Computer Methods and Programs in Biomedicine 213 (2022) 106520

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

# Gradient boosted trees with individual explanations: An alternative to logistic regression for viability prediction in the first trimester of pregnancy

Thibaut Vaulet[a,*], Maya Al-Memar[b], Hanine Fourie[b], Shabnam Bobdiwala[b], Srdjan Saso[b], Maria Pipi[b], Catriona Stalder[b], Phillip Bennett[b], Dirk Timmerman[c,d], Tom Bourne[b,c,d], Bart De Moor[a]

[a] *ESAT-STADIUS, Stadius Centre for Dynamical Systems, Signal Processing and Data Analytics (STADIUS), Leuven (Arenberg) Kasteelpark Arenberg 10 - box 2446, Leuven 3001, Belgium*
[b] *Tommy's National Early Miscarriage Research Centre, Queen Charlotte's and Chelsea Hospital, Imperial College, Du Cane Road, London W12 0HS, United Kingdom*
[c] *Department of Development and Regeneration, KU Leuven, Leuven, Belgium*
[d] *Department of Obstetrics and Gynecology, University Hospitals Leuven, Leuven, Belgium*

ARTICLE INFO

*Article history:*
Received 19 May 2021
Accepted 2 November 2021

*Keywords:*
Machine learning
First trimester viability
Logistic regression
Gradient boosted tree
Post-hoc interpretability
Shapley value

ABSTRACT

*Background:* Clinical models to predict first trimester viability are traditionally based on multivariable logistic regression (LR) which is not directly interpretable for non-statistical experts like physicians. Furthermore, LR requires complete datasets and pre-established variables specifications. In this study, we leveraged the internal non-linearity, feature selection and missing values handling mechanisms of machine learning algorithms, along with a post-hoc interpretability strategy, as potential advantages over LR for clinical modeling.

*Methods:* The dataset included 1154 patients with 2377 individual scans and was obtained from a prospective observational cohort study conducted at a hospital in London, UK, from March 2014 to May 2019. The data were split into a training (70%) and a test set (30%). Parsimonious and complete multivariable models were developed from two algorithms to predict first trimester viability at 11–14 weeks gestational age (GA): LR and light gradient boosted machine (LGBM). Missing values were handled by multiple imputation where appropriate. The SHapley Additive exPlanations (SHAP) framework was applied to derive individual explanations of the models.

*Results:* The parsimonious LGBM model had similar discriminative and calibration performance as the parsimonious LR (AUC 0.885 vs 0.860; calibration slope: 1.19 vs 1.18). The complete models did not outperform the parsimonious models. LGBM was robust to the presence of missing values and did not require multiple imputation unlike LR. Decision path plots and feature importance analysis revealed different algorithm behaviors despite similar predictive performance. The main driving variable from the LR model was the pre-specified interaction between fetal heart presence and mean sac diameter. The crown-rump length variable and a proxy variable reflecting the difference in GA between expected and observed GA were the two most important variables of LGBM. Finally, while variable interactions must be specified upfront with LR, several interactions were ranked by the SHAP framework among the most important features learned automatically by the LGBM algorithm.

*Conclusions:* Gradient boosted algorithms performed similarly to carefully crafted LR models in terms of discrimination and calibration for first trimester viability prediction. By handling multi-collinearity, missing values, feature selection and variable interactions internally, the gradient boosted trees algorithm, combined with SHAP, offers a serious alternative to traditional LR models.



* Corresponding author.
*E-mail address:* thibaut.vaulet@esat.kuleuven.be (T. Vaulet).

https://doi.org/10.1016/j.cmpb.2021.106520

**List of abbreviations**

| | |
|---|---|
| AUC | area under the curve |
| CI | confidence Interval |
| CRL | crown –rump length |
| FH | fetal heart |
| GA | gestational age (in days) |
| LGBM | light Gradient Boosted Machine |
| LMP | last menstruation period |
| LR | logistic regression |
| MSD | mean sac diameter |
| MYSD | mean yolk sac diameter |
| PUQE score | pregnancy-unique quantification of emesis and nausea |
| SHAP | SHapley additive exPlanations |

## 1. Introduction

First trimester miscarriage is the most common complication of early pregnancy. Although difficult to assess, its incidence in recognized pregnancies is estimated around 13–17% in recent studies [1–3]. These adverse events can be traumatizing and can cause psychological distress for several months following a loss [4–6]. Diagnostic uncertainty in early pregnancy is associated with increased anxiety [7], justifying the need for models that accurately predict the outcome of a pregnancy. In order to predict the risk of miscarriage, several models based on logistic regression (LR) have been developed over the years [8–13]. Despite good discriminative performance (AUC between 0.75 and 0.95), a significant number of miscarriages remain difficult to predict using these models.

The rise of artificial intelligence and machine learning (ML) in recent decades has led to the development of more complex algorithms which have demonstrated outstanding performance in numerous settings [14–16], including diagnosis performance similar to human medical-experts [14]. In comparison to LR, more sophisticated ML models are intrinsically nonlinear, avoiding the explicit formulation of interaction terms and/or nonlinear transformation of variables. In addition, some ML algorithms can also natively handle missing values, i.e. they can be trained on incomplete datasets whereas data imputation is needed before using LR [17]. Recently, advanced machine learning algorithms have been applied to various pregnancy-related conditions. For instance, Liu et al. demonstrated that tree-based ensembles outperformed traditional regression-based methods to predict early pregnancy loss after *in vitro* fertilization [18], although the evaluation and hyperparameters tuning procedures were not reported. Moreira et al. used averaged one-dependence estimators to predict the childbirth outcome of pregnancies with hypertensive disorders [19]. Bruno et al. applied Support Vector Machine to predict recurrent pregnancy losses [20]. Kuhle et al. compared logistic regression with advanced machine learning algorithms for the prediction of fetal growth abnormalities [21]. The early prediction of adverse pregnancy outcomes with efficient models offer the opportunity to prevent a range of future complications [22]. However, despite significant advantages, the use of advanced ML to develop clinical models is still relatively uncommon.

A common barrier to the adoption of more advanced ML models in clinical practice is often explained by their lack of transparency regarding predictions [23,24]. However, more recently, a model-agnostic framework based on Shapley values has emerged that explains individual predictions [25]. Methods based on Shapley values decompose each model's prediction as a collaboration of individual variables. It is therefore straightforward to perceive the contribution of each individual variable to the final prediction. This approach has a solid theoretical foundation derived from game theory to provide useful post-hoc model explanations and make ML models more interpretable.

In this study, we aimed to assess the utility of interpretable machine learning for first trimester viability prediction. We first compared the predictive and calibration performance of LR models and gradient boosted trees. We then derived meaningful explanations at the patient-level and compared the global behavior of both algorithms. Finally, we highlighted the potential benefits of machine learning with post-hoc interpretability strategy for clinical modeling.

The paper is organized as follows: in Section 2, we first introduce the data cohort and the sets of variables. We then describe the two models used to predict the first trimester viability, as well as the performance metrics and the validation strategy employed, before introducing the SHAP post-hoc interpretability framework. The Section 3 reports the results in terms of models performance and interpretability. In Section 4, we discuss the main findings of this study and we elaborate on the different levels of post-hoc interpretability and its application to clinical predictive modeling. Finally, we highlight the advantages and limitations of both predictive modeling approaches, before addressing our concluding remarks.

## 2. Materials and methods

### 2.1. Data and study design

The study was based on data derived from a prospective observational cohort study based at Queen Charlotte's & Chelsea Hospital, London, conducted between March 2014 and May 2019. The study had been approved by NHS National Research Ethics Service (NRES) Riverside Committee London (REC 14/LO/0199) and NHS North East – Newcastle and North Tyneside 2 Research Ethics Committee (17/NE/0121). All participants provided written informed consent. Details on the study design and recruitment criteria can be found in [26].

Women with intrauterine pregnancies (either a confirmed viable pregnancy or pregnancy of unknown viability) were recruited and followed up with serial ultrasound scans in the first trimester. Demographic, clinical and ultrasound scan data were collected. The main outcome was defined as the presence of viable pregnancy at 11–14 weeks of gestational age (GA). All scans when a diagnosis of miscarriage was made were excluded. Participants with unknown date of last menstrual period (LMP) were also excluded. Due to the progressive drop out of miscarriage patients from the cohort, data at more advanced GA are biased towards viable pregnancies. In the present dataset, 15.5% of samples are associated with a miscarriage before 70 days of gestational age, whereas this proportion drops to 6.3% after 70 days.

To avoid the algorithms learning that these pregnancies are at less risk of miscarriage, we focused on the first half of the first trimester: scans with GA greater than 70 days were therefore excluded.

