_pi\_right$	153 (62.96%)	0.87	0.86	0.26	0.91
tag_gu_edf	112 (46.09%)	1.11	1.0	0.62	6.12
tag_gu_loc	110 (45.27%)	3.61	2.0	2.22	1.53
tag_gu_cord	140 (57.61%)	1.06	1.0	0.44	8.1
$tag\_gu\_cord\_ins$	161 (66.26%)	3.28	4.0	1.28	-1.24
tag_apgar5	26 (10.7%)	9.25	10.0	1.5	-3.74
tag_smok	33 (13.58%)	1.41	1.0	1.08	2.35
tag_ivf	25 (10.29%)	1.87	2.0	0.46	-2.15
tag_bp_sys	105 (43.21%)	106.22	104.22	12.78	0.54
tag_bp_dias	105 (43.21%)	65.36	64.06	10.72	0.53
tag_bp_hr	118 (48.56%)	77.5	76.4	10.1	0.27
tag_volume_wb	220 (90.53%)	763487.03	683058.59	313741.75	0.59
$tag\_volume\_amniotic$	220 (90.53%)	460971.62	526734.02	242231.28	-0.32
$tag\_control\_at\_scan$	0 (0.0%)	0.62	1.0	0.49	-0.49
$tag\_gu\_mca\_cpr1$	138 (56.79%)	1.79	1.83	0.46	-0.22
tag_del_bwc	7 (2.88%)	47.22	48.04	31.14	-0.0
$plac\_t2s\_mean$	0 (0.0%)	57.92	59.69	19.11	-0.01
$plac\_t2s\_vol$	0 (0.0%)	366.89	340.31	173.32	0.77
$plac\_t2s\_lacu$	0 (0.0%)	19.47	19.08	6.16	2.53
$plac\_t2s\_skew$	0 (0.0%)	13.33	2.36	30.53	3.49
$plac\_t2s\_kurt$	0 (0.0%)	1.34	0.68	1.89	2.37
$lung\_t2s\_left\_mean$	159 (65.43%)	65.26	62.16	18.45	1.21

lung_t2s_left_vol	159 (65.43%)	13.88	11.08	8.71	1.4
lung_t2s_left_lacu	159 (65.43%)	20.06	18.43	8.72	1.58
lung_t2s_left_skew	159 (65.43%)	7.27	1.78	13.54	2.99
lung_t2s_left_kurt	159 (65.43%)	0.99	0.58	1.16	1.81
lung_t2s_both_mean	155 (63.79%)	67.06	62.62	19.26	1.17
lung_t2s_both_vol	155 (63.79%)	31.95	26.15	20.13	1.2
tag_vol_body	100 (41.15%)	1009.95	971.36	456.13	8.94
tag_cervix_length	26 (10.7%)	30.7	31.48	10.39	-0.71
brain_t2s_mean	101 (41.56%)	165.0	174.29	42.69	-0.56
brain_t2s_vol	101 (41.56%)	128.08	109.09	72.89	1.32
brain_t2s_lacu	101 (41.56%)	68.79	73.92	19.64	-0.77
brain_t2s_skew	101 (41.56%)	2.27	0.17	6.07	4.43
brain_t2s_kurt	101 (41.56%)	0.91	0.73	0.67	1.56
lung_t2s_right_mean	158 (65.02%)	67.97	63.24	19.56	1.07
lung_t2s_right_vol	158 (65.02%)	19.64	15.85	13.1	1.15
lung_t2s_right_lacu	158 (65.02%)	22.13	19.95	9.71	1.22
lung_t2s_right_skew	158 (65.02%)	7.66	2.42	14.4	4.1
lung_t2s_right_kurt	158 (65.02%)	1.14	0.82	1.19	2.08
t1_1	199 (81.89%)	1003.34	1016.46	243.97	-0.8
diff_1	140 (57.61%)	55.75	57.47	19.34	-0.22
diff_2	140 (57.61%)	0.0	0.0	0.0	1.14
diff_3	140 (57.61%)	0.0	0.0	0.0	2.41
diff_4	140 (57.61%)	0.0	0.0	0.0	0.0
diff_5	140 (57.61%)	0.01	0.0	0.01	3.3
diff_6	140 (57.61%)	0.0	0.0	0.0	0.0
diff_7	140 (57.61%)	52.51	52.21	32.07	0.11
diff_8	140 (57.61%)	0.03	0.03	0.01	1.02
diff_9	140 (57.61%)	76.01	75.25	22.84	0.99
diff_10	140 (57.61%)	0.0	0.0	0.0	0.62
diff_11	140 (57.61%)	0.33	0.3	0.17	0.68
tag_complete_id	0 (0.0%)	2400682.76	1000187.0	2864221.51	1.98

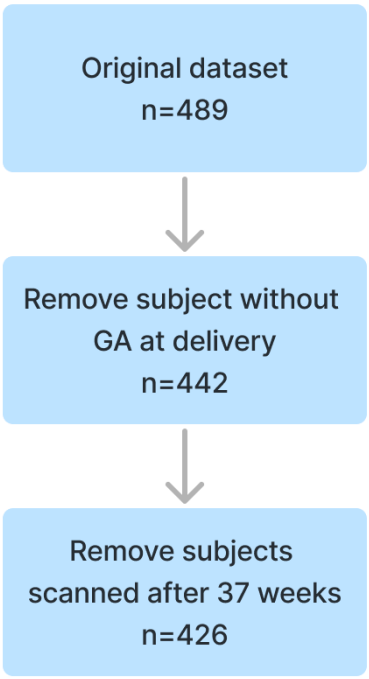
Appendix C. Hyperparameters of the base models.

Hyperparameters investigated by the grid search for each of the base models

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

Appendix D. Details of First Preprocessing Steps.

Number of subjects kept after each initial preprocessing step.



Appendix E. 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 $\hat{D}_{ij}$.

Impute D_{ij} with column means: $\hat{D}_{ij} = (\hat{d_{ij}}) \leftarrow \text{FillWithMeans}(D_{ij})$
$k \leftarrow 1$
while $k \leq$ MaxIter **do**
    for each column D_j in D_{ij} **do**
        Set imputations for D_j to missing in $\hat{D}_{ij}$
        $f_j \leftarrow \text{FitRegressionModel}(\hat{D}_{ij}, \hat{D}_j)$
        for each row D_i in D_{ij} **do**
            if $\hat{d_{ij}}$ is missing in $\hat{D}_{ij}$ **then**
                Impute missing value: $\hat{d_{ij}} \leftarrow f_j(\hat{D}_i \setminus \{\hat{d_{ij}}\})$
            end if
        end for
    end for
    $k \leftarrow k + 1$
end while
return $\hat{D}_{ij}$