International Journal of Data Mining and Bioinformatics

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

Plasma proteins related to the state of depression: a case-control study based on proteomics data of pregnant women

Yuhao Feng, Jinman Zhang, Zengyue Zheng, Chenyu Xing, Min Li, Guanghong Yan, Ping Chen, Dingyun You, Ying Wu

DOI: 10.1504/IJDMB.2025.10064226

Article History:

Received:	23 June 2023
Last revised:	08 February 2024
Accepted:	09 February 2024
Published online:	10 July 2025

Plasma proteins related to the state of depression: a case-control study based on proteomics data of pregnant women

Yuhao Feng

Department of Biostatistics,
School of Public Health,
Southern Medical University,
Guangzhou 510080, China
Email: yhfeng2435@gmail.com

Jinman Zhang

National Health Commission's Key Laboratory
for Healthy Births in Western China,
Department of Obstetrics and Gynecology,
First People's Hospital of Yunnan Province,
Kunming 650032, China
Email: 171887587@qq.com

Zengyue Zheng

Department of Biostatistics,
School of Public Health,
Southern Medical University,
Guangzhou 510080, China
Email: ZhengZengYue@outlook.com

Chenyu Xing, Min Li, Guanghong Yan and Ping Chen

NHC Key Laboratory of Periconception Health Birth in Western China,
School of Public Health,
Kunming Medical University,
Kunming 650500, China
Email: 18331507503@163.com
Email: 1171756127@qq.com
Email: 18468049813@163.com
Email: 2644672845@qq.com

Dingyun You

Institute of Biomedical Engineering,
Kunming Medical University,
Kunming, 650500, China
Email: youdingyun@qq.com

Ying Wu*

Department of Biostatistics,
School of Public Health,
Southern Medical University,
Guangzhou 510080, China
Email: wuying19890321@gmail.com
*Corresponding author

Abstract: Prenatal and postpartum emotional changes in pregnant women in early pregnancy are of great significance to the physical and mental health of mothers and infants. To identify factors related to this, we conducted this study to identify feature proteins that cause maternal depression. Boruta algorithm (BA), recursive partition algorithm (RPA), regularised random forest (RRF) algorithm, least absolute shrinkage and selection operator (LASSO) algorithm, and genetic algorithm (GA) were used to select features. Extreme gradient boosting (XGBoost), back propagation neural network (BPNN), support vector machine (SVM), random forest (RF), and logistic regression (LR) were selected to construct the predictive models. All models showed a good performance in predicting, with the mean AUC (the area under the receiver operating curve) exceeding 80%. Features will provide clues to prevent depression in pregnant women and improve the physical and mental health of mothers and babies.

Keywords: pregnant women; depression; proteomics; biomarkers; feature selection.

Reference to this paper should be made as follows: Feng, Y., Zhang, J., Zheng, Z., Xing, C., Li, M., Yan, G., Chen, P., You, D. and Wu, Y. (2025) 'Plasma proteins related to the state of depression: a case-control study based on proteomics data of pregnant women', *Int. J. Data Mining and Bioinformatics*, Vol. 29, No. 3, pp.313–337.

Biographical notes: Yuhao Feng is a third-year graduate student of Biostatistics at the Department of Public Health School at Southern Medical University. His research interests include causal inference, clinical trials, and statistical methods.

Jinman Zhang is a Professor and Doctor in Obstetrics and Gynecology at the National Health Commission's Key Laboratory for Healthy Births in Western China with a focus on the application of medical data analysis. She is also interested in bioinformatics.

Zengyue Zheng is a second-year graduate student in Biostatistics at Southern Medical University. His research interests include real-world research, observational studies, and causal inference.

Chenyu Xing is a graduate student at Kunming Medical University who majored in theories and applications of omics data, especially focusing on mining data from laboratory research and bioinformatics.

Min Li is a graduate student at Kunming Medical University who majored in theories and applications of omics data, especially focusing on mining data from laboratory research and bioinformatics.

Guanghong Yan is a graduate student at Kunming Medical University who majored in theories and applications of omics data, especially focusing on mining data from laboratory research and bioinformatics.

Ping Chen is a graduate student at Kunming Medical University who majored in theories and applications of omics data, especially focusing on mining data from laboratory research and bioinformatics.

Dingyun You is the Vice Director and researcher in the School of Public Health at Kunming Medical University. He serves on the Youth Committee of Clinical Epidemiology and Evidence-Based Medicine of the Chinese Medical Association and Director of the Chinese Hospital Rescue Association.

Ying Wu is an Associate Professor at Southern Medical University. She concurrently serves as the Secretary-General of the Guangdong Provincial Biostatistics Society, a medical device review consulting expert for the Guangdong Provincial Food and Drug Administration, and a member of the Medical and Health Systems Engineering Professional Committee of the Chinese Society of Systems Engineering. She participated in the formulation and writing of several statistical guiding principles issued by the State Food and Drug Administration.

1 Introduction

Postpartum depression (PPD) is a common problem after a child's birth and may influence the quality of life (QOL). Research into postpartum QOL and depression can be used for better care for mothers and to improve their well-being (Sadat et al., 2014). Many studies have discussed the relationship between depression and maternal (infant) health from different perspectives (Zhao et al., 2018; Redinger et al., 2020; Friedman et al., 2020). The distinction between those and this paper is we consider the maternal prenatal and postpartum emotional changes and related factors based on proteomics. We believe that finding and understanding these factors is not only vital for preventing mood changes in women's prenatal and postnatal but also vital for their babies physically and mentally.

Previous discoveries (before 2020) focused on factors associated with maternal depression, but from the view of proteomics is relatively rare (as of the time of this writing; Zhao et al., 2016; Zhang et al., 2018). Edvinsson et al. (2019) discussed the relationship between depression and proteins and the relationship between drugs and depression at the protein level. Brann et al. (2017) investigated whether inflammatory markers in third-trimester plasma samples could predict the presence of depressive symptoms at eight weeks postpartum. Recent research (after 2020) also examined this topic. Mao et al. (2021) found that betaine and succinic acid were significantly associated

with maternal depression in metabolomics. Redei et al. (2021) stated that ESR2 and mPRβ could identify depressive symptoms in pregnant women in transcriptomic. Redinger et al. (2020) emphasised that studies focusing only on late pregnancy may underestimate risk and pointed out that early identification is critical for prevention and treatment. Also, Pak et al. (2020) extended and reconsidered PPD from a perspective of social survey and mentioned the importance of relevant linguistics/sentiment that can contribute to PPD. Aparicio et al. (2020) claimed that higher psychosocial stress predicted breast milk's higher cortisol concentrations. Although this study explored the relationship between human milk composition and maternal mental health, its conclusions remind us that more work is needed to prove the relationship between depression and proteins.

Methods and frameworks for selecting features include but are not limited to the following. Haddow et al. (2011) centred on protein structure as a feature selection problem and compared results from various approaches and the standard method. Hadzic et al. (2010) developed an intelligent system based on data-mining technologies to prevent depression and claimed that this system could help all parties involved. Chattopadhyay (2013) attempted an innovative way of diagnosing depression, in which a mathematical model was used to help doctors assign appropriate class labels of depression flexibly.

The above research explored the relationship between maternal depression and proteins, which are essential for promoting maternal and infant health. However, there are not many such studies, and some only rely on the judgement of individual clinicians, which may be an obstacle for researchers who plan to survey maternal depression from proteomics. Given this, we conducted a case-control study to obtain clues about the state of depression and proteins, and we hope these clues may be used as a reference in this field for both scientific research and clinical practice.

The structure of this paper is as follows. In Section 2, we provide the details of the materials. Methods are separated into two parts in Section 3, i.e., the feature selection and predictive performance. Also, we provide the proofs for algorithmic modification in the supplementary materials for readability. Section 4 Results are presented and explained clearly by principles of algorithms and models. We discuss and conclude with a summary in Section 5.

2 Materials

2.1 Sample size and data source

This study was a case-control study that targeted investigating associations between plasma proteins and maternal depression. The subjects of this study were singleton pregnant women who underwent pregnancy examination and delivery in the Obstetrics Department of the First Affiliated Hospital of Kunming Medical University (Yunnan Province, China) from September 2019 to December 2020. These pregnant women were enrolled on this study randomly, and their venous peripheral blood samples were collected during the examination, and then these samples were stored in –80°C refrigerators. The baseline information was obtained through questionnaires, and the general epidemiological data of pregnant and lying-in women were collected, including age, weight, height, gestational age, education level, occupation, residence, etc. (in the

supplementary material). We compared these questionnaires with the records in the Obstetrics Department to ensure their accuracy and authenticity. Specifically, records contained foetal conditions, labour records, delivery records, postpartum records, delivery methods, delivery gestational weeks, complications, etc. We then checked maternal pregnancy outcomes and related information from these records to ensure authenticity and credibility.

The inclusion and exclusion criteria were these. For inclusion criteria:

1 singleton pregnancy

2 pregnant women who underwent examination during pregnancy and whose blood samples were preserved

3 have the ability to read and communicate and voluntarily participated in the survey

4 without symptoms of miscarriage and threatened premature delivery

5 spontaneous premature delivery, which including insufficient monthly delivery, preterm premature rupture of membrane (PPROM, in line with the diagnostic criteria for premature birth).

For exclusion criteria:

1 multiple pregnancies

2 those with severe medical and surgical diseases (gestational diabetes, gestational hypertension, preeclampsia, congenital heart disease, mental illness, etc.)

3 taking antidepressants and other psychotropic drugs

4 therapeutic premature births, stillbirths, miscarriages, a congenital defection, etc.