### 2.2. Variables and univariates analysis

Two sets of variables were used in the models. To limit the risk of overfitting, a restricted set of predefined variables was chosen based on expert opinion and previous published studies [8–13,15,27]. This parsimonious features set contained: *maternal age, number of previous miscarriages, worst bleeding score reported, difference in estimated GA between LMP and mean sac diameter (MSD), GA by LMP, the Pregnancy-Unique Quantification of Emesis and Nau-*

*sea (PUQE) score, crown-rump length (CRL), MSD, fetal heart (FH)* and *MSD*FH*. This last term models an interaction between MSD and the presence of FH. Since advanced ML algorithms should model such interaction without explicit formulation, this term was omitted with the gradient boosted trees algorithm.

To assess the internal feature selection mechanism of gradient boosted trees algorithm, a more complete set of variables was also used in parallel, independently of expert knowledge. This complete set includes the parsimonious set augmented by: *maternal ethnicity, gravida, parity, supplementation with folic acid, smoking status, certainty of LMP, previous cesarian section, bleeding score at presentation', number of bleeding days, pain score at presentation, no of days with pain, worst pain score, mean yolk sac diameter (MYSD), presence of amnion sign, GA by MSD, GA by CRL.* A detailed description of the symptom variables can be found in [26]. Univariate analysis of the cohort characteristics with regard to the main outcome were performed with the Student's t-test for continuous variables and the chi-square test for binary or categorical variables.

### 2.3. Internal validation

The initial dataset was split into a training (70%) and test set (30%), stratified according to the main outcome to preserve a similar outcome prevalence between both sets. To avoid data leakage, i.e. the contamination of the training set with information from the test set in case data are not independent, the ultrasound scans from the same patients were strictly allocated to either the training or the test set.

### 2.4. Predictive models

1. Logistic regression

Multivariable logistic regression was used as a baseline model against more advanced ML models. Logistic regression is a statistical model used to perform regression analyses on binary outcomes. More specifically, logistic regression is a generalized linear model defined as : $logit(p(y = 1)) = X\beta$, where $p \in [0, 1]$, $y$ is the dependent binary variable, $X$ is the matrix of independent predictors, also known as explanatory variables and $\beta$ is the vector of parameters, or coefficients, optimized during model training. The binary dependent variable $y$ is related to the linear model $X\beta$ through the logit link function defined as:

$$logit\,(p(y=1)) = log\left(\frac{p(y=1)}{1-p(y=1)}\right),$$

$$\text{where } p(y=1) = \frac{exp(X\beta)}{1+\exp(X\beta)}$$

As a generalized linear model, logistic regression does not require a normal distribution of the residuals. In addition, unlike ordinary linear regression, logistic regression models do not rely on homoscedasticity. The additive constraint of multivariable regression restricts the model capacity but facilitates the understanding of the prediction process.

To account for repeated measurements (e.g. clustered data), LR was trained using the cluster robust variance-covariance matrix. Multiple imputation was used to accommodate the presence of missing values. The training set was imputed 20 times using Multiple Imputation by Chained Equations and predictive mean matching [28]. Missingness was assumed to be at random.

2. Light gradient boosted machine (LGBM)

We used a gradient boosted trees algorithm as the advanced ML model. Gradient boosted trees is a tree-based ensemble algorithm that produce prediction by averaging a large number of individual decision trees predictions. The individual decision trees are constructed sequentially with the goal to reduce the error of the previous model at each iteration. With gradient boosting, the structure of the next tree to add to the current ensemble is determined through the optimization of an objective function $\mathcal{L}$ via its gradient [29]. However, converting a decision tree learning algorithm into an optimization problem is not straightforward as the gradient with respect to the model's parameters is not directly computable. With special formulations of $\mathcal{L}$, it is possible to optimize the construction of a new tree such that for each node split, the best split is chosen, taking into account the model complexity. Turning the objective function into a splitting criterion avoids the intractable problem of constructing all possible trees at each iteration.

If the function $f_i$ represents the structure of a single decision tree, each new tree is added to the previous fixed ensemble as follows:

$$\hat{y}_i^{(0)} = 0$$

$$\hat{y}_i^{(1)} = \hat{y}_i^{(0)} + f_1(X_i)$$

$$\hat{y}_i^{(2)} = \hat{y}_i^{(1)} + f_2(X_i)$$

... where $\hat{y}_i$ and $\mathbf{X}_i$ represent the prediction and the vector of explanatory variable for patient i, respectively.

The final predictions after adding t trees to the ensemble are given by:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(X_i) = \hat{y}_i^{(t-1)} + f_t(X_i)$$

Tree-based ensembles have demonstrated state-of-the-art performance in various settings, frequently outperforming neural networks in tabular datasets (e.g. [30]). They are often easier to optimize than neural networks which require additional architecture specifications. Tree-based ensembles also benefit from an existing implementation for the exact calculation of Shapley values (see below) in the SHAP package [31]. LightGBM [32] was used to implement the gradient boosted trees. The number of trees was chosen with early stopping. The optimization of the other hyperparameters was performed with tree-structured parzen estimators through 5-folds cross validation of the training set [33]. The list of hyperparameters tuned is reported in Table S2, other parameters were used with their default settings. Models are referred as parsimonious or complete following the dataset on which they were trained.

### 2.5. Performance

The predictive performance of the models was assessed on the test set. Overall performance was assessed with the Brier score which measure the accuracy of probabilistic predictions, the lower the score, the better [34]. Discriminative performance was assessed with the area under the curve (AUC) of ROC curves. Statistical comparisons between two models AUC was performed with the DeLong method [35]. Calibration was assessed with the calibration slope and the calibration-in-the-large [34]. For the calibration slope, a significant departure from the perfect calibration slope of 1 was assessed by the Wald-test [34].

We reported those metrics in two different forms: (1) raw metric evaluated on the whole test set, not adjusted for the presence of repeated measurements from the same patients: (2) longitudinal metrics: for each GA $t$, the corresponding metric was computed on the subset of samples included in a time window of 20 days, cen-

tered on $t$. In case of repeated measurements per patient within that time window, only the closest prediction to $t$ was included to compute the metric. All metrics are reported with a 95% confidence interval (CI).

### 2.6. Post-hoc interpretability

To derive explanations of the model's individual predictions, we used the SHAP framework: an additive feature attribution method [25]. These model-agnostic methods rely on explanation models to decompose each prediction as a sum of individual feature contributions. Following the notation of Lundberg and Lee [25], the explanation model $g$ of additive feature attribution methods takes the form:

$$g(z') = \phi_0 + \sum_{j=1}^{d} \phi_j z_j'$$

where $z' \in \{0, 1\}^d$ is a *simplified* version of z, represented by a binary vector of dimension d, which simulates any subset of predictors from z by indicating their presence or their absence. $z_j'$ represents therefore the presence (=1) or absence (=0) of feature j in z. $\varphi_j \in R$ represents the feature attribution of the $j^{th}$ variable of z.

In the SHAP method, those feature attributions $\varphi_j$ are represented by Shapley values, derived from the collaborative game theory where a game payout is fairly distributed among the players of a game, taking into account the possible combinations of players. In the predictive analytics context, the game payout is the prediction and the players are the variables. The computation of the Shapley values guarantees a fair decomposition of the final prediction among the set of variables values. The Shapley value of a variable value represents its contribution to the current prediction. With SHAP, this contribution is expressed as a relative contribution between the current prediction and a baseline prediction value often set to E[f(**X**)] Therefore, the SHAP values estimate the contribution of each variable to explain the difference between f(**X**) and E[f(**X**)], the averaged prediction of the model. As a result, the explanations provided under the SHAP frameworks are contrastive which makes them more intuitive to understand for non-expert users[36]. In addition, SHAP is the only additive attribution method that remains locally faithful to the black box model prediction.

$$y = f(x) = g(x') = \phi_0 + \sum_{j=1}^{d} \phi_j x_j'$$

where $\varphi_0$ corresponds to a baseline prediction, often set to E[f(**X**)]. A more detailed explanation with theoretical formulation can be found in [25,30].

In order to derive meaningful comparisons amongst different models, we set the common baseline value to the outcome prevalence in the training set (0.88), which estimates the overall prior probability of miscarrying in the study-population. The Shapley values can be obtained from two different methods. The interventional approach breaks the potential inter-variables dependency to compute the features SHAP values, referred in the text as independent SHAP, while the correlated approach relies on the conditional expectation, which takes into account inter-features correlations. Recent studies suggest [37,38] that the conditional expectation approach can be misleading as some variables that are not used explicitly by the model can receive credits if they correlate with some other important variables. Where appropriate, we used the training set as background dataset for feature perturbation to compute the SHAP values using the interventional approach. For tree-based ensemble models, first order interaction SHAP values [31] were also computed.

The aggregation of SHAP values from individual predictions provides global model explanations. Global feature importance was obtained as the mean of absolute SHAP value across all instances for each variable. Similarly to the longitudinal visualization for the model performances, longitudinal features importance was also computed.

### 2.7. Software

All analysis have been performed with Python 3.6.6.

