Methylation of inflammatory and stress related genes association with maternal depression and premature birth

1st Amisha Jindal 2nd Florina Asani 3rd Jocelyn Petitto 4th Sanika Patki

Abstract— **In this project we aim to analyze the association of methylation of inflammatory related genes, TNFα and FOXP3, the stress related genes, BDNF, FKBP5, Nr3c1, with maternal depression and how this affects premature birth. We investigate this association using Random Forest Trees and Gradient Boosting Algorithms by looking into the underlying features that are ranked as highly important in building these models.**

Keywords—epigenetics, depression, random forest, gradient boosting

I. Introduction

Physical health, stress, and mental health are intrinsically linked. As the acceptance of these interwoven relationships grow, the research into their manifestation within specific populations has increased. For example, the studies have shown maternal depression leads to higher risk of preterm birth, which is further compounded by factors such as socioeconomic status [1]. The effect of stressors on Latina women is an under researched domain whose needs are of particular importance as they are uniquely stressed because of immigration status, racism, and sexism. Their well-being is of greater public health importance than ever as they are also the population with the largest rise in birthrate and slowest decrease in premature births [2].

HPA axis dysregulation has been shown to be independently associated with both adverse life events as well as perinatal depression [3]. The Latina population is three to four times as likely to develop perinatal depression compared to the general population [4]. Increased expression of pro-inflammatory biomarkers such as TNFα are associated with preterm births [5], [6].

One mode of regulating gene expression is methylation, which can decrease or halts the transcription of a gene [7]. DNA methylation is measured as it occurs at a specific site, a single nucleotide, and has a binary value. Generally, multiple sites near the promotor region of a gene are methylated to turn expression of the gene off. Measuring the methylation at each site is done over an entire sample and is given as an average rather than a binary value.

Multiple studies have given evidence that OXTR methylation levels, which control the expression of the oxytocin receptor, are associated with a plethora of social behaviors and psychopathologies [8]. Psychological stress has been correlated with an increase in inflammatory markers such as TNFα in the bloodstream. [9] TNF-α and FOXP3 are cytokines associated with inflammatory response regulation. Demethylation of TNFα increases the expression of this gene. [10], [11] Expression of TNFα has been associated with an increased incidence of premature birth [6] while FOXP3 expression has been studied as a predictor of spontaneous abortion. [12]

This paper will specifically focus on analyzing how epigenetic markers and acculturation as well as acculturative stress affect and relate with depression and discrimination. It also analyzes how epigenetics markers and psychopathological conditions are associated with premature birth.

Determining cutoff points for the self-reported psychological surveys was difficult as multiple options existed [13]–[15]. This makes the use of any regression-based method difficult to interpret and adds measurement bias to the use of clustering algorithms. Additionally, the response bias that is inherent to self-reported data may be compounded by concerns unique to this cohort, specifically that their responses would affect their legal status in this country. Previous studies using this dataset adjusted the cutoff points in an attempt to account for this potential confounder [16].

Other challenges we are facing when addressing this problem stem from the sample size and distribution of the target within the population.

By completing this study, we hope to establish ways to look at small datasets often found in medical research using AI and machine learning methodologies. With respect to the results of this particular study, it is our goal to provide results that enable the planning of interventions in at risk communities. More specifically, that this will catalyze groups to address the needs of expecting Latina mothers by addressing their unique needs and promote their mental wellbeing.

