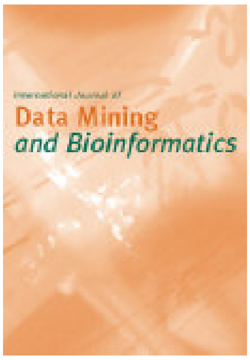


International Journal of Data Mining and Bioinformatics

ISSN online: 1748-5681 - ISSN print: 1748-5673
https://www.inderscience.com/ijdmb

Screening and sparsifying maternal immune features for predicting labour induction based on a glass-box model

Rong Hu, Jing Li, Yan Liu, Qianqian Zhang, Zhaoyi Bai, Zhongwei Zhao, Weiwei Guo, Rong Liu

DOI: 10.1504/IJDMB.2025.10073158

Article History:

Received:	22 February 2025
Last revised:	21 July 2025
Accepted:	28 July 2025
Published online:	02 September 2025

Screening and sparsifying maternal immune features for predicting labour induction based on a glass-box model

Rong Hu, Jing Li, Yan Liu, Qianqian Zhang, Zhaoyi Bai, Zhongwei Zhao, Weiwei Guo and Rong Liu*

Department of Obstetrics and Gynaecology,
Tianjin First Central Hospital,
#24 Fukang Road, Nankai District,
Tianjin 300192, China
Fax: +86-22-23626572
Email: rong_hu@outlook.com
Email: plumwain@126.com
Email: 30819007@nankai.edu.cn
Email: 691102127@qq.com
Email: iambaizhaoyi@163.com
Email: pingjing.good@163.com
Email: 15602178704@163.com
Email: liuronghch1999@sina.com
*Corresponding author

Abstract: Immune signatures strongly associate with the progression of labour towards the active phase of labour. However, the detailed relationship is still not clear. Herein, interpretable machine learning methods are implemented for mining complex immune data. Principal component analysis and covariance analysis are employed to achieve dimensionality reduction of the immune features (1,058) as input. Using 16 key immune features as input, RMSE decreased from 277 min to 214 min by Ridge model. Moreover, sure independence screening and sparsifying operator (SISSO) was implemented to establish a glass-box model for generating interpretable mathematical information format of key immune features associated with induced labour progression. The prediction accuracy was further improved by SISSO input with only 14 features ($R^2 = 0.9934$, RMSE = 42 min, MAE = 30 min), and the exact mathematical format of the model was obtained [equation (5)]. Reliable description of progression is established from labour induction until establishing active labour.

Keywords: interpretable machine learning; labour induction; compressed-sensing method; regression; pregnancy.

Reference to this paper should be made as follows: Hu, R., Li, J., Liu, Y., Zhang, Q., Bai, Z., Zhao, Z., Guo, W. and Liu, R. (2025) 'Screening and sparsifying maternal immune features for predicting labour induction based on a glass-box model', *Int. J. Data Mining and Bioinformatics*, Vol. 29, No. 6, pp.1–15.

Biographical notes: Rong Hu is an attending physician in the Department of Obstetrics and Gynaecology at Tianjin First Central Hospital with 12 years of clinical expertise. Driven by the need for precision medicine in obstetrics, she is conducting research at the intersection of clinical medicine and machine learning.

Jing Li is an attending physician in the Department of Obstetrics and Gynaecology of Tianjin First Central Hospital centring on evidence-based intrapartum care, including standardised labour induction protocols.

Yan Liu is a chief physician in the Department of Obstetrics and Gynaecology of Tianjin First Central Hospital specialising in managing high-risk pregnancies, particularly complex maternal comorbidities and obstetric complications.

Qianqian Zhang is an attending physician in the Department of Obstetrics and Gynaecology of Tianjin First Central Hospital committed to increasing vaginal delivery rate through optimised induction protocols.

Zhaoyi Bai is an attending physician in the Department of Obstetrics and Gynaecology of Tianjin First Central Hospital focusing on decision-making pathways for heterogeneous patients.

Zhongwei Zhao is an attending physician with 11 years of clinical experience in the Department of Obstetrics and Gynaecology of Tianjin First Central Hospital.

Weiwei Guo is an attending physician with nine years of clinical experience in the Department of Obstetrics and Gynaecology of Tianjin First Central Hospital.

Rong Liu is a chief physician in the Department of Obstetrics and Gynaecology of Tianjin First Central Hospital. As the Director of the Obstetrics Department at Tianjin First Central Hospital and the standing committee member of the Perinatal Medicine Branch of Tianjin Medical Association she has worked in developing real-time AI decision-support systems for labour management.

1 Introduction

Labour induction has an enormous impact on maternal and foetal well-being (NICE, 2021; Oladapo et al., 2018; Sande et al., 1983), which is usually recommended in the situation of deteriorating foetal or/and maternal status. The benefit and safety of induction, e.g., lower rates of caesarean delivery, were shown by Grobman et al. (2018), which can exist even in the absence of a biomedical indication. Labour inductions constitute around 40% of all deliveries now and are being increasingly performed in the USA (Ameri et al., 2024; Rydahl et al., 2019). During active labour, the local environment is inflammatory (Ando et al., 2021; Migale et al., 2016; Orsi and Tribe, 2008; Pique-Regi et al., 2019; Sivarajasingam et al., 2016). It is detectable and echoed in maternal circulating immune cells, e.g., increased frequencies of CD56^{+}CD16^{+} natural killer cells, neutrophils activation, and the increase of inflammatory cytokines (Yuan et al., 2009; Zhang et al., 2017). Currently, the reliable description of progression since labour induction until establishing active labour is still weak, and the prediction of complication and labour induction success is still not accurate enough. This is owing to the highly nonlinear correlation and intricate relationship between the 'biological determinants', (e.g., immune features) and the progression since labour induction (Migale et al., 2016; Orsi and Tribe, 2008; Pique-Regi et al., 2019; Sivarajasingam et al., 2016).