## 3. Results

1. Study cohort

Fig. 1 displays the study flowchart. Patients who underwent termination of pregnancy ($n = 28$), withdrew from the study ($n = 7$), and who were lost to follow up ($n = 40$) were excluded. Additionally, patients with unknown or missing LMP ($n = 50$) were also excluded. A total of 1154 patients (986 viable pregnancies and 168 miscarriage) were available for the analysis. On average, each patient underwent 2.05 (± 0.83 SD) ultrasound scans (miscarriages: 1.70 (± 0.73 SD); viable pregnancies: 2.65 (± 0.83 SD)). After stratified splitting, the training set and the test set consisted in 807 and 347 patients respectively. Table S1 summarizes the descriptive statistics of the cohort at the patient and ultrasound scan levels, including the number of missing values per variable.

2. Model performances

The overall performance metrics (with 95% CI) for parsimonious and complete models are reported in Table 1. Compared to the parsimonious LR, the parsimonious LGBM model had similar overall (Brier scores: 0.078 vs 0.076), discriminative (AUC 0.860 vs 0.885; $p$-value: 0.279) and calibration performance (calibration slope: 1.183, $p$-value: 1.222 vs 1.195, $p$-value: 0.098; calibration in the large: 0.001 vs 0.001). Furthermore, our results did not demonstrate the need to impute the training and testing data when using LGBM (Table 1). Therefore, all subsequent analysis and figures are based on LGBM without imputation. The models based on a preselected set of meaningful variables had similar discriminant performances as models based on the complete set of variables (LR models AUC: 0.886 vs 0.876, $p$-value: 0.348; LGBM models AUC: 0.885 vs 0.889, $p$-value: 0.574, Table 1). Moreover, the parsimonious LGBM had slightly better calibration performances than the complete LGBM (Slope: 1.195, $p$-value: 0.098 vs 1.298, $p$-value: 0.019, cal. in the large: 0.001 vs 0.010). Fig. 2 displays the parsimonious models performances longitudinally, based on the GA by LMP at the time of the scan. This longitudinal metrics assessment demonstrates similar behavior between LR and LGBM. Fig. S1 displays the same metrics for the complete models. Complete LGBM model demonstrated slightly worse calibration performance than complete LR model which can be explained by a greater flexibility of LGBM models. However, the discriminative performance of complete LGBM model was higher than complete LR, probably due to the built-in feature selection mechanism of LGBM.

In summary, LGBM models performed as good as LR approaches, without the need of missing values imputation and explicit specifications of variable interactions.

3. Post-hoc interpretability

The feature importance expressed as the mean of individual absolute SHAP values per variable are displayed in Fig. 3 for the parsimonious models and in Fig. S2 for the complete models. The preselected variables were mostly associated with a high feature importance in the complete models (Fig. S2). These results also reflect the univariate analysis as reported in Table S1. A notable exception

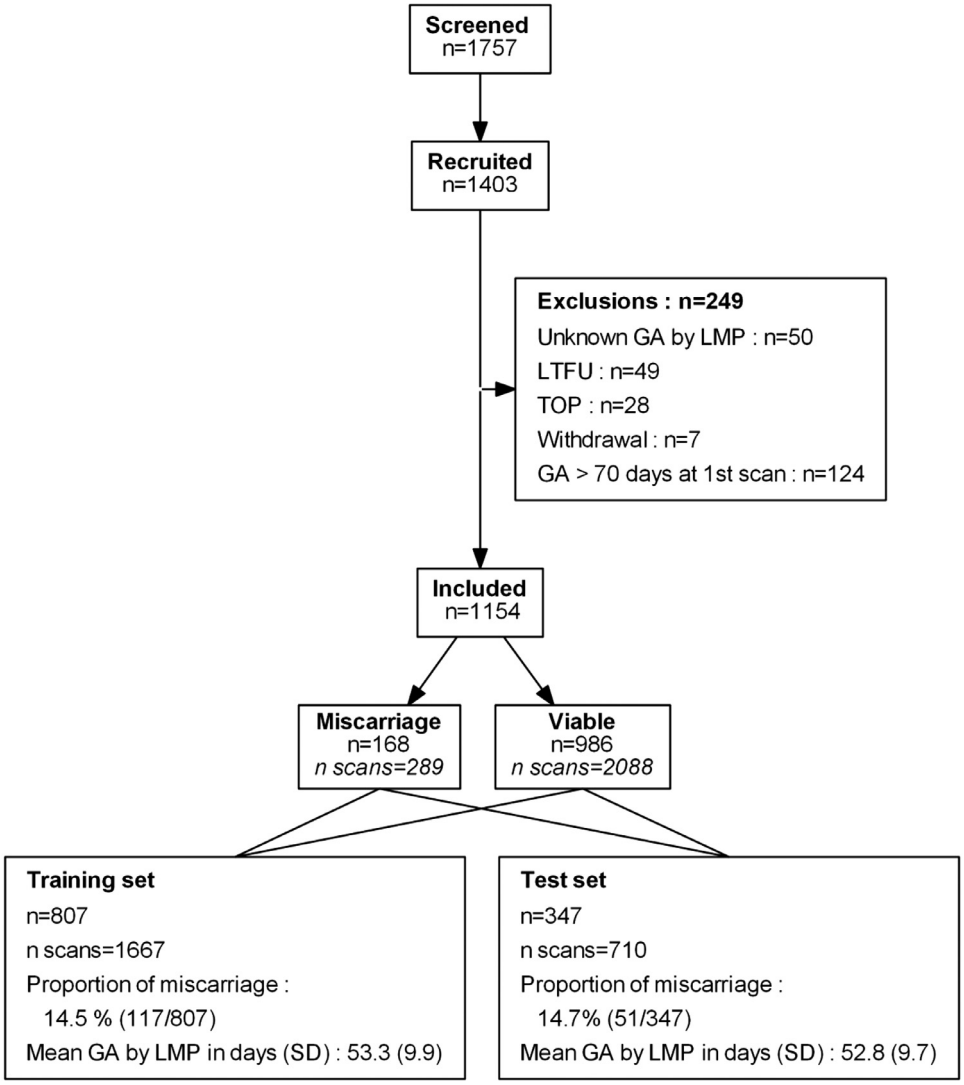


**Fig. 1.** Study flowchart *n* represents the number of unique patients and *n scans* the number of scans. All repeated scans of a given patient were strictly allocated to either the training or the test set. GA: gestational age; LMP: last menstruation period; LFTU: lost to follow-up, TOP: termination of pregnancy.

**Table 1**
Predictive performances of the models on the test se**t** – raw metrics with 95% CI. The Brier score assesses the accuracy of probabilistic predictions, the lower the score the better. AUC evaluates the discrimination performance. The calibration in the large evaluates the mean calibration and corresponds to the difference between the averaged binary outcome and the averaged prediction. The calibration slope summarizes how the predicted risks correspond to the observed risks. An ideal calibration slope is equal to 1 and departures from 1 indicate potential model miscalibration (e.g. due to over/underfitting). Overall, LR and LGBM performed similarly in terms of calibration and discrimination. Complete LGBM models demonstrated similar discriminative performances as parsimonious models but their calibration was slightly worse, probably resulting from a too large flexibility compared to LR models.

| | Parsimonious Models | | | Complete Models | | |
|---|---|---|---|---|---|---|
| | LR + MICE | LGBM with missing data | LGBM + MICE | LR + MICE | LGBM with missing data | LGBM + MICE |
| Brier score | 0.078 (0.065 0.093) | 0.076 (0.062 0.090) | 0.078 (0.063 0.092) | 0.080 (0.066 0.095) | 0.076 (0.062 0.090) | 0.076 (0.063 0.090 |
| AUC | 0.886 (0.852 0.919) | 0.885 (0.8480.922) | 0.881 (0.845 0.918) | 0.876 (0.841 0.911) | 0.889 (0.854 0.924) | 0.892 (0.852 0.926) |
| Calibration in the large | 0.001 (−0.022 0.019) | 0.001 (−0.021 0.020) | 0.002 (−0.023 0.019) | 0.005 (−0.026 0.016) | 0.010 (−0.031 0.010) | 0.014 (−0.035 0.006) |
| Calibration slope | 1.183 (0.950 1.415) | 1.195 (0.964 1.424) | 1.158 (0.934 1.380) | 1.030 (0.824 1.235) | 1.298 (1.048 1.547) | 1.344 (1.087 1.599) |

is the worst bleeding score variable. Although an important potential predictor for miscarriage, this variable was associated with poor feature importance in most of the models (Fig. 3).

The features importance as described above was not necessarily constant through time as reported in Fig. 4. The longitudinal assessment of features importance from the LR model demonstrates that, under the independent assumption, the interaction term MSD*FH was the determining variable throughout all gestational ages, far above the other predictors, although its importance decreases in the second half of the examined period. Taking into account correlated variables, it remained the first driving variable but its importance decreased as the credit was shared among other variables. On the LGBM model, this analysis demonstrated that the CRL globally stayed the most significant variable while the importance of the MSD variable decreased with GA. In the opposite, the difference in estimated GA between LMP and ultrasound measurements became more important in the second half of the examined period.

