

A Hybrid-Effects Gradient Boosting and SHAP-Interpretable XGBoost Framework for Personalized NIPT Timing and Multiclass Aneuploidy Classification in High-BMI Pregnancies

Jinghan Wang *, Guanze Lin

Dalian Foreign Languages University, Dalian, Liaoning, 116044, China

* Corresponding author: Jinghan Wang

Abstract: Background: Selecting an informative sampling time for non-invasive prenatal testing (NIPT) can be challenging in pregnancies with elevated body mass index (BMI), because fetal fraction and measurement reliability may vary across gestation. Methods: We analyzed 1,687 regional NIPT records and developed an integrated retrospective modeling framework. After gestational-age harmonization, missing-data handling, and quality screening, a gradient-boosting regression model with a mixed-effects correction was used to model male-fetus Y-chromosome fraction. Ward clustering and risk-constrained optimization produced BMI-stratified schedules; Cox modeling and bootstrap resampling assessed covariate-stratified schedules. For female fetuses, an XGBoost classifier with SMOTE balancing and SHAP interpretation was trained for five-class aneuploidy screening. Results: The male-fetus model achieved mean five-fold validation R^2 of 0.85 and threshold-classification accuracy of 92.3%. The optimized four-stratum schedule recommended testing at 10.7, 13.3, 17.0, and 20.8 weeks across increasing BMI strata, reducing the reported expected risk by 32% relative to the initial grouping. The female-fetus classifier achieved 94.2% accuracy and a multiclass AUC of 0.97; Z21, Z18, and Z13 were the dominant explanatory features. Conclusions: The framework links timing prediction, risk-aware stratification, and interpretable classification using one retrospective dataset. Prospective, multi-center validation is required before clinical deployment.

Keywords: Non-invasive Prenatal Testing; Fetal Fraction; Body Mass Index; Risk-aware Timing; XGBoost; SHAP.

1. Introduction

Non-invasive prenatal testing (NIPT) uses circulating cell-free DNA to screen for common fetal aneuploidies. In the present dataset, the male-fetus Y-chromosome fraction was used as an operational marker of whether a sample met the study threshold ($Y \geq 4\%$). A fixed sampling week can be suboptimal when maternal BMI and other covariates are associated with the timing of threshold attainment.

This retrospective study reorganizes the original modeling work into an integrated research article. Its objective was to develop a transparent decision-support framework that (i) models the nonlinear relation between gestational age, BMI, and Y-chromosome fraction; (ii) derives risk-aware sampling schedules; and (iii) classifies female-fetus aneuploidy patterns with feature-level explanations. The intended contribution is a reproducible analytic framework, not a replacement for diagnostic testing or clinical judgment.

NIPT is a screening procedure rather than a diagnostic test. Its analytical utility depends on both fetal fraction and the reliability of chromosome-specific signals; a non-informative result can therefore introduce repeat sampling, delay counselling, and uncertainty in pregnancy management. Maternal adiposity is particularly relevant because higher BMI is associated with lower fetal fraction in many cohorts, although the magnitude and timing of this relationship can vary by assay, population, and fetal-fraction estimator.

Most operational pathways apply a fixed gestational-age rule and then manage low-fetal-fraction results reactively. That workflow overlooks the fact that threshold attainment is a time-to-event problem with covariate-dependent uncertainty.

It also does not provide an interpretable connection between a timing recommendation for male pregnancies and the chromosome-signal classification task in female pregnancies.

The present work reframes the competition study as a retrospective modeling investigation. We asked whether a harmonized analytic pipeline could predict the male-fetus Y-chromosome fraction, transform predictions into risk-aware BMI-specific timing recommendations, and summarize female-fetus aneuploidy-screening signals using an explainable multiclass model. The aim is methodological: to document a decision-support framework and its evidence boundaries, not to establish a clinical standard of care.

The original sequential problem statements, algorithm descriptions, sensitivity checks, and model evaluations were merged into an integrated Methods-Results narrative so that each claim can be read alongside its model, reported validation evidence, and limitation.

2. Related Work

The clinical and analytical foundations of NIPT, fetal-fraction measurement, BMI-associated no-call risk, and aneuploidy screening have been extensively studied [1-12]. The present work does not claim a new clinical threshold; it assembles these considerations into a transparent retrospective timing-and-classification framework.

