Multi-Modal Modeling of Preterm Birth Risk in an Underrepresented Black/African-Ancestry Cohort

a thesis presented
by
Budhachandra Singh Yumkhaibam
to
The Department of Computer Science and Engineering

in partial fulfillment of the requirements
for the degree of
Master
in the subject of
Computer Science

University of Texas at Arlington
Arlington, Texas
December 2025

Thesis advisor: Dr. Jacob Luber, Dr. Jennifer Woo Budhachandra Singh Yumkhaibam

Multi-Modal Modeling of Preterm Birth Risk in an Underrepresented Black/African-Ancestry Cohort

Abstract

In this thesis, We present a machine learning model with logistic regression approach to predict preterm birth (PTB) in a vulnerable (non Hispanic Black/African ancestry cohort). Regularized, stratified 5 fold logistic regression showed the strongest performance from all features combined. In the best performing model (clinical/psychosocial + cytokine + TPM + SNPs), PTB risk was driven by hypertensive history, low vitamin D/VDBP, nicotine use, and lower social/SES support, with a few transcript signals also contributing. The work provides a reproducible pipeline and interpretable biomarkers for PTB in an understudied population.

***Index terms*—** Interpretability, Feature Attribution, Mutli-omics Integration, Biomarkers, Pathways, Health Disparities, Precision medicine

Contents

1

Introduction

1.1 Introduction

Preterm birth (PTB), defined as delivery before 37 weeks of gestation, remains a leading cause of neonatal morbidity and mortality worldwide. In the United States, PTB affects approximately one in ten births and disproportionately impacts non-Hispanic Black women, who experience nearly twice the rate observed in non-Hispanic White women. Despite decades of research, the biolog-

ical and social mechanisms underlying PTB remain incompletely understood, limiting effective early prediction and prevention strategies. PTB is not a single disease entity but a multifactorial syndrome arising from complex interactions among maternal and fetal genetic factors, clinical conditions, biomarkers, nutrition, and psychosocial stressors. This heterogeneity is further complicated by distinct PTB phenotypes, primarily spontaneous and medically indicated PTB, which may share overlapping pathways yet diverge in etiology and clinical presentation. Traditional clinical risk assessment tools—such as cervical length measurement and fetal fibronectin testing—offer limited predictive power, particularly for spontaneous PTB.

Advances in genomics and transcriptomics have expanded efforts to identify molecular predictors of PTB, including genome-wide association studies (GWAS), gene expression profiling, and polygenic risk scores. However, these approaches have yielded few reproducible signals, often due to small effect sizes, limited sample diversity, and insufficient integration with clinical and environmental factors. Notably, most large-scale genetic studies of PTB have focused on European ancestry populations, reducing generalizability and contributing to poor predictive performance in high-risk, underrepresented groups. Recent studies increasingly emphasize the importance of integrative, multi-modal approaches that jointly model genetic variation, gene expression, biomarkers, and psychosocial determinants. Such approaches better reflect the biological and social complexity of PTB but introduce new analytical challenges, including heterogeneous data types, missingness, limited sample sizes, and trade-offs between model performance and interpretability. Existing analytical pipelines often address these modalities in isolation or rely on highly complex models that are difficult to interpret in clinical settings.

To address these gaps, we present a methodological framework for multi-modal PTB risk modeling using a uniquely characterized cohort composed exclusively of non-Hispanic Black women. Our approach integrates clinical and psychosocial variables, inflammatory cytokines, nutritional biomarkers, genome-wide genotype data, and whole-blood transcriptomic profiles within an in-

terpretable machine-learning framework. Rather than optimizing for maximal predictive accuracy alone, we emphasize model stability, transparency, and reproducibility, enabling systematic evaluation of how each data modality contributes to PTB risk prediction. This work is positioned as a methods-focused contribution, providing a scalable and population-inclusive pipeline for PTB research. By explicitly addressing data harmonization, modality integration, and interpretability constraints, our framework lays the groundwork for future biological pathway analyses and translational applications aimed at reducing persistent disparities in preterm birth outcomes.

1.2 Clinical Conditions

Preterm birth (PTB) arises from multiple biological pathways and is commonly classified as either spontaneous or medically indicated. Spontaneous PTB includes preterm labor and premature rupture of membranes and is often associated with inflammation, infection, or uterine overdistension, making it particularly difficult to predict. Medically indicated PTB [3] typically results from maternal or fetal complications such as hypertensive disorders (e.g., preeclampsia [22]), fetal growth restriction, or placental insufficiency. Although these subtypes share some overlapping mechanisms, including immune activation and placental dysfunction, they also involve distinct biological processes. Analytically separating these phenotypes is therefore important to reduce signal dilution and improve biological interpretability.

Well-established clinical risk factors for PTB include a prior history of preterm delivery, extremes of maternal age, low pre-pregnancy body mass index (BMI), limited prenatal care, and exposure to psychosocial stressors [6] or substance use. Identifying individuals at elevated risk enables earlier intervention and more targeted clinical management, which are critical for improving maternal and neonatal outcomes.

Clinical assessment of PTB risk relies on a combination of symptoms, imaging, and biochemical

markers. Key diagnostic indicators include uterine contractions accompanied by cervical dilation or effacement before term. Transvaginal ultrasonography is commonly used to measure cervical length, with a shortened cervix serving as a strong predictor of preterm risk[5]. In addition, biochemical markers such as fetal fibronectin in cervicovaginal secretions provide complementary prognostic information.

1.3 Genetic Factors

Genetic variation plays an important role in preterm birth (PTB), and genome-wide association studies (GWAS) have been used to identify genetic contributors to gestational duration and PTB risk. However, GWAS findings have been limited by small effect sizes and low reproducibility. In addition, most existing studies are based on European ancestry populations, which limits generalizability and reduces the effectiveness of polygenic risk scores in non-Hispanic Black women[11].

Several single nucleotide polymorphisms (SNPs) associated with spontaneous PTB have been reported, particularly in genes related to inflammation, hormonal regulation, and placental function[15]. Both maternal and fetal genetic factors contribute to PTB risk, though their individual effects are difficult to disentangle, reflecting the polygenic and heterogeneous nature of the condition.