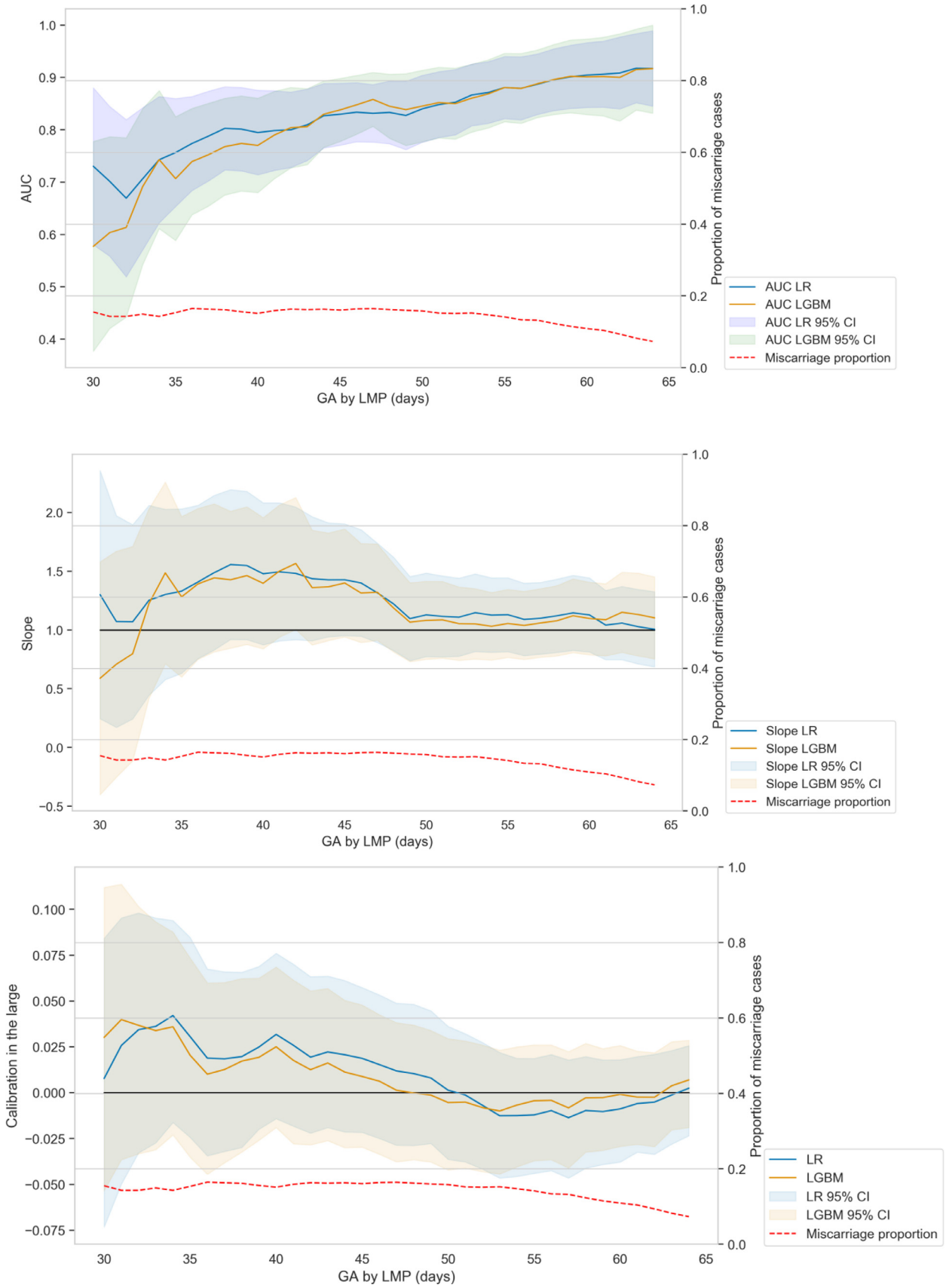


**Fig. 2.** Longitudinal assessment of the performance metrics for the parsimonious models. AUC, calibration slope and calibration in the large are displayed depending on the GA by LMP at the date of scan using a time window of 30 days around the GA. Both LR and LGBM display similar profiles in terms of discrimination and calibration performance. Note that the proportion of pregnancies remaining at risk of miscarriage naturally decreases with time, which partly explains the increase of AUC for higher GA. AUC: area under the curve; GA: gestational age; LGBM: light gradient boosted machine; LMP: last menstruation period; LR: logistic regression.

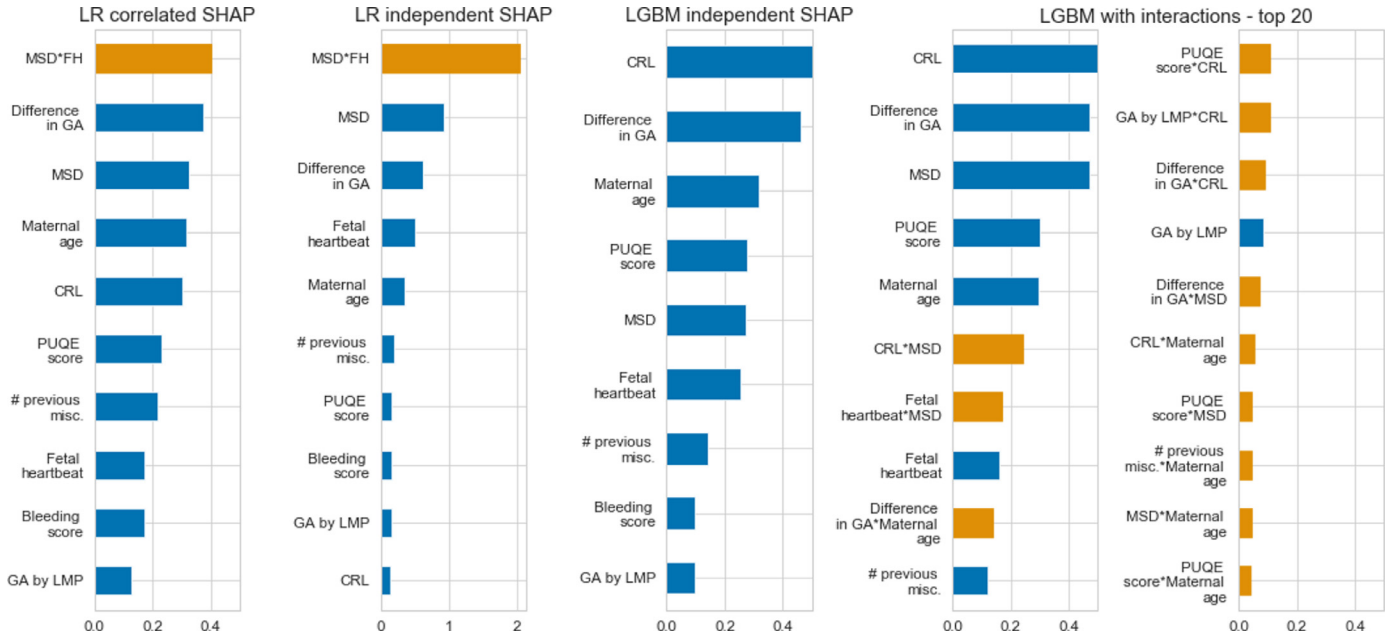


**Fig. 3.** Raw features importance measured with averaged absolute SHAP values per variable.
Variables that contribute significantly to the model's predictions for many patients have a high importance depicted as a large averaged absolute SHAP value. Main effects are colored in blue, interaction effects (first order variable interaction) are colored in orange. For LR models, the correlated SHAP approach (first column) takes variables collinearity into account when computing the SHAP values. The resulting feature importance is more balanced among correlated variables than the independent SHAP approach (second column) which directly reflects the LR coefficients. LGBM are reported with main effect only (third column) and with first order interactions (fourth and fifth columns, only the top 20 features). Despite similar performances, the algorithms have a different internal use of the same set of features. For example, the interaction MSD*FH is the main driving force of the LR model but appears as the 7th most important variable (second interaction term) in LGBM. CRL: crown-rump length; GA: gestational age; LGBM: light gradient boosted machine; LMP: last menstruation period; LR: logistic regression; MSD: mean sac diameter; PUQE: Pregnancy-Unique Quantification of Emesis.

The decision paths followed by the LR and LGBM parsimonious models for three patients with different predicted risks are displayed in Fig. 5. Finally, the three main first-order interaction effects modeled by LGBM were CRL*MSD, FH*MSD and Maternal Age * Difference in GA, and are displayed in Fig. 6.