Several methods of data analytics or machine learning were developed and applied to mine medical science data (Alber et al., 2019; Ando et al., 2021; Camacho et al., 2018; De Santiago and Polanski, 2022; Kaur et al., 2019; Peng et al., 2021; Shehab et al., 2022; Swanson et al., 2023; Zheng et al., 2023). Thus far, the works of machine learning in medical science mostly used black-box models to regress measurable biological properties (descriptors), such as immune features, that can be related to the medical process (e.g., labour) (Afrifa et al., 2023; Alsharif, 2023; Ando et al., 2021; Chen et al., 2024; Meyer et al., 2023). It is challenging to extract meaningful biological insights from a black-box model owing to its high complexity. The internal logic of the black-box models cannot be readily explainable. Interpretable methods of machine learning, which merge the biological interpretability of mathematics-based model with the prediction capacity of black-box models, provide an alternative to the conventional black-box models.

The reliable prediction of progression after labour induction is an important component of theoretical description and clinical decision-making for maternal and foetal well-being. These relationships should be revealed by one descriptor, including several parameters that can capture the underlying mechanism of labour process. The challenge is that tiny changes in biological determinants may cause a qualitative variation of the pregnant body. For instance, during labour, many different phenomena, processes and changes of feto-maternal physiology exist, including endocrine adaptations (McLean et al., 1995; Mendelson, 2009; Mesiano, 2007), infiltration of immune cells into the placenta and foetal membranes (Gomez-Lopez et al., 2014; Romero et al., 1989; Shynlova et al., 2013), foetal membrane rupture (Menon et al., 2019), uterine contractility augmentation, cervical dilation (Norwitz et al., 1999), and culminating in foetus delivery. After identification of the descriptor, essentially any learning approach, including regressions and classifications based on kernel function, artificial neural networks, and so on, can be straightforwardly applied. The key role of the descriptor has been identified explicitly in many pioneering works of catalysis (Andersen et al., 2019) and materials science (Ghiringhelli et al., 2017, 2015; Ouyang et al., 2018) using machine learning methods, while relatively less attention has been focused on medical science.

In this study, the peripheral immune data after labour induction have been regressed based on data dimensionality reduction and model hyperparameter adjustment to avoid overfitting and increase prediction accuracy. An interpretable machine learning method, named sure independent screening and sparsifying operator (SISSO), is implemented to describe the dynamic changes of the specific maternal immune system in the period between labour induction and active labour establishment. SISSO can even deal with billions of object features, and does not suffer from high correlation of the features. We establish a mathematical model, which is an analytic and explicit function of input immune features, rather than a specific biological model. We accept that the intricate processes, which compete and/or cooperate in induced labour, can not be describable necessarily by a rigorous and closed biological equation with complete pathways (Draxl and Scheffler, 2018). This work opens a new perspective of screening the biological determinants that affect the variability in induced labour, which can provide mathematical models for the establishment of new biological mechanism.

2 Data mining methods

2.1 Dataset

The data published by Ando et al. (2021) have been used. In their observational study, a time schedule of immunologic adaptations during labour was detected, in which functional changes and peripheral immune cell phenotype were serially analysed after medical induction of labour. In their file of 'preprocessed.csv' (Ando et al., 2021), the 'time' column and 'sampleID' column jointly mark the index of the sample. The values in the feature column corresponding to the same indicators of 'time' and 'sampleID' are treated as the same sample. Therefore, a vector with 1058 dimensions and a total of 48 vectors were obtained. In the regression, X is a 48 × 1,058 matrix, and the corresponding Y is a 48 dimensional vector with the values in the column of 'minsinceinduction'. We hope to establish an accurate and interpretable relationship between intracellular signalling marker or frequency of immune cell type and time since the induction of labour. During model training, the data were split into two sets randomly, including the training set (70%) and the testing set (30%).

2.2 Machine learning models

In this work, different machine learning models were employed for the regression of the data, including ridge (Tikhonov, 1943), least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), and sure independence screening (SIS) and sparsifying operator (SISSO) (Ouyang et al., 2018). The method of principal component analysis (PCA) (Lever et al., 2017) was employed to achieve dimensionality reduction of the input. Compared with SISSO, covariance analysis (Khammar et al., 2020) was also carried out for descriptor selection.

2.2.1 LASSO regression

LASSO regression is a multivariate linear regression added with 1-norm part as model parameter. In the LASSO model, a hyperparameter, λ, controls the relative importance of the regularisation term and the mean square error loss term. The value of λ was optimised firstly to increase the prediction accuracy. 5-fold cross validation was implemented during the optimisation to achieve a good generalisation ability.

$$L = \|y - X\beta\|^2 + \lambda\|\beta\|_1 \tag{1}$$

2.2.2 Ridge regression

Ridge regression is another multivariate linear regression model using the square of 2-norm as model parameter. In the Ridge model, the hyperparameter, λ, also determines the relative importance of the regularisation term and the mean square error loss term. The value of λ was optimised to increase the prediction accuracy of Ridge model. 5-fold cross validation was implemented during the optimisation to achieve a good generalisation ability.

$$L = \left\| y - X\beta \right\|^2 + \lambda \left\| \beta \right\|_2^2 \tag{2}$$

2.2.3 Principal component analysis

PCA is a robust tool of data analysis in many fields, ranging from material science to medical science (Shlens, 2014). Relevant information can be extracted from confusing data sets by PCA using a non-parametric technique. PCA can simplify the high-dimensional data with high complexity, in which the patterns and trends can be retained. The dimension of the data can be reduced by PCA through geometrically projecting onto lower dimensions, i.e., principal components (PCs). The best summary of the data can be determined by using a limited number of PCs.