Genetic analyses are further complicated by clinical heterogeneity, including differences between spontaneous and medically indicated PTB[3]. Accounting for phenotype-specific variation is therefore critical for improving biological interpretation and reducing signal dilution. Together, these limitations highlight the need for integrative approaches that combine genetic data with clinical and environmental context.

1.4 Biomarkers and Nutritional aspects

Maternal nutrition plays a crucial role in achieving favorable birth outcomes. Several studies have explored the relationship between the consumption of nutrients or specific foods and the risk of adverse birth outcomes.

Assessing Vitamin D levels and other biomarkers such as VDBP (Vitamin D Binding Protein), IL-6, IL-8, IL-10, TNF-Alpha, CRP become very essential to study the mechanisms and interplay affecting PTB and risks associated[5,21,4]. For example, adequate Vitamin D levels during pregnancy are linked to reduced risks of preterm birth[2]. It supports immune function, placental development, and anti-inflammatory responses, which may help maintain pregnancy to term.

For future studies, approaches to nutrition and particularly the supplementation of key micronutrients like zinc, magnesium, and calcium, which have been increasingly studied for their potential influence on preterm birth outcome need to be considered.

Overall, it is important to ensure adequate intake of these nutrients during pregnancy which can be a preventive strategy against preterm birth, especially in high-risk or nutrient-deficient populations.

1.5 Psychosocial Stressors and Factors

Psychosocial factors are well-established[9] contributors to preterm birth (PTB). Elevated stress[6], anxiety[1], depression[16,13,14], and limited social support during pregnancy have been associated with increased PTB risk. Chronic stress and sleep disturbance may influence hormonal regulation and inflammatory pathways that affect the timing of labor. Additional risk factors include exposure to domestic violence, financial hardship, and social isolation, which may also indirectly affect PTB risk through altered health behaviors such as poor nutrition or inconsistent prenatal care.

Social and environmental conditions further shape psychosocial risk.[18] Neighborhood-level stressors, including poverty, pollution, violence, and reduced access to healthcare, have been linked

to adverse pregnancy outcomes[7]. These stressors disproportionately affect non-Hispanic Black women, who are more likely to experience structural disadvantage and racial discrimination[8], both of which have been associated with increased PTB risk.

Together, these findings highlight the importance of incorporating psychosocial and environmental factors into PTB research, particularly in studies focused on high-risk and underrepresented populations.

1.6 Recent Work

Recent work has applied machine learning methods to PTB prediction using clinical and electronic health record data, often reporting improved predictive performance compared to conventional statistical models[17]. Other studies have explored biomarker-based or multi-omics approaches, integrating genomic, transcriptomic, or proteomic signals to capture underlying biological mechanisms[19]. However, most of these models are developed in mixed-ancestry or predominantly European-ancestry cohorts and prioritize overall predictive accuracy rather than interpretability or population-specific generalization.

A smaller body of work has examined PTB risk through demographic or race–aware analyses, highlighting the importance of psychosocial stressors and social determinants of health among non-Hispanic Black women[20]. These studies typically rely on clinical or psychosocial variables alone and do not integrate high-dimensional biological data. Consequently, there remains a gap in PTB modeling approaches that are both multimodal and explicitly designed for high-risk, underrepresented populations.

2

Results

2.1 Study Design and Cohort Description

This study employs a retrospective case–cohort design using data from the *Biosocial Impact on Black Births (BIBB)* study, a longitudinal cohort that recruited non-Hispanic Black pregnant women from Wayne State University (Detroit, MI) and The Ohio State University (Columbus, OH). All study procedures were approved by the respective Institutional Review Boards, and written informed

consent was obtained from all participants. The present work constitutes a secondary analysis of this cohort.

Eligible participants were adults aged 18 years or older with singleton pregnancies. Individuals were excluded if they reported endocrine disorders, prenatal corticosteroid exposure prior to recruitment, in vitro fertilization, major fetal or chromosomal anomalies, or substantial substance use during pregnancy. Due to differences in data availability and quality control requirements, the effective sample size varied across data modalities.

2.2 Overall model comparison

Model	Accuracy	ROC-AUC	PR-AUC
Model 1 (clinical + cytokine)	[0.65 $\pm$ 0.11]	[0.70$\pm$ 0.11]	[0.57 $\pm$ 0.17]
Model 2 (+ SNPs)	[0.72 $\pm$ 0.04]	[0.78 $\pm$ 0.10]	[0.67 $\pm$ 0.13]
Model 3 (+ RNA-seq)	[0.79 $\pm$ 0.10]	[0.92 $\pm$ 0.03]	[0.93 $\pm$ 0.04]

Table 2.1: Cross-validated performance across models (mean $\pm$ SD).

Table 2.1 summarizes cross-validated performance across the three models (mean $\pm$ SD across 5 folds). We compared three predictive models that progressively incorporate additional data modalities from the BIBB cohort. Model 1 uses clinical/psychosocial and biomarker features; Model 2 adds genotype-derived features; and Model 3 further adds transcriptomic features (gene-level TPM). All models were evaluated under the same stratified cross-validation protocol to ensure a fair comparison. Overall, we observe a consistent performance improvement as additional modalities are included. Relative to Model 1, Model 2 shows a modest gain in discriminative performance (e.g., AUROC from 0.70 to 0.78), suggesting that genotype features contribute complementary signal beyond clinical and biomarker data. Model 3 achieves the strongest performance across metrics (e.g., AUROC 0.92 and AUPRC 0.93), indicating that transcriptomic features provide additional predictive information not captured by the other modalities.

We also note that Model 3 exhibits more stable performance across folds (lower variance in Table 2.1), which is consistent with improved generalization when transcriptomic information is included. The ROC and precision–recall curves for each model are shown in Figures 2.5, which visually reinforce the ranking observed in the aggregate metrics.

2.2.1 SHAP

To interpret model predictions, we used SHapley Additive exPlanations (SHAP), a model-agnostic framework based on cooperative game theory. SHAP values quantify the contribution of each feature to an individual prediction by fairly distributing the model output among input features. This approach provides both local and global explanations and has been widely adopted for interpreting machine learning models. We report feature importance based on mean absolute SHAP values across cross-validation folds[12]. SHAP values were computed for each fold to assess feature importance, directionality, and stability across models. Particular attention was given to differences between the strongest and weakest performing folds to evaluate robustness.