## 4. Discussion

In this paper, we compared the predictive performance and the interpretability of an advanced ML algorithm over a carefully specified LR model. Overall, the different models demonstrated similar predictive performances. Both algorithms, although delivering different explanations, were intuitively explained by the SHAP framework at the local-level, i.e. for each individual patient prediction, and at the global or model-level. While LR models remain simple to implement and to train, gradient boosted trees algorithms demonstrated additional potential benefits for clinical modeling. A comparative on the use of LR vs LGBM for clinical modeling is reported in Table 2.

### 4.1. Predictive performance

For the specific problem of first trimester viability, and given the dataset available, more advanced models such as gradient boosted trees did not demonstrate outstanding benefit in terms of predictive performance over a simple linear model carefully crafted with an interaction term. The raw and longitudinal performance metrics on the test set were similar for both algorithms (Table 1 and Figs. 2, S1), with the exception of the complete LGBM model's calibration slope which was worse than with the complete LR model (calibration slope of 1.298, $p$-value: 0.019 vs 1.030, $p$-value: 0.776, respectively and Fig. 1b). This corroborates previous studies which found the absence of performance gain from advanced models compared to LR in clinical modeling [21,39,40].

### 4.2. Interpretability

#### 4.2.1. Model-level interpretability with SHAP feature importance

Although both algorithms displayed similar performances, parsimonious LR and LGBM demonstrated different behaviors when inspecting the models under the SHAP framework. The predictions from the LR model are mostly driven by the interaction term MSD*FH (Figs. 3,4), whereas LGBM predictions are mostly driven by CRL and the difference in estimated GA (Figs. 3,4). This phenomenon is known as the Rashomon effect [41], where multiple algorithms with similar performances can have completely different internal mechanism to derive their final predictions. The SHAP framework remains a method to derive individual explanations regarding a specific model. As a result, it should not be regarded as a way to derive absolute (causal) explanations. Model dependency should therefore be kept in mind when delivering post-hoc explanations to physicians.

### 4.3. Longitudinal SHAP feature importance and GA-dependence

Variables such as maternal age or the history of previous miscarriages are naturally independent of GA and display therefore a constant feature importance throughout the range of GA (Fig. 4). On the other hand, some variables demonstrate changes in their feature importance depending on the GA (Fig. 4). This phenomenon is partly explained by the specificities of the dataset and the encoding of the data. For instance, the fetal heartbeat is frequently absent (FH=0) on ultrasound scans performed at very early GA, irrespective of the future pregnancy outcome. Therefore, the interaction term MSD*FH of the LR model is encoded as zero, even if

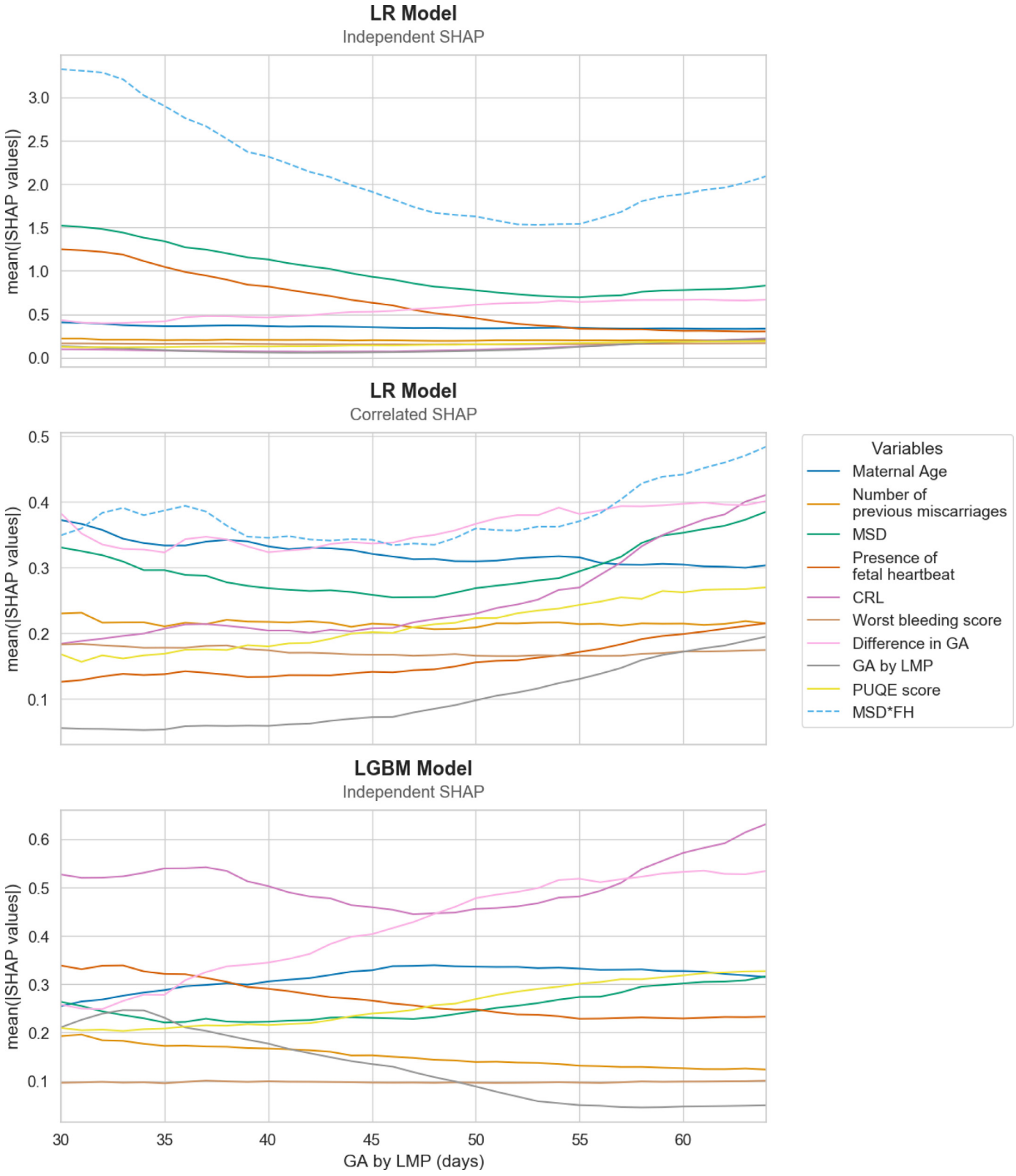


**Fig. 4.** Longitudinal features importance measured as averaged absolute SHAP values per variable. In the LR model, under the independent SHAP values computation, the interaction term MSD*FH is much more important in the beginning than in the second half of the examined period, although it remains from far the main driving variable throughout all gestational ages. Under the correlated approach, this interaction term remains the most important variable, but the credit is now shared among other correlated variables. In the LGBM model, the CRL variable stays the main variable while the importance of MSD decreases with GA. On the other hand, the difference in estimated GA becomes more important in the second half of the examined period. CRL: crown-rump length; FH: fetal heart; GA: gestational age; LGBM: light gradient boosted machine; LMP: last menstruation period; LR: logistic regression; MSD: mean sac diameter; PUQE: Pregnancy-Unique Quantification of Emesis.