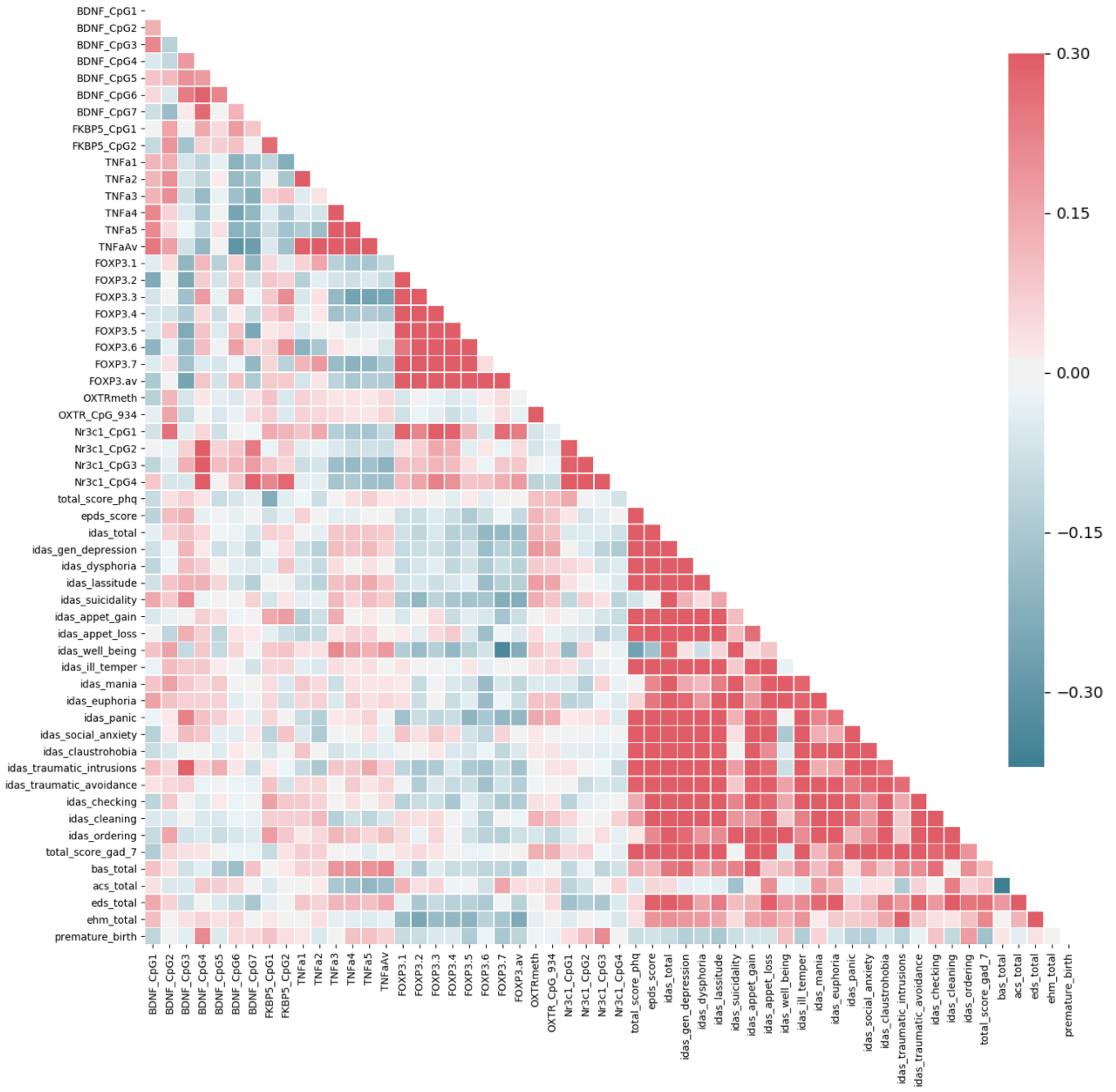

Fig. 1 Heatmap of Pearson correlation coefficients describing the relationships between combinations of epigenetic markers, psychopathological test scores, and premature birth. The darker regions show a higher correlation between individual CpG sites associated with inflammatory genes (*TNFα*, *FOXP3*) than with stress related genes (*BDNF*, *FKBP5*, *N3C1*).

II. Background

A. Machine Learning

Previous methods of analyzing this specific data looked at one-to-one correlations between methylated genes and survey scores. To better understand and model patterns, specifically using epigenetic data, we employed different machine learning techniques.

The machine learning methods that we decided to use are Gradient Boosting and Random Forest Trees. Gradient Boosting and Random Forest Trees are both ensemble learning methods that build by combining outputs from individual trees. In Gradient Boosting the trees are built sequentially, as each tree learns from the previous one. They are highly effective when there is imbalanced data. Random Forest Trees trains each tree independently with random samples of data, which makes it less prone to overfitting and is highly applicable for use in bioinformatics, such as for medical diagnosis.