We grouped features by shared clinical or biological context and used these groupings only for visualization. SHAP plots were color-coded by group to improve readability and to check for broad, group-level patterns. These groupings are not model inputs or constraints; they are purely interpretive (see table 2.2).

2.3 Model 1: Clinical + Cytokine

Performance. Figure 2.1 shows ROC and PR curves for all folds and the overall average;

Interpretability. Figures 2.2a and 2.2b report SHAP summaries for the best and worst folds.

Group	Features (readable labels)	Color label
Physical	Weight, BMI, Height	Blue
Vitamin	Vitamin D (25OHD), VDBP	Orange
Psychosocial	Social support, Discrimination, CESD	Green
Socioeconomic	Income, Education, Insurance, Work status, Marital status, Nicotine use	Red
Medical	Asthma, HTN history, Thyroid, Diabetes, Heart, Kidney, Pregnancy HTN, Age	Purple
Cytokine	IFN-γ, IL-6, IL-8, IL-10, TNF-α, MIF, CRP (T1/T2)	Brown
SNP	rs73160620, rs11980802 (dosage)	Pink
Gene expression	RN7SL1, RN7SK, RN7SL2, MIR3648-1, RMRP, RPPH1, MT-ATP8, MALAT1, MT-ND4L, ALAS2	Teal

Table 2.2: SHAP feature groupings and plot color labels.

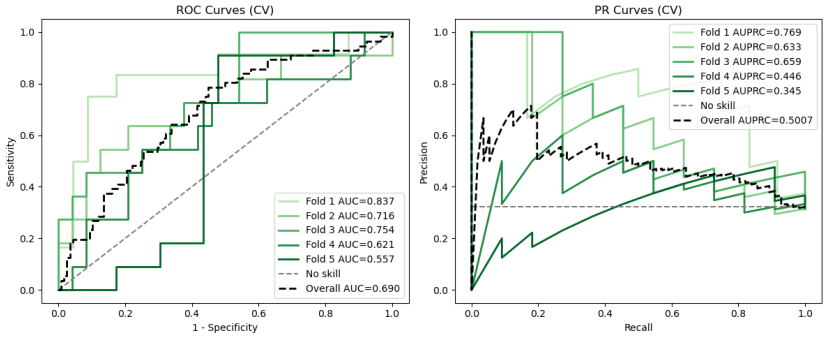
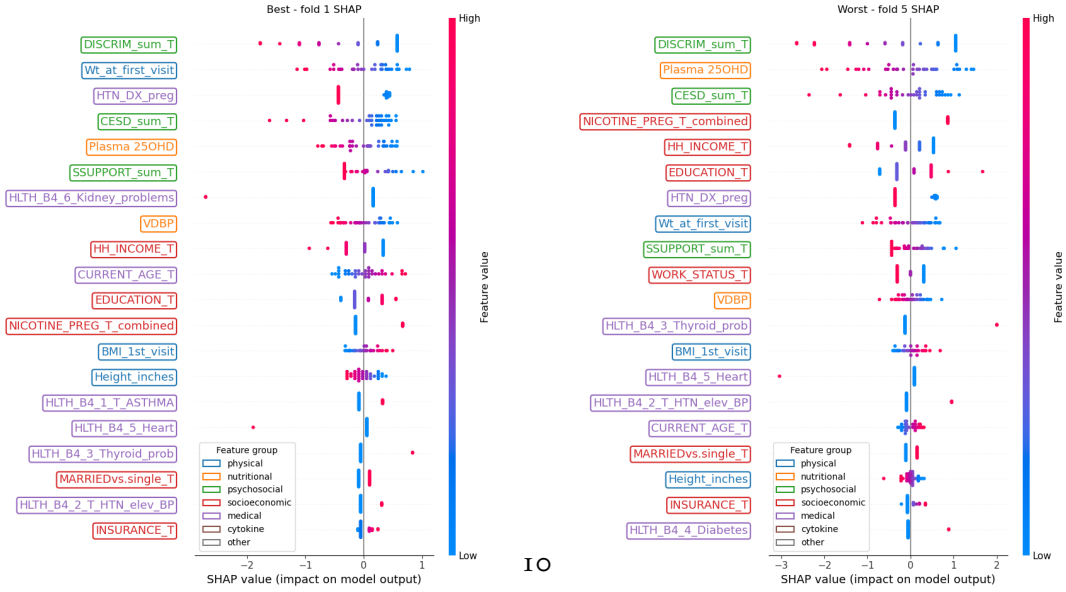


Figure 2.1: Model 1 ROC and PR curves (5-fold CV + overall).



(a) Best fold

(b) Worst fold

Figure 2.2: Model 1 SHAP summaries.

Interpretation (Model 1). SHAP highlights clinical and cytokine drivers, with psychosocial stress measures, vitamin D–related markers, and baseline inflammatory cytokines among the strongest contributors. The best-fold plot shows consistent directional effects for these features, while the worst-fold plot reflects greater dispersion, suggesting some instability in smaller subsets.

2.4 Model 2: + SNPs

Performance. Figure 2.3 (ROC) and (PR). **Interpretability.** Figures 2.4a and 2.4b.

