



A Methodological Evaluation of Machine Learning Models and Explainability Techniques for Spontaneous Abortion Using a Public Benchmark Dataset

Aswan Supriyadi Sunge^{1*}, Harco Leslie Hendric Spits Warnars², Maybin K. Muyebea³, Sri Rahayu⁴,
Selly Septina⁵, Arief Teguh Nugraha⁶

¹ Informatics Engineering Department, Faculty of Engineering, Pelita Bangsa University, Bekasi 17530, Indonesia

² Computer Science Department, BINUS Graduate Program, Doctor of Computer Science (DCS), Bina Nusantara University, Jakarta 11480, Indonesia

³ School of Science, Engineering & Environment (SEE), University of Salford, Salford M5 4WT, United Kingdom

⁴ Poltekkes Kemenkes Denpasar, Midwifery Department, Bali Province, Denpasar 80224, Indonesia

⁵ Department of Obstetri and Gynecology, Faculty of Medical, Yarsi University, Jakarta 10510, Indonesia

⁶ Department of Management, Faculty of Economics and Business, Pelita Bangsa University, Bekasi 17530, Indonesia

Corresponding Author Email: aswan.sunge@pelitabangsa.ac.id

Copyright: ©2026 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ijssse.160807>

ABSTRACT

Received: 22 June 2026

Revised: 5 August 2026

Accepted: 14 August 2026

Available online: 31 August 2026

Keywords:

spontaneous abortion, maternal health, Machine Learning, interpretable machine learning, statistical validation, multi-parameter data

Spontaneous abortion continues to pose serious challenges in maternal healthcare because reliable identification of women at increased risk remains difficult owing to the complexity of associated risk factors and limitations in available clinical data. This study evaluates the performance and interpretability of Machine Learning (ML) models using a publicly available benchmark dataset containing demographic, physiological, and behavioral maternal features, including age, BMI, body temperature, heart rate, stress levels, blood pressure, activity patterns, and previous pregnancy loss history. Rather than assuming high predictive performance, the analysis focuses on understanding model behavior and identifying the features that consistently receive high importance scores within the available dataset. Four ML models, namely Random Forests (RF), Gradient Boosting (GB), Support Vector Machines (SVM), and Neural Networks (NN), were evaluated using stratified cross-validation, confidence interval estimation, Friedman statistical testing, Nemenyi post-hoc comparison, and robustness analysis. The integration of Friedman statistical testing with Nemenyi post-hoc comparison represents a methodological contribution that remains uncommon in the spontaneous abortion ML literature, where model comparisons are often reported without formal statistical significance testing. Model interpretability was examined through SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), providing complementary global and local explanations of feature importance. The results indicate that AlcoholConsumption, Age, and heart rate consistently received the highest feature importance rankings according to SHAP and LIME. However, all evaluated models achieved predictive performance close to chance level, indicating that the publicly available benchmark dataset provides limited discriminative information for reliable spontaneous abortion prediction. Consequently, this study should be interpreted as a methodological evaluation and transparent benchmark of ML models and explainability techniques rather than as the development of a clinically applicable risk prediction tool. The findings also emphasize the importance of clinically verified datasets with more comprehensive obstetric and laboratory variables for future research.

1. INTRODUCTION

Spontaneous abortion, commonly known as miscarriage, remains one of the most common complications of early pregnancy, affecting approximately 10-20% of clinically recognized pregnancies worldwide [1-3]. In addition to its physical consequences, miscarriage often imposes substantial emotional and psychological burdens on women and their families, with effects that may persist long after the pregnancy loss [4, 5]. The condition is associated with a wide range of

maternal, fetal, environmental, and lifestyle-related factors, making its prediction and prevention particularly challenging. Maternal age is considered one of the most important risk factors, as the likelihood of chromosomal abnormalities and subsequent pregnancy loss increases significantly after the age of 35 years. Other factors that have been associated with miscarriage include a history of previous pregnancy loss, chronic medical conditions such as diabetes mellitus and thyroid disorders, obesity, uterine abnormalities, smoking, excessive alcohol consumption, and illicit drug use [6]. In

addition, both sporadic and recurrent miscarriages have been linked to genetic, anatomical, endocrine, immunological, and behavioral determinants, with chromosomal abnormalities accounting for a substantial proportion of first-trimester pregnancy losses. Elevated body mass index, excessive caffeine intake, psychosocial stress, and socio-cultural influences such as early marriage have also been reported as contributing factors [7, 8]. Given the diversity and complexity of these risk factors, early identification of women at higher risk is essential for supporting timely interventions, improving antenatal care, and reducing adverse pregnancy outcomes. However, achieving reliable risk prediction remains challenging because the performance of Machine Learning (ML) models depends heavily on the availability, completeness, and clinical relevance of the underlying data. Consequently, evaluating model performance and understanding the limitations imposed by available datasets are essential steps toward developing more reliable and clinically applicable prediction frameworks.

Clinical practice, however, has not always kept pace with this complexity. Conventional assessments still tend to rely on a limited set of observable parameters and clinician judgment, which leaves room for missed or delayed identification of at-risk pregnancies [9]. There is a growing recognition that structured data, if analyzed properly, could offer something more systematic. Electronic health records and wearable-derived measurements, for instance, carry signals that traditional evaluation methods simply cannot process at scale [10, 11].

ML has gradually entered this space, and for good reason. Its ability to handle mixed-type, high-dimensional data makes it well-suited for maternal health contexts where risk factors span across demographic, physiological, and behavioral domains [12-14]. Narrowly evaluated models tend to leave important aspects unaddressed, including consistency of performance across different data splits, the statistical meaningfulness of observed differences, and identification of which maternal features are actually driving the predictions [15-17].

These gaps motivated the present study. Rather than adding yet another accuracy comparison to the literature, this work takes a more deliberate approach that combines rigorous statistical evaluation with interpretability analysis to better understand both model behavior and feature relevance in spontaneous abortion risk detection. The analysis draws on a multi-parameter maternal dataset covering numerical and categorical features, and evaluates several ML classification models under a framework that includes stratified cross-validation, confidence interval estimation, Friedman statistical testing, Nemenyi post-hoc comparison, robustness analysis, and dual-layer explainability through SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME).

The contributions of this study can be summarized as follows. First, it provides a statistically grounded comparison of ML models that extends beyond conventional accuracy-based evaluation by incorporating confidence interval estimation, Friedman statistical testing, Nemenyi post-hoc comparison, and robustness analysis. Second, it combines global and local explainability through SHAP and LIME to examine feature importance and model behavior from complementary perspectives. Third, rather than proposing a clinically applicable risk prediction tool, this study establishes a transparent methodological benchmark for evaluating ML

models on a publicly available spontaneous abortion dataset, while highlighting the limitations of the available features and the importance of clinically verified datasets for future research.

2. SIMILAR PREVIOUS RESEARCH

2.1 Machine Learning in healthcare and maternal health

The use of ML across data-intensive fields has grown considerably over the past decade, driven largely by its capacity to model relationships that conventional statistical methods tend to oversimplify [18-20]. Machine learning has demonstrated utility in healthcare, particularly for risk prediction, health-related outcome assessment, and clinical risk stratification [21-23]. Broader reviews have also highlighted the opportunities, limitations, and challenges associated with the application of ML across healthcare and other data-intensive domains [24].

Within the broader healthcare literature, maternal health has emerged as a meaningful application area for ML. Studies have examined pregnancy outcome prediction [25-27], maternal risk factor assessment [28], and clinical decision support [29], with varying degrees of success. The findings from these works are generally encouraging in terms of technical performance, but they also reveal a pattern worth noting: many studies emphasize classification accuracy as a primary evaluation metric, while interpretability receives comparatively less attention. This is a meaningful gap in a domain where clinical trust and explainability are not peripheral concerns but central ones.

Within this landscape, several algorithmic approaches have been considered for prediction and classification tasks, including Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB), and Neural Network (NN). RF has been widely investigated as a tree-based learning approach [30], while SVM provides an alternative framework for classification through the construction of decision boundaries in feature space [31]. GB has also been applied to predictive tasks and can provide strong predictive performance [32]. NN can model complex nonlinear relationships and interactions, making it another relevant approach for comparative evaluation in prediction tasks.

A common thread across comparative machine learning studies is the need for careful evaluation of whether observed differences between models are statistically meaningful rather than attributable to sampling variation. In addition, SHAP and LIME have been increasingly introduced as complementary tools for improving model transparency through global and local explanations [33, 34]. However, the combined use of explainability techniques with formal statistical comparison remains less commonly emphasized in studies addressing spontaneous abortion prediction. The present study was designed with these methodological considerations in mind.

2.2 Comparative positioning of the present study

Situating the present study within this body of work requires attention to both methodological overlaps and meaningful differences. Previous studies using supervised ML with demographic, behavioral, and maternal health variables have reported challenges in achieving reliable miscarriage prediction. This observation is consistent with the pattern

observed in the present study and suggests that the modest classification performance may reflect limitations in the predictive information available in the input features rather than an isolated methodological shortcoming.

Studies working with verified clinical biomarkers have consistently reported stronger discriminative performance. RF and SHAP applied to a retrospective cohort of 523 RSA patients using routine metabolic and inflammatory indicators achieved an AUC of 0.85, with endometrial thickness, neutrophil percentage, and HOMA- β emerging as the most influential predictors [35]. An evaluation of ten classical ML algorithms alongside a Transformer-based tabular model using SMOTE, 5-fold cross-validation, RFE-based feature selection, and SHAP on a multidimensional RSA clinical dataset reached an AUC of 0.927 [36]. A study developing 28 prediction models from 565 immunologically abnormal pregnancies found that XGBoost achieved an AUC of 0.9209, with SHAP identifying pharmacological treatment variables as the primary contributors, though when applied under clinical risk scoring rules, the F1-score dropped to 0.2852, lower than the figures reported in the present study, underlining that strong AUC performance does not automatically translate into reliable multi-metric classification [37]. It was further demonstrated that ML models trained on routine blood data can meaningfully stratify threatened abortion risk in a clinical population, reinforcing the established connection between input feature quality and predictive ceiling [38].