B. Epigenetics – DNA Methylation

DNA methylation in the dataset was reported as the percent methylation at the site over the sample [17]. Methylation data has inherent disadvantages: at the micro level, it is a binary variable, but it is distributed unequally within a single subject, so it is not collected as a binary variable. The analysis in this paper are done using the percent methylation, often referred to as the β value.

C. Paper Organization

Our paper is organized in 5 sections. In Introduction we give an overview of the current state of the area of research that we address and its importance. We explain the importance of looking into the association between epigenetic markers as well as psychopathological conditions and the challenges and benefits that come with it.

In the Background section, we introduce the current used approaches to address such problems and the importance of using a machine learning model to overcome the previous disadvantages. We then explain the Methods we used to understand the underlying patterns and extract meaningful information from our models with respect to discrimination, depression and premature birth.

And in Results we explain our findings as well as compare different approaches used. And lastly, conclusion includes discussions upon obtained results and our future work.

III. Methodology

A. *Description of dataset*

The dataset for this study is the combination of the dataset from a 2018 study on discrimination and stress related gene methylation in Latina mothers [16] and an additional dataset from the same researchers collected from the same cohort of study participants with Hispanic background. The first dataset consists of 151 patients and 714 features, and the second dataset consists of 152 records and 727 features.

We merged the two datasets, then narrowed the new dataset down to a subset of variables that were appropriate to our assumptions and the associations that were being considered: epigenetic biomarkers, and the psychological related features with the associated scores such as the IDAS total score and subset scores, like general depression. The scores were calculated given self-reported survey administered to the patients and scored by the initial researchers.

B. *Initial Data Exploration*

Our approach to explore the presented hypothesis is to use classification and regression models to extract the underlying patterns and important information that leads the models to more accurate decisions. As a result, we will be better able to understand the relationships between epigenetics markers and psychological states. Based on the predicted due date and the actual delivery date, we calculated for premature birth and added it as one of our features. Initially, we looked at Pearson correlation coefficients to better understand the association between epigenetic and psychological data using heatmaps (Fig. 1). In the following plot, we identified the most relevant features to use for our models, which were *TNFα* average score, *FOXP3* average score. For the stress related genes (*BDNF, FKBP5, N3C1*), we used the specific sites instead of the averages because the low correlation between some of their sites could potentially affect our model performance.

To support the initial hypothesis, we plotted and analyzed the distribution and change in depression score before and after birth for patients who gave birth prematurely as well as for patients who did not give birth prematurely (Fig. 2). We observed an overall increase in depression for patients who gave birth prematurely.

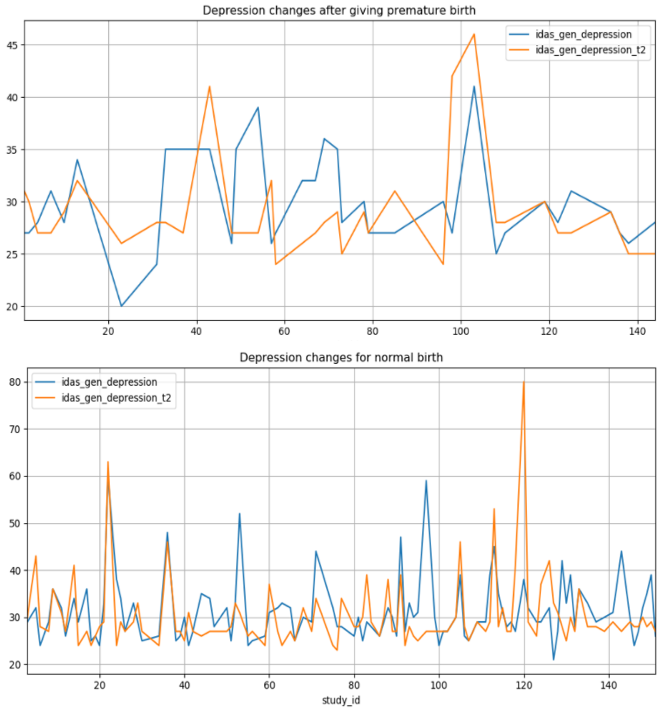

Fig. 2 Depression scores in Latina women before and after giving birth. The top graph shows scores for mothers who gave birth prematurely while the bottom graph shows scores for mothers who gave birth after carrying to term.