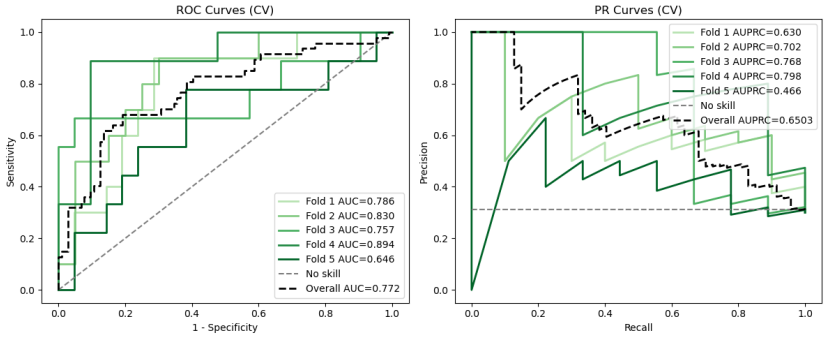
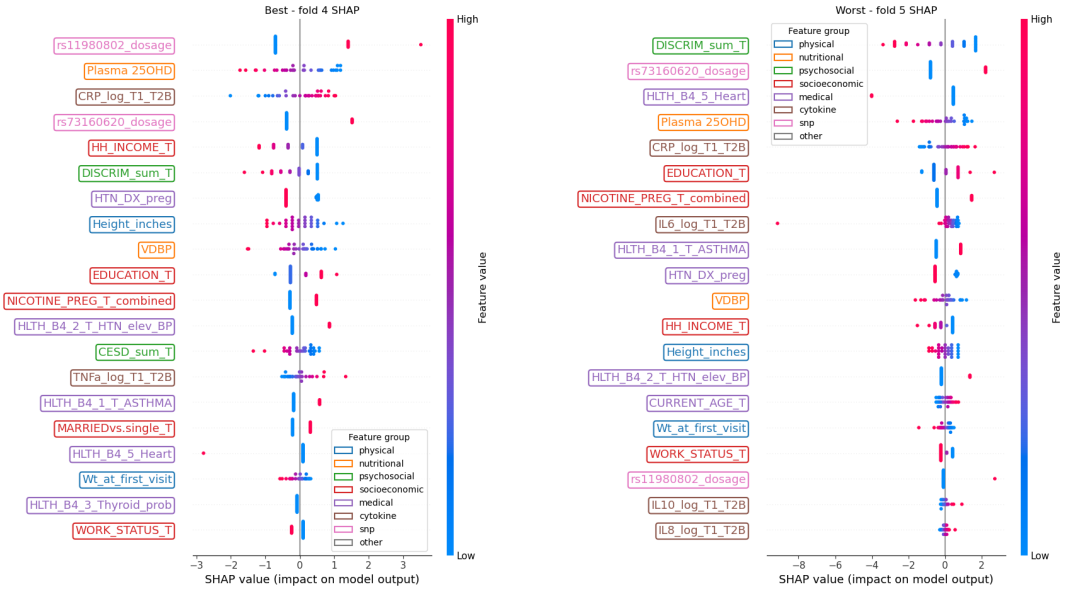


Figure 2.3: Model 2 ROC and PR curves (5-fold CV + overall).



(a) Best fold

(b) Worst fold

Figure 2.4: Model 2 SHAP summaries.

Interpretation (Model 2). After adding the two GWAS SNPs, SHAP shows whether genetic dosages enter the top-ranked features alongside clinical and cytokine signals. Overall patterns remain dominated by clinical/biomarker variables, indicating the SNP features provide incremental but limited explanatory shift in this cohort.

2.5 Model 3: + RNA-seq

Performance. Figure 2.5 (ROC) and (PR). **Interpretability.** Figures 2.6a and 2.6b.

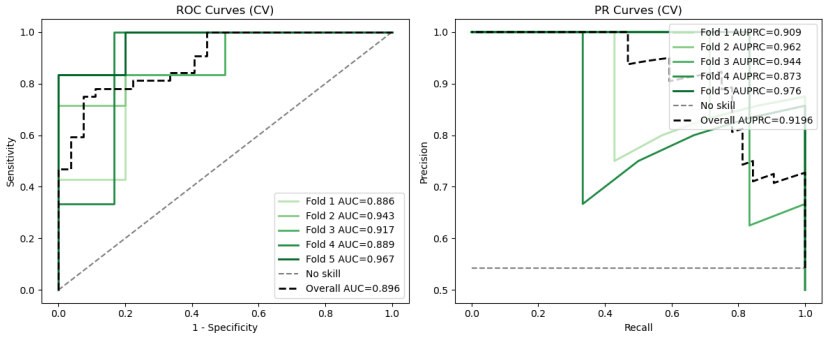
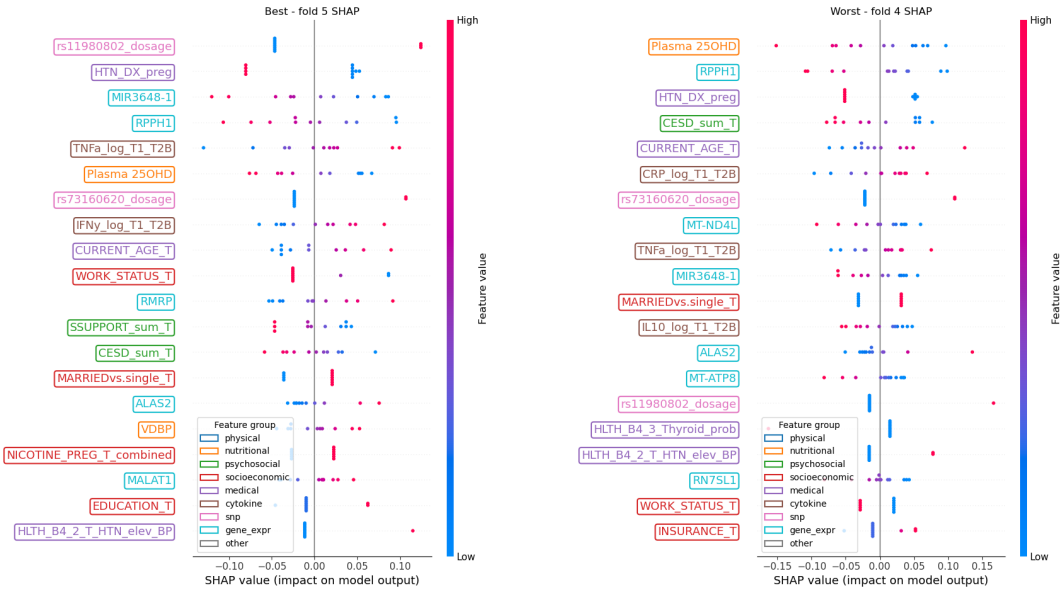


Figure 2.5: Model 3 ROC and PR curves (5-fold CV + overall).



(a) Best fold

(b) Worst fold

Figure 2.6: Model 3 SHAP summaries.

Interpretation (Model 3). With RNA-seq TPM features (top-variance k=10), SHAP shows a mixed set of transcriptomic and clinical features near the top of the ranking. The best-fold plot suggests added biological signal from expression features, while the worst-fold plot indicates variability consistent with the reduced sample size.

2.6 Exploratory Transcriptomic Analysis

[Permutation Enrichment; DAVID Clustering] For participants with available RNA sequencing data, exploratory analyses were conducted to examine gene-level signals associated with PTB. Differential expression and permutation-based enrichment analyses were performed to identify biologically relevant patterns. Functional annotation and clustering were carried out using established gene ontology resources.