A particularly relevant methodological contrast appears in a dual-center retrospective study of 1,664 blastocyst transfer cycles that began with over 40 candidate features before a panel of reproductive medicine experts guided the selection down to 32 clinically validated inputs using Mutual Information and Recursive Feature Elimination. The resulting Voting Classifier achieved an AUC of 0.836, illustrating how structured pre-modeling feature curation contributes

substantially to discriminative performance [39]. No equivalent expert-guided screening was applied in the present study, which likely represents one contributing factor to the performance ceiling encountered here and points to domain-informed feature selection as a concrete priority for future work.

Taken together, these studies establish a consistent pattern: ML performance in spontaneous abortion prediction is strongly conditioned on the clinical validity and domain specificity of input features. The present study deliberately operates outside the high-biomarker zone, working with behavioral and lifestyle attributes from a public repository to examine what ML classifiers can and cannot extract from such inputs-supported by an evaluation framework that incorporates statistical hypothesis testing and dual-layer interpretability analysis, elements that remain uncommon in the existing spontaneous abortion literature.

3. PROPOSED MODEL

This study follows a quantitative experimental design, drawing on theoretical frameworks in reproductive health and relevant literature to build an interpretable ML-based framework for spontaneous abortion risk assessment. The core objective is not simply to classify outcomes, but to identify which maternal attributes carry the most weight in driving those classifications, and to do so in a way that remains transparent enough for clinical interpretation. Feature contribution analysis sits at the heart of this framework, as understanding the relative importance of individual risk factors is arguably more actionable in a healthcare setting than a prediction label alone. The overall structure of the proposed framework is presented in Figure 1.

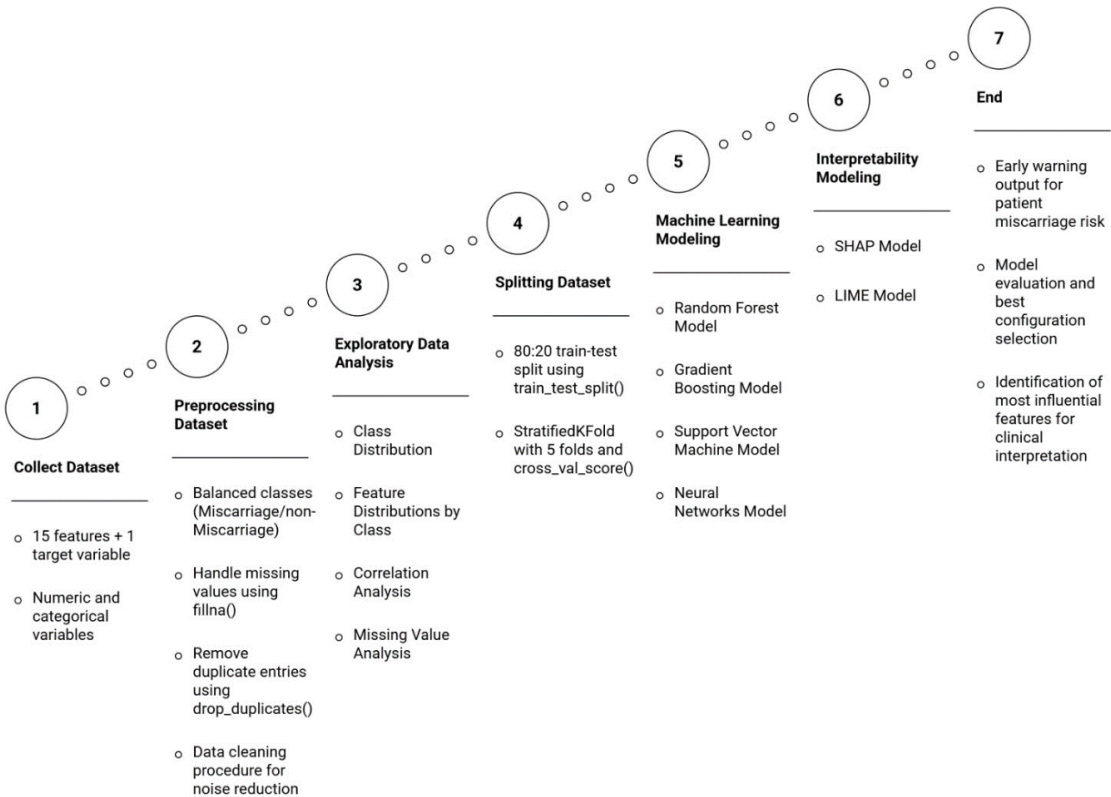


Figure 1. Miscarriage prediction framework

3.1 Collect dataset

The dataset used in this study was obtained from a publicly available repository and consists of 1,000,000 observations, 15 predictive features, and a binary target variable indicating spontaneous abortion outcomes [40]. The dataset was selected because it provides a large-scale benchmark for evaluating ML models using demographic, physiological, behavioral, and pregnancy-related variables. The original repository, however, does not provide comprehensive documentation regarding patient recruitment, healthcare institutions, ethical approval, or data acquisition procedures. Consequently, the clinical provenance of the records cannot be independently verified. For this reason, the dataset is treated as a publicly available benchmark for methodological evaluation rather than as a fully validated clinical dataset. The findings of this study should therefore be interpreted within the context of methodological evaluation and not as direct evidence of clinical effectiveness.

A detailed description of all predictive features is presented in Table 1. The variables cover multiple dimensions of maternal characteristics, including demographic, physiological, behavioral, and pregnancy-related information available in the original public repository. However, several clinically established predictors of spontaneous abortion, such

as gestational age, ultrasound findings, bleeding symptoms, progesterone level, β -hCG, chromosomal abnormalities, infection status, thyroid function, and detailed obstetric history, are not included in the benchmark dataset. Consequently, this study evaluates ML models using only the variables available in the publicly accessible benchmark dataset rather than developing a comprehensive clinical prediction model. Although the dataset has limitations regarding clinical completeness and documentation, it provides a publicly accessible benchmark for evaluating the comparative performance, statistical reliability, and interpretability of ML models under a reproducible experimental setting.

Several feature names are retained exactly as provided in the original repository to preserve consistency with the source dataset. The repository explicitly describes *Nmisc* as representing a history of previous miscarriage. However, it does not provide formal clinical definitions or measurement protocols for variables such as *Activity*, *Binking*, *Walking*, *Drinving*, *Sitting*, and *Alcohol Consumption*. Therefore, these variables are reported using their original repository terminology without introducing additional assumptions regarding their clinical interpretation.

Table 1. Dataset features and description

Features	Description	Types
Age	Age at pregnancy.	Ordinal
BMI	Weight during pregnancy.	Continuous
Nmisc	History of previous miscarriage (as described in the original repository).	Ordinal
Activity	Variable retained using the original repository terminology; the source repository does not provide a formal clinical definition or measurement protocol.	Binary
Binking	Variables retained using the original repository terminology; the source repository does not provide formal clinical definitions or measurement protocols.	Binary
Walking	Variables retained using the original repository terminology; the source repository does not provide formal clinical definitions or measurement protocols.	Binary
Drinving	Variables retained using the original repository terminology; the source repository does not provide formal clinical definitions or measurement protocols.	Binary
Sitting	Variables retained using the original repository terminology; the source repository does not provide formal clinical definitions or measurement protocols.	Binary
Location	Place where the woman spends most of her time (as described in the original repository).	Nominal
Temp	Body Temperature.	Ordinal
BPM	Heart Rate Variability.	Continuous
Stress	A physical and psychological response to pressure.	Ordinal
BP	Blood Pressure.	Ordinal
Alcohol Consumption	The amount of alcohol consumed by an individual.	Continuous
Drunk	Alcohol intoxication.	Ordinal
Miscarriage	Class.	Binary

3.2 Preprocessing dataset

Data preprocessing was carried out to ensure that the dataset was clean, consistent, and suitable for model development. Duplicate records were removed using the *drop_duplicates()* function to eliminate redundant observations. The dataset was subsequently inspected for missing values, and no missing values were identified; therefore, no data imputation was required. Predictor variables were retained using the representation provided in the original public benchmark dataset to preserve consistency and reproducibility. Continuous numerical features were normalized using the *MinMaxScaler* implementation from the *scikit-learn* library for the SVM and NN models, whereas RF and GB were trained using the original feature scales because tree-based algorithms

are generally insensitive to feature scaling. To prevent validation data leakage, the *MinMaxScaler* was fitted exclusively on the training data within each cross-validation fold and subsequently applied to the corresponding validation data. Following preprocessing, the dataset was partitioned into training and testing subsets before model evaluation.

3.3 Exploratory data analysis

Prior to model development, a structured exploratory data analysis was conducted to characterize the dataset and surface any properties that could influence modelling decisions.

3.3.1 Class distribution

The distribution of the target variable is presented in Figure

2. The dataset contains 500,571 samples in the non-miscarriage class (0) and 499,429 samples in the miscarriage class (1), representing 50.06% and 49.94% of the total respectively. This near-perfect balance between classes confirms that the dataset does not present a class imbalance problem, and resampling techniques such as SMOTE were therefore not applied. The balanced distribution also means that accuracy remains a meaningful metric alongside recall and F1-score, though the latter two are still prioritized given the clinical context of the classification task.

