Genetics and Population Analysis

# Enhanced Fetal Fraction Estimation in Non-Invasive Prenatal Testing via Advanced Imputation and Deep Learning Ensembles

## Matthias De Smet [*,1,2]

[1] Department of Biomolecular Medicine, Ghent University, Belgium
[2] Center for Medical Genetics Ghent, Ghent University Hospital, Belgium

[*] To whom correspondence should be addressed.

Associate Editor: Prof. Yvan Saeys

## Abstract

**Motivation:** The clinical reliability of Non-Invasive Prenatal Testing (NIPT) is fundamentally dependent on the accurate estimation of fetal fraction (FF). While shallow whole-genome sequencing (sWGS) is the standard for aneuploidy screening, estimating FF solely from low-pass coverage data remains challenging due to stochastic sparsity and signal noise.

**Methods:** We evaluated an architectural pipeline incorporating K-Nearest Neighbors (KNN) imputation to address missingness in 100kb genomic coverage bins derived from sWGS. The feature space was processed through Principal Component Analysis (PCA), and we systematically benchmarked an ensemble of homogeneous learners, including Support Vector Machines (SVM), XGBoost, and Neural Networks (NN), optimized via Bayesian search (Optuna). To ensure clinical readiness and interoperability, final ensembles were exported using the Open Neural Network Exchange (ONNX) format.

**Results:** Systematic benchmarking revealed that coverage-based genomic bins lack sufficient biological signal for robust FF estimation. Principal Component Analysis indicated that the first component captured only 1.4% of the total variance, reflecting a feature space dominated by stochastic noise. Consequently, while the KNN-imputed Neural Network ensemble achieved a Mean Absolute Error (MAE) of 4.66%, it yielded a negative coefficient of determination ($R^2 = -4.96$). This metric indicates that the model failed to outperform a null baseline and that KNN imputation on high-dimensional raw data did not successfully recover a predictive biological manifold.

**Conclusion:** Our results demonstrate the limitations of using sWGS read counts as the sole input for fetal fraction regression. We conclude that advanced imputation cannot compensate for the lack of other data modalities, and that future deep learning architectures for NIPT must integrate these to achieve clinical utility.

**Availability:** The updated PREFACE pipeline and test data are available at GitHub.

**Contact:** matthias.desmet@ugent.be.

**Supplementary information:** Supplementary data are available at request from the author.

## 1 Introduction

The clinical utility of Non-Invasive Prenatal Testing (NIPT) relies upon the precise quantification of the fetal fraction (FF), defined as the proportion of cell-free DNA (cfDNA) in maternal plasma originating from the placenta. Fetal fractions falling below a 4% threshold significantly increase the probability of false-negative results [2], often leading to inconclusive clinical findings that necessitate invasive follow-up procedures. Although various computational methods exist for FF estimation [9, 12, 6, 10] the PREFACE (PREdict FetAl ComponEnt) [18] framework established a significant precedent by utilizing shallow whole-genome sequencing (sWGS) and neural networks to predict FF in a gender-independent manner. As NIPT protocols transition toward lower-pass sequencing to increase throughput, the resulting data are increasingly characterized by stochastic sparsity. In typical sWGS (0.1x–1x coverage), genomic bins often exhibit high levels of missingness or extreme noise. Traditional handling methods, such as zero-filling or simple mean imputation, can inadvertently mask the biological variance required for accurate prediction. Principal Component Analysis (PCA)[16] is particularly sensitive to these shifts. If the underlying variance is flattened by simplistic imputation, the resulting principal components lose the subtle signals associated with cfDNA fragmentation patterns. This necessitates the adoption of advanced imputation strategies, such as K-Nearest Neighbors (KNN)[22], to preserve the manifold structure of the data prior to dimensionality reduction. In this context, the manifold refers to the assumption that high-dimensional genomic coverage data actually lies on a lower-dimensional subspace defined by biological constraints, such as nucleosome positioning and fragmentation signatures. By using KNN, we ensure that missing data points are estimated based on their proximity to other samples within this biological subspace, rather than being pulled toward a non-biological global average. Furthermore, the transition from single-model architectures to homogeneous ensemble learning represents a shift toward higher model reliability. Single neural networks are prone to overfitting in high-dimensional, low sample size (HDLSS) contexts, where they may memorize batch-specific artifacts. By employing a GroupShuffleSplit cross-validation framework[17], an ensemble of networks can be trained on disparate data folds defined by clinical groups, such as sequencing runs. This strategy leverages the Diversity Prediction Theorem [11, 14], ensuring that the final prediction generalizes across between-laboratory biases and provides a stable estimator for diverse clinical populations. Finally, shifting to a Python-based ecosystem facilitates the use of production-grade frameworks like PyTorch[15] and the ONNX[5] format. This ensures the updated PREFACE framework is not only more accurate but also fully interoperable with cloud-native clinical pipelines. In this study, we systematically benchmark 12 combinations of advanced imputers and regression learners to demonstrate the precision of KNN-imputed deep-learning ensembles in high-throughput prenatal diagnostics.