Based on the correlation information, we created and compared regression models to see how methylation of inflammation and stress related genes combined with acculturation and acculturative stress scores can help predict the depression score. We used Random Forest Regressor and Gradient Booster Regressor as our models, both of which produced marginally close models, but not the best.

- *Random Forest Regressor* – a machine learning ensemble technique capable to predict numerical values. It is a model that accurately predicts values by constructing a number of decision trees at training time and outputting the mean prediction of the individual trees.

- *Gradient Boosting Regressor* – a machine learning ensemble technique capable to predict numerical values by construction decision trees consecutively, with every consecutive tree learning from the previous models, most frequently decision trees.

We used the aforementioned models to find the underlying patterns and significance of our input variables by reading into the model's feature importance and the permutation feature importance to see which variables are considered more important for predicting the depression score.

Given the current political climate, due to concerns about reporting depression, it is possible the scores given for depression are not entirely reflective of the participant's experiences. After discussing it with collaborators, it was determined discrimination stress was most likely to be accurately reported.

Therefore, we created a similar model to see how stress genes, inflammatory related genes and psychological scores can help predict discrimination score and to look at the underlying feature importance.

After analyzing these relations, we wanted to further look into how depression, acculturative stress, discrimination, economic hardship alongside stress genes and inflammatory related genes affect or correlate with premature birth. For this purpose, we used Random Forest Classifier.

- *Random Forest Classifier* – a machine learning ensemble technique that uses a number of relatively uncorrelated trees for class prediction. Every tree gives out a prediction and the final output is the class with the most votes.

Our target variable distribution was imbalanced, as around 80% of our patients had a normal birth, so we used the class weight parameter to adjust the initial weights and give premature birth classes more weight.

C. *Assessment Protocol*

To assess our results, initially we looked into the feature importance ranking that took place in all of the models that we created, to figure out which were considered important with respect to our model associated targets.

To get the objective scores from our models, we split our data into training, testing and validation set; we used a 10-fold cross-validation on the testing and training part for training. We used the scores from our testing set to help adjust the hyperparameters for the models to get better results. Once we found what seemed to be the best models, we used the validation set to evaluate them.

For our regression model, we used the mean squared error, mean absolute error, the r-squared error and the variance score for evaluation.

For our classification model, because it is a binary classifier we used not only accuracy, but also precision, recall, area under curve and a confusion matrix.

The analysis using the aforementioned metrics were done using the scikit-learn library and the associated functions from their metrics module [18].

IV. RESULTS

Our final results are not conclusive and are subject to change.

A. Gradient Boosting Regressor:

1. Predicting Depression

When depression (`idas_gen_depression`) is predicted using Gradient Boosting Regressor, `total_score_gad_7` holds the highest feature importance and permutation importance in the prediction, followed by `total_score_phq`.

This means that anxiety and patient's health related questionnaire answers are significant and related to depression levels. This makes sense, because they are also highly correlated. Next in importance, are the acculturative stress score and stress related genes alongside the inflammatory related genes.

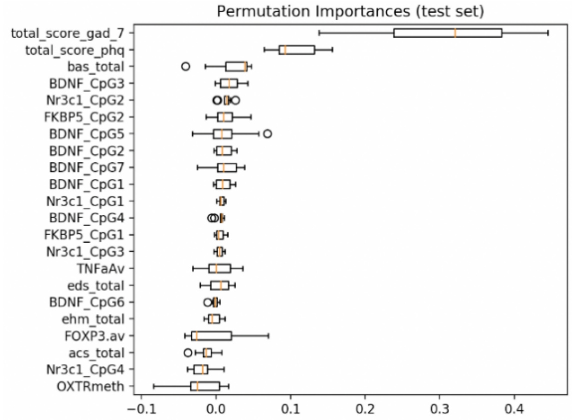

Fig. 3. Ranking of importance of the psychological and epigenetic features for predicting depression using a Gradient Boosting Regressor.