This study was approved by the Medical Ethics Committee of Kunming Medical University by strict review and complies with relevant ethical standards. All subjects voluntarily participated in the study and signed informed consent forms and patient information will be kept confidential.

From the perspective of external validity and statistical power, the larger the sample size, the better. However, considering the limitations of practicality and time constraints, the research subjects were divided into two groups according to whether they were depressed among the target pregnant women who met the inclusion criteria. The depressed group and the healthy group were matched by propensity score according to the age of the parturient and the gestational week of blood collection. These 24 pregnant women were assigned to each group, and 48 cases were included in the final study. The 48 blood samples were analysed by using proteomics. We realised that our sample size is not big enough, and the results drawn from this dataset may not be perfect. However, it is reasonably presumed that it still can serve as a reference in similar research, especially when the subjects are pregnant women.

2.2 Process for obtaining proteomics data and quality controlling

Taking the sample from the –80 centigrade refrigerator, centrifuged at 12,000 g for 10 minutes at 4 degrees Celsius to remove cell debris, and then transferred the supernatant to a new centrifuge tube. Next, high-abundance proteins were eliminated

through the high-abundance protein removal kit Pierce™ Top 14 abundant protein depletion spin columns kit (Thermo Scientific). Protein concentration was determined using the BCA kit. By performing the same enzymatic digestion of the protein of each sample and adjusting its volume to reach the same level in concentration as the lysate, a final concentration of 5 mM was obtained by adding dithiothreitol (DTT) and incubating at 56 degrees Celsius. The reduction was carried out for 30 minutes. Iodoacetamide (IAA) was then added to a final concentration of 11 mM and was incubated for 15 minutes at room temperature in the dark. Put the alkylated sample into an ultrafiltration tube, centrifuged at 12,000 g for 20 minutes at room temperature, then used 8M urea for three replacements. Later, the replacement buffer for three replacements was used, and finally, the 1:50 proportional trypsin was added, and enzymatic treatment was performed for one night. Peptides were centrifuged at 12,000 g for 10 minutes at room temperature, recovering once with ultrapure water, and the peptide solutions were combined twice.

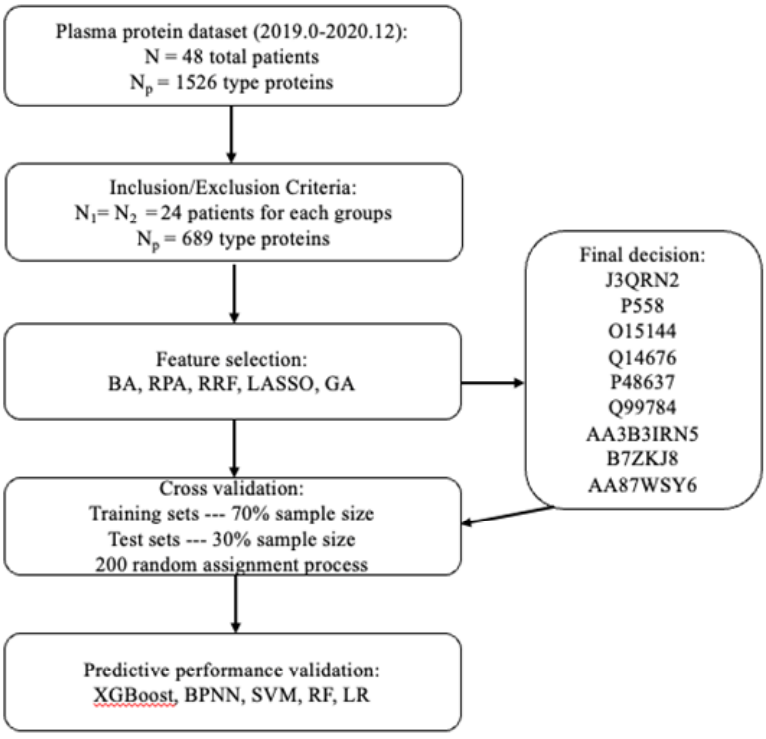
Peptides were separated using high-performance liquid chromatography (HPLC A) and Easy-nLC 1200 ultra-high-performance liquid separation system (names for kits and suppliers can be seen in the supplementary materials). Mobile phase A was an aqueous solution containing 0.1% formic acid and 2% acetonitrile; mobile phase B was an aqueous solution containing 0.1% formic acid and 90% acetonitrile. Liquid gradient setting as these: 0~68 min, 4~20% of B; 68~82 min, 20~32% of B; 82~86 min, 32~80% of B; 86~90 min, 80% of B, the flow rate maintained at 500 nL per minute. Peptides were injected into an NSI ion source (nano-spray-ionisation source) for ionisation once the peptides were separated by an ultra-high performance liquid phase system and then analysed by an Orbitrap Exploris™480 mass spectrometer. The ion source voltage was set to 2.3 KV, and the high-field asymmetric waveform ion mobility spectrometry (FAIMS) compensation voltage was set to –45 V and –70 V. The high-resolution Orbitrap was used to detect and analyse the polypeptide precursor and its secondary fragments. The scanning range of the primary mass spectrogram was set to 400–1,200 m/z (mass-to-charge ratio), and the scanning resolution was set to 60,000; the fixed starting point of the scanning range of the secondary mass spectrogram is 110 m/z, the secondary scanning resolution was set to 30,000, and TurboTMT (tandem mass tag) was set off. The data-dependent acquisition (DDA) program was used for data acquisition, i.e., the first 15 polypeptide precursor ions with the highest signal intensity were selected from the first scan and then projected into the higher energy collision-induced dissociation (HCD) collision cell with 27% energy, i.e., the secondary mass spectrometry was performed to obtain more accurate results. To better use mass spectrometry, we adjusted the parameters of automatic gain control (AGC) to 75%, the signal threshold to 1E4 ions per second, the maximum injection time to 100 milliseconds, and the dynamic exclusion time of tandem mass spectrometry scanning to 30 seconds, to reduce the repetitive scanning of precursor ions.

Secondary mass spectrometry data for this experiment were retrieved using Proteome Discoverer (V2.4.1.15). Set the search parameters by using the HOMO_SAPIENS_9606_PR_20210721.FastA (78120 sequences) as the database, then added an inverse library to estimate the false discovery rate (FDR) caused by random matching. Furthermore, a public contamination pool was added to counteract the effect of protein contamination in the results. Use Trypsin (full) as the method of enzyme digesting. Setting the number of missed cleavage sites to 2, the minimum peptide length to 6 amino acid residues, the maximum number of peptide modifications to 3, and the error for the mass of the primary precursor ions to 10 ppm (parts per million). Control the error of the

weight of the second fragment ions within 0.02 Da. Set aminomethylmethyl ester (C) as the fixed modification, oxidation (M), acetyl (N-terminal), methionine (M), and methionine with acetyl (M) as variables which can be modified. FDR is set to 1% for proteins, similarly for peptides and peptide spectrum match (PSM) identifications.

We reviewed all information of patients in the stage of collecting questionnaires and recording the raw data. Specifically, we checked the logical connection between the raw data and the records to guarantee no marked errors and modified the related unreasonable data.

Figure 1 An overview workflow of marker protein identification related to maternal depression



Note: BA – Boruta algorithm, RPA – recursive partition algorithm, RRF – regularised random forest, LASSO – least absolute shrinkage and selection operator algorithm, GA – genetic algorithm, XGBoost – extreme gradient boosting, BPNN – back propagation neural network, SVM – support vector machine, RF – random forest, LR – logistic regression. N, N1 and N2 indicate the total and group sample size, respectively. Np for types of proteins. In the final decision framework are the names of proteins.

2.3 Overview workflow of data analysis

The workflow process of data analysis is shown in Figure 1. Firstly, the original data (from mass spectrometry detection) were filtered, including removing the missing data and scaling the remaining. Secondly, five algorithms were used to screen for feature proteins. Thirdly, predictive models were built with states of depression as the dependent variable and nine features as independent variables. Finally, we evaluated the

performance of each model and chose AUC as the evaluation criterion. In detail, we discarded proteins with more than 50% missing values and then normalised the remaining proteins using the Z-scores method. Five algorithms took the state of depression as the dependent variable (binary outcome) and proteins as the independent variable (continuous variable). Proteins were selected as the feature were those that appeared at least twice in each algorithm. Remarkably, we decided them as features this way because we considered the principle of each algorithm to be slightly different, so the feature proteins found by these algorithms were not exactly consistent. As a result, nine proteins became the focus of the following analysis, which were helpful to robust results. Finally, we examined the predictive performance according to AUC. To ensure the results were informative and unbiased, we randomly split the dataset into the training set (70%) and the test set (30%) for constructing and validating predictive models.