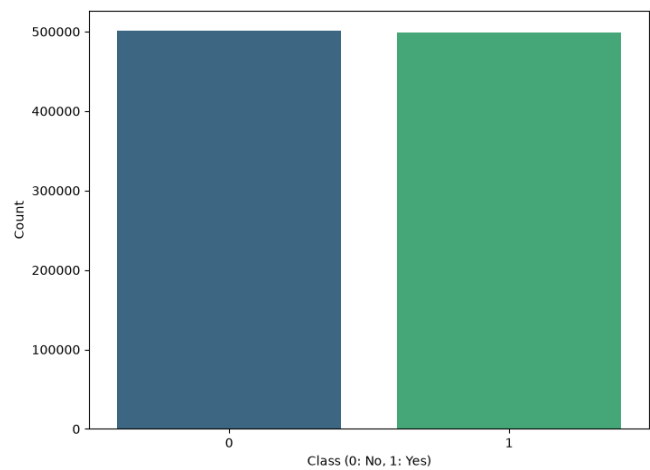


Figure 2. Distribution class miscarriage

3.3.2 Feature distributions by class

Kernel density estimates for eight features split by class label are presented in Figure 3. The most immediately striking observation across all plots is the near-complete overlap between the two class distributions, with the miscarriage (1) and non-miscarriage (0) curves being virtually indistinguishable across every feature examined. This pattern holds for both continuous features and ordinal variables. For Age, the distribution is multimodal with peaks recurring at regular intervals between 16 and 26, and both classes follow an identical pattern with no visible separation. Nmisc shows discrete peaks at integer values (0, 1, 2, 3), again with complete class overlap at each point. Temp similarly displays a multimodal structure with recurring peaks between 35 °C and 41 °C, with no differential distribution between classes. BPM spans a wide range from approximately 40 to 250, with multiple peaks throughout, but again both classes track each other precisely with no discriminative separation visible. For the ordinal features, Stress and BP both show discrete multi-level distributions with peaks at each ordinal category, and neither feature shows meaningful class separation. AlcoholConsumption displays a broad, near-uniform distribution across its range with both classes producing an almost flat and identical density curve, a pattern consistent with the feature containing little discriminative signal despite its high ranking in the SHAP analysis. Drunk shows a sharp spike at value 2, with both classes concentrated at the same point.

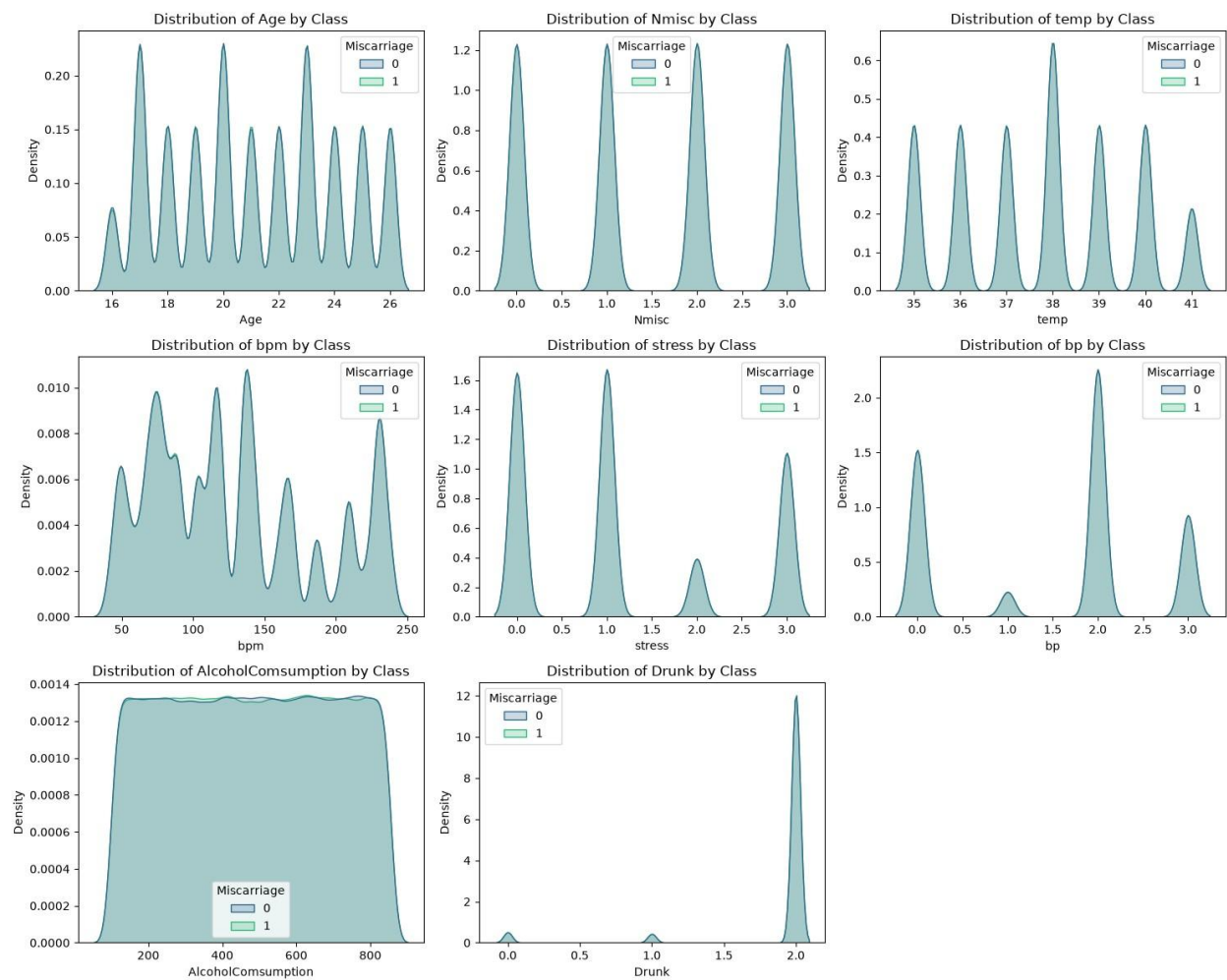


Figure 3. Feature distributions by class

3.3.3 Correlation analysis

A correlation matrix was computed across all 15 features to identify potential multicollinearity and to contextualize the feature importance rankings produced by SHAP in the later analysis. The feature correlation heatmap is presented in Figure 4. The most notable finding is the strong positive correlation between BPM and BP ($r = 0.90$), which indicates a high degree of linear dependency between these two physiological variables. A moderate correlation is also observed between BPM and stress ($r = 0.23$), which is clinically plausible given the known relationship between heart rate and stress response. Beyond these two pairs, all remaining inter-feature correlations are negligible, with values at or near zero across the matrix.

Correlations with the target variable (Miscarriage) are

uniformly weak across all features. The highest correlation with the outcome is Drinving at 0.0026, followed by Age at 0.0021, with all remaining features showing correlations below 0.001 in absolute value. These near-zero correlations confirm that no individual feature carries meaningful linear discriminative power with respect to the miscarriage outcome, which is consistent with the modest classification performance observed across all models.

The strong BPM-BP correlation ($r = 0.90$) warrants attention in the context of the SHAP analysis presented later, as multicollinearity between these two features may affect the stability of individual SHAP value attributions. However, given that BPM ranked among the top globally influential features while BP did not, the practical impact of this collinearity on the interpretability findings appears limited.

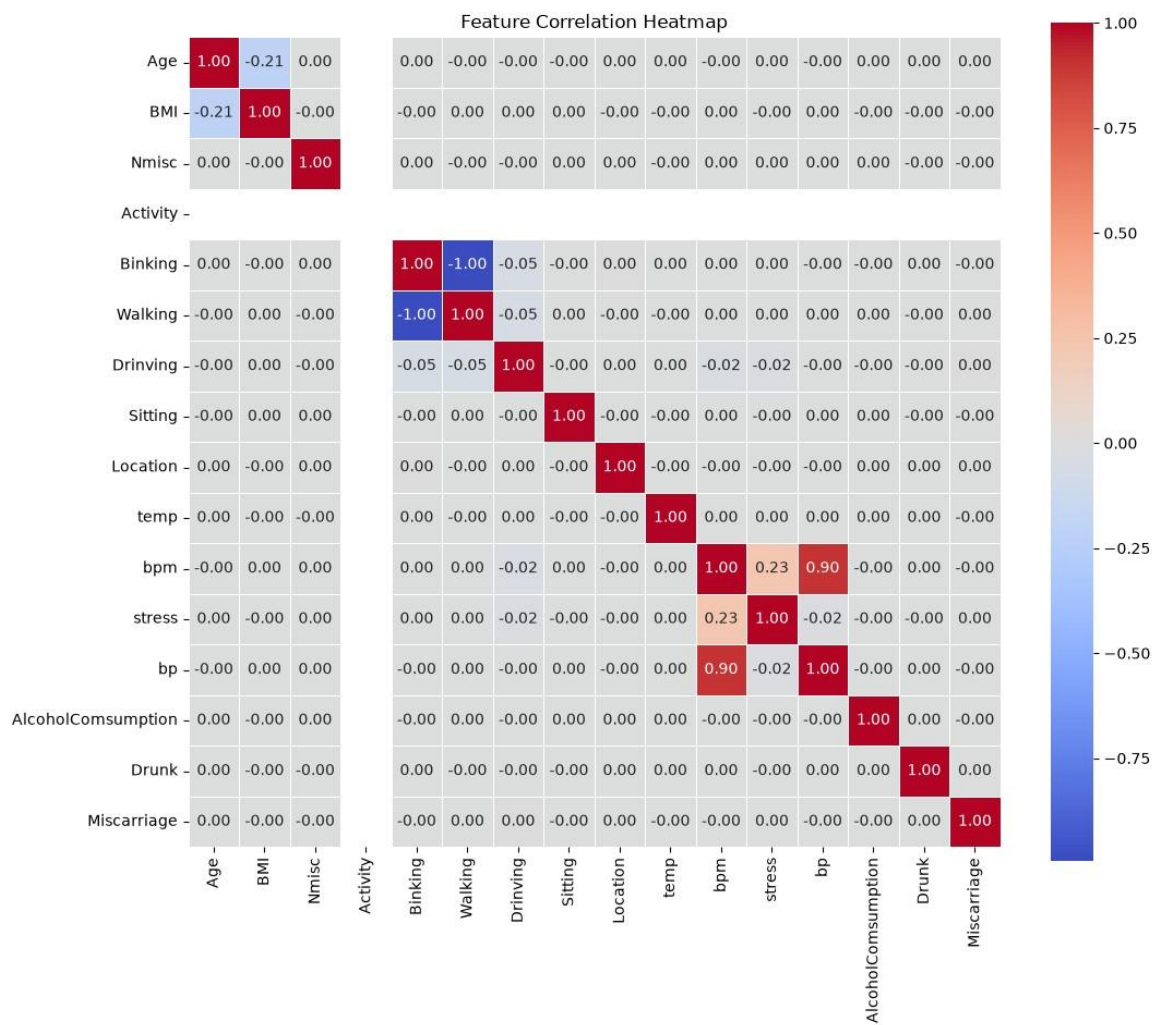


Figure 4. Correlation heatmap

3.3.4 Missing value analysis

Missing value analysis was conducted across all 15 features and the target variable prior to preprocessing. As presented in Table 2, no missing values were detected in any column, with each feature recording 0 missing entries and 0.0% missingness across the full dataset of 1,000,000 samples. This indicates that the dataset was either pre-cleaned prior to public release or generated under controlled conditions that precluded missing entries. Although the *fillna()* function was applied as a precautionary step during preprocessing, the analysis confirms that no imputation was actually performed on any feature. The absence of missingness also means that the pattern of missing

values, which can itself carry clinical information in real-world medical datasets, was not a factor in this study, and this represents one distinction between the present dataset and datasets sourced directly from clinical records.