2. Predicting Discrimination

When discrimination (`eds_total`) is predicted using Gradient Boosting Regressor, acculturation and economic hardship, anxiety and depression hold the highest importance. The third most important feature is acculturative stress which is associated with the stress genes and inflammatory related genes respectively. We can see this distribution and association in Fig. 4.

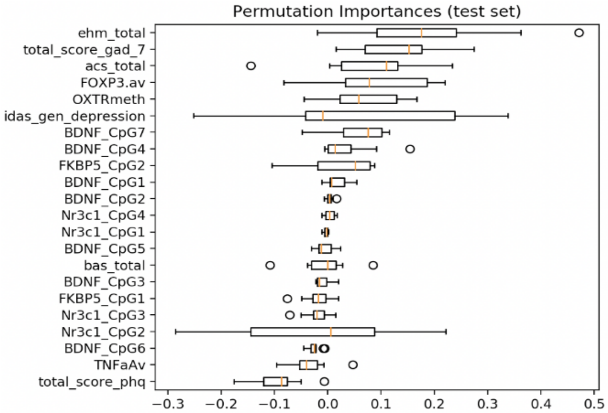

Fig. 4. Ranking of importance of the psychological and epigenetic features in predicting descrimination using a Gradient Boosting Regressor.

B. Random Forest Regressor:

1. Predicting Depression

When depression (`idas_gen_depression`) is predicted using Random Forest Regressor, anxiety, again, and personal health questionnaire hold the highest feature importance. Next in importance is the oxytocin receptor gene, *OXTR*, and the inflammatory related gene, *FOXP3*.

2. Predicting Discrimination

When discrimination (`eds_total`) is predicted using Random Forest Regressor, acculturation and economic hardship measure hold the highest feature importance and permutation importance, respectively, followed by anxiety and acculturative stress.

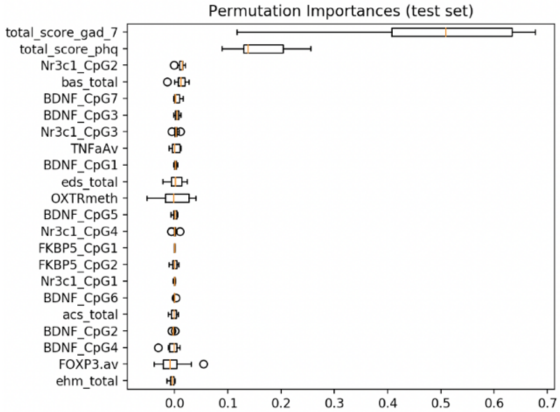

Fig. 5. The importance of the psychological and epigenetic features in predicting depression using Random Forest Regressor.

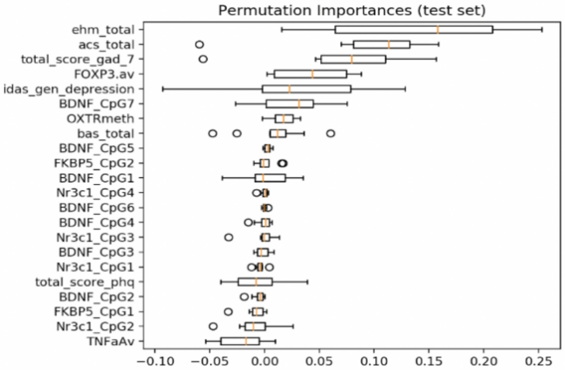

Fig. 6. The importance of the psychological and epigenetic features in predicting depression using Random Forest Regressor.

The results of comparative assessment are given in Table I.

TABLE I. EVALUATION METRIC RESULTS OF THE MODELS WHEN PREDICTING DEPRESSION AND DISCRIMINATION

Regressor	**Dependent Variable**	**Mean Squared Error**	**Mean Absolute Error**	**Variance**	**R-Score**
Gradient Boosting	Depression	28.088	4.048	0.427	0.407
	Discrimination	9.978	2.397	0.291	0.073
Random Forest	Depression	23.280	3.712	0.514	0.508
	Discrimination	10.008	2.454	0.308	0.070

C. Classification of Pre-term Birth using Random Forest Classifier:

The evaluation of the classification model produced the results shown in Table II.