## 2 Methods

### 2.1 Data preprocessing

The primary input data consisted of a sample registry containing metadata and corresponding genomic coverage files. Genomic coverage was derived from sWGS data processed via WisecondorX[19], which segments the genome into 100kb fixed-width bins. For each bin, a $\log_2$-transformed copy number ratio was calculated relative to a reference panel.

$$\text{Ratio}_{bin} = \log_2\left(\frac{\text{Observed Reads}}{\text{Expected Reads}}\right) \qquad (1)$$

A ratio of 0 indicates a normal diploid state, positive values indicate duplications, and negative values indicate deletions. For the purpose of FF estimation, the model uses these profile-corrected ratios as the raw features. Data integration was performed by constructing a feature matrix $X \in \mathbb{R}^{N \times M}$, where $N = 186$ represents the euploid samples (89 male, 97 female) and $M$ represents the genomic bins ($M = 3 \times 10^4$). To ensure the model learned features associated with constitutional fetal fraction rather than large-scale numerical abnormalities, bins residing on chromosomes 13 (Down syndrome), 18 (Edwards syndrome), 21 (Patau syndrome), X, and Y (Sex chromosomal aneuploidies (SCAs)) were excluded. This prevents confounding by common autosomal trisomies. Furthermore, genomic bins with a missingness rate exceeding 1% were removed to eliminate regions affected by systematic sequencing artefacts, ensuring the remaining feature set was suitable for downstream imputation.

### 2.2 Cross validation

To assess model generalizability and mitigate overfitting risks in a high-dimensional feature space, we employed a GroupShuffleSplit cross-validation strategy. Within this framework, groups were defined by percentile ranges of the FF target. This stratified shuffling approach was specifically designed to ensure robust model evaluation despite the constraints of a limited sample size ($N = 186$). By partitioning the data based on FF percentiles, we ensured that each test set remained a statistically representative subset of the original population's phenotypic distribution. This methodology prevents systemic bias inherent in random splits, where an accidental over-concentration of median cases in the test set can lead to artificial fluctuations in performance metrics. Furthermore, this approach addresses the instability often associated with standard $K$-fold partitions in small genomic cohorts. By utilizing GroupShuffleSplit to perform Monte Carlo resampling across multiple iterations, the architecture was subjected to a rigorous assessment across the entire spectrum of patient profiles. The resulting MAE and $R^2$ values provide stable, highly reproducible metrics of clinical utility.
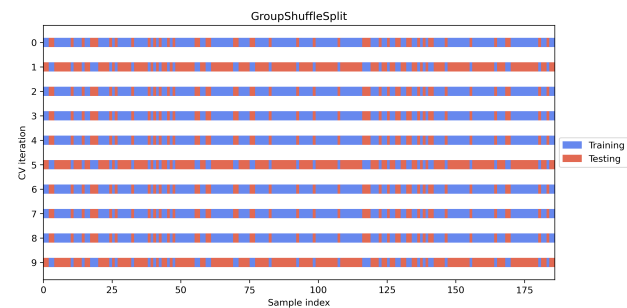