Beyond the absence of missing values, the dataset also exhibits a nearly balanced class distribution, weak feature-target correlations, and a restricted maternal age range (16-26 years). These characteristics differ from those commonly observed in routine clinical datasets, which typically contain more heterogeneous populations and incomplete records. Consequently, the experimental findings should be interpreted within the context of a publicly available benchmark dataset

intended for methodological evaluation rather than as evidence directly generalizable to real-world clinical settings.

Table 2. Missing value summary

Features	Missing Values	%
Age	0	0
BMI	0	0
Nmisc	0	0
Binking/Walking/Drinving/Sitting	0	0
Location	0	0
Temp	0	0
BPM	0	0
Stress	0	0
BP	0	0
AlcoholConsumption	0	0
Drunk	0	0
Miscarriage	0	0

3.3.5 Outlier identification

Outlier analysis was conducted on eight continuous features using the Interquartile Range (IQR) method, with results visualized through boxplots presented in Figure 5. Age shows a distribution spanning approximately 16 to 26 years with a median around 21, and whiskers extending to the full observed range without visible outlier points. Nmisc displays a median of 2 with an IQR between 1 and 3, and a lower whisker

extending toward 0 with no extreme outliers. Temp ranges from 35 °C to 41 °C with a median of approximately 38 °C and no outlying values detected. BPM shows the widest spread among continuous features, with a median around 120, an IQR approximately between 80 and 165, and whiskers extending to roughly 45 and 230-the broad range reflects the natural variability of heart rate data and does not appear to represent measurement error. Stress and BP both display ordinal-like distributions with medians at 1 and 2 respectively and whiskers reaching their maximum values of 3, consistent with their discrete categorical nature. AlcoholConsumption shows a wide IQR between approximately 300 and 650 with a median near 500 and whiskers extending to around 100 and 850, indicating substantial spread but no extreme outlier points. The most notable finding is in Drunk, where two isolated outlier points are visible at values of 0 and 1-marked as circles outside the whiskers. Given that Drunk is a binary or near-binary variable with the vast majority of values concentrated at 2, these points represent genuine distributional anomalies. However, their small number relative to the total dataset size of one million samples means their influence on model behavior is expected to be negligible.

Overall, the outlier analysis confirms that no feature contains extreme values severe enough to substantially distort model training, though the broad range of BPM warrants monitoring in distance-sensitive classifiers such as SVM.

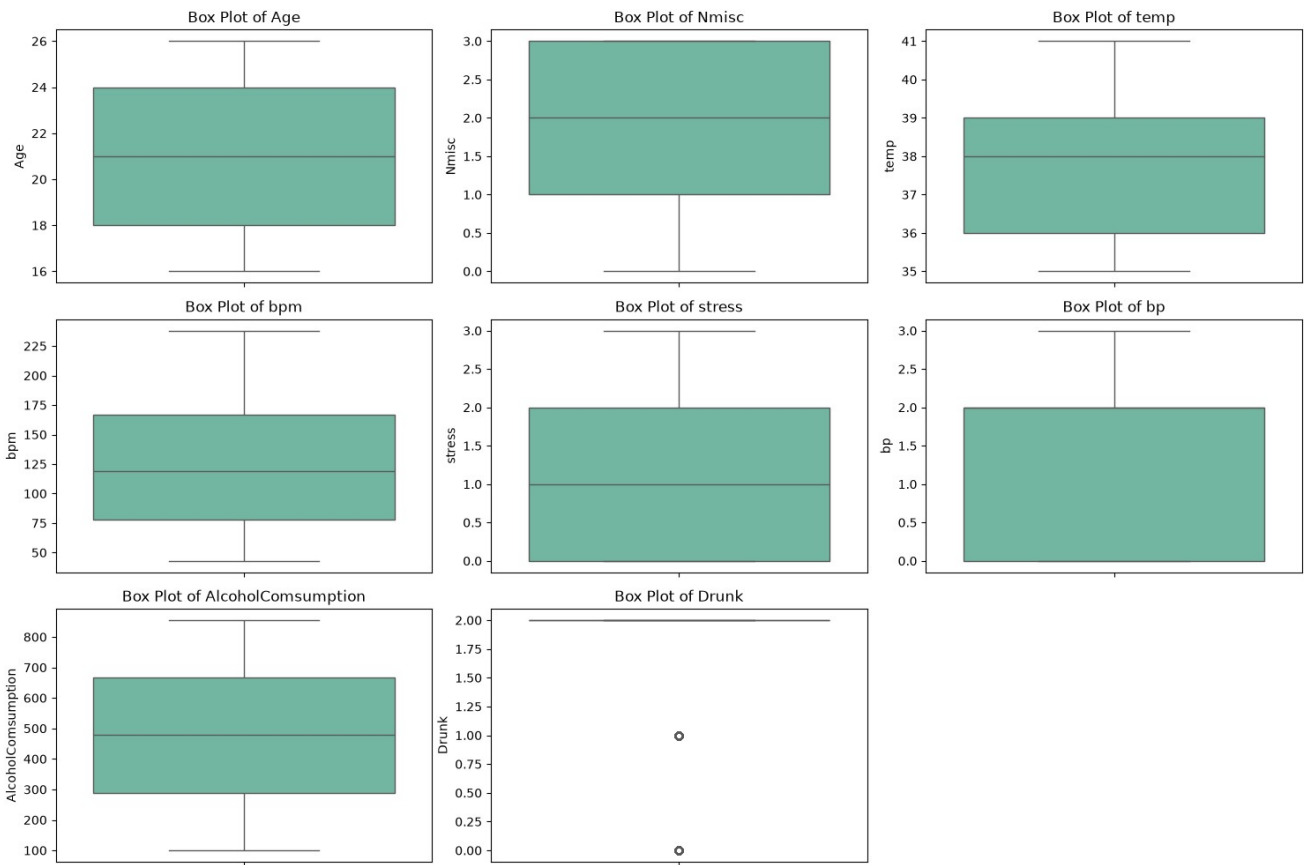


Figure 5. Outlier count

3.4 Splitting dataset

The dataset was divided into training and testing subsets using a 70:30 ratio through the *train_test_split()* function, with 70% of the data used for model training and the remaining 30% reserved for independent performance evaluation. To

obtain a more reliable assessment of model stability, five-fold StratifiedKFold cross-validation was additionally performed using the *cross_val_score()* function while preserving the original class distribution in each fold. To prevent information leakage, all preprocessing steps requiring parameter estimation, including feature scaling for the SVM and NN

models, were incorporated into the cross-validation workflow. Specifically, the MinMaxScaler was fitted exclusively on the training partition of each fold and subsequently applied to the corresponding validation partition. This ensured that no information from the validation data was used during preprocessing, thereby providing an unbiased estimate of model performance.

3.5 Machine Learning modeling

The selected features were input into ML models, including RF, GB, SVM, and NN. These models were chosen due to their capability to handle mixed-type data, capture complex non-linear relationships, and deliver robust predictions.

(1) RF model

An ensemble method that aggregates predictions from multiple decision trees built on bootstrap samples with random feature selection. The final prediction is defined as:

$$\hat{y} = mode\{h_t(x)\}_{t=1}^T \tag{1}$$

Node splitting is commonly based on Gini impurity:

$$G = 1 - \sum_i p_i^2 \tag{2}$$

(2) GB model

Constructs an additive model in a stage-wise manner by iteratively minimizing a differentiable loss function.

$$F(x) = \sum_{m=1}^M y_m h_m(x) \tag{3}$$

Each iteration updates the model using the negative gradient:

$$F_m(x) = F_{m-1}(x) + y_m h_m(x) \tag{4}$$

(3) SVM model

Aims to determine an optimal separating hyperplane that maximizes the margin between classes.

$$f(x) = w^T + b \tag{5}$$

$$\min \frac{1}{2} ||w||^2 s.t. y_i (w^T x_i + b) \geq 1 \tag{6}$$

For non-linear cases, a kernel function is used:

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j) \tag{7}$$

(4) NN model

Designed to capture complex non-linear relationships through multiple layered transformations.

$$\alpha = \sigma(W_x + b) \tag{8}$$

For binary classification, the loss function is:

$$L = -\frac{1}{N} \sum [y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})] \tag{9}$$

The network consists of one hidden layer containing 100

neurons with ReLU activation. The Adam optimizer was used with a constant learning rate of 0.001, and training was run for a maximum of 200 iterations. No regularisation beyond the default L2 penalty (alpha = 0.0001) was applied.

All four models were implemented using the scikit-learn library with predefined baseline hyperparameter settings, as summarized in Table 3. The selected hyperparameters were fixed across all experiments to ensure a consistent and reproducible comparison among classifiers. No hyperparameter optimization procedures, such as Grid Search, Random Search, or Bayesian Optimization, were performed because the primary objective of this study was to evaluate the baseline behavior of different ML algorithms on a publicly available benchmark dataset rather than to maximize predictive performance. The random_state parameter was fixed at 42 whenever applicable to ensure reproducibility. Future studies may investigate optimized hyperparameter configurations together with clinically informed feature selection using independently validated clinical datasets.