Table 1 Baseline information of the sample data

	Total (n = 48)	*Health (n = 24)*	*Control (n = 24)*	*P-value*
Age in years	29.6 (25.5, 33.7)	29.6 (25.5, 33.7)	29.7 (25.5, 33.9)	0.983
BMI	23.1 (19.6, 26.6)	23.1 (19.3, 26.9)	23.0 (19.8, 26.2)	0.813
Blood sampling time in weeks	14.9 (7.7, 22.1)	14.9 (7.6, 22.2)	14.9 (7.6, 22.2)	0.984
Delivery time in weeks	36.8 (33.6, 40)	39.4 (38.6, 40.2)	34.1 (31.7, 33.8)	<0.0001
Education				0.419[a]
Bachelor and above	33 (66.8%)	16 (66.7%)	17 (70.8%)	
Below bachelor	11 (22.9%)	7 (29.2%)	4 (16.7%)	
Missing*	4 (8.3%)	1 (4.2%)	3 (12.0%)	
Residence				0.370[a]
City	31 (64.6%)	15 (62.5%)	16 (66.7%)	
Village	15 (31.3%)	9 (37.5%)	6 (25.0%)	
Missing*	2 (4.2%)	0	2 (8.3%)	
Nation				0.787[a]
Han	35 (72.9%)	18 (75.0%)	17 (70.8%)	
Minority	9 (18.8%)	5 (20.8%)	4 (16.7%)	
Missing	4 (8.3%)	1 (4.2%)	3 (12.0%)	
Occupation				0.459
Enterprises	18 (37.5%)	11 (45.8%)	7 (29.2%)	
Individual business	15 (31.3%)	7 (29.2%)	8 (33.3%)	
Unemployed/others	15 (31.3%)	6 (25.0%)	9 (37.5%)	
Planned delivery				0.724[a]
Yes	38 (79.2%)	18 (75.0%)	20 (83.3%)	
No	10 (20.8%)	6 (25.0%)	4 (16.7%)	
Gender of infant				1
Male	23 (47.9%)	11 (45.8%)	12 (50.0%)	
Female	25 (52.1%)	13 (54.2%)	12 (50.0%)	

Notes: *Missing data in these variables, the miss proportions less than 12%.
[a]Fisher's exact test.

2.4 Pre-processing and baseline information

The data for this study were derived from a case-control study at Kunming Medical University, which aimed to investigate the association between plasma proteins and maternal depression. Concretely, the study collected plasma samples of 48 pregnant women ranging from 25 to 40 years old; the delivery time ranged from 28 to 39 weeks; the birth weight of infants ranged from 990 grams to 4,010 grams. The time for sampling of plasma samples was different according to the health situation of each pregnant woman. The earliest sampling time was in the seventh week, and the latest was in the 23rd week. Demographically, the average age was 29.6 years old, and the average BMI was 23.1. Among them, 35 were the Han Nationality, accounting for 72.9%, and the rest were national minorities. The demographic characteristics are shown in Table 1. There were no statistical differences in age, BMI (post-pregnancy), gestational age of blood collection, education level, place of residence, ethnicity, occupation, foetal gender, and whether this was a planned pregnancy in the two groups of patients. However, for pregnant women, the average gestational age in the control group was 34.1 weeks, and the average gestational age of the encounter in the healthy group was 39.4 weeks, which was statistically significant (Table 1).

Information was obtained after sampling the plasma proteins of these 48 observations. Each observation contained a total of 1,526 kinds of proteins in their sample. We subsequently normalised (Z-scores) the raw data for statistical analysis. However, due to uncontrolled factors in the sampling procedures, the information about all 1,526 proteins was not detected in the blood sample for all pregnant women. To ensure that all analyses were based on a complete dataset, we removed proteins with missing data, so the final dataset for analysis contained 689 proteins. Details are presented in the section of Results.

3 Methods

3.1 Algorithms for feature selection

We adapted the raw data to make it more reasonable and applicable for algorithms and models adopted in our exploration, for which we provide the details as a separate file for readability. There are many statistical methods for feature selection, and they aim for different scenarios (Remeseiro and Bolon-Canedo, 2019). This paper chose the five algorithms for two reasons. First, they are frequently used in which the outcome variable is dichotomous; second, they are consistent with the original data format, which ensures that further analysis and results are reasonable and consistent. The point is they are frequently used for feature selection, which implies they are already accepted in this field. Also, we realised that the methods chosen here might not be the latest. But, at the same time, we need to point out that the purpose of this study was not to compare the effects of different algorithms but to identify proteins associated with maternal depression. Using these frequently used algorithms to confirm the features ensures that the selected proteins are reliable. This provides a clue to preventing maternal depression and improving maternal and infant mental health. The following is a brief introduction to the five algorithms.

The first algorithm is Boruta, whose goal is to select all feature sets related to the outcomes of interests rather than select the feature set that can minimise the model cost function for a marker model. The significance of the Boruta algorithm is that it can help one more comprehensively understand the influencing factors of dependent variables to perform better and more efficient feature selection. For the details of its principle derivation and R implementation, please refer to Kursa and Rudnicki (2010).

The second and third algorithms were RPA and RRF. These two algorithms are also high-frequency used for feature selection in machine learning. The principles and implementation steps of these two methods are similar. Specifically, according to a unique feature, the data is divided into several sub-regions (subtrees), and then the sub-regions are recursively divided until a particular condition is met. Then, the division is stopped and used as a leaf node. If the condition is not met, the recursive division continues. For more details about these two algorithms, please refer to Rajaguru and Chakravarthy (2019) and Liu et al. (2014).

The LASSO algorithm is an example of the regularisation of regression algorithms. LASSO is performed using the 'glmnet' package in R. The features selected by LASSO bear a greater biological significance and are named pivotal genes before being used in machine learning-based model validation (Hai et al., 2022).

GA is a computational model of the biological evolution process that simulates the natural selection and genetic mechanism of Darwin's theory of biological evolution and is a method to search for the optimal solution by simulating the natural evolution process. Its main feature is that it directly deals with objects, has no limitations of derivation and function continuity and has inherent implicit parallelism and a better ability to search for optimal solutions globally. Ghaheri et al. (2015) provided introductions for the application of GA in medicine.

3.2 Predictive models using the feature proteins.

The reasons for the five models employed here to validate the predictive performance of these proteins are: first, these models fit the research data well, i.e., the outcomes are binary; second, they are widely used for prediction, and they can help avoid false positive errors and ensure that the findings are robust.

For our dataset, the main advantages and disadvantages of the five models are as follows. For XGBoost, the benefits are:

1 the complexity of the tree model is added to the regular term

2 it introduces feature subsampling.

Both 1 and 2 can avoid overfitting; the drawback is that it is time-consuming when the data volume is large because it does pre-ranking of the features of the nodes before iterations and traverses to select the optimal split point. For BPNN, advantages are:

1 it achieves a nonlinear mapping of inputs and outputs so that it is very suitable for multi-dimensional feature construction

2 it can use a variety of different transfer functions that can be adapted to a variety of different data; the disadvantage is that the convergence speed is slow.

For SVM, the pros are:

1 it is very effective in solving classification and regression problems with high-dimensional features and still has good results when the feature dimension is larger than the number of samples

2 when the sample size is not massive, the classification accuracy is high, and the ability of generalisation is strong; the con is that it is computationally overloaded when the sample size is very large.

For RF, the advantages are:

1 the importance of each feature for the output can be given

2 due to the random sampling, the variance of the trained model is small, and the ability of generalisation is strong; the disadvantage is that it tends to fall into overfitting on certain sample sets with relatively large noise.

For LR, the values are:

1 it is computationally inexpensive and easy to understand and implement

2 it has good robustness to small data noise for it does not receive the effects of minor multi-collinearity; the weakness is it is prone to under-fitting.

Note that the descriptions of the strengths and weaknesses of these models are not to compare the quality of each other but to emphasise the practicality of these methods in this situation. Because the properties of these models are consistent with the properties of the raw dataset, they are suitable for analysis. The intention of applying these models here is because they can help judge whether the proteins are the features that can be represented for the state of depression, not focus on picking a model that surpasses others.

The assignment procedure was that 48 observations were randomly separated into training sets and test sets. The training set contains 70% of the subjects, and the test set contains 30%. For training sets, we used these nine feature proteins to explore the outcomes of interest through XGBoost, BPNN, SVM, RF, and LR models. Then, we evaluated the predictive performance through the test data. Here, we simulated this procedure 200 times for each model. AUC was used as the criterion to assess the performance of the model. We chose AUC as the criterion for two reasons. First, AUC is a universally used criterion to evaluate the predictive performance of biomarkers. Second, in this research, AUC is one of the most suitable criteria because the ability of a diagnostic biomarker to discriminate between subjects who develop the disease (cases) and subjects who do not (controls) is often measured by the area under the receiver operating characteristic curve (Rosner et al., 2015). Again, we acknowledge the dataset was not considerable. However, the collective process of the data for this study was challenging, and the time and effort behind this were expensive (because the subjects were pregnant women). For clinical and practical purposes, we conducted this study with this valuable data to provide clues to improve maternal mental health. Constructions and validations of all models were performed on the Rstudio platform (version 4.1.2).

4 Results

4.1 The selected feature proteins

A total of 3,282,329 secondary spectra were obtained through mass spectrometry analysis. Compared with the Universal Protein Database, 755,944 matched-spectrums were labelled, and the effective utilisation rate of the spectra was 23.03%. Altogether 13,338 peptide sequences were resolved from the matched-spectrums, and 11,701 were unique peptide sequences. 1,526 proteins were identified and analysed through unique peptide segments, and 1,387 proteins were quantified. Unique peptides were the only signals identified for a specific protein. Targeted identification of proteins based on unique peptides can significantly improve the accuracy. In this project, there were 11,701 Unique peptides, accounting for 87.73% of all identified peptides, indicating that the accuracy of the proteins identified this time is high (Table 2).

Table 2 Results from mass spectrometry test results.

Total spectrums	*Matched spectrums*	*Peptides*	*Unique peptides*	*Identified proteins*	*Quantifiable proteins*
3,282,329	755,944	13,338	11,701	1,526	1,387

Notes: Total spectrums – the number of secondary spectra generated by mass spectrometry detection. Matched spectrums – the number of peptide sequences parsed from the matching results. Peptides – the number of identified segments, i.e., the number of peptide sequences were analysed from the matching results. Unique peptides – number of unique peptides identified; the number of unique peptides identified for the related protein. Identified proteins – the number of identified proteins, the number of proteins analysed through specific peptide segments. Quantifiable proteins – the number of proteins quantified through unique peptide segments.