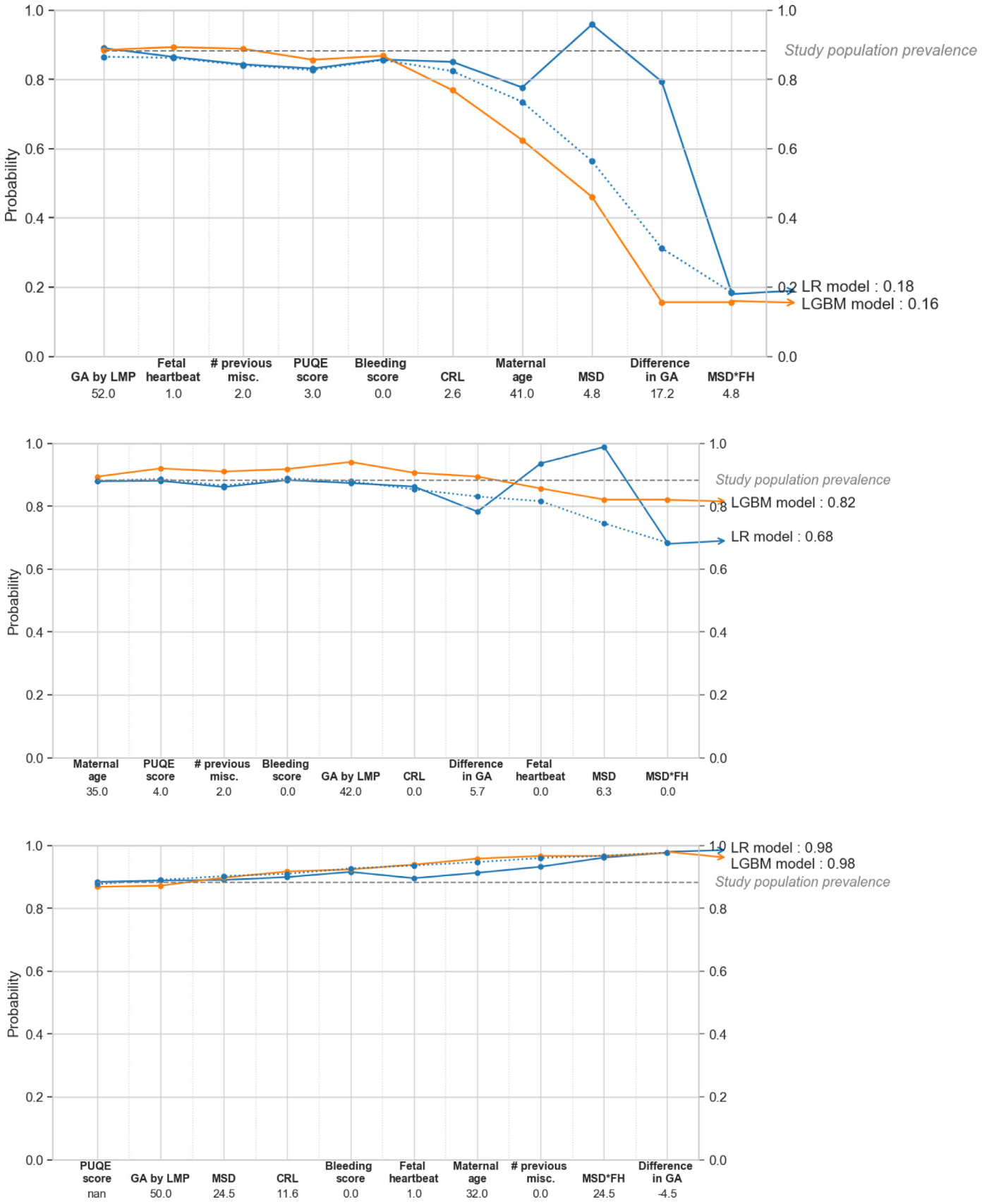


**Fig. 5.** *Decision path plots*. Example of individual prediction explained with the SHAP values for 3 instances of the test set (dashed line = correlated SHAP, plain line = independent SHAP). The SHAP values attributed to each variable fixed at their current value (as indicated below the x-axis) are gradually summed from left to right to explain the departure of the current prediction from the study population prevalence (*gray dashed line*), acting as a baseline viability probability. The bigger the magnitude of the SHAP value, the larger the variable contribution to the final prediction. The variables are ordered from lower to higher importance at the prediction level. For example, in the top graph, small values of CRL (2.6 mm) and MSD (4.8 mm) in combination with an older maternal age (41 years) and a large difference in estimated GA (17.2 mm) produce a low chance of viability from both models. CRL: crown-rump length; GA: gestational age; LGBM: light gradient boosted machine; LMP: last menstruation period; LR: logistic regression; MSD: mean sac diameter; Nan: missing; PUQE: Pregnancy-Unique Quantification of Emesis.

**Table 2**
A comparison of logistic regression models and gradient boosted trees for clinical modeling.

| | Logistic Regression | Gradient boosted trees |
|---|---|---|
| Predictive performance | | |
| Discrimination | Comparable if models correctly specified | |
| Calibration | Comparable if models correctly specified | |
| Interpretability | | |
| *Raw interpretability* | Based on model's coefficients. Subject to: multicollinearity, various unit scales, logit transform,... Not straightforward for physicians | Complex, possibility to extract feature importance. Not straightforward for physicians |
| *Under the SHAP framework* | | |
| Model-level | SHAP feature importance | |
| Patient-level | Decision path plots *Individual predictions explained as a sum of SHAP values* | |
| Collinearity | Handled by correlated SHAP | Less affected by collinearity |
| Variable interactions | Based on the pre-specified interactions | Possible with SHAP interaction values |
| Specificities | | |
| Missing values | Require complete datasets | Handle missing data internally |
| Feature selection | Possible with l1-regularization | Internal |
| Non-linearity (e.g. interactions) | Pre-specified | Internal |
| Optimization | Careful variable specifications | Require hyper-parameters tuning |

MSD is observed. As the pregnancy evolves, it becomes unlikely not to observe a fetal heartbeat. As a result, for more advanced GA, a smaller proportion of pregnancies displays a null value for the interaction term FH*MSD. Because the overall feature importance is expressed as a mean of absolute values, the highly negative SHAP values associated with MSD*FH=0 are counterbalanced by the large positive SHAP values when both MSD and FH are observed in the second half of the period (Fig. 4A). When the credit is shared among correlated variables, this phenomenon is significantly reduced (Fig. 4B). A similar phenomenon explains why the importance of the CRL variable increases with higher GA in the LGBM model (Fig. 4C). For early GA, the embryo is often not visible. Therefore, the proportion of pregnancies where the CRL variable is encoded as zero is higher for early GA compared to advanced pregnancies. The positive discrepancy in GA estimation, a strong predictor of miscarriage, increases with GA. Hence, the large positive differences in GA, associated with large negative SHAP values, are mostly observed in the second half of the period explaining the constant increase in variable importance for the discrepancy in GA (Fig. 4C).

These considerations highlight a drawback of the SHAP framework: the measure of feature importance, computed as mean(|SHAP values|), is directly dependent on the composition of the sample. The data used to compute the feature importance should be representative of the targeted population. The aggregation of individual SHAP explanations under the absolute operator might also obfuscate complex patterns of variable importance.

### 4.4. Interpretability under multicollinearity

Interestingly, CRL had a very low importance in the parsimonious LR model under the independent SHAP approach (Fig. 3, coefficient's $p$-values = 0.532 from Table S3). This is mostly explained by the collinearity with other variables (especially MSD) and highlights the limitations of LR in presence of correlated variables which disturbs the relationship between independent and dependent variables. Although this problem does not necessarily impact the prediction performance, it infringes the interpretability of LR models. Thoughtful variables selection *a priori*, dimensionality reduction or regularization can alleviate this phenomenon [42]. However, the correlated method for SHAP values presents an interesting alternative to display interpretable feature importance in the presence of collinearity as it shares credits among correlated variables, even if not explicitly used by the model. Under this method, the CRL variable importance drastically increased and reflected a more realistic view of this ultrasound parameter importance (Fig. 3). On the other hand, LGBM, due to its boosted nature, is more robust to the multicollinearity problem, as depicted in the feature importance analysis even under the independent SHAP approach (Fig. 3).

### 4.5. Interpretable individual predictions with decision path plots

At the individual prediction level, the SHAP framework decomposes each prediction into a sum of Shapley values. This sum explains the departure of the current prediction from a baseline prediction. The SHAP values attributed to each variable value can be organized into a meaningful visualization plot to derive the decision path followed by the model to reach the current prediction. Examples of such decision paths plots are reported in Fig. 5. Those plots constitute a meaningful way to translate complex algorithms decisions into interpretable predictions. As a model-agnostic explainer relying on the original variables additively, it allows for meaningful comparisons between different algorithms.

### 4.6. Exploring interaction effects

The SHAP values from non-linear models can be computed taking first order interactions into account [31]. The interaction plots, based on Shapley values, provide a clever alternative to partial dependence plots. In the parsimonious LGBM, the interaction effect between MSD and FH ranked 6th in terms of feature importance and constitutes the 2nd most important interaction term (out of 36 possible combinations) which corroborates its use in the LR model (and in previous study [27]). The interaction between CRL and MSD was the most important interaction effect and bears similar interpretation as the interaction between MSD and FH: intrauterine pregnancy without visible embryo is at higher risk of miscarriage when MSD increases. In the third interaction effect, a large discrepancy in GA appears to be modeled as a protective variable in young women while it becomes a risk factor in older women. Such interaction has never been reported and its clinical relevance remains uncertain as it might result from spurious findings based on the specificity of the training set.