**Fig. 1.** Example of dataset splits by GroupShuffleSplit

### 2.3 Dimensionality reduction

The imputed feature matrix contains approximately $3 \times 10^4$ genomic bins ($M$) for 186 samples ($N$). Direct application of machine learning algorithms on this $M \gg N$ dataset introduces statistical challenges inherent to high-dimensional data, potentially leading to severe overfitting. To compress the feature space while retaining biological signal, we employed PCA. While non-linear manifold learning techniques such as t-SNE or UMAP are powerful for visualization, they often distort global distance metrics and feature linearity. This makes them less suitable for regression tasks where preserving the continuous, additive magnitude of the signal is crucial. PCA linearly transforms the correlated genomic bin counts into a set of orthogonal principal components. The top components typically capture systematic biological variation correlating with the fetal

fraction, while lower-order components represent stochastic sequencing noise. We selected the top $k$ components that explained 95% of the total variance.

## 2.4 Missing Data and Imputation

In high-dimensional genomic datasets, missing values often arise from technical filtering or stochastic dropout. Understanding the mechanism of this missingness is prerequisite to selecting an appropriate imputation strategy. The statistical literature typically classifies missingness into three categories[20]:

- Missing Completely at Random (MCAR): The probability of a data point being missing is unrelated to any values in the dataset. In sequencing, this might correspond to a random bubble in a flow cell or a transient optical error.
- Missing at Random (MAR): The missingness is related to observed variables. For instance, if sequencing depth (an observed variable) is low, missingness in certain bins increases.
- Missing Not at Random (MNAR): The missingness is related to the unobserved value itself. In genomics, this is common. Hard-to-sequence regions (e.g., high GC content) are more likely to be missing because of their sequence composition.

We evaluate imputation strategies under the assumption that sWGS missingness is a hybrid of MAR and MNAR. The stochastic nature of low-pass sequencing introduces random dropout (MAR), while sequence-specific biases introduce systematic dropout (MNAR). We tested several imputations strategies:

- Constant Imputation (Zero-filling): This method replaces missing values with 0.0. In the context of $\log_2$ copy number ratios, a value of 0.0 represents a balanced diploid state ($\log_2(\text{Obs}/\text{Exp}) = 0$). Consequently, this imputation strategy erroneously conflates missing data (absence of signal) with normal biological state (expected signal). By replacing missing bins with 0.0, the model effectively treats technical dropout as biological normality, artificially suppressing the variance required for anomaly detection.
- Global Mean/Median Imputation: This method replaces a missing value in bin $j$ with the mean or median of bin $j$ across all $N$ samples. While this preserves the first moment (mean) of the distribution, it severely distorts the second moment (variance) and the covariance structure. It essentially pulls the data towards the centroid of the population, reducing the spread of the data and making samples look more similar than they actually are. In the context of PCA, this variance deflation is particularly damaging because PCA seeks to maximise variance. Features with artificially reduced variance contribute less to the principal components.
- K-Nearest Neighbours (KNN) Imputation estimates missing values by averaging the values of the $k$-nearest samples in the feature space using Euclidean distance. For a given sample $x_i$ with a missing value in bin $j$, the algorithm identifies the $k$ most similar samples (neighbours) in the dataset based on Euclidean distance calculated using the observed features. The missing value is then imputed as a weighted average of the values in bin $j$ from these neighbours.

$$\hat{x}_{ij} = \frac{\sum_{n \in \mathcal{N}_k(x_i)} w_n x_{nj}}{\sum_{n \in \mathcal{N}_k(x_i)} w_n} \qquad (2)$$

where $\mathcal{N}_k(x_i)$ is the set of nearest neighbors and $w_n$ is a weight (often $1/d$ where $d$ is distance).

Although it incurs a computational penalty during inference, KNN Imputation was selected as the default strategy due to its ability to preserve local data structure and provide greater stability in high-dimensional spaces.

## 2.5 Model selection

We benchmarked three distinct algorithms for the regression task.