2.2.4 Descriptor selection based on covariance

As a widely used statistical method, covariance analysis deals with quantitative data from experimental studies in many fields, including medical science. Given an n-dimensional random variable, $X = (x_1, x_2 \ldots x_n)$, there exists covariance between its different components (features), which can form a covariance matrix, $C_{n\times n}$, as shown in equation (3).

$$C_{ij} = E\left[\left(x_i - E\left(x_i\right)\right)\left(x_j - E\left(x_j\right)\right)\right] \tag{3}$$

where E represents expectation. The positive value of the element C_{ij} in the covariance matrix indicates that the features (x_i and x_j) are positively correlated. If the value is negative, it indicates that x_i and x_j are negatively correlated. The larger the absolute value, the stronger the correlation. If C_{ij} is zero, then x_i and x_j are independent.

2.2.5 SISSO

To construct the feature spaces of Φ_1, Φ_2 and Φ_3, the set of functional/algebraic operators given in equation (4) was used.

$$\hat{H}^m \overset{\text{def}}{=} \left\{+, -, \times, \exp, \exp-, {}^{-1}, {}^{-2}, {}^{-3}, sqrt\right\} \tag{4}$$

m, the superscript, describes that a dimensional analysis was performed when we applied $\hat{H}^m$ to primary features (φ_1 and φ_2). Only the combinations that are physically meaningful have been retained. Only the primary features possessing the same unit have been subtracted or added.

In a small feature subspace which is selected by SIS, the sparsifying l_0 constraint has been applied. The subspace size is equal to a SIS value (user-defined) times the descriptor dimension.

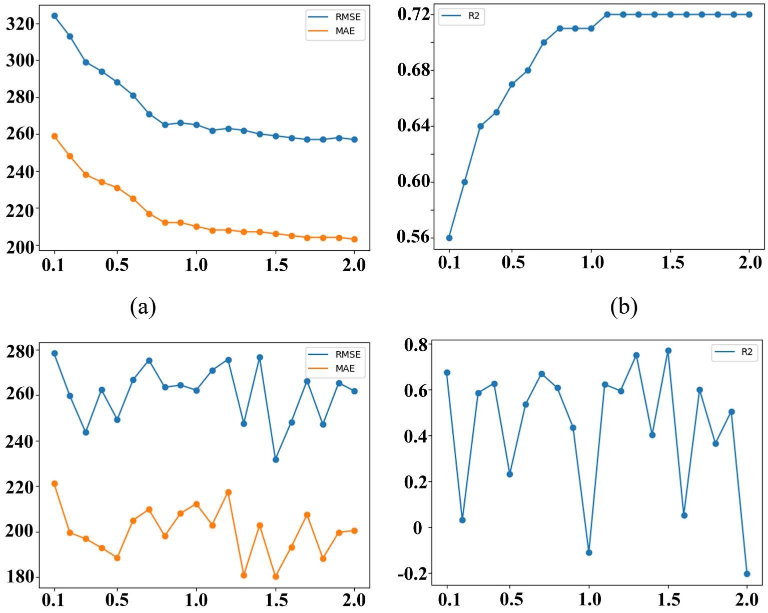
In our work, the dimension of the descriptor was set to 4; maximal feature complexity was also set to 4. Number of features in each of the SIS-selected subspace was set to 50 (Ouyang et al., 2018). The feature was removed when its number is smaller than 10^{-3} and larger than 10^5.

3 Result

3.1 LASSO regression

In the LASSO model, the value of hyperparameter (λ) was optimised firstly to increase the prediction accuracy. The results are shown in Figures 1(a)–1(b). The value of R^2 increased constantly with the increase of λ from 0.1 to 1.1. When λ was higher than 1.1, R^2 was nearly unchanged. The LASSO model performs best at λ = 2.0, where RMSE = 257 min, MAE = 203 min and R^2 = 0.72. Therefore, the value of λ was set to be 2.0 in the subsequent investigation and discussion. Our results based on LASSO model is consistent with those reported by Ando et al. (2021) (RMSE = 277 min).

Figure 1 Influence of hyperparameter (λ) on RMSE (blue) and MAE (orange) values of (a) LASSO regression and (c) Ridge regression. Influence of hyperparameter (λ) on R^2 of (b) LASSO regression and (d) Ridge regression (see online version for colours)



(a)

(b)

(c)

(d)

3.2 Ridge regression

In the Ridge model, the value of λ was also adjusted, and the results are shown in Figures 1(c)–1(d). The prediction accuracy of Ridge model exhibited severe fluctuations by changing the value of λ from 0.1 to 2.0, which is very different from the trend for LASSO regression. According to our results, Ridge regression model has the best performance when λ = 1.5, where R^2 = 0.77, RMSE = 232 min, and MAE = 180 min.