The hyperparameter settings presented in Table 3 were adopted as baseline configurations to ensure a consistent and reproducible comparison across all evaluated classifiers. No hyperparameter optimization procedures, such as Grid Search, Random Search, or Bayesian Optimization, were performed because the primary objective of this study was to evaluate the baseline behavior of different ML algorithms on a publicly available benchmark dataset rather than to maximize predictive performance. Consequently, the experimental design emphasizes methodological consistency across models. Future studies should investigate optimized hyperparameter configurations together with clinically informed feature selection using validated clinical datasets to determine the extent to which predictive performance can be improved.

Table 3. Hyperparameter settings for all evaluated models

Model	Hyperparameter	Value
Random Forest (RF)	n_estimators	100
	max_depth	None
	min_samples_split	2
	min_samples_leaf	1
	max_features	sqrt
Gradient Boosting (GB)	random_state	42
	n_estimators	100
	learning_rate	0.1
	max_depth	3
	min_samples_split	2
Support Vector Machine (SVM)	min_samples_leaf	1
	random_state	42
	kernel	rbf
	C	1.0
	gamma	scale
Neural Network (NN)	probability	True
	hidden_layer_sizes	(100,)
	activation	relu
	solver	adam
	learning_rate	constant
	max_iter	200
	random_state	42

3.6 Interpretability modeling

To enhance transparency and clinical applicability, the proposed framework incorporates interpretability techniques that explain the contribution of individual maternal features to miscarriage risk predictions. Explainable ML methods,

including SHAP and LIME, are employed to quantify feature importance and provide both global and local explanations of the model's predictions.

(1) SHAP model

An explainability method based on cooperative game theory that quantifies the contribution of each feature to a model's prediction. It assigns an importance value (Shapley value) to each feature by considering all possible feature combinations. The explanation model is defined as:

$$f(x) = \phi_0 + \sum_{i=1}^M \phi_i \quad (10)$$

where, ϕ_i represents the contribution of feature i , and ϕ_0 is the baseline prediction. The Shapley value is computed as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|! (M - |S| - 1)!}{M!} [f(S \cup \{i\}) - f(S)] \quad (11)$$

This formulation ensures that feature contributions are fairly distributed based on all possible feature interactions.

(2) LIME model

Provides interpretability by approximating the complex model locally around a specific instance using a simpler surrogate model. Instead of explaining the global behavior of the model, LIME focuses on local fidelity by perturbing the input data and learning an interpretable model g that approximates the original model f in the neighborhood of an instance x .

$$g^* = \arg \min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g) \quad (12)$$

This approach allows LIME to identify the most influential features contributing to a specific prediction in a local region of the data space.

The final stage of the proposed framework generates classification outputs together with model-specific interpretability results obtained using SHAP and LIME. These outputs are intended to facilitate the comparative evaluation of ML models, feature importance analysis, and methodological assessment within a reproducible benchmark framework. The accompanying interpretability results improve transparency by illustrating how individual features influence model predictions. Given the chance-level predictive performance observed in this study, the generated outputs are intended solely for research and methodological evaluation and should not be interpreted as clinical decision-support or diagnostic recommendations.

4. RESULTS AND DISCUSSION

4.1 Results

Table 4 presents the classification performance of all evaluated models alongside the Majority Class Baseline. Across all metrics, accuracy values were strikingly similar, falling within a narrow range of 49.8473% to 50.0570%, which is essentially equivalent to random chance. GB returned the highest accuracy at 50.0457% and precision at 49.9879%, while RF led on recall at 47.7825% and F1-score at 48.7614%. GB followed with a recall of 47.0503% and F1-score of

48.4747%, whereas SVM recorded a recall of 40.2245% and F1-score of 44.4813%. NN produced the lowest recall at 25.3062% and F1-score at 33.5653%. The Majority Class Baseline, despite achieving the highest raw accuracy at 50.0570%, produced zero values across precision, recall, and F1-score, confirming that it failed entirely to identify positive cases. This contrast between the baseline and the trained models, however modest the gap, underscores why accuracy alone is an insufficient basis for evaluating classifier performance in this context.

Table 5 presents the advanced performance metrics for all evaluated models alongside the Majority Class Baseline. Sensitivity values follow the same ranking observed in Table 2, with RF recording the highest at 47.7825% and NN the lowest at 25.3062%, while specificity shows an inverse pattern-NN achieved the highest at 74.5763% and RF the lowest at 51.9075%, reflecting the tendency of NN to classify the majority of instances as negative. GB returned a balanced position with sensitivity at 47.0503% and specificity at 53.0342%, and SVM recorded sensitivity of 40.2245% and specificity of 59.4569%. The Majority Class Baseline produced a sensitivity of zero and a perfect specificity of 100%, confirming that it identified no positive cases whatsoever. Matthews Correlation Coefficient (MCC) and Cohen's Kappa values are the most direct indicators of overall discriminative ability, with GB producing the only marginally positive values at 0.0008 for both measures, while RF, SVM, and NN all returned slightly negative figures-confirming that none of the classifiers achieved meaningful discrimination beyond chance, consistent with the near-random accuracy figures reported in Table 4.

Figure 6 presents the confusion matrix results for all evaluated models. Looking at RF first, the model correctly classified 77,950 negative cases and 71,592 positive cases, while misclassifying 72,221 negative cases as positive and 78,237 positive cases as negative. GB followed a similar pattern, producing 79,642 true negatives and 70,495 true positives, with 70,529 false positives and 79,334 false negatives. Moving to the discriminative classifiers, SVM achieved the highest number of true negatives at 89,287 but identified considerably fewer true positives at 60,268 compared to both RF and GB. NN showed the most pronounced imbalance of all, recording 111,992 true negatives alongside only 37,916 true positives, with 38,179 false positives and 111,913 false negatives. Taken together, these patterns suggest that RF and GB maintained a more balanced recognition of both classes, while SVM and NN leaned heavily toward negative classification, capturing fewer positive instances in the process.

Figure 7 presents the learning curves of all evaluated models. One pattern that stands out across all four is the cross-validation score, which remained relatively stable at around 0.50 regardless of how many training examples were added, suggesting that increasing the training set size did not meaningfully improve generalization. The training scores told a somewhat different story. RF and NN both started higher with smaller training sets, then gradually declined as more samples were introduced, pointing to a degree of overfitting in the early stages that weakened as the data grew. GB and SVM behaved more consistently throughout, with training scores staying within a range of approximately 0.51 to 0.57 across different training set sizes.

Table 6 presents the stratified cross-validation results across all evaluated models. GB came out slightly ahead on all four

metrics, recording a mean accuracy of 0.5008 ± 0.0008 , precision of 0.5003 ± 0.0008 , recall of 0.4755 ± 0.0176 , and F1-score of 0.4874 ± 0.0090 . RF produced comparable figures, with accuracy at 0.4997 ± 0.0010 , precision at 0.4990 ± 0.0011 , recall at 0.4562 ± 0.0088 , and F1-score at 0.4766 ± 0.0051 , suggesting that the two ensemble methods behaved similarly across folds. SVM and NN fell further behind on recall and F1-score specifically, with SVM returning 0.3740 ± 0.0155 and 0.4274 ± 0.0100 respectively, and NN producing the lowest values of all at 0.3396 ± 0.2399 and 0.3563 ± 0.1742 . The notably large standard deviations observed for NN on these two metrics also point to inconsistent behavior across folds, a factor worth bearing in mind while interpreting its overall performance. GB and RF demonstrated greater

stability and slightly stronger mean performance than SVM and NN throughout the cross-validation process.

Table 7 presents the RF model performance metrics alongside their 95% confidence intervals. The accuracy settled at 0.4997, with a confidence interval spanning 0.4983 to 0.5011, while precision came in at 0.4990 with a range of 0.4974 to 0.5006. Recall reached 0.4562, bounded between 0.4440 and 0.4685, and the F1-score was recorded at 0.4766 with an interval of 0.4695 to 0.4837. The relatively narrow intervals observed across all four metrics suggest that the RF model behaved consistently throughout the evaluation rather than fluctuating across different samples, lending some confidence to the reported figures even if the performance values themselves remain modest.

Table 4. Model comparison with Majority Class Baseline

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest (RF)	49.8473	49.7813	47.7825	48.7614
Gradient Boosting (GB)	50.0457	49.9879	47.0503	48.4747
Support Vector Machine (SVM)	49.8517	49.7458	40.2245	44.4813
Neural Network (NN)	49.9693	49.8272	25.3062	33.5653
Majority Class Baseline	50.057	0	0	0

Table 5. Model comparison with sensitivity, specificity, Matthews Correlation Coefficient (MCC), and Cohen's Kappa

Model	Sensitivity (%)	Specificity (%)	MCC (%)	Cohen Kappa (%)
Random Forest (RF)	47.7825	51.9075	-0.0031	-0.0031
Gradient Boosting (GB)	47.0503	53.0342	0.0008	0.0008
Support Vector Machine (SVM)	40.2245	59.4569	-0.0032	-0.0032
Neural Network (NN)	25.3062	74.5763	-0.0014	-0.0012
Majority Class Baseline	0	100	0	0