- Support Vector Machines (SVM)[4] with Linear Kernels were used as a baseline. They are considered effective in high-dimensional spaces but may miss complex interactions between genomic regions.
- XGBoost (Extreme Gradient Boosting)[3] is an ensemble tree-based method, where it builds decision builds sequentially, with each new tree correcting the errors of the previous ones. It handles non-linearities well but requires intensive tuning to prevent overfitting on small datasets.
- Neural Networks [8] were utilized for their capacity to model highly complex, non-linear relationships.

When dealing with small sample sizes ($N = 186$), single neural networks are prone to high variance[7]. Their weights can vary wildly based on the specific composition of the training batch, leading to overfitting. To counteract this, we utilised an ensemble of homogeneous learners, multiple instances of the same network architecture trained on different subsets of the data. This approach exploits the Diversity Prediction Theorem which states:

Collective Error = Average Individual Error − Prediction Diversity

This theorem mathematically guarantees that if the individual models in an ensemble make diverse errors (i.e., their errors are uncorrelated), the error of the ensemble average will be lower than the average error of the individual models. By training on different folds of the data (generated via GroupShuffleSplit), we ensure that the individual networks view the problem from slightly different angles, thereby maximising prediction diversity and stabilising the final output.

## 2.6 Hyperparameter Tuning

Hyperparameter optimization was conducted to identify configurations that minimized validation error. Grid search, the traditional method for hyperparameter tuning, performs an exhaustive search over a manually specified subset of the hyperparameter space. This is computationally inefficient and often misses optimal configurations located between grid points. We utilized Optuna[1], a define-by-run hyperparameter optimization framework, rather than standard GridSearchCV. Optuna employs a Tree-structured Parzen Estimator (TPE)[23].

$$EI(x) = \frac{l(x)}{g(x)} \qquad (3)$$

where $l(x)$ and $g(x)$ represent the probability density functions of the hyperparameters given that the objective function value is lower than (good) or higher than (bad) a threshold $y^*$, respectively. This Bayesian optimization algorithm models the probability of hyperparameters yielding lower validation errors, allowing the search to intelligently prune unpromising areas and converge on optimal settings more efficiently. We constructed a comprehensive grid of parameters for each model type to be evaluated as shown in tables 2.6, 2.6, and 2.6.

## 2.7 Model Evaluation

Model performance was evaluated using standard regression metrics [13], including R-squared ($R^2$), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). $R^2$ represents the proportion of variance explained by the model, while MAE provides a direct interpretation of the expected

Table 1. Hyperparameter search space for Neural Network models.

| Parameter | Range | Step Size / Scaling |
|---|---|---|
| Number of layers | $1 - 3$ | 1 |
| Learning rate | $10^{-4} - 10^{-2}$ | Log-scaled |
| Hidden size | $16 - 128$ | 16 |
| Dropout rate | $0.1 - 0.5$ | 0.1 |
| Batch size | $8 - 64$ | 8 |

Table 2. Hyperparameter search space for Support Vector Machine models.

| Parameter | Range | Step Size / Scaling |
|---|---|---|
| C | $10^-3 - 10^2$ | Log-scaled |

Table 3. Hyperparameter search space for XGBoost models.

| Parameter | Range | Step Size / Scaling |
|---|---|---|
| Number of Estimators | $100 - 800$ | 10 |
| Max Depth | $1 - 10$ | 1 |
| Learning rate | $10^{-4} - 10^{-2}$ | Log-scaled |
| Subsample Ratio | $0.5 - 1$ | 0.1 |
| Subsample Ratio | $0.5 - 1$ | 0.1 |
| Colsample By Tree | $0.5 - 1$ | 0.1 |

error in fetal fraction percentage points. RMSE penalizes large errors more heavily, making it critical for detecting outliers. High performance is characterized by an $R^2 > 0.90$ and low RMSE/MAE. A significant divergence between RMSE and MAE indicates the presence of large outlier errors, suggesting a failure to generalize to specific "hard" samples.