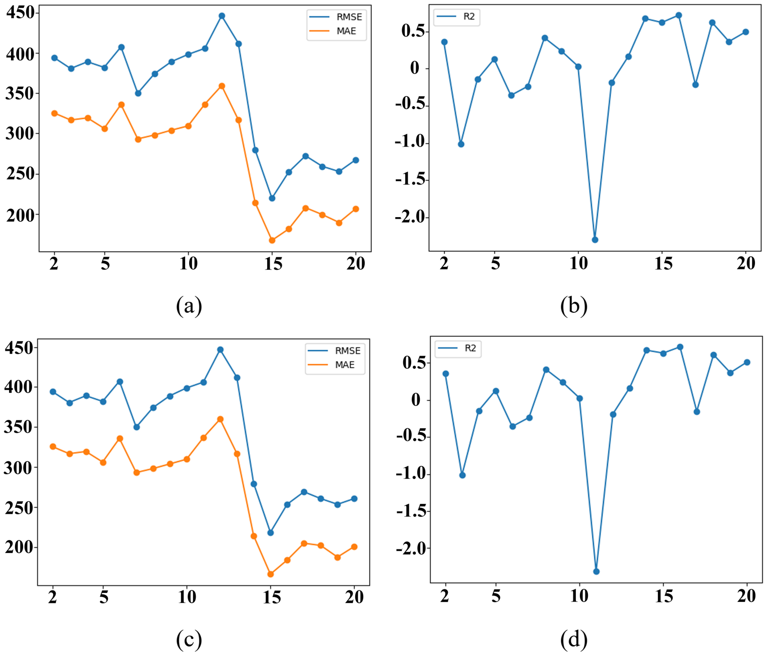
3.3 Interpretability brought by dimensionality reduction

Multiple linear regression models with regularisation terms (LASSO and Ridge) achieved good regression results. But the input vector of each sample was as high as 1,058 dimensions. The excessively high input dimension inevitably leads to a large amount of redundancy in the model, and overfitting can easily occur. We hope to make the model more concise by using dimensionality reduction technology, so that the knowledge learned by machine learning models from the data can be more easily understood by researchers.

3.3.1 Dimensionality reduction by PCA

Firstly, the dimensionality of the input data is reduced by PCA technology. Then, the performance of the linear models (LASSO and Ridge) using the input generated by PCA dimensionality reduction was tested.

Figure 2 Influence of the dimension of input data after PCA processing on RMSE (blue) and MAE (orange) of (a) LASSO model and (c) Ridge model. Influence of the dimension of input data after PCA processing on R^2 of (b) LASSO model and (d) Ridge model (see online version for colours)



(a)

(b)

(c)

(d)

We investigated the changes in model performance of LASSO when the input vector was reduced to different dimensions (2–20 dimensions) by PCA technique. As shown in Figures 2(a)–2(b), when the input dimension was lower than 12, both RMSE and MAE values were large. RMSE and MAE were constantly decreased by increasing the input dimension from 12 to 15. RMSE and MAE had the minimum values of 219 min and 167 min, respectively, when the input dimension was 15. At this time, the R^2 value is 0.62. When the input dimension was higher than 15, both RMSE and MAE did not change much. The maximum goodness of fit ($R^2 = 0.72$) was obtained when the input dimension was 16. Under this condition, the corresponding values of RMSE and MAE were 252 min and 181 min, respectively.

After PCA dimensionality reduction, the performance change of Ridge model is almost the same with that of LASSO model. As shown in Figures 2(c)–2(d), the minimum values of RMSE and MAE were 218 min and 166 min, respectively, when the input dimension was 15 and the corresponding value of R^2 was 0.63. When the input dimension of the model was 16, the maximum R^2 was achieved, which is 0.72, and corresponding RMSE and MAE were 253 min and 184 min, respectively. Both Ridge and LASSO models exhibited similar performance when the input dimension is reduced by PCA.

3.3.2 Descriptor selection based on covariance

In order to increase interpretability of the model, the covariance method was employed for dimensionality reduction, in which correlation analysis was achieved based on covariance values. We calculated the covariance between different features and the target property (time since induction), sorted them based on the absolute values of the covariance, and selected the features possessing the highest covariance with the target property as the descriptor. The 30 features selected based on covariance values that have the strongest correlation with the target properties are listed in Table 1. Here, we tested the prediction performances of LASSO and Ridge models using the top 20 features as the input descriptor.

Figures 3(a)–3(b) show the impact of the top N ($N = 2$–20) features as input on model performance of LASSO. These features possess higher absolute values of the covariance with the target property. When the top 19 features were selected as input, the smallest RMSE and MAE were obtained, which are 227 min and 155 min, respectively. But the goodness of fit was only 0.37. When the top 16 features were selected, the goodness of fit was 0.74, while RMSE and MAE are 231 min and 179 min, respectively. Thus, LASSO model can achieve the best performance when input with the top 16 features.

The pattern of Ridge model is almost identical to that of LASSO. Figures 3(c)–3(d) shows that the minimum RMSE (218 min) and MAE (151 min) values were achieved when Ridge model was input with the top 19 features possessing the highest correlation with the target property. But the goodness of fit was only 0.42. When the top 16 features were selected, the goodness of fit was 0.77, while RMSE and MAE were 214 min and 166 min, respectively. At this point, the goodness of fit is consistent with the performance of a model that directly uses all features as input.