The modeling design draws on established methods for imputation, boosting, survival analysis, resampling, imbalance handling, attribution, and prediction-model evaluation [13-25]. These sources motivate the workflow and reporting standards, while the quantitative findings remain specific to the supplied regional dataset.

3. Methods

3.1. Dataset and Preprocessing

The source dataset contained 1,687 records (1,082 male-fetus and 605 female-fetus records) with 31 maternal, sequencing-quality, and chromosome-related variables. Gestational age was standardized to decimal weeks and analyses were restricted to the study window of 10.0-25.0 weeks. Low-rate missingness was imputed using multivariate imputation by chained equations, while implausible values and low-quality sequencing records were screened using statistical and clinical-logic rules. The post-screening analytic sample comprised 1,655 records (1,064 male-fetus and 591 female-fetus records).

3.2. Timing Models and Risk Optimization

For male fetuses, gradient boosting regression modeled Y-chromosome fraction using gestational age, BMI, age, and the gestational-age $\times$ BMI interaction [1]. A random intercept was then used to account for repeat measurements within individuals. Predictor selection combined correlation screening, recursive feature elimination, and variance-inflation-factor assessment. The reported tuned settings were a tree depth of 4, learning rate of 0.08, 150 estimators, and a minimum leaf size of 8.

BMI strata were first obtained by Ward hierarchical clustering and were subsequently adjusted to minimize a reported expected-risk objective while maintaining a $\geq 90\%$ probability of reaching $Y \geq 4\%$. Measurement uncertainty was propagated with 1,000 Monte Carlo draws using the model residual standard deviation (0.78%). A covariate-stratified Cox model [2] and 1,000 bootstrap replications were additionally used to evaluate the stability of schedules that included corrected BMI, age group, and in-vitro fertilization status.

3.3. Female-fetus Aneuploidy Classification

Female-fetus records were assigned to normal, trisomy 13 (T13), trisomy 18 (T18), trisomy 21 (T21), or multiple-anomaly classes. Chi-square filtering, recursive feature elimination, and SHAP confirmation reduced the candidate set to Z21, Z18, Z13, ZX, and GC21. Synthetic minority oversampling (SMOTE) was applied within the modeling workflow before fitting an XGBoost multiclass classifier [3]. SHAP values were used to summarize feature-level contributions [4].

3.4. Validation and Reporting

Five-fold cross-validation was used throughout the reported analyses. Model performance is reported as the values supplied in the source study; no re-estimation from patient-level data was performed in preparing this manuscript. All findings should therefore be interpreted as retrospective model evidence.

Study design and analysis population. This was a retrospective secondary analysis of a regional NIPT dataset. The source file contained maternal descriptors, gestational timing, sequencing-quality indicators, chromosome-related measurements, and outcome labels. Records were partitioned by fetal sex because the Y-chromosome signal was used only as an operational fetal-fraction marker in male pregnancies, whereas female pregnancies were analyzed through chromosome-specific Z-score patterns. No participant-level re-estimation was performed while preparing this manuscript;

numerical results are those supplied in the source analysis.

Data harmonization. Gestational age was transformed to decimal weeks to avoid inconsistencies between week-day representations. Variables with low-rate missingness were imputed using multivariate imputation by chained equations, and quality screening removed implausible values and records that did not satisfy the study sequencing-logic checks. The resulting analytic set contained 1,655 records, including 1,064 male-fetus and 591 female-fetus records. The source analysis reported that mean missingness of key variables decreased from 11.36% to 0.85% after preprocessing.

Predictor specification. Candidate variables were evaluated in three stages: univariable association screening, recursive feature elimination, and inspection of collinearity using variance-inflation factors. Gestational age, BMI, maternal age, and the gestational-age-by-BMI interaction were retained for the principal male-fetus prediction model. Height and weight were not simultaneously retained with BMI where redundancy was substantial. This specification prioritizes interpretability and stability rather than exhaustive feature inclusion.

Male-fetus fraction model. Gradient boosting regression was selected to accommodate nonlinear covariate effects and their interaction. The source configuration used a maximum tree depth of 4, learning rate of 0.08, 150 estimators, and minimum leaf size of 8. A random intercept was then included to account for correlation among repeated samples from the same individual. Model performance was summarized by five-fold cross-validation with R^2 , root mean square error, mean absolute error, and accuracy for classifying whether the operational Y-chromosome threshold was attained.