These analyses were exploratory and intended to support biological interpretation rather than definitive pathway inference.

3

Discussion

From a technical perspective, we designed a sequence of nested models to test whether incrementally adding modalities improves predictive performance. Specifically, we compared models using (i) clinical/psychosocial and biomarker variables, then adding (ii) genotype-derived features, and finally adding (iii) RNA-seq gene expression summarized as TPM. Although performance tended to improve as modalities were added, this pattern should be interpreted cautiously because additional modalities can also introduce noise, missingness, and higher dimensionality. To prioritize

interpretability and reduce overfitting risk in a small-sample, high-dimensional setting, we used regularized logistic regression and interpreted feature contributions using SHAP. As expected, Model 1 provides the most straightforward and stable attributions, since it uses fewer features and includes the largest sample size (approximately 180 participants) among the three models. Model 1 achieved an accuracy of 0.65, which is informative but should be interpreted alongside AUROC/AUPRC given potential class imbalance. In Model 1, self-reported CESD discrimination measures emerged as an important predictor, consistent with [prior work] linking psychosocial stressors to adverse birth outcomes. Additional high-impact features included maternal weight/BMI-related variables, hypertensive disorders, and vitamin D–related markers (e.g., plasma 25(OH)D and VDBP), along with indicators of prior pregnancy history and comorbid conditions. To assess stability and improve transparency, we also examined the lowest-performing fold and confirmed that the highest-ranked predictors were broadly consistent between the best and worst folds.

Models 2 and 3 incorporate additional modalities and derived features beyond the baseline feature set. Both models achieve higher predictive performance, and notably Model 3 also shows higher and more consistent PR-AUC across folds, as reflected by a PR-AUC of 0.93. A trade-off of increased model complexity and feature dimensionality is that SHAP-based rankings can appear less consistent across folds, with the relative ordering of top predictors shifting compared to Model 1. These can also be attributed to feature correlation, added modalities, and smaller n. Importantly, many of the high-impact predictors from Model 1 remain influential, although their relative contributions change once genomic and transcriptomic information is included. Because Models 2 and 3 are trained on a smaller subset of participants (due to modality availability), improvements in performance should be interpreted alongside the possibility of increased overfitting risk and less stable attribution estimates. We reduced this risk by using regularized logistic regression, strictly separating training and evaluation splits, and ensuring that all preprocessing steps were fit on training data only to prevent data leakage. [In addition, we repeated the evaluation with multiple random seeds and

observed consistent performance trends.] [Only claim this if you can report the spread; otherwise, maybe soften to "we also verified..."]

This study makes a distinct contribution by focusing on a historically underrepresented and potentially vulnerable population, which remains understudied in genomic and multimodal PTB research. Studies in this setting are challenging because well-phenotyped multimodal datasets are limited in availability, particularly for ancestry-specific cohorts. We therefore aimed to maximize the value of the available data by conducting a comprehensive set of analyses while maintaining consistent evaluation procedures across model variants. Overall, the resulting models are broadly consistent with prior PTB literature, and the multimodal approach achieves competitive performance relative to prior studies in comparable ancestry-specific or small-cohort settings.

3.1 Comparison

Look at the literature and write comparisons to other recent PTB studies of PR [Maybe its not required since I have background and related works ?]

3.2 Limitations

A primary limitation of this study is incomplete modality availability across participants, which required restricting Models 2 and 3 to smaller analytic subsets. As a result, the three models are not directly comparable in a strictly controlled sense, because they were trained and evaluated on different sample sizes (and potentially different participant subsets).

The overall cohort size is modest, which increases the risk of overfitting and contributes to higher variance in cross-validation performance and feature attributions. In addition, Models 2 and 3 include [derived/engineered] features (e.g., modality-specific summaries), which may be more sensitive to preprocessing choices than directly measured clinical variables.

To reduce bias from missingness, we made conservative decisions to exclude features or participants with substantial missing data rather than relying on aggressive imputation or externally derived values. While this improves internal validity, it further reduces the effective sample size and may limit generalizability. Finally, multi modal datasets in historically underrepresented populations remain difficult to acquire at scale, and therefore external validation in independent cohorts is an important next step.

3.3 Implications

Model 3 achieved strong discriminative performance (AUROC ≈ 0.97). However, the model relies on high-burden multimodal inputs (e.g., genotyping and RNA-seq), which may limit near-term clinical feasibility due to cost, turnaround time, and data availability across settings.

Accordingly, the primary contribution of this study is not to propose an immediately deployable clinical tool, but to identify candidate correlates and potential biomarkers associated with PTB risk in this cohort. Many of the highest-impact predictors are consistent with prior PTB literature, supporting the biological plausibility of the learned signals. In addition, several candidate SNPs and top-ranked transcriptomic features (genes) emerged as influential in Models 2 and 3. These candidates appear to have limited prior evidence in PTB-specific studies and therefore warrant follow-up validation in independent cohorts and, where feasible, functional characterization.

4

Methods

4.1 Data Modalities

Multiple data modalities were incorporated, including clinical, psychosocial, biomarker, genomic, and transcriptomic data. Clinical variables included maternal demographics, obstetric history, and prenatal characteristics. Psychosocial measures captured depressive symptoms and experiences of discrimination. Biomarker data consisted of inflammatory cytokines and vitamin D–related mea-

sures. Genomic data were derived from SNP array genotyping, while transcriptomic data were obtained from whole-blood RNA sequencing for a subset of participants.

A summary of data modalities and sample sizes is provided in Table 4.1.

Data modality	Sample size	Number of features
Clinical / psychosocial	N = *184*	*36*
Biomarkers / cytokines	N = *183*	*75*
Genomic (SNPs / PCs)	N = *158*	*2*
Transcriptomic (RNA-seq)	N = *70*	*10*

Table 4.1: Summary of data modalities and sample sizes. "Number of features" is available but not all of them are used in the model. Genomic has 2 features meaning we identified 2 SNPs for analysis.