In the screening process of feature proteins, five algorithms were used. The features selected by them were slightly different (Table 3). We consider that different assumptions of different algorithms should be a reasonable explanation for the results in Table 3. Specifically, proteins selected by BA can be seen in Figure 2, in which the columns in green were ‘confirmed’, the columns in red were ‘refused’, and the ones in yellow indicated ‘tentative’. There were a couple of blue bars representing ShadowMax and ShadowMin. They were not actual features but were used by the Boruta algorithm to decide whether a variable is crucial. Features from RPA and RRF are shown in Figure 3 and Figure 4, respectively. It is clear which variables contributed to the outcome and how important they were. Figure 5 shows the number of features that should be included in the analysis. Lasso suggested that 7 to 10 features may represent the rest of them, so we picked eight features (Table 3). Figure 6 is the result of GA, which proved that internal fitness was desirable and external fitness was acceptable after 200 iterations (normally, 100 iterations is enough), and this is not the same as simulations for assignments. More explanations and details of these results can be referenced in the supplementary material and section ‘Algorithms for feature selection’. Eventually, we selected nine proteins that appeared at least twice in five algorithms as features. These methods have proven their performance, so it is reasonable to view these nine proteins as features this way practically and theoretically.

Table 3 The details of the proteins selected by five algorithms

Algorithm	*Features*												
Boruta	J3QRN2	P558	O15144	Q14676	P48637				AA87WSY6	P742	Q1518		
Recursive partition	J3QRN2	P558	O15144	Q14676	P48637	Q99784	AA3B3IRN5	B7ZKJ8					AA87X89
Regularised random forest		P558	O15144		P48637				AA87WSY6	Q8NBJ4	P27918	Q9UHG3	
LASSO	J3QRN2	P558	O15144		P48637	Q99784	AA3B3IRN5	B7ZKJ8	AA87WSY6				
Genetic	J3QRN3	P558		Q14676	P48637					D64D58			
Final decision	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N	N	N

Notes: The decision in the table means whether the protein is selected. Y indicates the protein is selected, and N the protein is not selected.

Figure 2 Proteins selected by BA (see online version for colours)

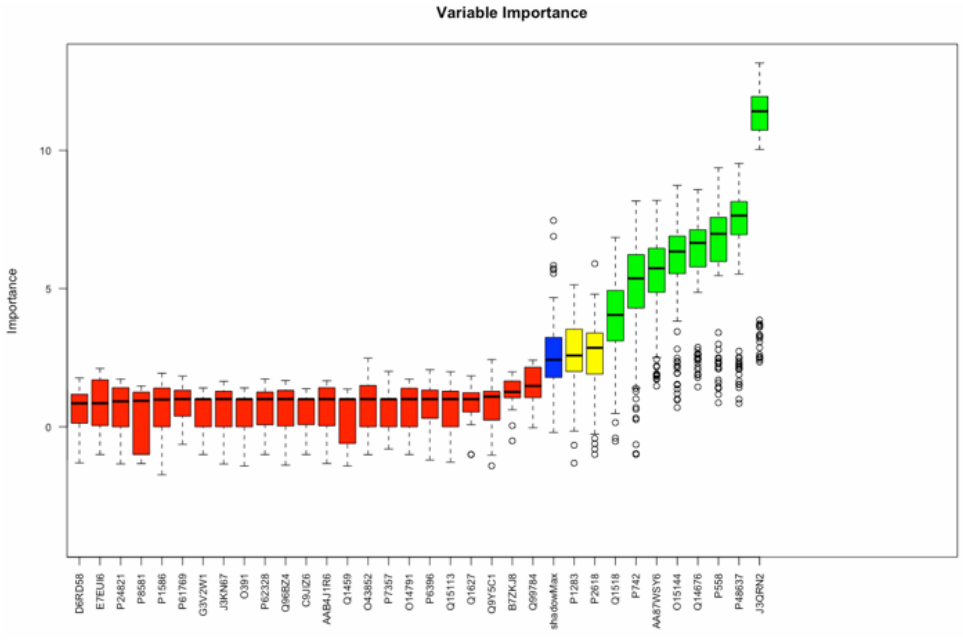
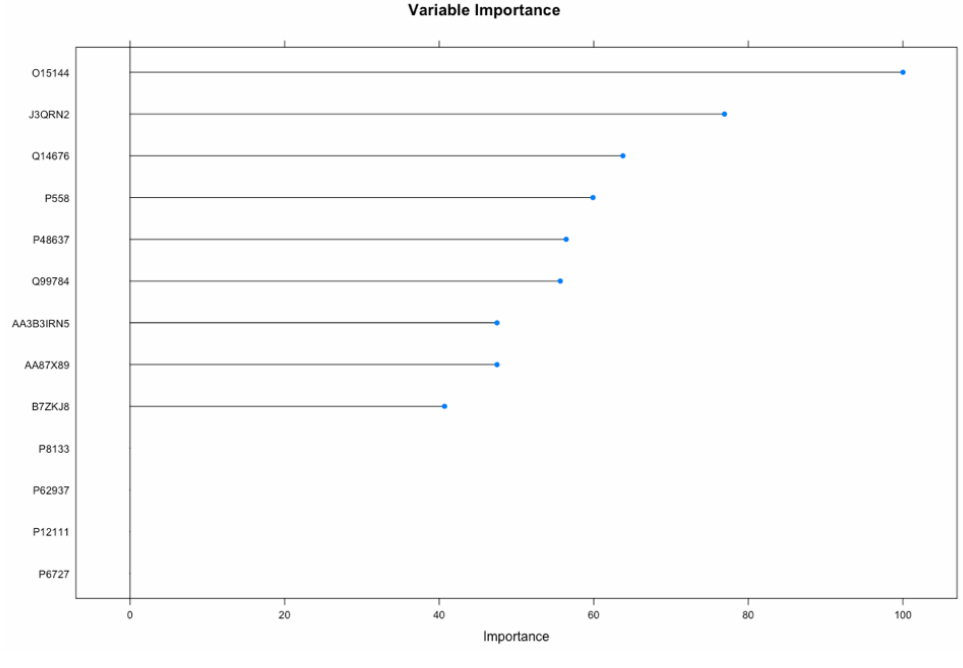


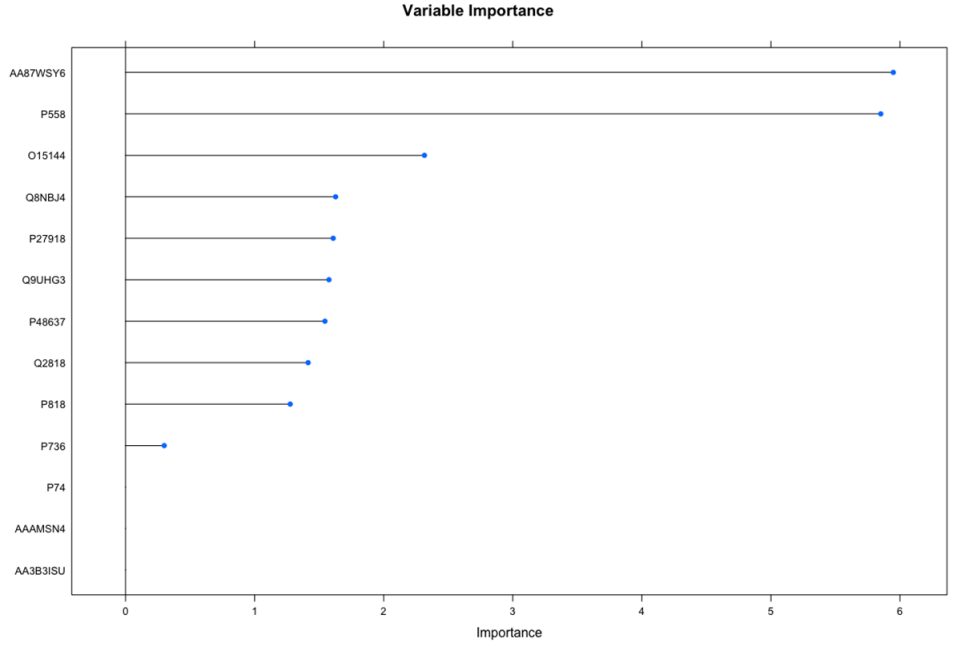
Figure 3 Proteins selected by RPA (see online version for colours)



We also explored the distribution of contents of these features in depressed and non-depressed pregnant women (Figure 7). Contents of proteins 'Q14676',

'AA3B3IRN5', and 'O15144' in depressed pregnant women were higher than those in non-depressed. While contents of proteins 'Q99784', 'P558', 'B7ZKJ8', 'P48637', 'AA87WSY6', and 'J3QRN2' were lower in depressed pregnant women than those non-depressed. We further estimated the Spearman correlation coefficient between features and outcomes and conducted hypothesis tests to identify them (Figure 7 and Table 4). Results showed that proteins such as 'Q14676' and 'O15144' had a significantly positive correlation with the outcome, and others such as 'B7ZKJ8' and 'P48637' had a significant negative correlation with depression.

Figure 4 Proteins selected by RRF (see online version for colours)



To gain an overall picture of the predictability of the features for depression, we built a univariate model for them separately and then validated their performance (Figure 8). We found that the predictive performance of 'Q14676', 'J3QRN2', 'AA3B3IRN5', and 'P558' were better compared to other proteins. The result is slightly different from the above, which we consider high correlations do not always mean strong predictability as a cause, especially when the number of trials is small. To solve this problem, we conducted 200 simulations, as detailed in the next section.