### 4.7. SHAP framework for LR models

While LR models are often labeled as interpretable models, it is yet to demonstrate that every clinician fully understands the intricacy of such models, especially in the presence of non-linear terms

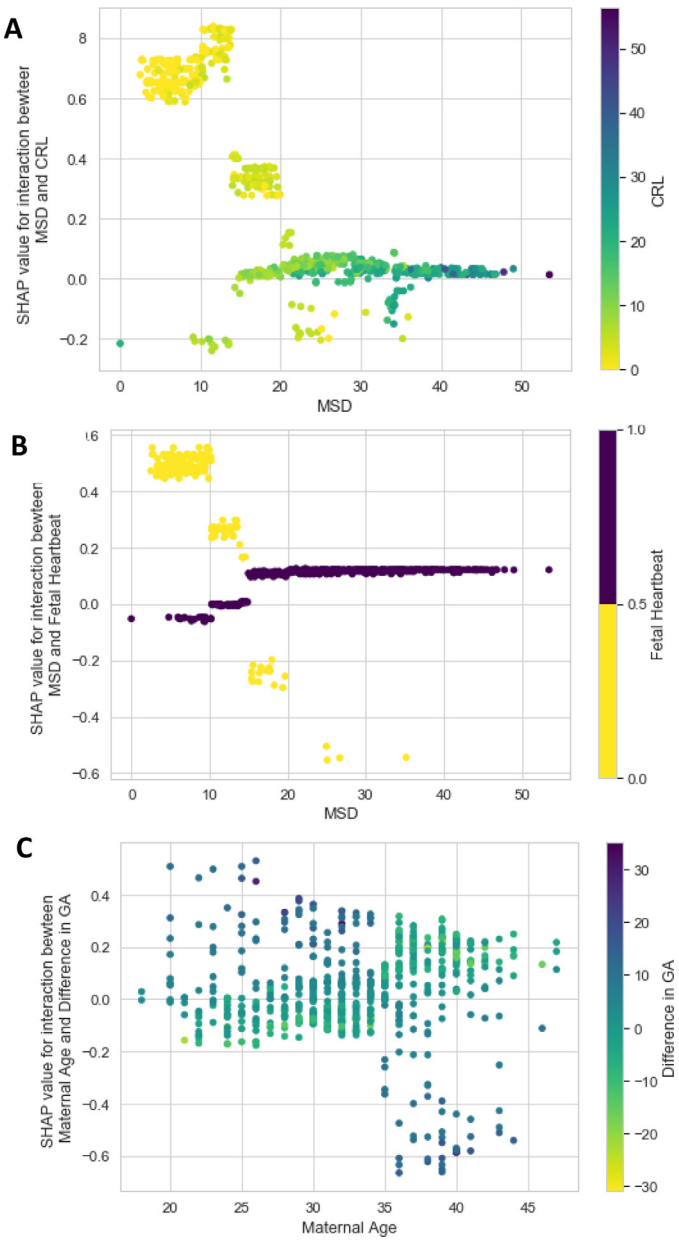


**Fig. 6.** *Top three of interaction effects learned from LGBM and measured with the SHAP framework.* The x-axis variable represents the first variable of the interaction. The color bar on the right indicates the value of the second variable present in the interaction. The SHAP value of the interaction is reported on the y-axis. **(A). MSD*CRL**, intrauterine pregnancies without visible embryo (CRL equals to zero) have a higher predicted risk of miscarriage (depicted as negative SHAP values) when MSD increases; (**B). MSD*FH**, pregnancies with high MSD without FH present have a higher predicted risk of miscarriage than with FH (see the negative SHAP values of yellow dots for MSD greater to 15 mm); and (**C). Maternal Age * difference in GA:** a large discrepancy in GA is modeled as a protective variable in young women while it becomes a risk factor in older women. CRL: crown-rump length; FH: fetal heartbeat, GA: gestational age; MSD: mean sac diameter.

and multicollinearity. We have demonstrated that the SHAP framework provides straightforward visualization of the feature importance and individual prediction explanations with decision path plots, irrespective of the variables' scales and collinearity. Intuitive decision path plots are easy to understand and do not require deep knowledge of LR formulation. Therefore, we believe that, as a post-hoc explanation method, the SHAP framework can also benefit simple clinical models such as LR models in complement to traditional coefficients analysis.

### *4.8. Gradient boosted trees advantages*

Besides models' performance and interpretability, we also demonstrated that LGBM produced similar results on incomplete datasets compared to imputed data. The internal handling of missing values by LGBM constitutes therefore a potential advantage over LR where (multiple) imputation should be carefully performed beforehand. Furthermore, the inherent non-linearity of LGBM algorithms facilitates the development of efficient models as it does not require explicit interaction terms like LR models.

Finally, while a clever preselection of meaningful variables by expert knowledge is often recommended to prevent unnecessarily complicated models with an increased risk of overfitting [43], without prior knowledge it can be difficult to establish such predefined set of variables. Despite its flexibility, LGBM models maintained good discriminative performance even with a large set of variables on the complete dataset. This demonstrates the efficient internal feature selection/weighting mechanism of gradient boosted trees. The high ranking of the pre-specified variables within the complete LGBM model also reflected this feature selection mechanism.

### *4.9. Strengths and limitations*

To the best of our knowledge, this study is one of the first to apply interpretable ML to first trimester viability prediction. The models were trained on a qualitative dataset from a well-defined prospective study using of validated symptom scores from early on in the first trimester. The models development included proper missing values imputations and hyper-parameters tuning. Unlike many previous comparative studies, this paper provides a rigorous models comparison through an extensive performances assessment beyond simple discriminative performance, including calibration and longitudinal visualizations of the performance metrics based on the GA.

One of the limitations of this study is the absence of a proper external validation set. However, we should note that the focus of this paper is not on building the ultimate predictive model but rather to demonstrate the potential of ML with post-hoc interpretability methods for early pregnancy predictive analytics. Secondly, our models use an estimation of GA by LMP, which is, however, not always available or accurate in practice [44]. Lastly, we would like to point out some practical limitations of the Shapley values approach. Because of its feature perturbation nature, the computation of Shapley values often need access to a background dataset (unless using the specific approach for tree ensembles [31]), which might impinge its use for model deployment. Moreover, depending on the dataset dimensionality, this perturbation step can be computationally expensive due to the combinatory nature of the Shapley value computation.

## 5. Conclusion

In this paper, we have demonstrated and assessed the use of machine learning enhanced by a post-hoc interpretability method for first trimester viability prediction. Gradient boosted algorithms performed as good as carefully crafted LR models in terms of discrimination and calibration. Furthermore, gradient boosted trees algorithms present several advantages over traditional LR models, such as the handling of missing values and the internal modeling of non-linearity, making them serious candidates for future works on first trimester prediction. Finally, we showed that the understanding of clinical models, including traditional LR models, can be improved by the use of additive feature attribution frameworks.

## Declaration of Competing Interest

The authors declare that no conflict of interest exists

## Acknowledgments/Funding

KU Leuven: Research Fund (projects C16/15/059, C3/19/053, C24/18/022, C3/20/117), Industrial Research Fund (Fellowships 13–0260, IOF/16/004) and several Leuven Research and Development bilateral industrial projects; Flemish Government Agencies: FWO: EOS Project no G0F6718N (SeLMA), SBO project S005319N, Infrastructure project I013218N, TBM Project T001919N; PhD Grants (SB/1SA1319N, SB/1S93918, SB/1S1319N), EWI: the Flanders AI Research Program VLAIO: Baekeland PhD (HBC.20192204) and Innovation mandate (HBC.2019.2209), CoT project 2018.018 European Commission: European Research Council under the European Union's Horizon 2020 research and innovation program (ERC Adv. Grant grant agreement No 885682); Other funding: Foundation 'Kom op tegen Kanker', CM (Christelijke Mutualiteit)

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2021.106520.