3.4 SISSO

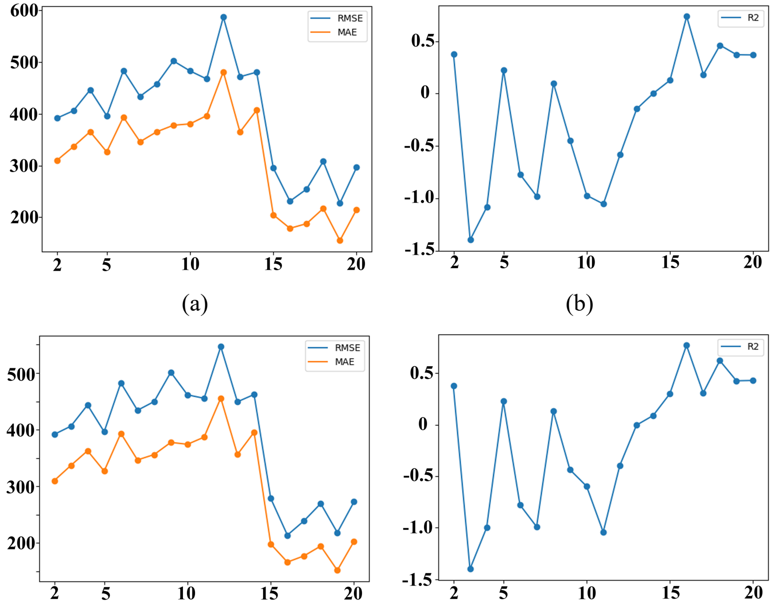
Covariance analysis can achieve good interpretability through selecting the most relevant properties as input, and the model performance is similar to those input using all features (1,058). However, we still hope to develop glass-box models with interpretable and superior performance. SISSO was chosen because the priority of this work is to derive compact, interpretable, and physically meaningful equations from a huge pool of candidate features in medical applications. In comparison, decision trees and SHAP serve different purposes: either as general-purpose classifiers (trees) or model explainers (SHAP).

Table 1 Top 30 features determined by covariance analysis

No.	*Features*[a]	*No.*	*Features*[a]	*No.*	*Features*[a]
1	CMCs_Frequency	11	Bcells_Frequency	21	mDCsHLADRlo_pSTAT3
2	CCR2poscMCs_Frequency	12	CD4Tem_Frequency	22	CD4Tcells_pSTAT3
3	CD4Tcells_Frequency	13	CD62LposCD8Tnaive_Frequency	23	NK_Frequency
4	Granulocytes_Frequency	14	CD4Tcm_Frequency	24	CD4Tef-154Sm_pSTAT3
5	CD8Tcells_Frequency	15	CD62LposCD4Tnaive_pSTAT3	25	CD62LposCD8Tnaive_pSTAT3
6	CD4Tmem_Frequency	16	CD4posTnaive_pSTAT3	26	MDSCs_pSTAT3
7	MDSCs_Frequency	17	CD8Tef_Frequency	27	Th1_pSTAT3
8	CD4posTnaive_Frequency	18	mDCsHLADRhi_pSTAT3	28	Granulocytes_pSTAT3
9	CD8Tnaive_Frequency	19	mDCs_pSTAT3	29	CD56loCD16posNK_Frequency
10	CD62LposCD4Tnaive_Frequency	20	CD8Tmem_Frequency	30	CCR2poscMCs_pSTAT3

Note: [a]For the detailed meaning of feature abbreviations, please refer to the paper of Ando et al. (2021).

Figure 3 RMSE (blue) and MAE (orange) values of (a) LASSO model and (c) Ridge model input with different amounts of features as descriptor. R^2 values of (b) LASSO model and (d) Ridge model input with different amounts of features as descriptor (see online version for colours)

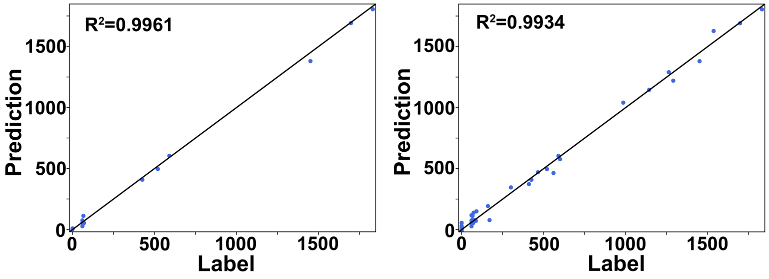


(a) (b)

(c) (d)

Note: The features with higher absolute value of covariance with the target property (time since induction) were preferably selected.

Figure 4 Regression results of complicated [(a), $R^2 = 0.9961$, RMSE = 38 min, MAE = 26 min] and all [(b), $R^2 = 0.9934$, RMSE = 42 min, MAE = 30 min] labour induction processes by the as-developed SISSO model (see online version for colours)



(a) (b)

A SISSO model has been built for mining medical data, and all the top 30 features determined by covariance analysis were tested. The best 4D descriptor of Φ3 was identified by SISSO, in which 14 features are included. Four immune features that strongly associated with time since induction are CD4Tcells_pSTAT3, mDCs_pSTAT3, MDSCs_pSTAT3, CCR2poscMCs_pSTAT3. Ten frequency features are also included in the model, including CD8Tcells_Frequency, CD4Tcm_Frequency, CD56loCD16posNK_