Risk-aware timing formulation. For each BMI stratum, the predicted Y-chromosome fraction was mapped to the earliest gestational week satisfying $Y \geq 4\%$ with an estimated probability of at least 90%. Ward hierarchical clustering provided an initial BMI partition, after which boundaries were adjusted to minimize the reported expected-risk objective. Higher weights were assigned to later testing, formalizing the trade-off between repeat testing after an early draw and delayed clinical action after a late draw.

Uncertainty and subgroup analyses. The residual standard deviation from the prediction model (0.78%) was propagated through 1,000 Monte Carlo draws. A twofold-residual-error scenario was included as a stress test. In parallel, a covariate-stratified Cox model treated threshold attainment as the event and observations that did not attain the threshold within the 10-25-week study window as right censored. Corrected BMI, maternal age group, and in-vitro fertilization status were assessed in the source framework; 1,000 bootstrap replications summarized schedule stability.

Female-fetus classification. Female-fetus records were categorized as normal, trisomy 13, trisomy 18, trisomy 21, or multiple anomalies. A sequence of chi-square filtering, recursive feature elimination, and SHAP-based confirmation retained Z21, Z18, Z13, ZX, and GC21. Because outcome classes were imbalanced, synthetic minority oversampling (SMOTE; five nearest neighbours) was applied within the modeling workflow before XGBoost fitting. The model was evaluated with five-fold cross-validation using accuracy, precision, recall, F1 score, and one-versus-rest area under the receiver operating characteristic curve.

Interpretability and reporting. SHAP values were used to summarize the direction-independent magnitude of each retained feature's contribution. The attribution analysis is

descriptive: it identifies which model inputs drove predictions, but it does not demonstrate biological causality. All results are consequently reported as retrospective model evidence. The study did not include a held-out external cohort, prospective clinical outcomes, or an independent recalibration of the operating threshold.

4. Results and Discussion

4.1. Data Quality and Male-fetus Prediction

Preprocessing reduced the mean missingness of key

variables from 11.36% to 0.85% and increased the reported data-consistency score from 82.1% to 95.3%. Gestational age was positively correlated with Y-chromosome fraction ($r = 0.32$, $p < 0.001$), whereas BMI ($r = -0.28$, $p < 0.001$) and age ($r = -0.21$, $p < 0.001$) were negatively correlated. The mixed-effects gradient-boosting model obtained $R^2 = 0.85 \pm 0.03$, RMSE = $0.78 \pm 0.06\%$, and threshold-classification accuracy = $92.3 \pm 1.5\%$ in five-fold validation.

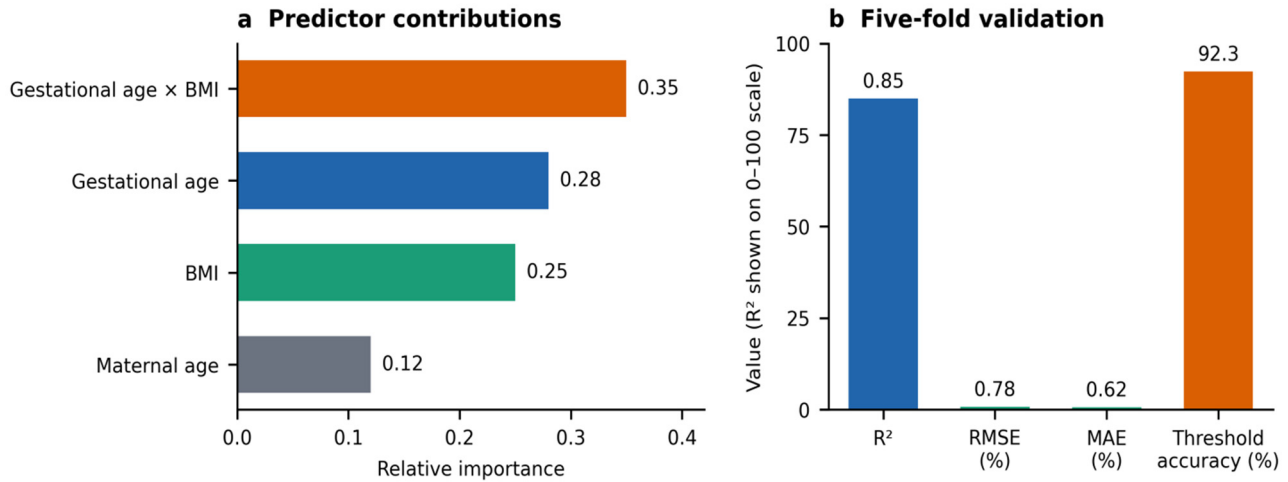


Figure 1. Predictors and validation evidence for the male-fetus fetal-fraction model. Panel A shows the reported relative predictor importance; panel B summarizes five-fold validation metrics. R^2 is plotted on a 0-100 scale solely for visual comparability.