4.2 Validation for feature proteins

At this stage, we built predictive models for the determined proteins and ran 200 simulations to achieve a robust conclusion. As mentioned above, the training set contained 70% of the subjects, and the test set 30%. We randomly assign observations to both the training and test sets. Because of randomisation, mothers allocated to the training set (test set) were different each time, which means the predictive results were

also affected. To obtain a compelling conclusion, we conducted the procedure with 200 simulations.

Figure 5 Numbers of proteins suggested by LASSO (see online version for colours)

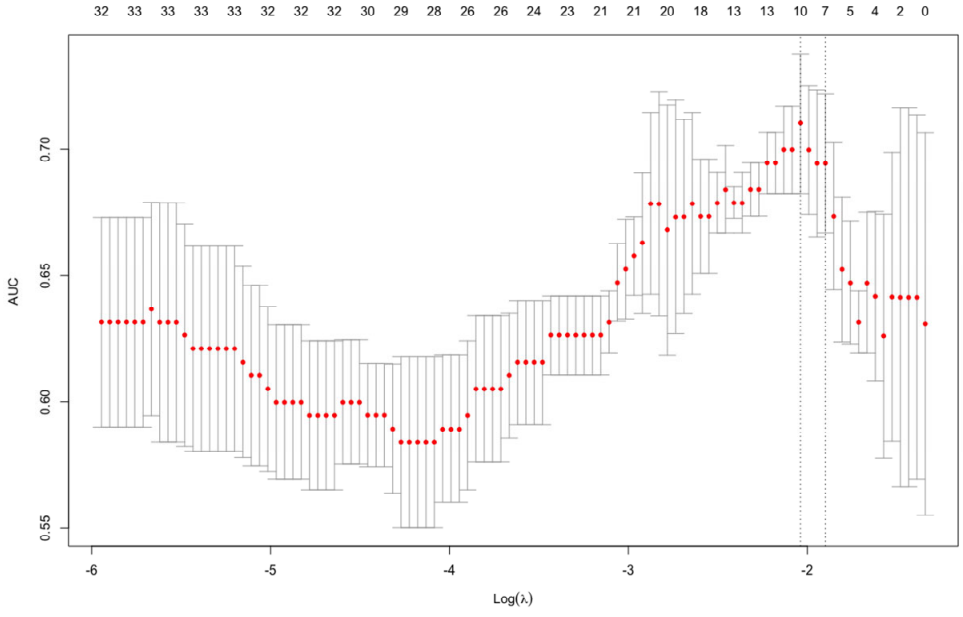
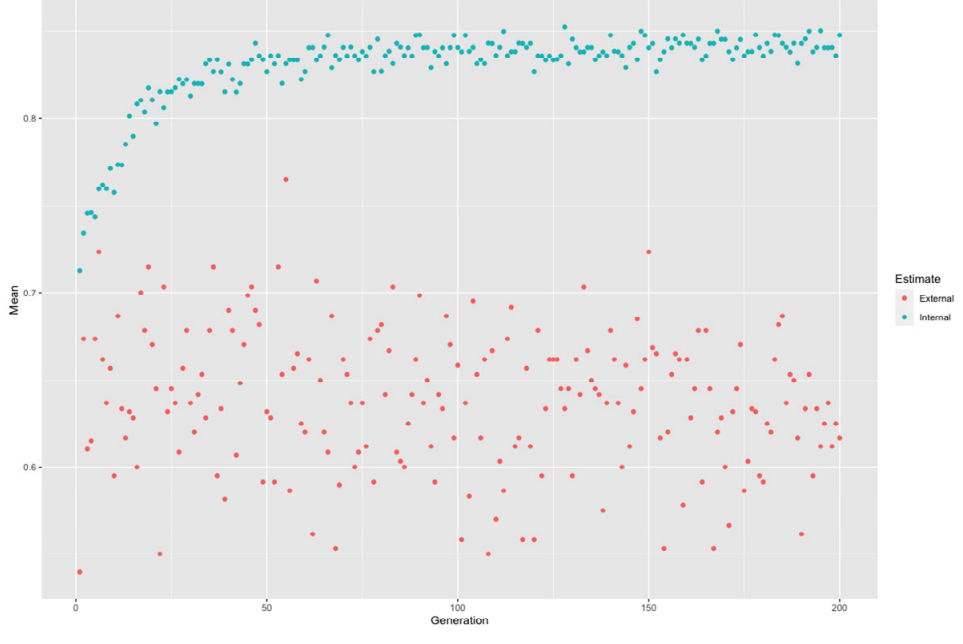


Figure 6 The result of iterations by GA (see online version for colours)

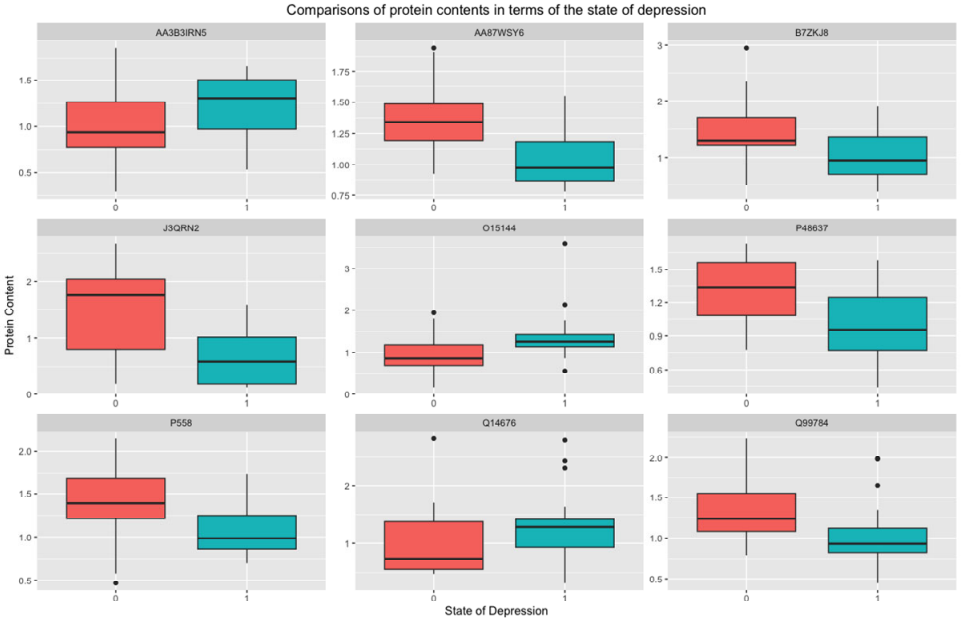


The simulation steps were as follows. First, randomly assigned subjects to the training set and test set. Second, modelling the state of depression and features using training sets. Third, validate the predictive performance of markers in test sets. Finally, set AUC as the criterion to appraise the performance of proteins (Figure 9). Here, we developed a function (codes were attached in the supplementary materials) that determines the number of random assignments according to the factual situation. This is key to the simulation because it hugely enhances efficiency. At the same time, it also provides a basis for achieving forceful results. All conclusions in this section were based on 200 simulations.

Table 4 Spearman correlation coefficients and P-values between proteins and maternal depression

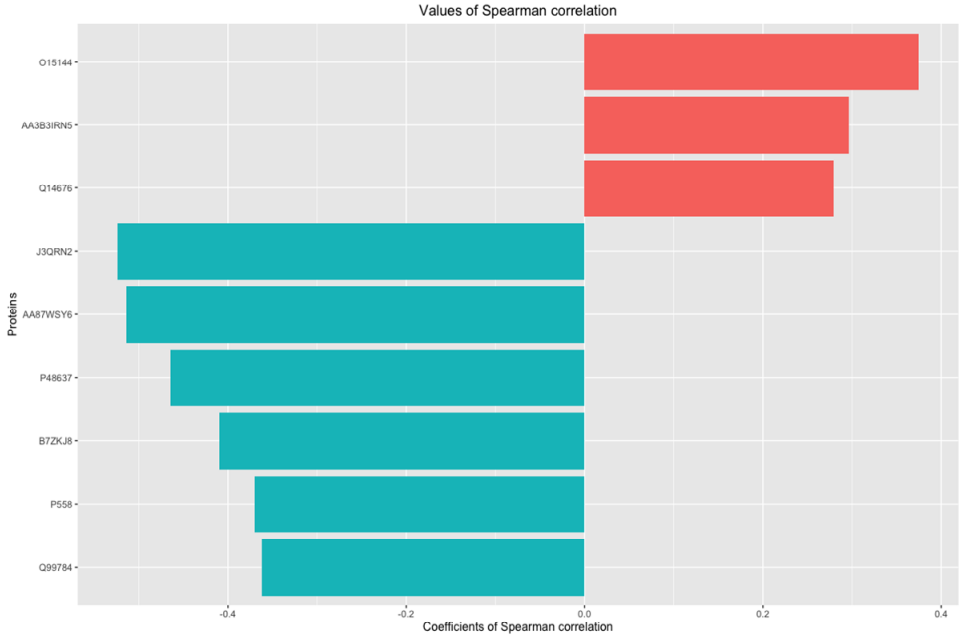
Protein	*Spearmean_corr.coeff*	*p.value*
P558	–0.3705	0.0095
P48637	–0.4648	0.0009
Q14676	0.2790	0.0408
AA87WSY6	–0.5138	0.0002
O15144	0.3744	0.0087
J3QRN2	–0.5236	0.0001
Q99784	–0.3618	0.0115
B7ZKJ8	–0.4095	0.0038
AA3B3IRN5	0.2966	0.0406