4.2 CLINICAL DATA

Clinical data were obtained from the BIBB cohort and included records for N=184 participants. The clinical feature set included vitamin D–related markers (Plasma 25OHD, VDBP), anthropometrics (weight, BMI, height), psychosocial measures (social support, discrimination, depressive symptoms), socioeconomic indicators (income, education, insurance, work status, marital status, nicotine use), and medical history variables (asthma, hypertension-related markers, thyroid disease, diabetes, heart and kidney problems, pregnancy hypertension, current age). All continuous variables were standardized prior to modeling, and missing values were handled.

4.3 CYTOKINES DATA

As part of the BIBB study, cytokine measurements were collected at multiple pregnancy timepoints corresponding to first-, second-, and third-trimester study visits. Plasma inflammatory cytokine biomarkers, including IFN-γ, IL-6, IL-8, IL-10, TNF-α, MIF, and CRP, were measured for $N =$

183 participants. All continuous variables were standardized prior to modeling, and missing values were handled.

4.4 Genomic data

Genotype data were available for $N = 158$ participants and stored as VCF/PLINK files (hg38). The dataset includes genome-wide SNP genotypes per individual, with sample IDs aligned to the clinical cohort. GWAS summary outputs (Manhattan/QQ plots and top-hit tables) were generated from these genotypes.

Preprocessing. Genome-wide association analysis used cohort-wide genotypes with PTB case–control status. Standard QC filtered low-quality variants/samples (call-rate, MAF, HWE), and ancestry PCs were included as covariates. Logistic regression tested SNP associations with PTB, producing ORs/p-values and Manhattan/QQ plots. Suggestive SNPs (e.g., rs73160620, rs11980802) were carried forward as candidate features.

SNP dosage was computed from VCF genotypes as the alternate-allele count: 0 = homozygous reference, 1 = heterozygous, 2 = homozygous alternate; missing calls were coded as NaN.

4.5 RNA-seq data

RNA sequencing data were available for a subset of the cohort $N = 70$. Whole-blood samples were processed to generate paired-end FASTQ files, and gene expression was quantified at the transcript and gene level for downstream analysis.

4.5.1 PCs and TPM

Genotype **principal components (PCs)** were computed from the quality controlled (QC) SNP data to capture population structure. After standard GWAS QC (call rate/MAF/HWE filters),

we used PLINK's PCA routine to derive eigenvectors from the genotype matrix. The resulting `PCA.eigenvec` file provides per-participant PC scores; we retained the first five PCs as covariates to adjust for ancestry-related variation. When we included these PCs in the third model, the model consistently overfit in our small cohort. We did not identify a clear cause despite additional checks, and the issue persisted even when using only 2 PCs. Given this instability, we prioritized other patient-level features.

RNA-seq gene expression was quantified using Salmon on paired-end FASTQ files with a GRCh38/Gencode transcriptome index. Salmon produced transcript-level abundances per sample, which were summarized to gene-level expression with `tximport`, yielding a TPM (**Transcripts Per Million**) matrix (`gene_tpm.tsv`). TPM normalizes for gene length and library size, making expression values comparable across genes and samples. In our experiments, TPM-based features produced more stable cross-validated metrics across runs, so we used TPM in the final model. A fixed top-variance RNA-seq panel (k=10) was used, and performance was assessed with stratified 5-fold CV so each validation fold was fully held out and had a similar case/term mix.

Additional exploratory analyses that are not directly related to the primary objectives, but may inform future work, are reported in Section 2.6.

Preprocessing. Raw reads were quantified against the hg38 reference transcriptome, and transcript-level abundances were aggregated to gene-level expression. Expression values were normalized as transcripts per million (TPM) to enable cross-sample comparison. For modeling, gene-level TPM matrices were used, and a reduced feature set was constructed from the most variable genes to mitigate dimensionality and the limited sample overlap with other modalities. More information is provided in 4.5.1

4.6 Missing Values

During data evaluation, missing values (NaN or null entries) were encountered across multiple variables. Several approaches exist for handling missing data, including imputation; however, given that these are medical records and are inherently noisy, imputation may introduce additional bias. Instead, samples containing missing values were removed. Although this decision reduced an already limited sample size, it allowed for a more reliable evaluation using complete observations without introducing imputation-related noise.

4.7 Outcome Definition

The primary outcome of interest was preterm birth (PTB). Gestational age at delivery was obtained from clinical records. For modeling purposes, outcomes were treated as a binary classification problem distinguishing preterm and term births.

4.8 Modeling Framework

We adopted a stepwise modeling strategy to evaluate the incremental contribution of each data modality. Three primary models were constructed:

- Model 1: clinical, psychosocial, and biomarker features
- Model 2: Model 1 augmented with genomic features
- Model 3: Model 2 further augmented with transcriptomic features

All models were implemented using L2-regularized logistic regression, selected for its balance between predictive performance and interpretability. The probability of preterm birth was modeled as:

$$P(Y = 1 \mid \mathbf{x}) = \frac{1}{1 + \exp\left(-(\mathbf{w}^T\mathbf{x} + b)\right)} \tag{4.1}$$

where $\mathbf{x}$ denotes the feature vector, $\mathbf{w}$ the learned coefficients, and b the intercept. Regularization strength was fixed across experiments to promote model stability.

4.8.1 MODEL 1

Model 1 uses the clinical and cytokine data described in Sections 4.2 and 4.3. These inputs require minimal preprocessing beyond standard normalization for comparable scaling. We treated this as the cleanest, low-dimensional model and expected it to be the most stable baseline, with added modalities potentially reducing stability. This assumption did not hold without additional tuning for the expanded models. Raw metrics are reported in the Results section.

After ID harmonization and missing-value filtering, the modeling frame comprised 174 participants and 29 predictors.

4.8.2 MODEL 2

Model 2 extends Model 1 by adding genomic features from the GWAS summary (see Section 4.4). Two suggestive SNPs (rs73160620, rs11980802) were selected as candidate variants and encoded as alternate-allele dosages (0/1/2; missing calls as NaN). Genotype sample IDs were cleaned to align with clinical PTIDs before merging, and the added genomic features reduced the usable cohort due to missing overlap. The resulting modeling frame comprised 151 participants with 32 predictors, capturing clinical, cytokine, and targeted genetic signals. Evaluation followed the shared procedure described in the Model Evaluation subsection 4.8.4.