Frequency, CD8Tmem_Frequency, CD62LposCD4Tnaive_Frequency, CCR2poscMCs_ Frequency, CD8Tef_Frequency, CD4Tem_Frequency, CD8Tnaive_Frequency, MDSCs_ Frequency. The exact prediction model of time since induction using these features is given in equation (5).

$$
\begin{aligned}
y = {} & 3{,}535.69 \times \frac{\left[\mathrm{CD4Tcells_{pSTAT3}}\right]}{\dfrac{\left[\mathrm{CD8Tcells_{Frequency}}\right]}{\left[\mathrm{CD4Tcm_{Frequency}}\right]} + \dfrac{\left[\mathrm{CD56loCD16posNk_{Frequency}}\right]}{\left[\mathrm{CD8Tmem_{Frequency}}\right]}} \\
& -5{,}303.47 \times \frac{\dfrac{\left[\mathrm{mDCs_{pSTAT3}}\right]}{\left[\mathrm{CD62LposCD4Tnaive_{Frequency}}\right]}}{\left[\mathrm{mDCs_{pSTAT3}}\right] \times \left[\mathrm{CD8Tcells_{Frequency}}\right] - \left[\mathrm{CCR2poscMCs_{Frequency}}\right]} \\
& -15.05 \times \frac{\left[\mathrm{MDSCs_{pSTAT3}}\right]}{\dfrac{\left[\mathrm{CD8Tef_{Frequency}}\right]}{\left[\mathrm{CD4Tem_{Frequency}}\right]} - \left[\mathrm{mDCs_{pSTAT3}}\right] \times \left[\mathrm{CCR2poscMCs_{pSTAT3}}\right]} \\
& +27.55 \times \frac{\left[\mathrm{CCR2poscMCs_{pSTAT3}}\right]}{\dfrac{\left[\mathrm{CD62LposCD4Tnaive_{Frequency}}\right]}{\left[\mathrm{CD8Tnaive_{Frequency}}\right]} - \dfrac{\left[\mathrm{CD8Tef_{Frequency}}\right]}{\left[\mathrm{MDSCs_{Frequency}}\right]}} + 4.07
\end{aligned}
\tag{5}
$$

The regression results of the as-developed model are shown in Figure 4. High accuracy ($R^2 > 0.99$, RMSE < 42 min, MAE < 31 min) was obtained on both the complicated data and the whole database. Therefore, an interpretable model with clear mechanism and superior performance has been established.

4 Discussion

Currently, it is still hard to describe labour onset and establishment, particularly induced labour, by a biologically founded model with an analytical and closed expression. Because the labour processes are determined by substantial intricate and multilevel theoretical concepts. A LASSO model was introduced by Ando et al. (2021) and a RMSE of 277 min was obtained. In the model, 1,058 features were considered, but only 48 samples were input. Thus, the equation system is an underdetermined system, resulting in an infinite number of solutions. The model possessed excessive hypothesis space, which will easily lead to overfitting. But the accuracy of regression was still low (RMSE = 277 min), indicating that the LASSO model can be further optimised. Herein, we firstly optimised the linear models (LASSO and Ridge) by adjusting hyperparameter (λ). The prediction accuracy was only slightly increased (Ridge model, $R^2 = 0.77$, RMSE = 232 min, MAE = 180 min). Moreover, PCA was employed to reduce the number of the features that input into the prediction model. By changing the dimension of the input from 2 to 20, the improvement in the regression results is still limited. In addition, the biological meanings of the new descriptors after PCA dimensionality reduction are unclear based on these traditional black-box models.

PCA is a mature dimensionality reduction technique. However, the physical meaning of the new descriptor after PCA dimensionality reduction is unclear. In order to increase interpretability of the model, the covariance method was employed for dimensionality reduction, in which correlation analysis was achieved based on covariance values. We calculated the covariance between different features and the target property (time since induction), sorted them based on the absolute values of the covariance, and selected the features with the highest covariance with the target property as the descriptors. The top 20 features determined by covariance analysis were tested in LASSO and Ridge models. The values of RMSE and MAE were decreased to 214 min and 166 min on Ridge model, respectively, where $R^2 = 0.77$. But the results are still not ideal. It can be concluded that the hypothesis spaces established by current linear models (LASSO and Ridge) is not suitable for the data of labour induction. It turns out that the linear approach breaks down due to the unstructured feature space with correlated candidate features of induced labour. The problems of large feature space and small training set have been solved by SISSO. SISSO exhibited superior advantages for dealing with the correlated immune features of induced labour. SISSO autonomously screens the best features and removes irrelevant features from the combination of 1,058 immune features as a new descriptor, so that the feature space has been efficiently optimised. The regression accuracy has been greatly increased ($R^2 = 0.9934$, RMSE = 42 min, MAE = 30 min). Given the available feature space, SISSO also identifies the accurate relationship between immune features and labour progress in terms of an analytical equation [equation (3)]. The 14 features selected by SISSO, which are most prominently in the STAT3 pathway, are biologically plausible markers for labour progression (Papatheodorou et al., 2013; Sivarajasingam et al., 2016). Previous reports (Vega-Sanchez et al., 2010; Yuan et al., 2009) showed that IL-6 gene expression in circulating leukocytes was enhanced during active labour, which well matches our conclusion that active labour associates STAT3 signalling.