TABLE II. EVALUATION METRICS RESULTS OF THE MODEL WHEN CLASSIFAYING PREMATURE BIRTH

Accuracy Score	0.769
Precision for preterm birth	0.50
Recall for preterm birth	0.17
F1-Score for preterm birth	0.25
Support for preterm birth	6
ROC score	0.55833

The parameters and the score for the best random forest estimator are as follows:

- Best parameters for the best Random Forest Estimator: {'min_samples_leaf': 1, n_estimators: 1000, 'min_samples_split': 2, 'class_weight': {0: 1, 1: 9}, 'max_depth': 80}
- Best score for the best Random Forest Estimator: 0.72

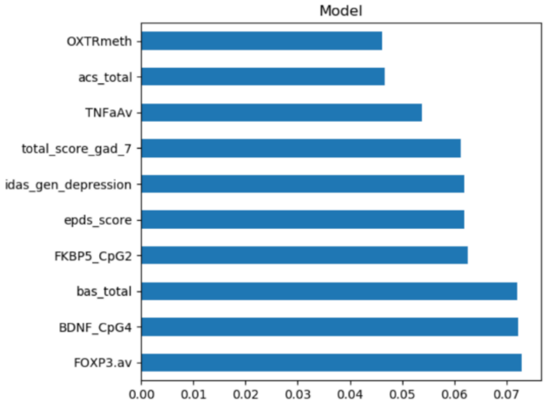

Fig. 7. The importance of psychological and epigenetic markers when predicting premature birth

The stress related gene *BDNF*, specifically at the 4th methylation site, the average of *FOXP3* methylation scores and the acculturation level score are the features affecting premature birth the most, as shown in Fig. 7.

V. Conclusion

The advantage of our approach is rooted in our choice of models. Those used are easy to interpret and extract meaningful relations from the underlying features. Random Forest Classifiers are widely used in the biomedical field for discovering rules. Due to their high predictive accuracy they are considered reliable models. This extends to high-dimensional problems with highly correlated features, such as our dataset.

In our current setting, a disadvantage is the distribution of values for variables in our dataset. It is difficult to analyze and gain clinically meaningful answers with these models given 152 patients with similarly distributed depression scores and highly imbalanced premature birth values. Even though we obtained mediocre results compared to what was expected, we believe that, given enough data, the accuracy of our model can improve, and our initial claim can be further be supported.

In our results, we managed to find significant associations between acculturation and acculturative stress with the methylation scores of stress and inflammatory related genes, as well as their importance in affecting depression and discrimination. We also noticed the depression scores increase after birth, with the effect emphasized in mothers giving birth preterm.

In the future, given more data, we would like to improve upon our existing models and strengthen our claim. Additionally, we would like to focus more broadly on the given epigenetic markers and their psychopathological associations. In the absence of additional data, we would like to explore what happens when different statistical methods are applied to the epigenetic data, such as converting the scores to M-scores [19].

.

Acknowledgment

We would like to thank Dr. Benjamin Nephew for providing us with the data necessary to do this study.

References

[1] N. K. Grote, J. A. Bridge, A. R. Gavin, J. L. Melville, S. Iyengar, and W. J. Katon, “A meta-analysis of depression during pregnancy and the risk of preterm birth, low birth weight, and intrauterine growth restriction,” *Arch. Gen. Psychiatry*, vol. 67, no. 10, pp. 1012–1024, Oct. 2010.

[2] March of Dimes, “March of Dimes Special Report: Maternal and Infant Health in US Hispanic Populations: Prematurity and Related Health Indicators,” 2014.

[3] S. Lara-Cinisomo, K. M. Grewen, S. S. Girdler, J. Wood, and S. Meltzer-Brody, “Perinatal Depression, Adverse Life Events, and Hypothalamic–Adrenal–Pituitary Axis Response to Cold Pressor Stress in Latinas: An Exploratory Study,” *Women’s Heal. Issues*, vol. 27, no. 6, pp. 673–682, Nov. 2017.