## 2.8 Model Distribution and Deployment

To facilitate seamless integration into clinical bioinformatics pipelines, the final trained models were exported using the ONNX format. While native Python serialization formats like `pickle` are frequently used, they present significant barriers to clinical deployment due to security vulnerabilities and strict dependencies on specific library versions. Such files are often tightly coupled to their creation environment, leading to environment-specific reproduction failures. In contrast, ONNX defines a framework-agnostic computation graph that decouples the model architecture from the training environment. This adoption provides critical benefits, including interoperability with existing clinical reporting tools, hardware acceleration to minimize inference latency, and rigorous reproducibility through "frozen" binary models.

## 3 Results

### 3.1 Dimensionality Reduction and Variance Analysis

To evaluate the underlying structure of the feature space, a Principal Component Analysis (PCA) was performed on the entire dataset. Analysis of the resulting scree plot reveals that the first principal component (PC1) accounts for approximately $1.4\%$ of the total variance. This suggests a high-dimensional feature space where information is broadly distributed rather than concentrated. The variance profile is characterized by an extremely shallow curve that plateaus rapidly after the first few components. This profile indicates that the dataset lacks a dominant signal, pointing toward high feature independence or a significant signal-to-noise ratio. Notably, this lack of dimensionality was found to be similar across all four imputation strategies. The relationship between the extracted features and the predicted fetal fraction ($FF$) was visualized by projecting the samples onto the first two principal axes ($PC1$ vs. $PC2$), with a color gradient representing $FF$ levels. The resulting plots show a complete absence of spatial clustering or gradient formation. High $FF$ (red) and low $FF$ (blue) samples are stochastically intermingled across the coordinate space, confirming that the primary axes of variation do not align with $FF$ fluctuations. Consistent with the scree analysis, the distribution of $FF$ values remained invariant across all imputation techniques. These findings suggest that the natural variation within the feature set is significantly more pronounced than any biological signal associated with $FF$ and indicate that no single bin or bins have a large influence on the predicted $FF$.

### 3.2 Model Performance and Predictive Stability

The predictive performance of the machine learning models was evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination ($R^2$). As summarized in Table 4, the XGBoost (Mean) permutation achieved the lowest overall error (MAE $= 4.61 \pm 4.11\%$). However, the Neural (KNN) model demonstrated superior stability, yielding a nearly identical mean error (MAE $= 4.66\%$) with a significantly lower standard deviation (SD $= 2.88$). A critical observation was the failure of neural architectures to converge when utilizing naive imputation strategies. While KNN-based imputation preserved model stability, the introduction of zero-filling (Neural Zero) resulted in a performance collapse, with the $R^2$ mean falling to $-18.77 \pm 16.69$. This catastrophic metric indicates that the model predictions were uncalibrated and uncorrelated with the target variable. An $R^2$ of this magnitude signifies that the optimization process failed to find a valid solution in the feature space, performing significantly worse than a null regressor that simply predicts the mean. Conversely, the SVM models exhibited complete invariance to the choice of imputation, maintaining a consistent but high error rate (MAE $= 8.39 \pm 2.73\%$), which suggests that the support vectors
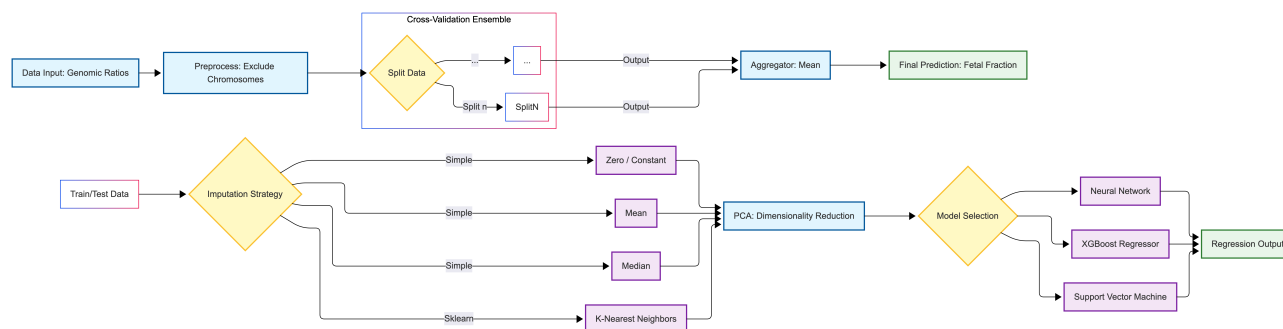