5 Conclusions

In summary, the regression of peripheral immune data of labour induction is systematically investigated. The input and model parameter (λ) of two black models (LASSO and Ridge) were optimised. The prediction accuracy of LASSO model and Ridge model is hardly improved by adjusting λ and reducing input dimensionality by PCA and covariance analysis, which demonstrates the inherent defect of both linear models. The hypothesis space established by the linear models (LASSO and Ridge) is not suitable for the complex immune data after labour induction. In comparison, SISSO can establish a suitable hypothesis space based on only 14 features for properly dealing with the complex immune data of induced labour. As a complementary method to the commonly-used black-box approaches, SISSO can identify and translate the hidden patterns into detailed mathematic forms, which can form testable theories and hypotheses. Our results demonstrate that the interpretable machine learning methods pave the way to significantly advance scientific understanding in medical domain.

Declarations

All authors declare that they have no conflicts of interest.

References

Afrifa, S., Varadarajan, V., Appiahene, P. et al. (2023) 'A novel artificial intelligence techniques for women breast cancer classification using ultrasound images', *Clinical and Experimental Obstetrics & Gynecology*, Vol. 50, No. 12, pp.271–284.

Alber, M., Buganza Tepole, A., Cannon, W.R. et al. (2019) 'Integrating machine learning and multiscale modelling – perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences', *NPJ Digital Medicine*, Vol. 2, No. 1, p.115.

Alsharif, W.M. (2023) 'The utilization of artificial intelligence applications to improve breast cancer detection and prognosis', *Saudi Medical Journal*, Vol. 44, No. 2, p.119.

Ameri, A., Jafariazar, Z., Annabi, M. et al. (2024) 'Effect of misoprostol versus oxytocin on delivery outcomes after labour induction in pregnant women: a systematic review and meta-analysis of randomized controlled trials', *European Journal of Obstetrics & Gynecology and Reproductive Biology*, Vol. 292, pp.75–88.

Andersen, M., Levchenko, S.V., Scheffler, M. et al. (2019) 'Beyond scaling relations for the description of catalytic materials', *ACS Catalysis*, Vol. 9, No. 4, pp.2752–2759.

Ando, K., Hédou, J.J., Feyaerts, D. et al. (2021) 'A peripheral immune signature of labor induction', *Frontiers in Immunology*, Vol. 12, p.725989.

Camacho, D.M., Collins, K.M., Powers, R.K. et al. (2018) 'Next-generation machine learning for biological networks', *Cell*, Vol. 173, No. 7, pp.1581–1592.

Chen, Y., Han, K., Liu, Y. et al. (2024) 'Identification of effective diagnostic genes and immune cell infiltration characteristics in small cell lung cancer by integrating bioinformatics analysis and machine learning algorithms', *Saudi Medical Journal*, Vol. 45, No. 8, p.771.

De Santiago, I. and Polanski, L. (2022) 'Data-driven medicine in the diagnosis and treatment of infertility', *Journal of Clinical Medicine*, Vol. 11, No. 21, p.6426.

Draxl, C. and Scheffler, M. (2018) 'NOMAD: the FAIR concept for big data-driven materials science', *MRS Bulletin*, Vol. 43, No. 9, pp.676–682.

Ghiringhelli, L.M., Vybiral, J., Ahmetcik, E. et al. (2017) 'Learning physical descriptors for materials science by compressed sensing', *New Journal of Physics*, Vol. 19, No. 2, p.023017.

Ghiringhelli, L.M., Vybiral, J., Levchenko, S.V. et al. (2015) 'Big data of materials science: critical role of the descriptor', *Physical Review Letters*, Vol. 114, No. 10, p.105503.

Gomez-Lopez, N., StLouis, D., Lehr, M.A. et al. (2014) 'Immune cells in term and preterm labor', *Cellular & Molecular Immunology*, Vol. 11, No. 6, pp.571–581.

Grobman, W.A., Rice, M.M., Reddy, U.M. et al. (2018) 'Labor induction versus expectant management in low-risk nulliparous women', *New England Journal of Medicine*, Vol. 379, No. 6, pp.513–523.

Kaur, H., Pannu, H.S. and Malhi, A.K. (2019) 'A systematic review on imbalanced data challenges in machine learning: applications and solutions', *ACM Computing Surveys*, Vol. 52, No. 4, p.79.

Khammar, A., Yarahmadi, M. and Madadizadeh, F. (2020) 'What is analysis of covariance (ANCOVA) and how to correctly report its results in medical research?', *Iran J. Public Health*, Vol. 49, No. 5, pp.1016–1017.

Lever, J., Krzywinski, M. and Altman, N. (2017) 'Principal component analysis', *Nature Methods*, Vol. 14, No. 7, pp.641–642, https://doi.org/10.1038/nmeth.4346.

McLean, M., Bisits, A., Davies, J. et al. (1995) 'A placental clock controlling the length of human pregnancy', *Nature Medicine*, Vol. 1, No. 5, pp.460–463.

Mendelson, C.R. (2009) 'Minireview: fetal-maternal hormonal signaling in pregnancy and labor', *Molecular Endocrinology*, Vol. 23, No. 7, pp.947–954.

Menon, R., Richardson, L.S. and Lappas, M. (2019) 'Fetal membrane architecture, aging and inflammation in pregnancy and parturition', *Placenta*, Vol. 79, pp.40–45.

Mesiano, S. (2007) 'Myometrial progesterone responsiveness', *Seminars in Reproductive Medicine*, Vol. 25, No. 1, pp.5–13.

Meyer, R., Weisz, B., Eilenberg, R. et al. (2023) 'Utilizing machine learning to predict unplanned cesarean delivery', *International Journal of Gynecology & Obstetrics*, Vol. 161, No. 1, pp.255–263.