## References

[1] M.C. Magnus, A.J. Wilcox, N.-.H. Morken, C.R. Weinberg, S.E. Håberg, Role of maternal age and pregnancy history in risk of miscarriage: prospective register based study, BMJ 364 (2019), doi:10.1136/bmj.l869.
[2] L.M. Rossen, K.A. Ahrens, A.M. Branum, Trends in risk of pregnancy loss among US Women, 1990–2011, Paediatr. Perinat. Epidemiol. 32 (2018) 19–29, doi:10.1111/ppe.12417.
[3] L. Foo, S. Johnson, L. Marriott, T. Bourne, P. Bennett, C. Lees, Peri-implantation urinary hormone monitoring distinguishes between types of first-trimester spontaneous pregnancy loss, Paediatr. Perinat. Epidemiol. 34 (2020) 495–503, doi:10.1111/ppe.12613.
[4] P.A. Geller, D. Kerns, C.M. Klier, Anxiety following miscarriage and the subsequent pregnancy: a review of the literature and future directions, J. Psychosom. Res. 56 (2004) 35–45, doi:10.1016/S0022-3999(03)00042-4.
[5] J. Farren, M. Jalmbrant, N. Falconieri, N. Mitchell-Jones, S. Bobdiwala, M. Al-Memar, et al., Posttraumatic stress, anxiety and depression following miscarriage and ectopic pregnancy: a multicenter, prospective, cohort study, Am. J. Obstet. Gynecol. 222 (2020) 367e1–367e22, doi:10.1016/j.ajog.2019.10.102.
[6] J. Farren, N. Mitchell-Jones, J.Y. Verbakel, D. Timmerman, M. Jalmbrant, T. Bourne, The psychological impact of early pregnancy loss, Hum. Reprod. Update 24 (2018) 731–749, doi:10.1093/humupd/dmy025.
[7] A. Richardson, N. Raine-Fenning, S. Deb, B. Campbell, K Vedhara, Anxiety associated with diagnostic uncertainty in early pregnancy, Ultrasound Obstet. Gynecol. 50 (2017) 247–254, doi:10.1002/uog.17214.
[8] L. Detti, L. Francillon, M.E. Christiansen, I. Peregrin-Alvarez, P.J. Goeske, Z. Bursac, et al., Early pregnancy ultrasound measurements and prediction of first trimester pregnancy loss: a logistic model, Sci. Rep. 10 (2020) 1545, doi:10.1038/s41598-020-58114-3.
[9] S. Choong, L. Rombauts, A. Ugoni, S. Meagher, Ultrasound prediction of risk of spontaneous miscarriage in live embryos from assisted conceptions, Ultrasound Obstet. Gynecol. 22 (2003) 571–577 Off J Int Soc Ultrasound Obstet Gynecol, doi:10.1002/uog.909.
[10] J. Elson, R. Salim, A. Tailor, S. Banerjee, N. Zosmer, D. Jurkovic, Prediction of early pregnancy viability in the absence of an ultrasonically detectable embryo, Ultrasound Obstet. Gynecol. 21 (2003) 57–61 Off J Int Soc Ultrasound Obstet Gynecol, doi:10.1002/uog.1.
[11] K. Lautmann, M. Cordina, J. Elson, J. Johns, K. Schramm-Gajraj, J.A. Ross, Clinical use of a model to predict the viability of early intrauterine pregnancies when no embryo is visible on ultrasound, Hum. Reprod. 26 (2011) 2957–2963 Oxf Engl, doi:10.1093/humrep/der287.
[12] S. Guha, V. Van Belle, C. Bottomley, J. Preisler, V. Vathanan, A. Sayasneh, et al., External validation of models and simple scoring systems to predict miscarriage in intrauterine pregnancies of uncertain viability, Hum. Reprod. 28 (2013) 2905–2911 Oxf Engl, doi:10.1093/humrep/det342.
[13] T. Bignardi, G. Condous, E. Kirk, B. Van Calster, S. Van Huffel, D. Timmerman, et al., Viability of intrauterine pregnancy in women with pregnancy of unknown location: prediction using human chorionic gonadotropin ratio vs. progesterone, Ultrasound Obstet. Gynecol. 35 (2010) 656–661 Off J Int Soc Ultrasound Obstet Gynecol, doi:10.1002/uog.7669.
[14] A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, et al., Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (2017) 115–118, doi:10.1038/nature21056.
[15] A.W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, et al., Improved protein structure prediction using potentials from deep learning, Nature 577 (2020) 706–710, doi:10.1038/s41586-019-1923-7.
[16] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, et al., Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc, 2020, pp. 1877–1901.
[17] T. Chen, C. Guestrin, XGBoost: a Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794, doi:10.1145/2939672.2939785. 16.
[18] L. Liu, Y. Jiao, X. Li, Y. Ouyang, D. Shi, Machine learning algorithms to predict early pregnancy loss after *in vitro* fertilization-embryo transfer with fetal heart rate as a strong predictor, Comput. Methods Programs Biomed. 196 (2020) 105624, doi:10.1016/j.cmpb.2020.105624.
[19] M.W.L. Moreira, J. Rodrigues, V. Furtado, N. Kumar, V.V. Korotaev, Averaged one-dependence estimators on edge devices for smart pregnancy data analysis, Comput. Electr. Eng. 77 (2019) 435–444, doi:10.1016/j.compeleceng.2018.07.041.
[20] V. Bruno, M. D'Orazio, C. Ticconi, P. Abundo, S. Riccio, E. Martinelli, et al., Machine learning (ML) based-method applied in recurrent pregnancy loss (RPL) patients diagnostic work-up: a potential innovation in common clinical practice, Sci. Rep. 10 (2020) 7970, doi:10.1038/s41598-020-64512-4.
[21] S. Kuhle, B. Maguire, H. Zhang, D. Hamilton, A.C. Allen, K.S. Joseph, et al., Comparison of logistic regression with machine learning methods for the prediction of fetal growth abnormalities: a retrospective cohort study, BMC Pregnancy Childbirth 18 (2018) 333, doi:10.1186/s12884-018-1971-2.
[22] L.C. Poon, H.D. McIntyre, J.A. Hyett, E.B. da Fonseca, M. Hod, The first-trimester of pregnancy-a window of opportunity for prediction and prevention of pregnancy complications and future life, Diabetes Res. Clin. Pract. 145 (2018) 20–30, doi:10.1016/j.diabres.2018.05.002.
[23] J. He, S.L. Baxter, J. Xu, J. Xu, X. Zhou, K. Zhang, The practical implementation of artificial intelligence technologies in medicine, Nat. Med. 25 (2019) 30–36, doi:10.1038/s41591-018-0307-0.
[24] Towards trustable machine learning, Nat. Biomed. Eng. 2 (2018) 709–710, doi:10.1038/s41551-018-0315-x.
[25] S.M. Lundberg, S.I Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017) https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.
[26] M. Al-Memar, T. Vaulet, H. Fourie, G. Nikolic, S. Bobdiwala, S. Saso, et al., Early-pregnancy events and subsequent antenatal, delivery and neonatal outcomes: prospective cohort study, Ultrasound Obstet. Gynecol. 54 (2019) 530–537, doi:10.1002/uog.20262.
[27] C. Bottomley, V. Van Belle, E. Kirk, S. Van Huffel, D. Timmerman, T Bourne, Accurate prediction of pregnancy viability by means of a simple scoring system, Hum. Reprod. 28 (2013) 68–76, doi:10.1093/humrep/des352.
[28] M.J. Azur, E.A. Stuart, C. Frangakis, P.J. Leaf, Multiple imputation by chained equations: what is it and how does it work? Int. J. Methods Psychiatr. Res. 20 (2011) 40–49, doi:10.1002/mpr.329.
[29] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. 29 (2001) 1189–1232, doi:10.1214/aos/1013203451.
[30] S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, et al., From local explanations to global understanding with explainable AI for trees, Nat. Mach. Intell. 2 (2020) 56–67, doi:10.1038/s42256-019-0138-9.
[31] S.M. Lundberg, G.G. Erion, S.I. Lee, in: Consistent individualized feature attribution for tree ensembles. Proceedings of the 2017 ICML Workshop on Human Interpretability in Machine Learning (WHI 2017), 2017, pp. 15–21.
[32] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, LightGBM, et al., A Highly Efficient Gradient Boosting Decision Tree, Advances in Neural Information Processing Systems 30 (2017) 3146–3154.
[33] J.S. Bergstra, R. Bardenet, Y. Bengio, B. Kégl, Algorithms for hyper-parameter optimization, in: Proceedings of the 24th International Conference on Neural Information Processing Systems, 2011, pp. 2546–2554.
[34] E.W. Steyerberg, Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating, 2nd edition, Springer, 2019.
[35] E.R. DeLong, D.M. DeLong, D.L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, Biometrics 44 (1988) 837–845, doi:10.2307/2531595.
[36] T. Miller, Explanation in artificial intelligence: insights from the social sciences, Artif. Intell. 267 (2019) 1–38, doi:10.1016/j.artint.2018.07.007.
[37] D. Janzing, L. Minorics, P. Blöbaum, Feature relevance quantification in explainable AI: a causal problem, Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, PMLR (2020) 2907–2916 https://proceedings.mlr.press/v108/janzing20a.html.
[38] M. Sundararajan, A. Najmi, The many shapley values for model explanation, in: Proceedings of the International Conference on Machine Learning, 2020, pp. 9269–9278. http://proceedings.mlr.press/v119/sundararajan20b.html. PMLR Available.
[39] E. Christodoulou, J. Ma, G.S. Collins, E.W. Steyerberg, J.Y. Verbakel, B.V. Calster, A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models, J. Clin. Epidemiol. 110 (2019) 12–22, doi:10.1016/j.jclinepi.2019.02.004.
[40] A.L. Lynam, J.M. Dennis, K.R. Owen, R.A. Oram, A.G. Jones, B.M. Shields, et al., Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: application to the discrimination between type 1 and type 2 diabetes in young adults, Diagn. Progn. Res. 4 (2020) 6, doi:10.1186/s41512-020-00075-2.

[41] L. Breiman, Statistical modeling: the two cultures (with comments and a rejoinder by the author), Stat. Sci. 16 (2001) 199–231, doi:10.1214/ss/1009213726.
[42] C.F. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, et al., Collinearity: a review of methods to deal with it and a simulation study evaluating their performance, Ecography 36 (2013) 27–46, doi:10.1111/j.1600-0587.2012.07348.x.
[43] F.E. Harrell, Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, Springer, 2015.
[44] G. Wegienka, D.D. Baird, A comparison of recalled date of last menstrual period with prospectively recorded dates, J. Womens Health 14 (2002) 248–252 2005, doi:10.1089/jwh.2005.14.248.