Figure 7 (a) Box plot of nine marker proteins in depressed and non-depressed groups (b) Spearman correlation coefficient between nine marker proteins and maternal depression (see online version for colours)



(a)

Figure 7 (a) Box plot of nine marker proteins in depressed and non-depressed groups (b) Spearman correlation coefficient between nine marker proteins and maternal depression (continued) (see online version for colours)



(b)

Figure 8 Performance of predictability of proteins (see online version for colours)

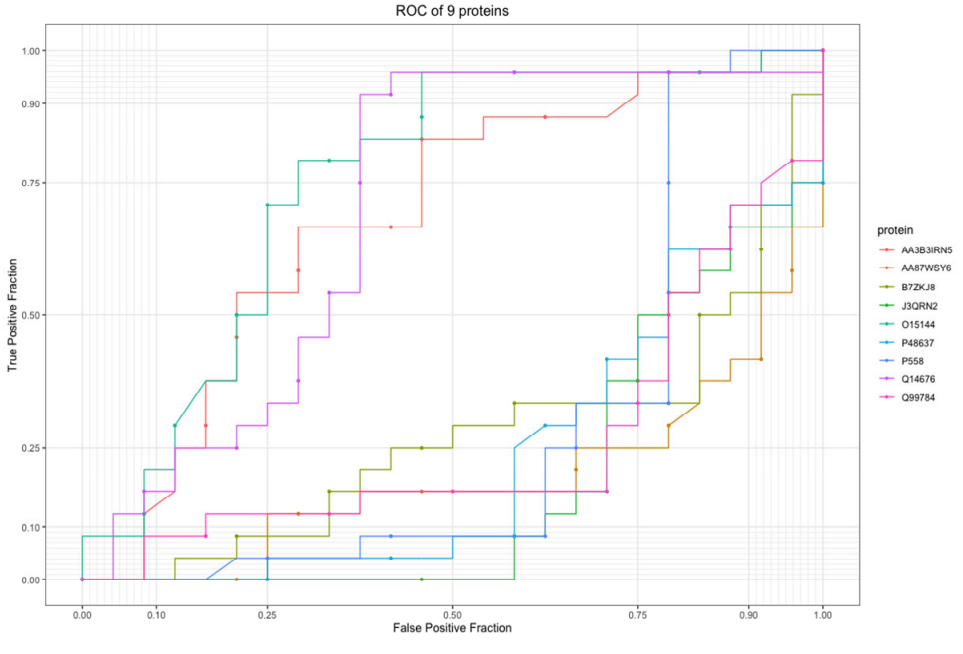
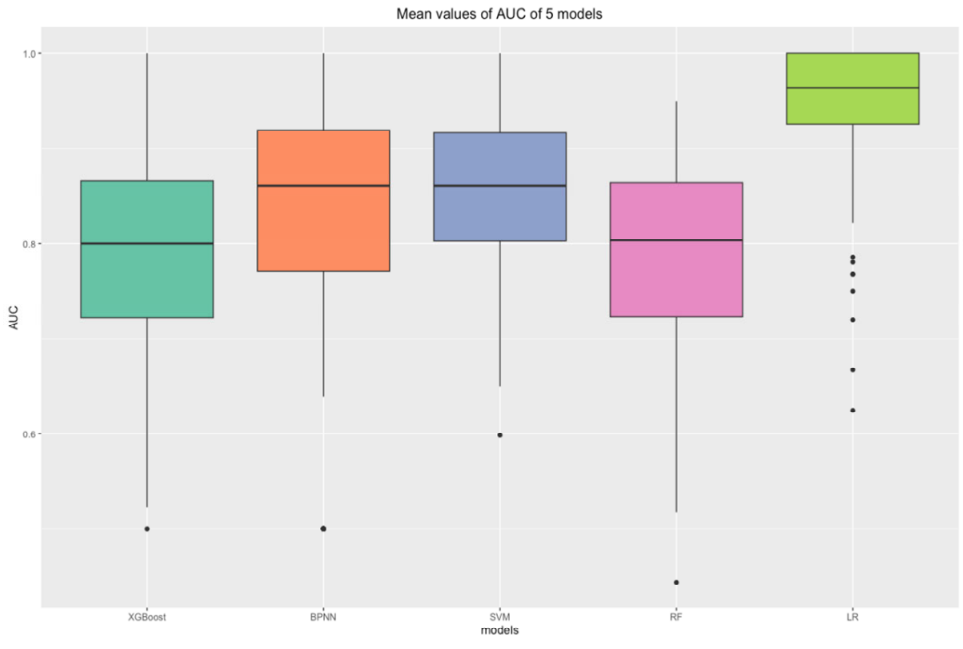


Figure 9 Performance of prediction of proteins (see online version for colours)



Note: XGBoost – extreme gradient boosting, BPNN – back propagation neural network, SVM – support vector machine, RF – random forest, LR – logistic regression.

In these simulations, all models achieved a relatively good predictive performance, and the mean values of the AUC of all models exceeded 80%. Precisely, LR believed that the average AUC of these features for predicting the state of depression was almost 100%. The results of BPNN and SVM were close, and both deemed that these proteins can accurately predict the state of depression with a mean AUC close to 90%, but SVM was less conservative than the BPNN model. XGBoost and RF provided a parallel suggestion, i.e., they indicated that the mean AUC exceeded 80%. Furthermore, if the purpose is not to make a clinical diagnosis in practice but an exploratory study of feature selection, an AUC of 80% and above is highly valuable for research.

It can be seen from the results that the predictive performance of these selected proteins is satisfactory. One thing that needs to be stressed is the purpose of comparing various models is not to show the superiority of one approach to another but to understand how these approaches work better in identifying and evaluating features. Also, we aim to provide hints for researchers on different choices so that doctors and researchers can decide on the best-matched models according to their own needs in practice.

5 Discussion and conclusions

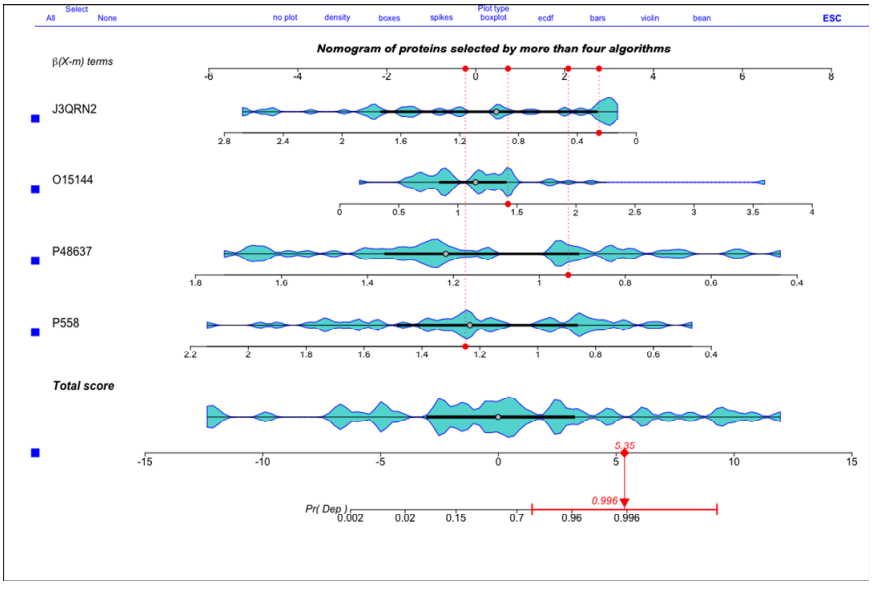
The state of health in pregnant women is getting more and more attention because it is not only related to the quality of life of a pregnant woman but also involved in the well-being of a family (Degirmenci and Yilmaz, 2020; Francis et al., 2021; Lagadec

et al., 2018). As far as the mental health of pregnant women is concerned, there are many articles on the related factors of maternal depression (Rastad et al., 2021; Nelson et al., 2018). However, as of the current research, few published papers centre on maternal depression and proteomics, so we have reasons to believe that it is meaningful to study maternal depression from a proteomics perspective.