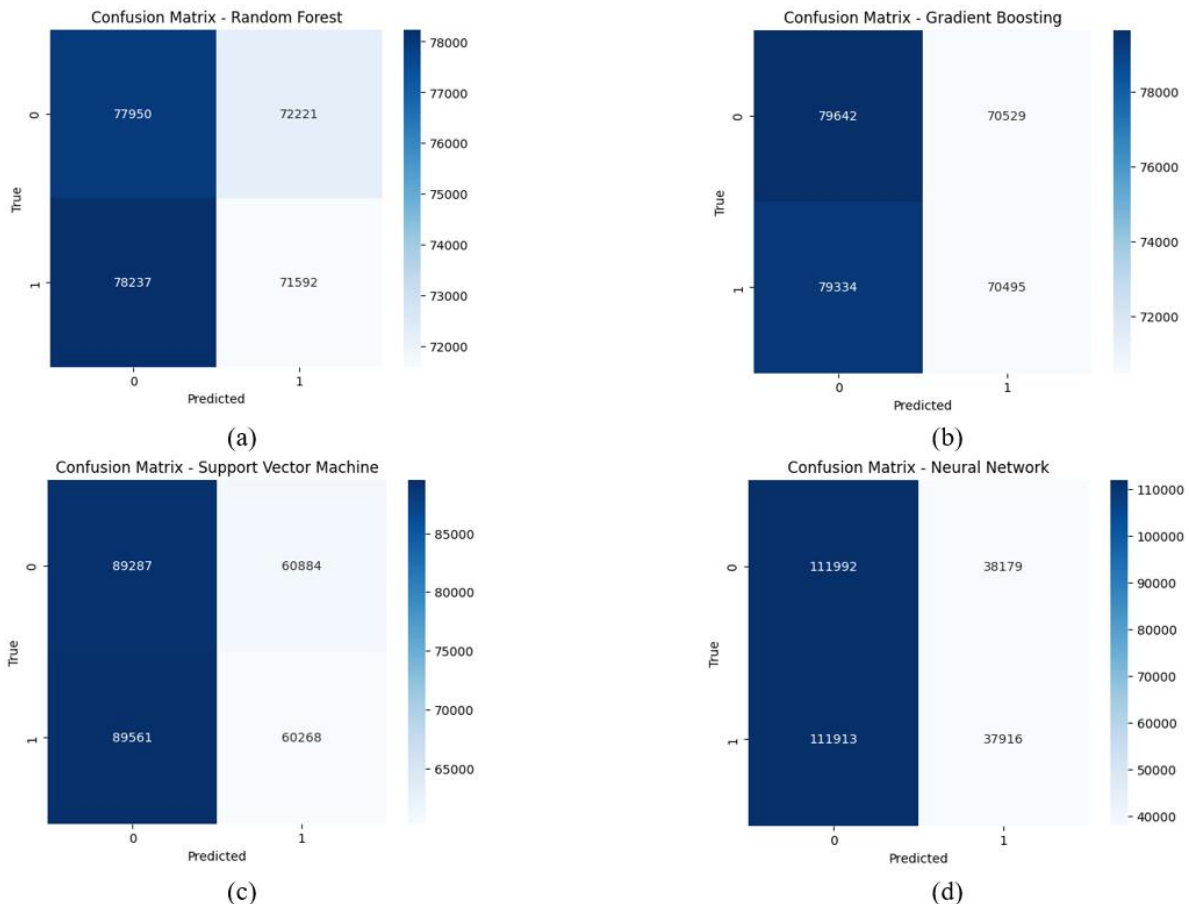


Figure 6. Confusion matrix all model

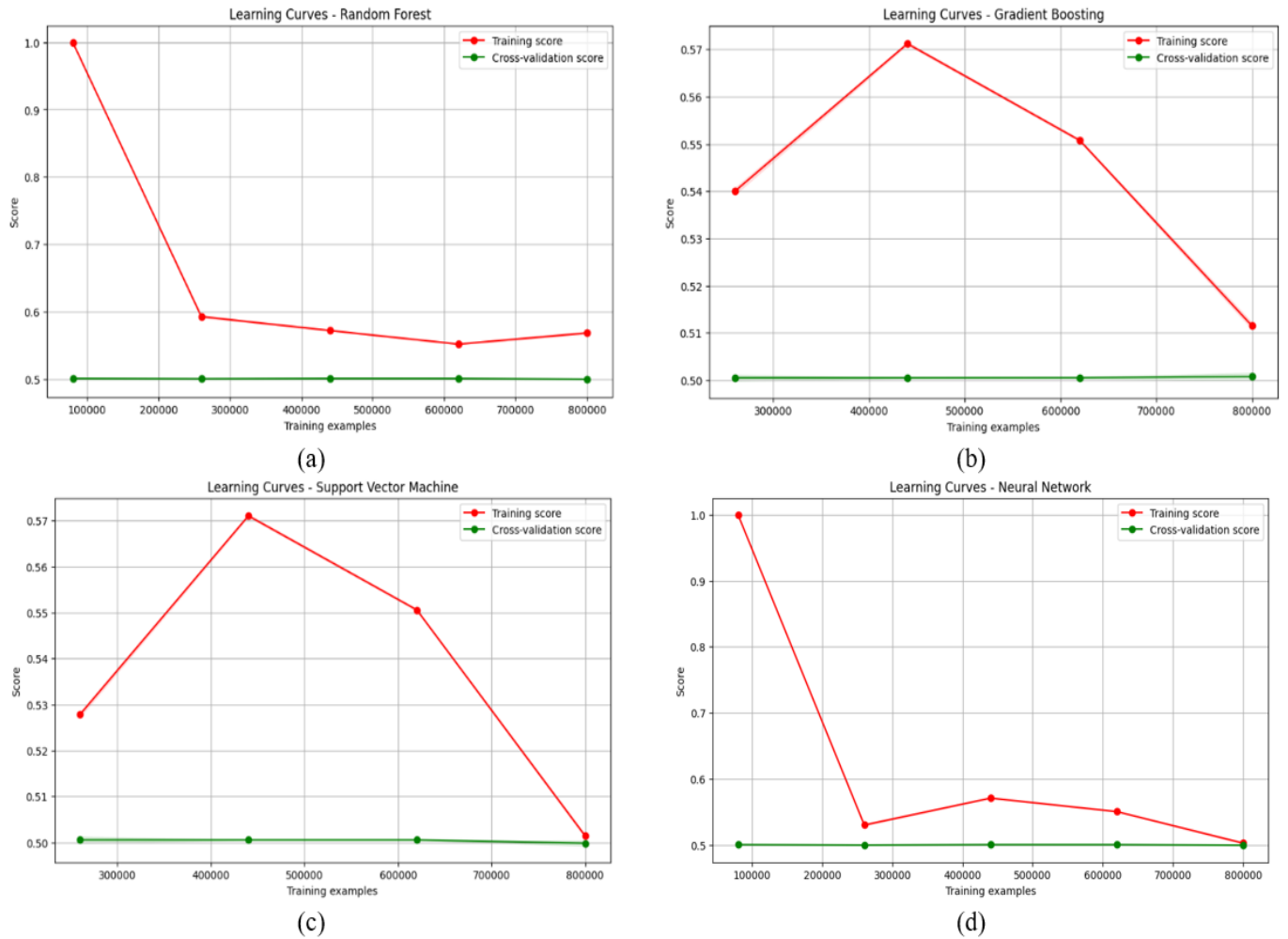


Figure 7. Learning curves of all evaluated models

Table 6. Cross-validation metrics comparison

Model	Stratified CV (Mean $\pm$ Std)			
	Accuracy	Precision	Recall	F1-Score
Random Forest (RF)	0.4997 $\pm$ 0.0010	0.4990 $\pm$ 0.0011	0.4562 $\pm$ 0.0088	0.4766 $\pm$ 0.0051
Gradient Boosting (GB)	0.5008 $\pm$ 0.0008	0.5003 $\pm$ 0.0008	0.4755 $\pm$ 0.0176	0.4874 $\pm$ 0.0090
Support Vector Machine (SVM)	0.4999 $\pm$ 0.0005	0.4991 $\pm$ 0.0007	0.3740 $\pm$ 0.0155	0.4274 $\pm$ 0.0100
Neural Network (NN)	0.4998 $\pm$ 0.0008	0.4991 $\pm$ 0.0009	0.3396 $\pm$ 0.2399	0.3563 $\pm$ 0.1742

Table 7. Random Forests (RF) performance metrics with 95% confidence intervals

Metrics	Value (%)	95% CI
Accuracy	0.4997	0.4983-0.5011
Precision	0.4990	0.4974-0.5006
Recall	0.4562	0.4440-0.4685
F1-Score	0.4766	0.4695-0.4837

Table 8 presents the Friedman test results conducted to assess whether performance differences among the evaluated models were statistically meaningful. For accuracy and precision, the test returned p-values of 0.1940 and 0.1777 respectively, both falling above the 0.05 significance threshold, which suggests that the differences observed between models on these two metrics were not substantial enough to rule out sampling variation as an explanation. Recall and F1-score told a different story, both returning a p-value of 0.0406, crossing the significance boundary and indicating that the models did differ in a statistically meaningful way on these particular metrics. This finding reinforces the earlier

observation that accuracy and precision alone would have painted an incomplete picture of model behavior, while recall and F1-score proved more sensitive in capturing genuine differences among the evaluated classifiers.

Table 9 presents the Nemenyi post-hoc test results examining pairwise differences among the evaluated models. Despite the Friedman test identifying significant differences for recall and F1-score, none of the individual pairwise comparisons reached statistical significance at the 0.05 level. The p-values recorded were 0.8831 for RF vs GB, 0.3159 for RF vs SVM, 0.4559 for RF vs NN, 0.0681 for GB vs SVM, 0.1219 for GB vs NN, and 0.9948 for SVM vs NN. The GB vs SVM comparison came closest to the significance boundary at 0.0681, though it still fell short. Taken as a whole, these results suggest that the performance variation detected by the Friedman test was distributed across models rather than being driven by any single classifier consistently outperforming the others.

Table 10 presents the robustness analysis results across all evaluated models. RF achieved the highest mean accuracy

50.0948% with a standard deviation of 0.0715%, whereas GB obtained a mean accuracy of 50.0570% with a standard deviation of 0.1144%. SVM and NN both produced a mean accuracy of 50.0570% with a standard deviation of 0.0000%. Although the zero standard deviation indicates identical accuracy values across repeated evaluations, it may also reflect limited variability in model predictions rather than superior robustness. Overall, the mean accuracies were closely clustered around 50%, indicating consistently near-chance predictive performance across all evaluated models. Consequently, the robustness analysis should be interpreted as describing the consistency of model behavior under the experimental protocol rather than evidence of reliable predictive capability.