4.8.3 MODEL 3

Model 3 augments Model 2 by adding transcriptomic features from the RNA-seq subset (see Section 4.5). Gene-level TPM matrices were aligned to the clinical/cytokine/SNP cohort through PTID harmonization. To avoid overfitting given the high dimensionality of expression data, a fixed panel of the most variable genes across the RNA-seq samples was selected and appended as expression features. This integration substantially reduced the usable cohort because RNA-seq was available for fewer participants, yielding a final modeling frame of 59 participants with 42 predictors spanning clinical, cytokine, genomic, and transcriptomic signals. Evaluation followed the shared procedure described in the Model Evaluation subsection 4.8.4.

4.8.4 MODEL EVALUATION

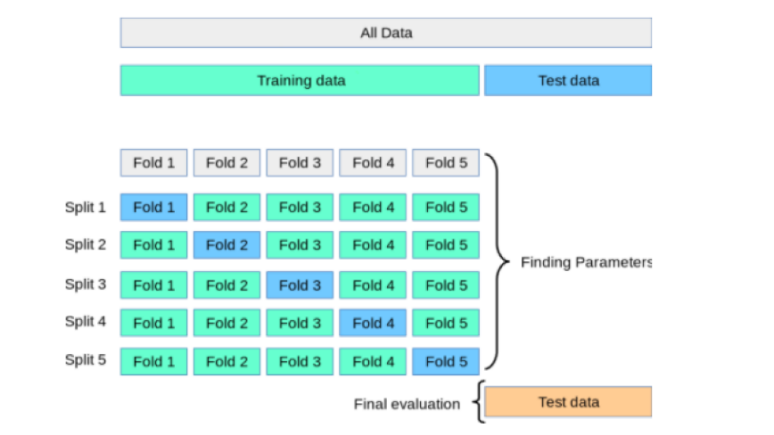


Figure 4.1: Schematic of the 5 fold cross-validation

Model performance was assessed using stratified 5-fold cross-validation to preserve outcome class balance across folds. Due to limited sample sizes, no separate hold-out test set was used. Evaluation metrics included accuracy, area under the receiver operating characteristic curve (ROC-AUC), and area under the precision–recall curve (PR-AUC)

An overview of the evaluation procedure is illustrated in Figure 4.1.

4.9 Experiment Design

Our experiment design compared three nested models to quantify how each modality alters predictive performance and interpretability. Model 1 served as the baseline using low-dimensional clinical and cytokine data. Model 2 added targeted genetic features derived from GWAS top hits, and Model 3 incorporated transcriptomic features selected from RNA-seq. The same modeling pipeline and evaluation protocol were applied across models to ensure comparability. We used SHAP to examine feature importance and robustness, focusing on best and worst folds to assess stability.

Model performance was evaluated using accuracy, ROC-AUC, and PR-AUC, which are standard metrics for binary classification. PR-AUC was emphasized due to class imbalance in the dataset, as it more directly reflects performance on the minority class and penalizes false positives and false negatives. For this reason, PR-AUC was used as the primary metric for comparison with prior work.

SNP features were limited to a small set of suggestive GWAS hits (e.g., rs73160620, rs11980802) to balance biological relevance with sample size and model stability. Each SNP was encoded as dosage (0/1/2 alternate-allele count), providing a compact genetic signal without introducing thousands of sparse predictors. This targeted approach reduces overfitting risk in a small cohort while allowing us to test whether top GWAS signals add incremental predictive value.

5

Conclusion

5.1 Conclusion

In this work, we presented a multi-modal framework for preterm birth (PTB) risk modeling that integrates clinical, psychosocial, biomarker, genomic, and transcriptomic data within a single analytical pipeline. By adopting a stepwise modeling strategy, we evaluated the incremental contribution of each data modality and demonstrated improved predictive performance with the inclusion of

genomic and transcriptomic features. Importantly, the use of interpretable models enabled examination of feature-level contributions while maintaining model stability in a limited-sample setting.

This study emphasizes the importance of considering PTB as a multifactorial condition influenced by biological, clinical, and social determinants. Focusing on a cohort of non-Hispanic Black women, a population that is both underrepresented in prior genomic studies and disproportionately affected by PTB, allowed us to address important gaps in existing literature. Together, our findings support the value of integrative, population-aware modeling approaches for PTB risk assessment.

5.2 Challenges and Limitations

Several limitations should be considered when interpreting the results of this study. First, sample size constraints, particularly for genomic and transcriptomic data, limited statistical power and necessitated the use of cross-validation rather than an independent hold-out test set. Differences in data availability across modalities also resulted in varying effective sample sizes, which may influence model comparisons.

Second, PTB is a clinically heterogeneous condition, and although phenotype-aware considerations were incorporated where possible, residual heterogeneity may still dilute biological signals. In addition, genomic analyses were limited by modest effect sizes and the complexity of disentangling maternal and fetal genetic contributions.

Finally, while interpretable modeling was prioritized, biological interpretation of certain features—particularly transcriptomic signals—remains challenging. The findings from enrichment and pathway analyses should therefore be viewed as exploratory and hypothesis-generating rather than definitive.

5.3 Future Directions

Future work will focus on expanding this framework through larger and more diverse cohorts to improve generalizability and model robustness. Increasing sample size, particularly for transcriptomic data, will enable more stable estimation of feature effects and facilitate independent validation of model performance.

Additional modeling strategies, including alternative regularization schemes and non-linear approaches, may be explored to capture more complex interactions while maintaining interpretability. Further refinement of phenotype stratification, such as separating spontaneous and medically indicated PTB, may also enhance biological relevance.

Finally, deeper investigation of genomic and transcriptomic signals through targeted pathway analysis and replication studies will be essential for translating predictive findings into biological insight. Ultimately, this work aims to contribute toward clinically meaningful and equitable risk assessment tools for preterm birth.

5.4 Acknowledgments

This work was supported by the Cancer Prevention and Research Institute of Texas (CPRIT) Recruitment of First-Time, Tenure-Track Faculty Members Grant (RR220015) (JML) and the University of Texas System Rising STARs Award (JML).

6

Declarations and Availability

6.1 Data availability

The data analyzed in this study are not publicly available due to privacy and institutional restrictions. Access may be granted by the original study team under appropriate approvals and data use agreements. Requests for data access should be directed to the custodians of the BIBB study. Additional details about participant recruitment, data collection, and governance are described in the

original BIBB study publication(s)[10].