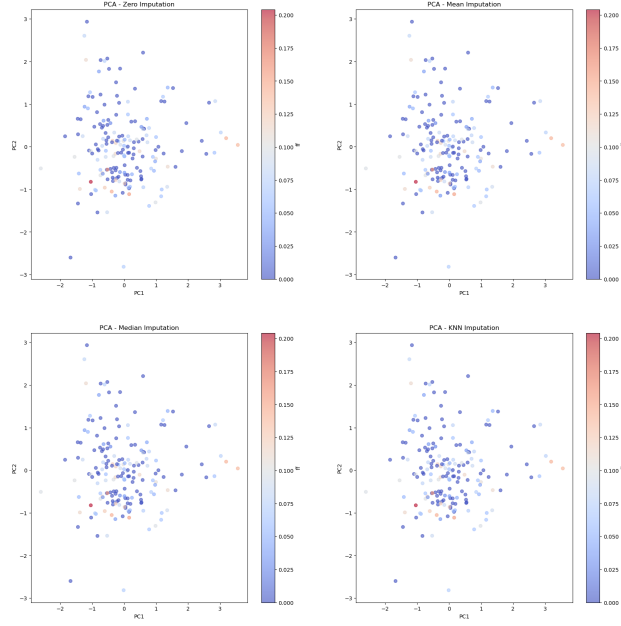**Fig. 2.** Diagram of the PREFACE workflow

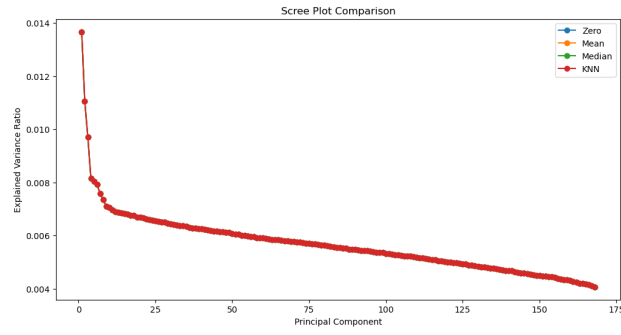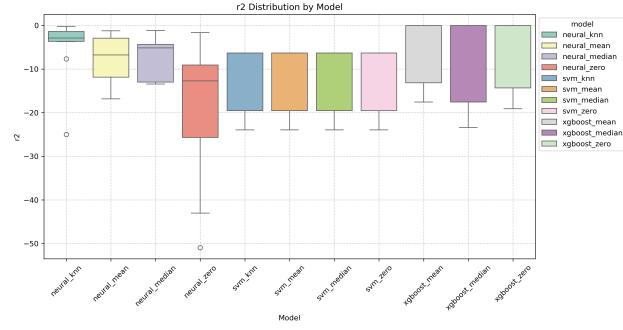**Fig. 3.** PCA Projection by Predicted Fetal Fraction



**Fig. 4.** Explained Variance Ratio per Principal Component

were unable to capture a meaningful signal within the high-dimensional noise.

Table 4. Predictive Performance Metrics with Variability. Results are reported as Mean $\pm$ Standard Deviation (SD) across permutations.

| Permutation | MAE (%) | RMSE (%) | $R^2$ |
|---|---|---|---|
| *Neural Network* | | | |
| KNN | $4.66 \pm 2.88$ | $5.73 \pm 3.12$ | $-4.96 \pm 7.35$ |
| Mean | $6.58 \pm 3.00$ | $7.41 \pm 2.69$ | $-7.83 \pm 5.56$ |
| Median | $6.20 \pm 2.56$ | $7.33 \pm 2.42$ | $-7.48 \pm 4.96$ |
| Zero | $10.26 \pm 5.04$ | $10.79 \pm 4.82$ | $-18.77 \pm 16.69$ |
| *SVM* | | | |
| All Methods* | $8.39 \pm 2.73$ | $8.82 \pm 2.61$ | $-11.62 \pm 8.50$ |
| *XGBoost* | | | |
| Mean | $4.61 \pm 4.11$ | $5.12 \pm 3.96$ | $-5.26 \pm 8.48$ |
| Median | $5.12 \pm 4.89$ | $5.60 \pm 4.74$ | $-7.02 \pm 11.31$ |
| Zero | $4.77 \pm 4.32$ | $5.25 \pm 4.18$ | $-5.73 \pm 9.23$ |

*SVM results were identical across all imputation strategies.