The ablation study results presented in Table 11 provide additional insight into the influence of the lowest-ranked features identified by the global SHAP importance analysis. Specifically, the three features with the lowest SHAP importance scores (Walking, Drinving, and Activity) were removed, and the classifiers were retrained using the same experimental protocol adopted in the primary experiments. The Full Features results are therefore identical to those reported in Table 4, ensuring that any observed differences are attributable solely to feature removal. As shown in Table 11, removing these three features resulted in only marginal changes in predictive performance across all evaluated classifiers. RF, GB, and SVM exhibited only negligible differences in Accuracy, Precision, Recall, and F1-score, while the NN showed a modest reduction in Recall and F1-score. Because all evaluated models demonstrated discrimination performance close to chance level, these small variations should be interpreted cautiously and should not be regarded as conclusive evidence regarding the predictive importance of the removed features. Instead, the ablation study indicates that removing these low-ranked features produced minimal changes in overall model behavior within the benchmark dataset under the adopted experimental setting.

The ROC curves presented in Figure 8 reveal that all four models produced an AUC of 0.50, placing them precisely at the diagonal line that represents random classification. This result is consistent across RF, GB, SVM, and NN, and is indistinguishable from the Majority Class Baseline. An AUC of 0.50 confirms that none of the classifiers were able to rank positive cases above negative cases at any threshold, meaning that the models' outputs carry no discriminative signal with respect to the miscarriage outcome. This finding aligns directly with the near-zero MCC and Cohen's Kappa values reported in Table 3 and reinforces the conclusion that the behavioral and lifestyle features in this dataset do not provide a sufficient basis for probabilistic miscarriage risk stratification. Rather than representing a failure of the modeling approach, this result is consistent with findings from

comparable studies using self-reported preconception data, where similar performance ceilings have been observed and attributed to the limited discriminative validity of behavioral proxies relative to verified clinical biomarkers.

The Precision-Recall curves presented in Figure 9 further confirm the absence of meaningful discriminative signal across all models. RF, GB, and NN produced near-flat curves at a precision level of approximately 0.50 across the full recall range, indistinguishable from the Majority Class Baseline which sits at the same horizontal level. All four models recorded an Average Precision (AP) of 0.50, consistent with the AUC values reported in Figure 8. SVM displayed a visually distinct descending curve, beginning at high precision near 1.0 at very low recall before declining sharply, which reflects its tendency to predict the positive class only at high confidence thresholds, resulting in few but initially precise positive predictions. However, this behavior did not translate into a higher AP score, as the area under its curve remained at 0.50. Taken together, the ROC and Precision-Recall analyses confirm that none of the evaluated models extracted usable probabilistic signal from the dataset, and that the classification outputs are effectively equivalent to random assignment across all threshold settings.

Table 8. Friedman test results for performance metrics

Metrics	p-Value	Significant ($\alpha = 0.05$)
Accuracy	0.1940	Not Significant
Precision	0.1777	Not Significant
Recall	0.0406	Significant
F1-Score	0.0406	Significant

Table 9. Nemenyi post-hoc test results

Comparison Model	p-Value	Significant ($\alpha = 0.05$)
RF vs GB	0.8831	Not Significant
RF vs SVM	0.3159	Not Significant
RF vs NN	0.4559	Not Significant
GB vs SVM	0.0681	Not Significant
GB vs NN	0.1219	Not Significant
SVM vs NN	0.9948	Not Significant

Note: Random Forests (RF), Gradient Boosting (GB), Support Vector Machines (SVM), Neural Networks (NN).

Table 10. Robustness analysis results

Model	Mean Accuracy (%)	Std Accuracy (%)
Random Forest (RF)	50.0948	0.071469
Gradient Boosting (GB)	50.0570	0.114358
Support Vector Machine (SVM)	50.0570	0.000000
Neural Network (NN)	50.0570	0.000000

Table 11. Ablation study results

Feature Set	Model	Accuracy	Precision	Recall	F1-Score
Full Features	Random Forest	49.8473	49.7813	47.7825	48.7614
Full Features	Gradient Boosting	50.0457	49.9879	47.0503	48.4747
Full Features	Support Vector Machine	49.8517	49.7458	40.2245	44.4813
Full Features	Neural Network	49.9693	49.8272	25.3062	33.5653
Reduced Features	Random Forest	49.8403	49.7807	49.2421	49.5100
Reduced Features	Gradient Boosting	49.9580	49.8957	47.4181	48.6253
Reduced Features	Support Vector Machine	49.8470	49.7451	41.0288	44.9685
Reduced Features	Neural Network	49.8910	49.6180	21.5886	30.0866

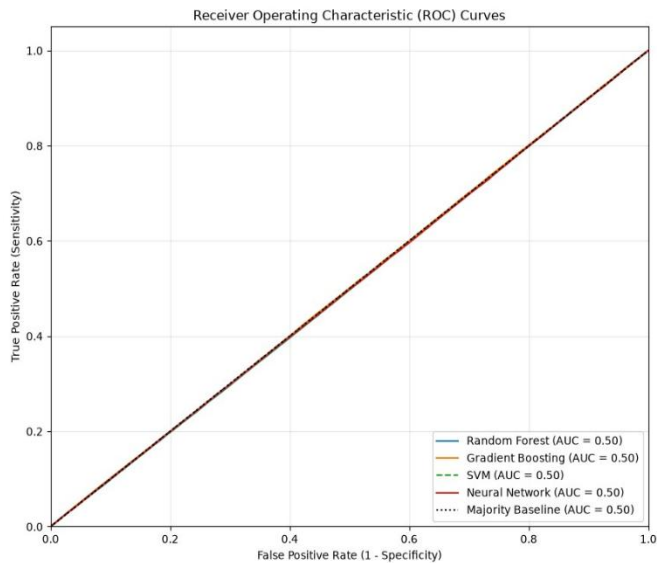


Figure 8. Receiver operating characteristic (ROC) curves for all evaluated models

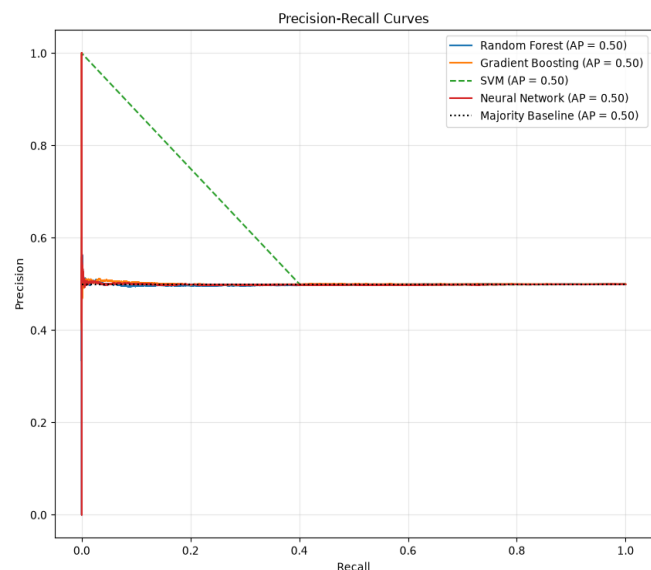


Figure 9. Precision-recall curves

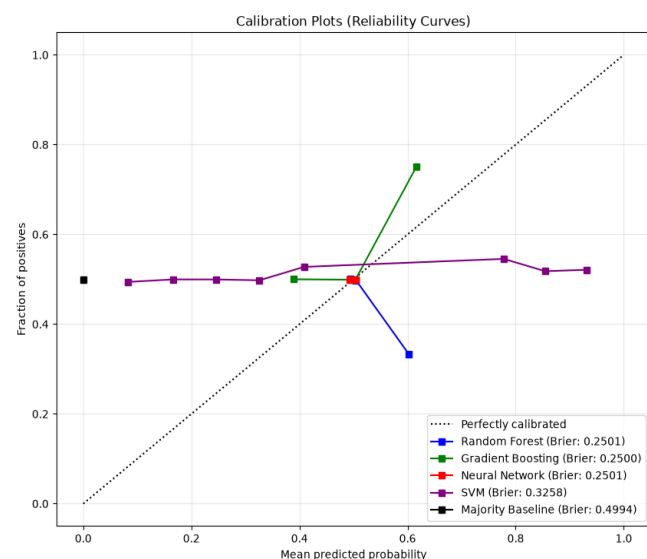


Figure 10. Calibration plots (reliability curves)

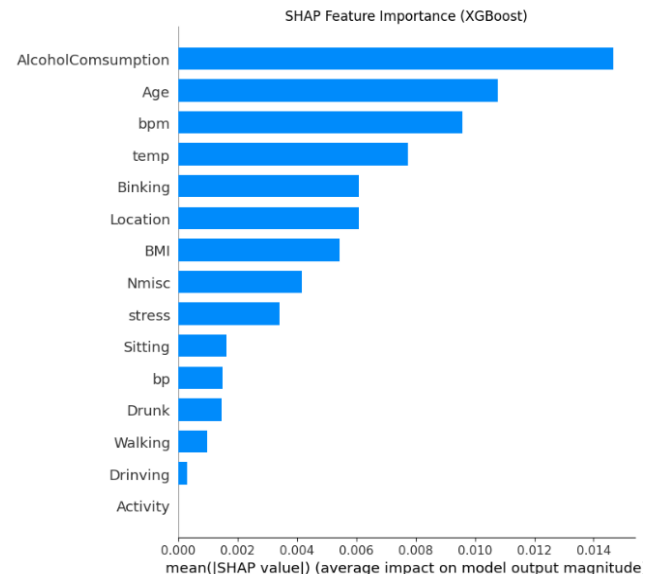


Figure 11. Result SHapley Additive exPlanations (SHAP) testing

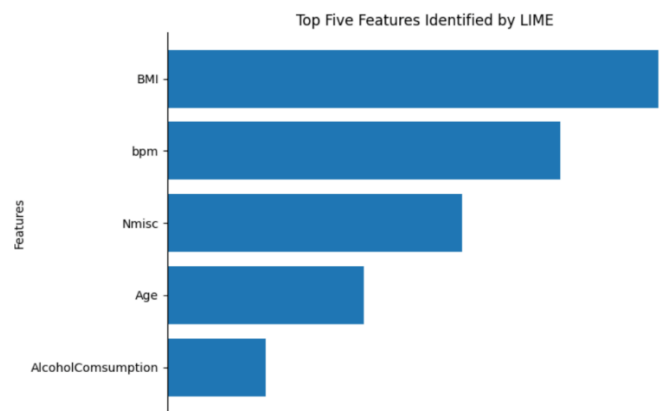


Figure 12. Result Local Interpretable Model-agnostic Explanations (LIME) testing

The calibration plots presented in Figure 10 illustrate differences in probability estimation among the evaluated models. RF, GB, and NN achieved Brier Scores of approximately 0.2500, which were lower than that of the Majority Class Baseline (0.4994). These results indicate differences in the calibration of predicted probabilities within the benchmark dataset. However, all evaluated models exhibited discrimination performance close to chance level, as reflected by their comparable AUC and AP scores. Therefore, the observed calibration characteristics should be interpreted only as descriptive properties of the probability estimates rather than evidence of clinically useful predictive performance. Among the evaluated models, GB followed the reference calibration line more closely across most probability intervals, although localized deviations were observed around the 0.6 probability threshold. RF and NN exhibited similar deviations at higher probability levels, while SVM showed greater dispersion around the reference line and a higher Brier Score (0.3258). These findings describe differences in probability estimation behavior across the evaluated algorithms but do not alter the overall conclusion that the benchmark dataset provides limited discriminative information for spontaneous abortion prediction.