6.2 Code availability

All code used for preprocessing, model training, evaluation, and figure generation is available at: `[GITHUB_LINK]`. The repository includes scripts and documentation sufficient to reproduce the main results reported in this thesis, subject to data access restrictions.

6.3 Acknowledgments

I would like to thank my advisor, Dr. Jacob Luber, and my co-advisor, Dr. Jennifer Woo, for their guidance and support throughout this project. I also thank the BIBB study team for making the dataset available for research use, and the members of my thesis committee for their feedback. Finally, I am grateful to the participants whose contributions made this research possible.

6.4 Contribution

6.5 Ethics

This work is a secondary analysis of an existing human-subjects dataset. All analyses were conducted in accordance with institutional policies and applicable ethical guidelines. The BIBB study obtained informed consent from participants and received approval from the appropriate institutional review board(s) as reported in the original publication(s)[?] . This thesis used de-identified data provided under institutional authorization and data use requirements. *[IRB #[IRB_NUMBER], University of Texas at Arlington.]*

References

[1] Accortt, E. et al. (2022). Association between diagnosed perinatal mood and anxiety disorders and adverse perinatal outcomes. *The Journal of Maternal-Fetal & Neonatal Medicine*, 35, 9066–9070.

[2] Amegah, A. K., Klevor, M. K., & Wagner, C. L. (2017). Maternal vitamin d insufficiency and risk of adverse pregnancy and birth outcomes: a systematic review and meta-analysis of longitudinal studies. *PLOS ONE*, 12, e0173605.

[3] An, H. et al. (2022). Impact of gestational hypertension and pre-eclampsia on preterm birth in china: a large prospective cohort study. *BMJ Open*, 12(9), e058068.

[4] Bodnar, L. M., Platt, R. W., & Simhan, H. N. (2015). Early-pregnancy vitamin d deficiency and risk of preterm birth subtypes. *Obstetrics & Gynecology*, 125(2), 439–447.

[5] Creswell, L. et al. (2023). Preterm birth: Screening and prediction. *International Journal of Women's Health*, 15, 1981–1997.

[6] Dove-Medows, E. et al. (2025). A qualitative exploration of perceptions of the aetiology of preterm birth among pregnant black women. *Midwifery*, 145, 104365.

[7] Ghimire, U. et al. (2021). Depression during pregnancy and the risk of low birth weight, preterm birth and intrauterine growth restriction- an updated meta-analysis. *Early Human Development*, 152, 105243.

[8] Giurgescu, C., Engeland, C. G., & Templin, T. N. (2015). Symptoms of depression predict negative birth outcomes in african american women: A pilot study. *Journal of Midwifery & Women's Health*, 60(5), 570–577.

[9] Giurgescu, C. et al. (2022). Neighborhoods, racism, stress, and preterm birth among african american women: A review. *Western Journal of Nursing Research*, 44, 101–110.

[10] Giurgescu, C., Zhang, L., Price, M., Dailey, R., Frey, H. A., Walker, D. S., Zenk, S. N., Engeland, C. G., Anderson, C. M., & Misra, D. (2020). Prenatal cigarette smoking as a mediator between racism and depressive symptoms: The biosocial impact on black births (bibb) study. *Public Health Nursing*, 37(5), 740–749.

[11] Jain, V. G., Monangi, N., Zhang, G., & Muglia, L. J. (2022). Genetics, epigenetics, and transcriptomics of preterm birth. *American Journal of Reproductive Immunology*, 88(4), e13600.

[12] Lundberg, S. M. & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)* (pp. 4765–4774).: Curran Associates, Inc.

[13] Natamba, B. K. et al. (2014). Reliability and validity of the center for epidemiologic studies-depression scale in screening for depression among HIV-infected and -uninfected pregnant women attending antenatal services in northern Uganda: a cross-sectional study. *BMC Psychiatry*, 14, 303.

[14] Noroña-Zhou, A., Aran, Ö., Garcia, S. E., Haraden, D., Perzow, S. E. D., Demers, C. H., Hennessey, E.-M. P., Melgar Donis, S., Kurtz, M., Hankin, B. L., & Davis, E. P. (2022). Experiences of discrimination and depression trajectories over pregnancy. *Women's Health Issues*, 32(2), 147–155.

[15] Nowak, A. L. et al. (2022). Dna methylation patterns of glucocorticoid pathway genes in preterm birth among black women. *Biological Research for Nursing*, 24(4), 493–502.

[16] Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*. Some metadata (e.g., issue/pages) may require manual confirmation if your BibTeX style demands it.

[17] Sharifi-Heris, Z., Laitala, J., Airola, A., Rahmani, A. M., & Bender, M. (2022). Machine learning approach for preterm birth prediction using health records: Systematic review. *JMIR Medical Informatics*, 10(4), e33875.

[18] Slaughter-Acey, J. C. et al. (2016). Racism in the form of micro aggressions and the risk of preterm birth among black women. *Annals of Epidemiology*, 26(1), 7–13.e1.

[19] Tarca, A. L., Pataki, B. Á., Romero, R., & The DREAM Preterm Birth Prediction Challenge Consortium (2021). Crowdsourcing assessment of maternal blood multi-omics for predicting gestational age and preterm birth. *Cell Reports Medicine*, 2(6), 100323.

[20] Weber, A., Darmstadt, G. L., Gruber, S., Foeller, M. E., Carmichael, S. L., Stevenson, D. K., & Shaw, G. M. (2018). Application of machine-learning to predict early spontaneous preterm birth among nulliparous non-hispanic black and white women. *Annals of Epidemiology*, 28(11), 783–789.

[21] Woo, J. et al. (2023). Vitamin d status as an important predictor of preterm birth in a cohort of black women. *Nutrients*, 15(21), 4637.

[22] Woo, J. et al. (2025). Gene expression differences based on low total 25(oh)d and low vdbp status with a preterm birth. *International Journal of Molecular Sciences*, 26(10), 4475.

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. The above illustration, "Science Experiment 02", was created by Ben Schlitter and released under CC BY-NC-ND 3.0. A template that can be used to format a PhD thesis with this look and feel has been released under the permissive MIT (X11) license, and can be found online at github.com/suchow/Dissertate or from its author, Jordan Suchow, at suchow@post.harvard.edu.