**Fig. 5.** Boxplot of $R^2$ variability across permutations

## 4 Discussion

The primary finding of this study is that count-based genomic features derived from sWGS, when binned at 100kb resolution, lack sufficient biological signal to drive robust fetal fraction ($FF$) estimation. In the $PC1$ vs. $PC2$ coordinate space, samples with high and low $FF$ levels were stochastically intermingled, with no visible clustering or gradient formation. This observation, coupled with the flat profile observed in the scree analysis, indicates that the variance in the dataset is dominated by stochastic noise rather than the biological signal of fetal aneuploidy or fraction. Consequently, the negative $R^2$ values observed across all regression permutations should be interpreted as a failure of the feature engineering limits rather than a model convergence failure. The models effectively performed worse than a null regressor because they attempted to fit patterns in what was essentially stochastic noise. A significant methodological limitation identified in this workflow is the application of K-Nearest Neighbors (KNN) imputation directly on the high-dimensional feature space ($M \gg N$). In such high-dimensional settings, Euclidean distance metrics converge, rendering the identification of nearest neighbors biologically meaningless due to the curse of dimensionality. This likely introduced noise amplification during the imputation step, corrupting the covariance structure necessary for PCA. To address this, future work must evaluate multivariate iterative imputation strategies, such as Multiple Imputation by Chained Equations (MICE)[24] or Random Forest-based imputation (MissForest[21]). Unlike simple distance-based metrics, these iterative imputers model the conditional distribution of missing variables based on the observed data structure. Implementing regularized versions of MICE capable of handling high-dimensional input pre-PCA could potentially recover the latent biological manifold that was lost in the current KNN-based approach.

## 5 Conclusion

This study underscores the fundamental limits of using shallow whole-genome sequencing (sWGS) coverage bins as the sole predictor for fetal fraction estimation. Contrary to our initial hypothesis, our benchmarking reveals that advanced imputation strategies, such as K-Nearest Neighbors (KNN), cannot recover a predictive manifold from high-dimensional count data when the underlying biological signal is obscured by stochastic sequencing noise. The persistent negative $R^2$ values across all ensemble architectures demonstrate that 100kb coverage bins lack the necessary information density for robust regression in this context. We conclude that future improvements in NIPT bioinformatics cannot rely on increasingly complex ensemble models applied to sparse coverage data. Instead, the focus must shift to feature engineering that captures multi-modal biological signals. Specifically, future workflows should integrate

fragment size distributions (fragmentomics) and methylation signatures, which have been shown to correlate more strongly with fetal DNA fraction than read counts alone. Additionally, we recommend replacing distance-based imputation with multivariate iterative strategies, such as Multiple Imputation by Chained Equations (MICE), to better handle the sparsity inherent in low-pass sequencing. Finally, reliance on high-depth chromosome Y counting in male fetuses remains the gold standard for ground-truth generation, and expanding these reference cohorts is essential for reducing label noise in training datasets.

## Acknowledgments

## Funding

## References

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase et al. Optuna: A Next-generation Hyperparameter Optimization Framework, 2019.

[2] Jacob A. Canick, Glenn E. Palomaki, Edward M. Kloza et al. The impact of maternal plasma DNA fetal fraction on next generation sequencing tests for common fetal aneuploidies. *Prenat Diagn*, 33 (7):667–674, July 2013. ISSN 1097-0223. doi: 10.1002/pd.4126.

[3] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, August 2016. doi: 10.1145/2939672.2939785.