4.2. Risk-aware Schedules

The optimized four-stratum scheme reduced the source study's total expected risk from 4.2 to 3.0 (32%). Recommended sampling weeks increased monotonically

across BMI strata. Under a twofold residual-error scenario, the predicted probability of $Y \geq 4\%$ declined most strongly in the highest BMI stratum, motivating an adjusted recommendation of 21.5 weeks in that scenario.

Table 1. Optimized BMI-stratified sampling schedule and uncertainty sensitivity.

BMI stratum (kg/m ²)	n	Recommended week	P ($Y \geq 4\%$) at recommendation	Expected risk	P at $2\times$ error
27.5-<30.0	265	10.7	92.5%	1.0	88.0%
30.0-<33.5	338	13.3	91.8%	4.8	86.2%
33.5-<37.0	302	17.0	90.5%	5.0	84.5%
≥ 37.0	149	20.8	90.2%	5.0	80.4%

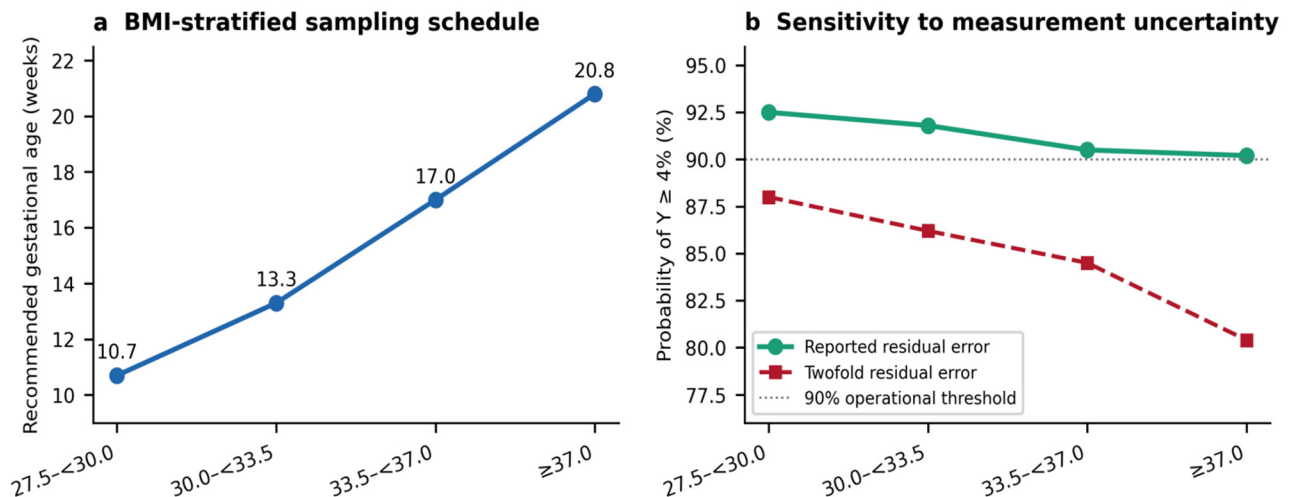


Figure 2. BMI-stratified timing recommendations and sensitivity of threshold attainment to measurement uncertainty. The dotted line denotes the 90% operational threshold used in the source optimization.