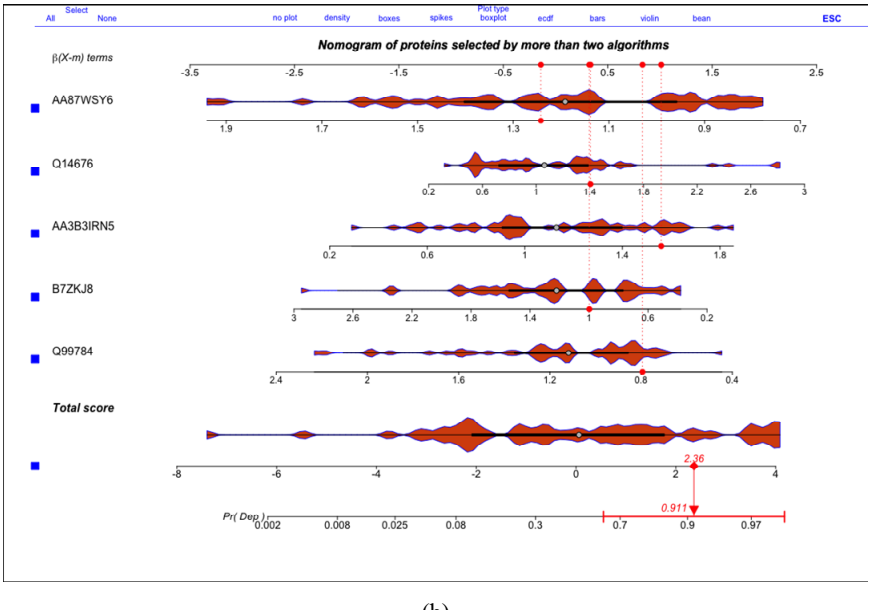
Maybe the analytical methods used here are not up-to-the-minute, as we aim to use reliable and widely accepted methods for proteomic studies of maternal depression. From the review of the relevant literature, the algorithms we chose have been proven to perform well in practice and still play a vital role in research (Liu et al., 2021; Hill et al., 2003; Hindson, 2022; Li et al., 2021). Detailly speaking, Reyaz-Ahmed et al. (2010) attempted to solve the problem of protein model assessment using SVM and claimed that the results from SVM were better than other machine learning techniques. Langlois et al. (2005) presented an SVM-based method for recognising a protein's fold from sequence information alone, which showed better prediction accuracy. Bardsiri and Eftekhari (2014) compared an approach based on the decision tree and suggested that the GA weighting fusion method achieved the best performance. Sumathi and Padmavathi (2019) made comparisons of Boruta, Enet, GA and consistency-based subset feature selections using cancer datasets, and they pointed out that Boruta is a ranking and feature selection algorithm used to identify the importance of variables in prediction. Chen et al. (2019) reported an intelligent prediction approach based on the XGBoost model and big data, and their results showed the accuracy of this model had a better performance than the BP neural network model. Pacheco et al. (2023) evaluated the performance of different models for predicting three types of fraudulent behaviour in a novel dataset with imbalanced data, in which they provided a thorough description of LR and RF with a conclusion that those models shared a similar performance. To better understand and visualise the predictive performance of the features, we offered the nomogram (Figure 10). We separated proteins into two parts mainly for readability, and details on the nomogram can be seen in the Supplementary materials and Graesslin et al. (2010) and Lo et al. (2020). Here, the nomogram shows the LR model and is dynamic since one can decide the state of depression by setting the precise values of features. Also, the nomogram of all proteins is given (Figure 11). Nomograms of other models also can be seen in the Supplementary materials.

We investigated maternal depression from the perspective of proteomics, which we believe is a novel angle of thinking to explore the influencing factors of depression at the molecular level. We found that the biological signalling pathways in which these proteins are mainly involved include leukocyte activation, participation in or mediating immune responses, lipid binding and other signalling pathways. Results from metabolic pathway enrichment further indicated that these proteins are widely involved in glycolysis or gluconeogenesis metabolism-related pathways such as metabolism, amino acid metabolism, and fatty acid degradation. This suggests that metabolic abnormalities in pregnant women may be an essential factor in the occurrence of the outcome. The possibility is that these proteins are jointly involved in infection, immune responses, complex binding, lipid metabolism, and energy metabolism. Changes in levels of the peripheral blood of pregnant women may lead to an imbalance of signalling pathways, alter the physiological environment of the body, and eventually trigger the outcome.

Figure 10 Nomograms to predict the probability of the state of depression (i.e., Pr (Dep)), (a) features selected by all algorithms (b) by more than two (see online version for colours)



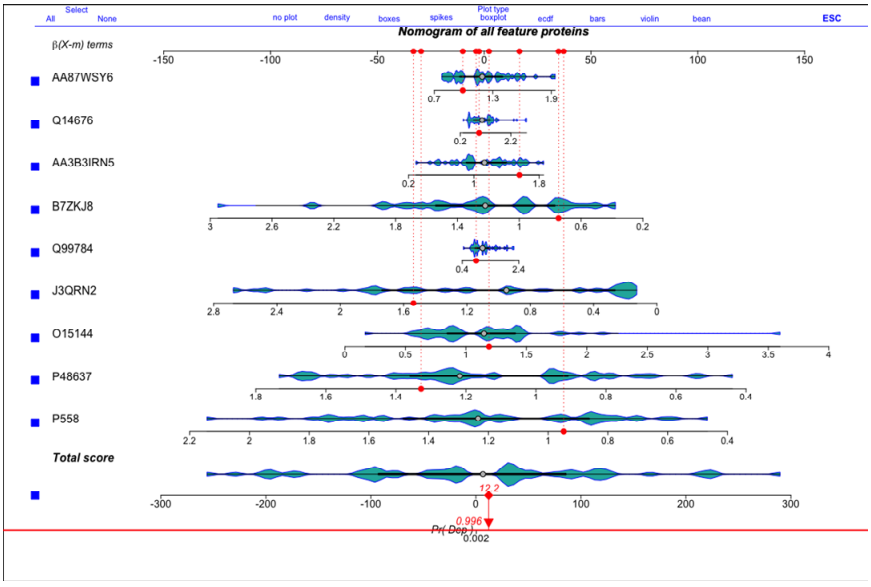
(a)



(b)

Note: This nomogram only shows the predictive performance of LR using features.

Figure 11 Nomogram to predict the probability of the state of depression using all features (see online version for colours)



Note: This nomogram only shows the predictive performance of LR using features.

Several limitations exist in this study. For example, we did not use all feature selection methods for protein selection and did not construct all predictive models to verify the performance. In addition, due to uncontrollable factors in the sampling process, it was unavoidable to delete proteins with missing data, and there may also be proteins with predictability among these censored proteins.

We identified features associated with maternal depression, of which nine were the most significant. We filtered the initial 689 proteins by five popular and well-performed algorithms. The proteins selected by each algorithm were not identical, which we believe the possibility is the different assumptions of algorithms. Proteins that were finally identified perform well in prediction. All features were selected by two algorithms at least, and some by all. In addition, we also designed a questionnaire to better understand the psychological state of pregnant women in the first trimester. The contents of the questionnaire have four parts: the first is the basic demographic information of pregnant women; the second is the information related to pregnancy (i.e., the history of previous pregnancy and response status of this pregnancy); the third is the information about nutrition (i.e., daily diet and sleep status); and the last includes investigations belonging to the field of psychology. Part contents of the questionnaire can be seen in the supplementary material. We believe that by combining the questionnaire information and the results of proteomics analysis, the state of depression in pregnant women can be diagnosed more accurately.

In summary, we provided information about the feature proteins that can predict the state of maternal depression and verify their predictive performance. The results show that these proteins achieve impressive predictability. It is sensible to believe that they can predict the state of depression of a pregnant woman. We drew meaningful conclusions that can provide innovative ideas for improving maternal depression as soon as possible

by analysing this valuable dataset. More importantly, we wish this work could provide clues for further research.

Supplementary materials

The supplementary materials include the codes for feature selection, performance validation, simulations, and the questionnaire, which can be seen online (https://github.com/vitaminc121/Data-and-Simulation).

Availability of data and materials

The datasets used and analysed during the current study are available from the corresponding author upon reasonable request.

Acknowledgements

We thank all mothers for consenting to participate in this study and providing their information. Also, we thank all the reviewers who helped us improve our work.

Yuhao Feng and Jinman Zhang contributed equally to this work.

References

Aparicio, M., Browne, P.D., Hechler, C., Beijers, R., Rodriguez, J.M., De Weerth, C. and Fernandez, L. (2020) 'Human milk cortisol and immune factors over the first three postnatal months, relations to maternal psychosocial distress', *PLoS One*, 21 May, Vol. 15, No. 5, p.e0233554.

Bardsiri, M.K. and Eftekhari, M. (2014) 'Comparing ensemble learning methods based on decision tree classifiers for protein fold recognition', *International Journal of Data Mining and Bioinformatics*, Vol. 9, No. 1, pp.89–105.

Brann, E., Papadopoulos, F., Fransson, E., White, R., Edvinsson, A., Hellgren, C., Kamali-Moghaddam, M., Bostrom, A., Schioth, H.B., Sundstrom-Poromaa, I. and Skalkidou, A. (2017) 'Inflammatory markers in late pregnancy in association with postpartum depression – a nested case-control study', *Psychoneuroendocrinology*, Vol. 79, pp.146–159, DOI: 10.1016/j.psyneuen.2017.02.029.

Chattopadhyay, S. (2013) 'Mathematical modelling of doctors' perceptions in the diagnosis of depression, a novel approach', *International Journal of Biomedical Engineering and Technology*, Vol. 11, No. 1, pp.1–17.

Chen, J.X., Zhao, F., Sun, Y.G. et al. (2019) 'Prediction model based on XGBoost for mechanical properties of steel materials', *International Journal of Modelling, Identification and Control*, Vol. 33, No. 4, pp.322–330.

Degirmenci, F. and Yilmaz, D.V. (2020) 'The relationship between psychosocial health status and social support of pregnant women', *J. Psychosom. Obstet. Gynaecol.*, Vol. 41, No. 4, pp.290–297, DOI: 10.1080/0167482X.2019.1678021.

Edvinsson, A., Hellgren, C., Kunovac Kallak, T., Akerud, H., Skalkidou, A., Stener-Victorin, E., Fornes, R., Spigset, O., Lager, S., Olivier, J. and Sundstrom-Poromaa, I. (2019) 'The effect of antenatal depression and antidepressant treatment on placental tissue, a protein-validated gene expression study', *BMC Pregnancy Childbirth*, 5 December, Vol. 19, No. 1, p.479, DOI: 10.1186/s12884-019-2586-y.

Francis, E.C., Zhang, L., Witrick, B. and Chen, L. (2021) 'Health behaviors of American pregnant women, a cross-sectional analysis of NHANES 2007–2014', *J. Public Health (Oxf)*, Vol. 43, No. 1, pp.131–138, DOI: 10.1093/pubmed/fdz117..

