Explaining Large-for-Gestational-Age Births: A Random Forest Classifier with a Novel Local Interpretation Method  
Yuhan Du¹, Anthony R Rafferty¹, Fionnuala M McAuliffe¹, Catherine Mooney¹  
¹University College Dublin, Dublin, Ireland.

Abstract
We proposed a novel local interpretation method for a random forest classifier based on feature occurrence frequency in trees that give the same prediction as the random forest classifier. The method shows promising results when applied to our random forest classifier for large-for-gestational-age births. Further validation of the method is required.

Introduction
Explainability is important for machine learning-based clinical decision support systems, because it helps to gain trust from users and it is essential to comply with the “right to explanation” mandated by the General Data Protection Regulation (GDPR) in the EU.

In our previous work [1], we developed a random forest classifier for the prediction of large-for-gestational-age births based on baseline maternal characteristics, baseline blood biomarkers and ultrasound scan findings at 20 weeks gestation. The overall workflow of the modelling process is shown below.

Method (cont.)
Consider a random forest classifier consisting of n number of decision trees $T = \{\text{tree}_1, \text{tree}_2, \text{tree}_3, ..., \text{tree}_n\}$ with k input features $F = \{f_1, f_2, ..., f_k\}$, as shown in the figure below.

Each decision tree is trained on a subset of features $F_i$, $F_i \subset F$, $i \in \{1, 2, ..., n\}$. For a particular prediction of the random forest classifier, m number of decision trees $T_s$ give the same prediction as the random forest (i.e. majority when decision threshold = 0.5). $T_s = \{\text{tree}_1, \text{tree}_2, ..., \text{tree}_m\} \subset T$, $T_s \leftarrow m$. $T_s \leftarrow T$.

For this particular prediction, it is the $T_s$ and the features included in $T_s$ that lead to the predicted class, instead of other decision trees that give opposite prediction. Therefore, we could estimate the contribution of each feature to a particular prediction by the occurrence frequency of the feature in $T_s$.

However, features with higher global importance are more often selected in internal nodes in the decision trees. To eliminate the effect of the global importance, we subtracted the occurrence frequency of a feature in all trees T from the occurrence frequency of the feature in $T_s$ (see equation below).

$$\text{Contribution}(f_i) = \frac{|\{ \text{tree}_s \in T_s, f_i \in T_s \}| - |\{ \text{all trees } T, f_i \in T \}|}{m}, \quad i \in \{1, 2, ..., k\}$$

The “contribution” indicates how much a feature contributes to a particular prediction without the impact of global predictive power. A larger contribution value indicates that the feature contributes more to the prediction, and vice versa.

Results & Discussion
To test the performance of our novel local interpretation method, we randomly generated 100 hypothetical cases, which have one numerical feature with an extreme value (defined as <25th or >75th percentile) and all other numerical features were within the normal range (defined as 25th-75th percentile).

We explained the predictions of these hypothetical cases by our model using our proposed local interpretation method, and ranked the features by their contribution values.

The features with extreme values were often ranked high. An overall decreasing trend was observed in the histogram of the rankings of these features (see figure above). The histogram goes up at the end (i.e. 12) because the features with extreme values contributed to the opposite prediction in some cases.

Therefore, we conclude that our local interpretation method gives promising results for our model. Further validation is required.

References