Migale, R., MacIntyre, D.A., Cacciatore, S. et al. (2016) 'Modeling hormonal and inflammatory contributions to preterm and term labor using uterine temporal transcriptomics', *BMC Medicine*, Vol. 14, No. 1, p.86.

NICE (2021) *Inducing Labour NICE Guideline*, pp.1–42, NICE, National Institute for Health and Care Excellence, London, UK [online] https://www.nice.org.uk/guidance/ng207 (accessed 12 February 2025).

Norwitz, E.R., Robinson, J.N. and Challis, J.R.G. (1999) 'The control of labor', *New England Journal of Medicine*, Vol. 341, No. 9, pp.660–666.

Oladapo, O., Vogel, J. and Gülmezoglu, A. (2018) *WHO Recommendations: Induction of Labour at or Beyond Term*, pp.1–39, Geneva, WHO [online] https://apps.who.int/iris/bitstream/handle/10665/277233/9789241550413-eng.pdf?ua=1 (accessed 12 February 2025).

Orsi, N.M. and Tribe, R.M. (2008) 'Cytokine networks and the regulation of uterine function in pregnancy and parturition', *Journal of Neuroendocrinology*, Vol. 20, No. 4, pp.462–469.

Ouyang, R., Curtarolo, S., Ahmetcik, E. et al. (2018) 'SISSO: a compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates', *Physical Review Materials*, Vol. 2, No. 8, p.083802.

Papatheodorou, D.C., Karagiannidis, L.K., Paltoglou, G. et al. (2013) 'Pulsatile Interleukin-6 leads CRH secretion and is associated with myometrial contractility during the active phase of term human labor', *The Journal of Clinical Endocrinology & Metabolism*, Vol. 98, No. 10, pp.4105–4112.

Peng, G.C.Y., Alber, M., Buganza Tepole, A. et al. (2021) 'Multiscale modeling meets machine learning: what can we learn?', *Archives of Computational Methods in Engineering*, Vol. 28, No. 3, pp.1017–1037.

Pique-Regi, R., Romero, R., Tarca, A.L. et al. (2019) 'Single cell transcriptional signatures of the human placenta in term and preterm parturition', *eLife*, Vol. 8, p.e52004.

Romero, R., Brody, D.T., Oyarzun, E. et al. (1989) 'Infection and labor: III. Interleukin-1: a signal for the onset of parturition', *American Journal of Obstetrics and Gynecology*, Vol. 160, No. 5, Part 1, pp.1117–1123.

Rydahl, E., Eriksen, L. and Juhl, M. (2019) 'Effects of induction of labor prior to post-term in low-risk pregnancies: a systematic review', *JBI Evidence Synthesis*, Vol. 17, No. 2, pp.170–208.

Sande, H.A., Tuveng, J. and Fønstelien, T. (1983) 'A prospective randomized study of induction of labor', *International Journal of Gynecology & Obstetrics*, Vol. 21, No. 4, pp.333–336.

Shehab, M., Abualigah, L., Shambour, Q. et al. (2022) 'Machine learning in medical applications: a review of state-of-the-art methods', *Computers in Biology and Medicine*, Vol. 145, p.105458.

Shlens, J. (2014) *A Tutorial on Principal Component Analysis*, ArXiv, Vol. abs/1404.1100,

Shynlova, O., Lee, Y-H., Srikhajon, K. et al. (2013) 'Physiologic uterine inflammation and labor onset: integration of endocrine and mechanical signals', *Reproductive Sciences*, Vol. 20, No. 2, pp.154–167.

Sivarajasingam, S.P., Imami, N. and Johnson, M.R. (2016) 'Myometrial cytokines and their role in the onset of labour', *Journal of Endocrinology*, Vol. 231, No. 3, pp.R101–R119.

Swanson, K., Wu, E., Zhang, A. et al. (2023) 'From patterns to patients: advances in clinical machine learning for cancer diagnosis, prognosis, and treatment', *Cell*, Vol. 186, No. 8, pp.1772–1791.

Tibshirani, R. (1996) 'Regression shrinkage and selection via the Lasso', *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 58, No. 1, pp.267–288.

Tikhonov, A.N. (1943) 'On the stability of inverse problems', *Proceedings of the USSR Academy of Sciences*, Vol. 39, pp.195–198.

Vega-Sanchez, R., Gomez-Lopez, N., Flores-Pliego, A. et al. (2010) 'Placental blood leukocytes are functional and phenotypically different than peripheral leukocytes during human labor', *Journal of Reproductive Immunology*, Vol. 84, No. 1, pp.100–110.

Yuan, M., Jordan, F., McInnes, I.B. et al. (2009) 'Leukocytes are primed in peripheral blood for activation during term and preterm labour', *Molecular Human Reproduction*, Vol. 15, No. 11, pp.713–724.

Zhang, J., Shynlova, O., Sabra, S. et al. (2017) 'Immunophenotyping and activation status of maternal peripheral blood leukocytes during pregnancy and labour, both term and preterm', *Journal of Cellular and Molecular Medicine*, Vol. 21, No. 10, pp.2386–2402.

Zheng, X., Ma, H., Dong, Y. et al. (2023) 'Immune-related biomarkers predict the prognosis and immune response of breast cancer based on bioinformatic analysis and machine learning', *Functional & Integrative Genomics*, Vol. 23, No. 3, p.201.