Friedman, L.E., Gelaye, B., Sanchez, S.E. and Williams, M.A. (2020) 'Association of social support and antepartum depression among pregnant women', *J. Affect Disord.*, Vol. 264, pp.201–205, DOI: 10.1016/j.jad.2019.12.017.

Ghaheri, A., Shoar, S., Naderan, M. and Hoseini, S.S. (2015) 'The applications of genetic algorithms in medicine', *Oman Med. J.*, Vol. 30, No. 6, pp.406–416, DOI: 10.5001/omj.2015.82.

Graesslin, O., Abdulkarim, B.S., Coutant, C., Huguet, F., Gabos, Z., Hsu, L., Marpeau, O., Uzan, S., Pusztai, L., Strom, E.A., Hortobagyi, G.N., Rouzier, R. and Ibrahim, N.K. (2010) 'Nomogram to predict subsequent brain metastasis in patients with metastatic breast cancer', *J. Clin. Oncol.*, Vol. 28, No. 12, pp.2032–2037, DOI: 10.1200/JCO.2009.24.6314.

Haddow, C., Perry, J. and Durrantm, F.J. (2011) 'Predicting functional residues of protein sequence alignments as a feature selection task', *Int. J. Data Min. Bioinform.*, Vol. 5, No. 6, pp.691–705, DOI: 10.1504/ijdmb.2011.045417.

Hadzic, M., Hadzic, F. and Dillon, T.S. (2010) 'Mining of patient data, towards better treatment strategies for depression', *International Journal of Functional Informatics and Personalised Medicine*, Vol. 3, No. 2, pp.122–143.

Hai, B., Song, Q., Du, C., Mao, T., Jia, F., Liu, Y., Pan, X., Zhu, B. and Liu, X. (2022) 'Comprehensive bioinformatics analyses reveal immune genes responsible for altered immune microenvironment in intervertebral disc degeneration', *Mol. Genet. Genomics*, Vol. 297, No. 5, pp.1229–1242, DOI: 10.1007/s00438-022-01912-3.

Hill, J.L., Brooks-Gunn, J. and Waldfogel, J. (2003) 'Sustained effects of high participation in an early intervention for low-birth-weight premature infants', *Dev. Psychol*., Vol. 39, No. 4, pp.730–744, DOI: 10.1037/0012-1649.39.4.730.

Hindson, J. (2022) 'Proteomics and machine-learning models for alcohol-related liver disease biomarkers', *Nat. Rev. Gastroenterol. Hepatol.*, Vol. 19, No.8, p.488, DOI: 10.1038/s41575-022-00655-1.

Kursa, M.B. and Rudnicki, R. (2010) 'Feature selection with the Boruta package', *Journal of Statistical Software*, Vol. 36, pp.1–13, DOI: 10.18637/JSS.V036.I11.

Lagadec, N., Steinecker, M., Kapassi, A., Magnier, A.M., Chastang, J., Robert, S., Gaouaou, N. and Ibanez, G. (2018) 'Factors influencing the quality of life of pregnant women, a systematic review', *BMC Pregnancy Childbirth*, Vol. 18, No. 1, p.455, Published 23 November, DOI: 10.1186/s12884-018-2087-4.

Langlois, R.E., Diec, A., Perisic, O. et al. (2005) 'Improved protein fold assignment using support vector machines', *International Journal of Bioinformatics Research and Applications*, Vol. 1, No. 3, pp.319–334.

Li, J., Zhou, K. and Mu, B. (2021) 'Machine learning for mass spectrometry data analysis in proteomics', *Current Proteomics*, Vol. 18, No. 5, pp.620–634.

Liu, S., Dissanayake, S., Patel, S. et al. (2014) 'Learning accurate and interpretable models based on regularized random forests regression', *BMC Syst. Biol.*, Vol. 8, No. Suppl 3, p.S5, DOI: 10.1186/1752-0509-8-S3-S5.

Liu, Y., Bai, F., Tang, Z., Liu, N. and Liu, Q. (2021) 'Integrative transcriptomic, proteomic, and machine learning approach to identifying feature genes of atrial fibrillation using atrial samples from patients with valvular heart disease', *BMC Cardiovasc. Disord.*, 28 January, Vol. 21, No. 1, p.52, DOI: 10.1186/s12872-020-01819-0.

Lo, S.N., Ma, J., Scolyer, R.A. et al. (2020) 'Improved risk prediction calculator for sentinel node positivity in patients with melanoma: the Melanoma Institute Australia Nomogram', *J. Clin. Oncol.*, Vol. 38, No. 24, pp.2719–2727, DOI: 10.1200/JCO.19.02362.

Mao, Q., Tian, T., Chen, J., Guo, X., Zhang, X. and Zou, T. (2021) 'Serum metabolic profiling of late-pregnant women with antenatal depressive symptoms', *Front Psychiatry*, 8 July, Vol. 12, p.679451, DOI: 10.3389/fpsyt.2021.679451.

Nelson, C., Lawford, K.M., Otterman, V. and Darling, E.K. (2018) 'Mental health indicators among pregnant aboriginal women in Canada, results from the Maternity Experiences Survey' [Indicateurs de santé mentale chez les femmes autochtones enceintes au Canada , résultats de l'Enquête sur l'expérience de la maternité], *Health Promot. Chronic Dis. Prev. Can.*, Vol. 38, Nos. 7–8, pp.269–276, DOI: 10.24095/hpcdp.38.7/8.01.

Pacheco, J., Chela, J. and Salomé, G. (2023) 'Fraud detection with machine learning, model comparison', *International Journal of Business Intelligence and Data Mining*, Vol. 22, No. 4, pp.434–450.

Pak, J., Kim, H.S. and Rhee, E.S. (2020) 'Characterising social structural and linguistic behaviours of subgroup interactions, a case of online health communities for postpartum depression on Facebook', *International Journal of Web Based Communities*, Vol. 16, No. 3, pp.225–248.

Rajaguru, H. and Chakravarthy, S.R.C. (2019) 'Analysis of decision tree and K-nearest neighbor algorithm in the classification of breast cancer', *Asian Pac. J. Cancer Prev.*, 1 December, Vol. 20, No. 12, pp.3777–3781, DOI: 10.31557/APJCP.2019.20.12.3777.

Rastad, Z., Golmohammadian, M., Jalali, A., Kaboudi, B. and Kaboudi, M. (2021) 'Effects of positive psychology interventions on happiness in women with unintended pregnancy, randomized controlled trial', *Heliyon*, 17 August, Vol. 7, No. 8, p.e07789.

Redei, E.E., Ciolino, J.D., Wert, S.L., Yang, A., Kim, S., Clark, C., Zumpf, K.B. and Wisner, K.L. (2021) 'Pilot validation of blood-based biomarkers during pregnancy and postpartum in women with prior or current depression', *Transl. Psychiatry*, Vol. 11, No. 1, p.68.

Redinger, S., Pearson, R.M., Houle, B., Norris, S.A. and Rochat, T.J. (2020) 'Antenatal depression and anxiety across pregnancy in urban South Africa', *J. Affect Disord.*, Vol. 277, pp.296–305, DOI: 10.1016/j.jad.2020.08.010.

Remeseiro, B. and Bolon-Canedo, V. (2019) 'A review of feature selection methods in medical applications', *Comput. Biol. Med.*, Vol. 112, p.103375, DOI: 10.1016/j.compbiomed. 2019.103375.

Reyaz-Ahmed, A., Harrison, R. and Zhang, Y.Q. (2010) 'Protein model assessment via machine learning techniques', *International Journal of Functional Informatics and Personalised Medicine*, Vol. 3, No. 3, pp.215–227.

Rosner, B., Tworoger, S. and Qiu, W. (2015) 'Correcting AUC for measurement error', *J. Biom. Biostat.*, Vol. 6, No.5, p.270, DOI: 10.4172/2155-6180.1000270.

Sadat, Z., Abedzadeh-Kalahroudi, M., Kafaei Atrian, M., Karimian, Z. and Sooki, Z. (2014) 'The impact of postpartum depression on quality of life in women after child's birth', *Iran Red. Crescent Med. J.*, Vol. 16, No. 2, p.e14995, DOI: 10.5812/ircmj.14995.

Sumathi, C.P. and Padmavathi, M.S. (2019) 'An experimental approach of applying Boruta and elastic net for variable selection in classifying breast cancer datasets', *International Journal of Knowledge Engineering and Data Mining*, Vol. 6, No. 4, pp.356–375.

Zhang, Y., Muyiduli, X., Wang, S., Jiang, W., Wu, J., Li, M., Mo, M., Jiang, S., Wang, Z., Shao, B., Shen, Y. and Yu, Y. (2018) 'Prevalence and relevant factors of anxiety and depression among pregnant women in a cohort study from south-east China', *J. Reprod. Infant Psychol.*, Vol. 36, No. 5, pp.519–529.

Zhao, Y., Kane, I., Mao, L., Shi, S., Wang, J., Lin, Q. and Luo, J. (2016) 'The prevalence of antenatal depression and its related factors in chinese pregnant women who present with obstetrical complications', *Arch. Psychiatr. Nurs.*, Vol. 30, No. 3, pp.316–321.

Zhao, Y., Munro-Kramer, M.L., Shi, S., Wang, J. and Zhu, X. (2018) 'A longitudinal study of perinatal depression among Chinese high-risk pregnant women', *Women Birth*, Vol. 31, No. 6, pp.e395–e402, DOI: 10.1016/j.wombi.2018.01.001.