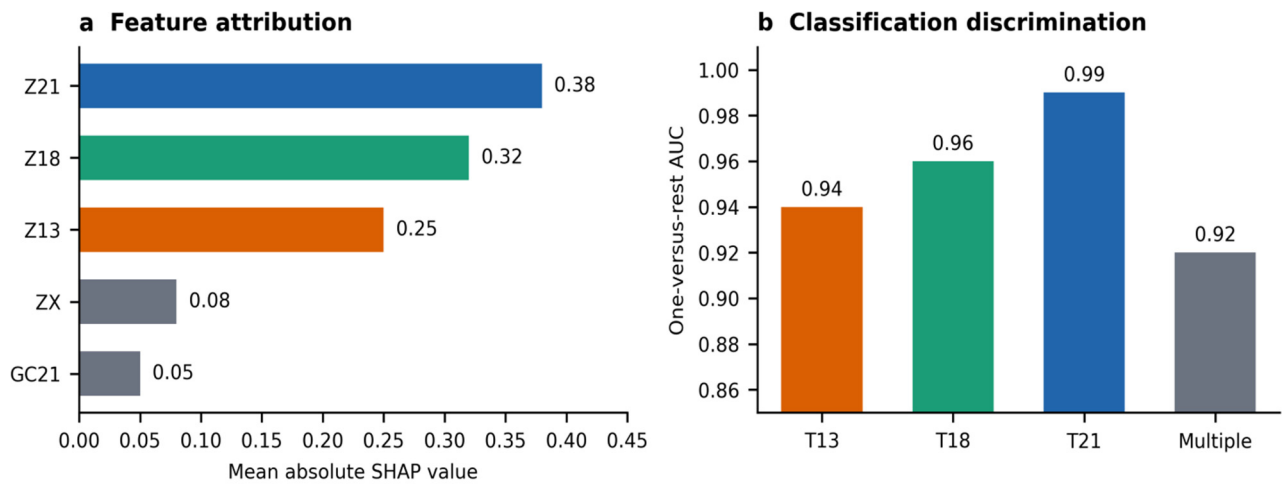


Figure 3. Explainable multiclass classification of female-fetus screening records. Panel A ranks the five retained features by mean absolute SHAP value; panel B reports class-specific one-versus-rest AUC.

Among 591 female-fetus records, 68 were labeled as anomalous in the source data. The XGBoost-SHAP model achieved 94.2% overall accuracy, 88.7% precision, 91.3% recall, and multiclass AUC of 0.97. Z21, Z18, and Z13 accounted for 88.1% of the mean absolute SHAP contribution, with T21 showing the strongest class-specific discrimination (AUC 0.99).

Cohort and data quality. The original study included 1,687 records (1,082 male-fetus and 605 female-fetus records)

across 31 variables. Following preprocessing and quality screening, 32 records were excluded. Reported associations in the male-fetus subset showed a positive relation between gestational age and Y-chromosome fraction ($r = 0.32$, $p < 0.001$) and negative relations for BMI ($r = -0.28$, $p < 0.001$) and maternal age ($r = -0.21$, $p < 0.001$). These associations motivated an explicitly nonlinear, interaction-aware timing model rather than a single fixed sampling week.

Table 2. Key classification results for the female-fetus XGBoost-SHAP model.

Metric	Overall	T13	T18	T21	Multiple anomalies
Accuracy (%)	94.2	88.5	92.3	98.1	86.7
Precision (%)	88.7	82.1	87.5	93.8	78.3
Recall (%)	91.3	85.7	90.0	96.8	83.3
F1 score	0.90	0.84	0.89	0.95	0.81
AUC	0.97	0.94	0.96	0.99	0.92

Male-fetus prediction performance. In five-fold cross-validation, the mixed-effects gradient-boosting model achieved $R^2 = 0.85 \pm 0.03$, $RMSE = 0.78 \pm 0.06\%$, $MAE = 0.62\%$, and threshold-classification accuracy of $92.3 \pm 1.5\%$. The gestational-age-by-BMI interaction had the greatest reported relative importance (0.35), followed by gestational age (0.28), BMI (0.25), and maternal age (0.12). This pattern indicates that the predicted timing shift was not attributable to BMI alone, but to its joint behavior with gestational progress.

BMI-stratified timing recommendations. The optimized four-stratum solution recommended testing at 10.7, 13.3, 17.0, and 20.8 weeks for BMI ranges $27.5 < 30.0$, $30.0 < 33.5$, $33.5 < 37.0$, and ≥ 37.0 kg/m², respectively. At those weeks, the reported probability of attaining $Y \geq 4\%$ ranged from 90.2% to 92.5%. In the source objective, aggregate expected risk decreased from 4.2 to 3.0, corresponding to a reported 32% reduction relative to the initial grouping.

Sensitivity to residual uncertainty. The Monte Carlo stress test indicated that timing recommendations in lower BMI strata were relatively robust to modeled residual error, whereas the upper BMI stratum showed the largest reduction in threshold-attainment probability. Under the twofold-residual-error scenario, probability at the recommended week declined from 88.0% in the first stratum to 80.4% in the highest stratum. The source analysis consequently proposed a

later adjusted recommendation of 21.5 weeks for the highest-BMI group under that stressed scenario.