[4] H. Santos, E. I. Fried, J. Asafu-Adjei, and R. Jeanne Ruiz, “Network structure of perinatal depressive symptoms in latinas: Relationship to stress and reproductive biomarkers,” *Res. Nurs. Heal.*, vol. 40, no. 3, pp. 218–228, 2017.

[5] A. Conde-Agudelo, A. T. Papageorghiou, S. H. Kennedy, and J. Villar, “Novel biomarkers for the prediction of the spontaneous preterm birth phenotype: A systematic review and meta-analysis,” *BJOG An Int. J. Obstet. Gynaecol.*, vol. 118, no. 9, pp. 1042–1054, 2011.

[6] K. Ghadimi and A. Danaeiniya, “Comparing of pro-inflammatory cytokines in the woman with preterm and term labors,” *J. Gynecol. Neonatal Biol.*, vol. 4, no. 1, pp. 22–26, 2018.

[7] C. Mitchell, L. M. Schneper, and D. A. Notterman, “DNA methylation, early life environment, and health outcomes,” *Pediatr. Res.*, vol. 79, no. 1–2, pp. 212–219, 2016.

[8] C. Maud, J. Ryan, J. E. McIntosh, and C. A. Olsson, “The role of oxytocin receptor gene (OXTR) DNA methylation (DNAm) in human social and emotional functioning: A systematic narrative review,” *BMC Psychiatry*, vol. 18, no. 1, pp. 1–13, 2018.

[9] A. L. Marsland, C. Walsh, K. Lockwood, and N. A. John-Henderson, “The effects of acute psychological stress on circulating and stimulated inflammatory markers: A systematic review and meta-analysis,” *Brain. Behav. Immun.*, vol. 64, pp. 208–219, Aug. 2017.

[10] K. E. Sullivan *et al.*, “Epigenetic Regulation of Tumor Necrosis Factor Alpha,” *Mol. Cell. Biol.*, vol. 27, no. 14, pp. 5147–5160, 2007.

[11] J. Huehn, J. K. Polansky, and A. Hamann, “Epigenetic control of FOXP3 expression: The key to a stable regulatory T-cell lineage?,” *Nat. Rev. Immunol.*, vol. 9, no. 2, pp. 83–89, 2009.

[12] W. Hou *et al.*, “Correlation between protein expression of FOXP3 and level of FOXP3 promoter methylation in recurrent spontaneous abortion,” *J. Obstet. Gynaecol. Res.*, vol. 42, no. 11, pp. 1439–1444, Nov. 2016.

[13] D. Watson *et al.*, “Further Validation of the IDAS: Evidence of Convergent, Discriminant, Criterion, and Incremental Validity,” *Psychol. Assess.*, vol. 20, no. 3, pp. 248–259, 2008.

[14] S. M. Stasik-O’Brien *et al.*, “Clinical Utility of the Inventory of Depression and Anxiety Symptoms (IDAS),” *Assessment*, vol. 26, no. 5, pp. 944–960, Jul. 2019.

[15] D. Watson, “Supplemental Material for Development and Validation of the Inventory of Depression and Anxiety Symptoms (IDAS),” *Psychol. Assess.*, vol. 3, pp. 253–268, 2007.

[16] H. P. Santos *et al.*, “Discrimination exposure and DNA methylation of stress-related genes in Latina mothers,” *Psychoneuroendocrinology*, vol. 98, no. August, pp. 131–138, 2018.

[17] R. England and M. Pettersson, “Pyro Q-CpG$^{\text{TM}}$: Quantitative analysis of methylation in multiple CpG sites by Pyrosequencing®,” *Nat. Methods*, vol. 2, no. 10, 2005.

[18] G. Varoquaux, L. Buitinck, G. Louppe, O. Grisel, F. Pedregosa, and A. Mueller, “Scikit-learn,” *GetMobile Mob. Comput. Commun.*, vol. 19, no. 1, pp. 29–33, 2015.

[19] D. Li, Z. Xie, M. Le Pape, and T. Dye, “An evaluation of statistical methods for DNA methylation microarray data analysis,” *BMC Bioinformatics*, vol. 16, no. 1, Jul. 2015.