[4] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach Learn*, 20(3):273–297, September 1995. ISSN 1573-0565. doi: 10.1007/BF00994018.

[5] ONNX Runtime Developers. ONNX Runtime, 2021.

[6] Véronique Duboc, David Pratella, Marco Milanesio et al. NiPTUNE: An automated pipeline for noninvasive prenatal testing in an accurate, integrative and flexible framework. *Brief Bioinform*, 23(1):bbab380, January 2022. ISSN 1477-4054. doi: 10.1093/bib/bbab380.

[7] Stuart Geman, Elie Bienenstock, and René Doursat. Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, 4(1):1–58, January 1992. ISSN 0899-7667, 1530-888X. doi: 10.1162/neco.1992.4.1.1.

[8] Simon S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, Upper Saddle River, NJ, 2. ed., [nachdr.] edition, 1999. ISBN 978-0-13-273350-2.

[9] Dongin Kim, Ji Yeon Sohn, Jin Hee Cho et al. KF-NIPT: K-mer and fetal fraction-based estimation of chromosomal anomaly from NIPT data. *BMC Bioinformatics*, 26:133, May 2025. ISSN 1471-2105. doi: 10.1186/s12859-025-06127-y.

[10] Sung K. Kim, Gregory Hannum, Jennifer Geis et al. Determination of fetal DNA fraction from the plasma of pregnant women using sequence read counts. *Prenat Diagn*, 35(8):810–815, August 2015. ISSN 1097-0223. doi: 10.1002/pd.4615.

[11] Anders Krogh and Jesper Vedelsby. Neural Network Ensembles, Cross Validation, and Active Learning. In *Advances in Neural Information Processing Systems*, volume 7. MIT Press, 1994.

[12] Tom Mokveld, Zaid Al-Ars, Erik A. Sistermans et al. WisecondorFF: Improved Fetal Aneuploidy Detection from Shallow WGS through Fragment Length Analysis. *Diagnostics (Basel)*, 12(1):59, December 2021. ISSN 2075-4418. doi: 10.3390/diagnostics12010059.

[13] Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to Linear Regression Analysis*. Wiley Series in Probability and Statistics. Wiley, Hoboken, New Jersey, sixth edition edition, 2021. ISBN 978-1-119-57872-7.

[14] Scott Page. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton Classics Series. Princeton University Press, Princeton, 1st ed edition, 2025. ISBN 978-0-691-27175-0.

[15] Adam Paszke, Sam Gross, Francisco Massa et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library, December 2019.

[16] Karl Pearson. LIII. *On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, November 1901. ISSN 1941-5982, 1941-5990. doi: 10.1080/14786440109462720.

[17] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort et al. Scikit-learn: Machine Learning in Python, June 2018.

[18] Lennart Raman, Machteld Baetens, Matthias De Smet et al. PREFACE: In silico pipeline for accurate cell-free fetal DNA fraction prediction. *Prenat Diagn*, 39(10):925–933, September 2019. ISSN 1097-0223. doi: 10.1002/pd.5508.

[19] Lennart Raman, Annelies Dheedene, Matthias De Smet et al. WisecondorX: Improved copy number detection for routine shallow whole-genome sequencing. *Nucleic Acids Res*, 47(4):1605–1614, February 2019. ISSN 1362-4962. doi: 10.1093/nar/gky1263.

[20] Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/63.3.581.

[21] Daniel J. Stekhoven and Peter Bühlmann. MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28 (1):112–118, January 2012. ISSN 1367-4811. doi: 10.1093/bioinformatics/btr597.

[22] O. Troyanskaya, M. Cantor, G. Sherlock et al. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, June 2001. ISSN 1367-4803. doi: 10.1093/bioinformatics/17.6.520.

[23] Shuhei Watanabe. Tree-Structured Parzen Estimator: Understanding Its Algorithm Components and Their Roles for Better Empirical Performance, 2023.

[24] Ian R. White, Patrick Royston, and Angela M. Wood. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med*, 30(4):377–399, February 2011. ISSN 1097-0258. doi: 10.1002/sim.4067.