Covariate-stratified stability. Cox and bootstrap analyses inspected whether the schedule remained stable after incorporating corrected BMI, age, and assisted-reproduction status. The source study reported bootstrap 95% confidence-interval coverage of $95.2 \pm 2.3\%$ and an estimated risk-reduction stability improvement of 25%. These values should be interpreted as internal resampling evidence rather than evidence of transportability, because the model was not tested on an independent external cohort.

Female-fetus screening performance. Among 591 female-fetus records after quality screening, 68 were labeled anomalous in the source dataset. The XGBoost-SHAP model yielded 94.2% overall accuracy, 88.7% precision, 91.3% recall, and multiclass AUC of 0.97. The T21 class showed the strongest one-versus-rest discrimination (AUC 0.99), followed by T18 (0.96), T13 (0.94), and multiple-anomaly labels (0.92).

Feature attribution. Chromosome-specific Z values were the dominant retained signals: Z21, Z18, and Z13 together accounted for 88.1% of the mean absolute SHAP contribution in the reported analysis. ZX and GC21 made smaller additional contributions. The ranking is consistent with the classifier's intended use as a screening model that distinguishes chromosome-signal patterns, but it should not

be read as a causal decomposition of aneuploidy risk.

Integrated evidence. Across the analytic modules, performance was strongest for the internal male-fetus prediction and female-fetus classification tasks. The scheduling component adds decision relevance by linking prediction to an operational 90% threshold, whereas the stress test makes its vulnerability visible. The consolidated evidence therefore supports feasibility of the framework within the source cohort, while identifying residual uncertainty, class imbalance, and external validity as the key constraints.

5. Discussion

The results support the central premise that sampling-time

selection can be formulated as a joint prediction-and-risk problem rather than a fixed-week rule. The reported interaction importance for gestational age and BMI, together with the monotonic shift in recommended timing across BMI strata, provides a quantitative rationale for individualized scheduling within this dataset.

The framework also integrates timing support for male fetuses with interpretable female-fetus screening. The classifier's strongest features were chromosome-specific Z values, which is consistent with the feature-selection pathway encoded in the study design. SHAP summaries improve auditability by showing which model inputs were most influential, but they do not establish causality or replace confirmatory diagnostic procedures.

Table 3. Consolidated validation evidence across the analytic modules.

Module	Primary endpoint	Reported validation result	Interpretation
Male-fetus fraction model	R^2 ; RMSE	0.85 ± 0.03 ; $0.78 \pm 0.06\%$	Strong retrospective fit
BMI timing optimization	Expected risk; timing CI width	3.0 ± 0.2 ; 0.3 ± 0.1 weeks	Risk-reduced schedule
Cox timing model	Bootstrap CI coverage	$95.2 \pm 2.3\%$	Stable within source analysis
Female classifier	Accuracy; multiclass AUC	$94.2 \pm 1.5\%$; 0.97 ± 0.02	High retrospective discrimination

Several limitations require emphasis. The dataset was regional and retrospective, class counts were small for some anomaly categories, and the preprocessing and SMOTE steps were reported from the original analysis rather than independently repeated here. Moreover, the operational $Y \geq 4\%$ threshold and risk weights are study constraints; their transportability should be evaluated prospectively with external cohorts, sequencing platforms, and prespecified clinical endpoints.

The principal contribution is the connection of three tasks often considered separately: prediction of a fetal-fraction proxy, scheduling under asymmetric timing risk, and explainable female-fetus screening. The linked framework prevents the timing recommendation from being treated as a static clinical rule. Instead, it is conditional on reported covariate effects and an explicit probability threshold.

The results should not be interpreted as showing that the reported weeks are universally optimal. Fetal fraction is assay- and cohort-dependent, and even studies that agree on an inverse association with BMI differ in sampling policy, fetal-fraction estimator, and clinical handling of non-informative results. The schedule generated here is therefore best understood as a cohort-specific decision-support output whose clinical value requires prospective assessment against standard care.

The uncertainty analysis is particularly important for interpretation. The same model that supports individualized timing also predicts lower threshold-attainment probability in the highest BMI stratum when residual error is enlarged. Rather than weakening the study, this result highlights a clinically useful boundary: a scheduling recommendation should carry a quantified uncertainty statement and should not obscure the possibility of repeat testing or alternative follow-up.

The female-fetus model demonstrates high internal discrimination and provides a transparent feature ranking. However, SMOTE-based balancing and five-fold validation do not substitute for a temporally or geographically independent test set. Calibration, decision-curve performance, false-negative consequences, and the workflow for confirmatory diagnostic testing are not established by the source data. These issues are essential before any model-

facing output could be used in a clinical pathway.

Several design limitations remain. The dataset was regional and retrospective; outcome classes were small for some anomaly categories; preprocessing choices and risk weights were inherited from the original analysis; and no patient-level data were available to independently reproduce the reported metrics. In addition, the $Y \geq 4\%$ threshold was treated as an operational study constraint rather than a universal assay requirement. These limitations bound the manuscript's claims to proof-of-concept model evidence.

Future work should prospectively validate the pipeline across laboratories, sequencing platforms, and demographic settings. A preregistered evaluation could compare a model-informed draw-time recommendation with the local standard pathway, report calibration and clinical utility, and define safety-net procedures for uncertainty, no-call results, and abnormal screening results.

6. Conclusion

This study develops an interpretable NIPT decision-support framework for high-BMI pregnancies, linking a mixed-effects gradient-boosting model for risk-aware timing optimization with an XGBoost-SHAP classifier for multiclass aneuploidy screening. While retrospective results justify the analytic approach, immediate clinical implementation is not warranted. Prior to deployment, the model strictly requires external prospective validation, rigorous calibration assessments, and established clinical governance for informed consent, abnormal findings, and non-informative tests to ensure safety and clinical utility.

References

- [1] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. <https://doi.org/10.1007/978-0-387-84858-7>.
- [2] Kleinbaum, D. G., & Klein, M. (2020). *Survival analysis: A self-learning text* (4th ed.). Springer. <https://doi.org/10.1007/978-3-030-37367-9>.
- [3] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD*

- International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. <https://doi.org/10.1145/2939672.2939785>.
- [4] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*. Curran Associates.
 - [5] Wang, E., et al. (2018). Fetal fraction evaluation in non-invasive prenatal screening. *Prenatal Diagnosis*.
 - [6] Gerson, K. D., et al. (2015). Factors affecting levels of circulating cell-free fetal DNA in maternal plasma and their implications for noninvasive prenatal testing. *Prenatal Diagnosis*.
 - [7] Beulen, L., et al. (2017). Comparing methods for fetal fraction determination and quality control of NIPT samples. *Prenatal Diagnosis*, 37, 769–773.
 - [8] Huang, Y., et al. (2023). Maternal and fetal factors influencing fetal fraction: A retrospective analysis of 153,306 pregnant women undergoing noninvasive prenatal screening. *Frontiers in Genetics*.
 - [9] van Prooyen Schuurman, L., et al. (2024). Insights into non-informative results from non-invasive prenatal screening through gestational age, maternal BMI, and age analyses. *Prenatal Diagnosis*.
 - [10] American College of Obstetricians and Gynecologists. (2020). Screening for fetal chromosomal abnormalities. *Obstetrics & Gynecology*, 136, e48–e69.
 - [11] Gregg, A. R., et al. (2016). Noninvasive prenatal screening by next-generation sequencing. *Genetics in Medicine*, 18, 1056–1065.
 - [12] Kilpatrick, S. J., et al. (2015). Current guidance on prenatal screening. *Obstetrics & Gynecology*, 125, 653–662.
 - [13] van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45, 1–67.
 - [14] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29, 1189–1232.
 - [15] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34, 187–220.
 - [16] Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC Press.
 - [17] Chawla, N. V., et al. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.
 - [18] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. <https://doi.org/10.1145/2939672.2939785>.
 - [19] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*. Curran Associates.
 - [20] Collins, G. S., et al. (2015). TRIPOD statement. *Annals of Internal Medicine*, 162, 55–63.
 - [21] Wolff, R. F., et al. (2019). PROBAST. *Annals of Internal Medicine*, 170, 51–58.
 - [22] Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a ROC curve. *Radiology*, 143, 29–36.
 - [23] DeLong, E. R., et al. (1988). Comparing areas under correlated ROC curves. *Biometrics*, 44, 837–845.
 - [24] Steyerberg, E. W., et al. (2010). Assessing prediction-model performance. *Epidemiology*, 21, 128–138.
 - [25] Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis. *Medical Decision Making*, 26, 565–574.