Figure 11 presents the SHAP feature ranking based on average absolute SHAP values. AlcoholConsumption consistently received the highest global feature importance score, followed by Age, BPM, and body temperature, indicating that these variables contributed most strongly to the model outputs within the benchmark dataset. A second tier of contributors included Binking, Location, BMI, and Nmisc, whose SHAP values were present but comparatively lower. At the other end of the spectrum, Walking, Drinving, and Activity registered the smallest SHAP values, suggesting that these behavioral indicators had only a limited influence on the model outputs. Overall, the distribution of feature importance was concentrated in a small subset of variables, while the remaining features contributed relatively little to the model behavior observed in the evaluated benchmark dataset.

Figure 12 presents the five most influential features identified by the LIME local explanation for a representative prediction. The analysis identified BMI, bpm, Nmisc, Age, and AlcoholComsumption as the features that contributed most to the model's local decision for the selected instance. Rather than displaying raw feature values or automatically generated threshold intervals, the revised visualization summarizes the ranking of the locally influential features to improve readability and avoid potential misinterpretation. The LIME analysis complements the global feature importance obtained from SHAP by providing an instance-specific explanation of model behavior. As with the SHAP analysis, these results should be interpreted as explanations of model behavior within the benchmark dataset rather than as evidence of clinically validated risk factors.

4.2 Discussion

The overall findings indicate that the evaluated classifiers were unable to extract a reliable discriminative signal for spontaneous abortion from the available benchmark dataset. Although RF and GB showed somewhat more balanced classification behavior than SVM and NN, the results across multiple evaluation procedures consistently converged toward chance-level discrimination. This suggests that the observed differences between algorithms were relatively modest and should not be interpreted as evidence that one classifier substantially outperformed the others. More importantly, the consistency of this pattern across conventional metrics, cross-validation, statistical comparison, robustness analysis, and discrimination curves indicates that the limitation lies primarily in the predictive information available in the dataset rather than in the choice of a particular classifier.

The comparison with previous research further highlights the importance of the feature set and data provenance. Studies based on clinically verified biomarkers have generally reported stronger predictive performance than approaches relying primarily on demographic, behavioral, or self-reported variables. This difference suggests that clinically meaningful biological information may be more important for spontaneous abortion prediction than increasing model complexity alone. Consequently, improving the quality and clinical relevance of input variables may be a more productive direction than simply optimizing the existing classifiers.

The statistical and robustness analyses provide an additional methodological perspective. The absence of consistent pairwise differences between classifiers indicates that small variations in performance should be interpreted cautiously. At the same time, the stability of the overall pattern across

different evaluation procedures strengthens the conclusion that the benchmark provides limited predictive information. Thus, the study demonstrates the value of combining conventional performance metrics with statistical and robustness analyses when evaluating ML models on challenging healthcare datasets.

The interpretability analysis provides complementary information about model behavior. SHAP indicated that AlcoholConsumption, Age, BPM, and body temperature were among the most influential features at the global level, whereas LIME identified BMI, BPM, Nmisc, Age, and AlcoholComsumption as the most influential features for the representative local instance. The difference between the global and local rankings is expected because SHAP summarizes model behavior across observations whereas LIME explains an individual prediction. However, because the underlying models exhibited limited discriminative ability, these explanations should be regarded as descriptions of model behavior within the benchmark rather than evidence of clinically validated risk factors.

The ablation analysis further suggests that the lowest-ranked behavioral features contributed little to overall model behavior. Removing Walking, Drinving, and Activity did not substantially alter the performance of the evaluated classifiers, although a minor change was observed for NN. This finding is consistent with the broader interpretability results, but it should not be interpreted as definitive evidence that these variables have no predictive relevance, particularly given the limited discriminative performance of the underlying models.

Taken together, the findings support the interpretation of this study as a methodological benchmark rather than a clinically deployable prediction system. The main contribution is therefore not the identification of a high-performing classifier, but the demonstration that model performance, statistical significance, robustness, and explainability should be evaluated jointly before predictive claims are made. Future investigations should prioritize clinically validated datasets and clinically informed feature selection, together with broader model evaluation and external validation.

5. LIMITATIONS

Several limitations of this study should be acknowledged. First, the analysis relied on a single publicly available benchmark dataset whose clinical provenance could not be independently verified because the original repository does not provide detailed information regarding patient recruitment, healthcare institutions, ethical approval, or data acquisition procedures. Consequently, the findings should be interpreted within the context of methodological evaluation rather than as evidence directly generalizable to routine clinical practice.

Second, the dataset exhibits several characteristics that differ from those commonly observed in real-world clinical datasets, including a nearly balanced class distribution, the absence of missing values, a restricted maternal age range of 16-26 years, and generally weak feature-target correlations. In addition, several clinically established predictors of spontaneous abortion, such as gestational age, ultrasound findings, bleeding symptoms, progesterone level, β -hCG, chromosomal abnormalities, infection status, thyroid function, and detailed obstetric history, were not available in the public dataset. Furthermore, the original repository provides limited

documentation for several variables, including their semantic definitions and coding schemes, which may restrict their clinical interpretation and preprocessing. These limitations may have reduced the discriminative information available to the evaluated models and therefore constrained predictive performance.

Third, only four ML classifiers were evaluated. Although these models represent widely used classification approaches, alternative algorithms or ensemble strategies may demonstrate different behavior on the same dataset. Likewise, SHAP and LIME were the only explainability techniques investigated, and the identified feature importance has not been clinically validated.

Finally, no modifications were made to the original dataset to artificially improve model performance. The reported results therefore reflect the inherent characteristics of the publicly available benchmark dataset used in this study. Future research should validate the proposed framework using independently collected and clinically verified hospital datasets with broader demographic diversity, while also exploring additional ML models and explainability techniques to improve robustness, generalizability, and clinical applicability.

6. CONCLUSION AND FUTURE WORK

This study evaluated the performance and interpretability of four ML classifiers for spontaneous abortion prediction using a publicly available benchmark dataset containing multi-parameter maternal information. RF and GB produced more balanced classification outcomes than SVM and NN, particularly on recall and F1-score, though the differences were modest rather than dramatic. Across cross-validation, confidence interval estimation, Friedman statistical testing, Nemenyi post-hoc comparison, and robustness analysis, the four models converged toward similar overall performance levels, with RF standing out for the consistency of its results across different evaluation conditions. However, the overall predictive performance of all evaluated models remained close to chance level, indicating that the available benchmark dataset contains limited discriminative information for reliable spontaneous abortion risk prediction. The advanced metrics including sensitivity, specificity, MCC, and Cohen's Kappa confirmed that none of the classifiers achieved meaningful discrimination beyond chance, with all models recording MCC and Kappa values near zero. ROC-AUC and Precision-Recall analyses further reinforced this finding, with all four models producing an AUC and AP of 0.50, identical to the Majority Class Baseline. Calibration analysis revealed that RF, GB, and NN produced well-calibrated probability estimates despite their near-random discriminative ability, with Brier Scores of approximately 0.25 compared to the baseline value of 0.4994.

The interpretability analysis complemented the quantitative evaluation by providing insights into model behavior beyond conventional performance metrics. SHAP identified AlcoholConsumption, Age, bpm, and temp as the features with the strongest global influence on model predictions. Complementing this global perspective, the LIME analysis identified BMI, bpm, Nmisc, Age, and AlcoholConsumption as the five most influential features for a representative prediction, providing an instance-specific explanation of the model's decision. The ablation study further demonstrated that

removing the three lowest-ranked features (Walking, Drivning, and Activity) resulted in only minor performance changes across all evaluated models, supporting the limited contribution of these variables within the benchmark dataset. Together, SHAP and LIME provided complementary global and local perspectives on model interpretability. However, given the overall predictive performance close to random chance, these explanations should be interpreted as describing model behavior within the benchmark dataset rather than as evidence of clinically validated risk factors.

These conclusions should be interpreted in light of the characteristics of the publicly available benchmark dataset used in this study. Although the dataset contains one million observations, its nearly balanced class distribution, absence of missing values, restricted maternal age range (16-26 years), weak feature-target correlations, and limited availability of clinically established predictors distinguish it from routine clinical datasets. In addition, the original repository does not provide sufficient information regarding patient recruitment, healthcare institutions, ethical approval, or data acquisition procedures, preventing independent verification of the clinical provenance of the records. Consequently, the reported findings should be regarded as a methodological evaluation of ML models and explainability techniques on a benchmark dataset rather than as evidence supporting a clinically applicable spontaneous abortion risk prediction tool.

Future work in this area would benefit from validation against larger and clinically verified datasets with confirmed provenance, the application of expert-guided feature selection prior to modeling, inclusion of a broader range of classification algorithms including transformer-based tabular models, and the application of additional explainability approaches beyond SHAP and LIME. Such studies are necessary to determine whether reliable and clinically applicable spontaneous abortion prediction models can be achieved using independently collected and clinically verified real-world datasets.

AUTHOR CONTRIBUTIONS

All authors contributed significantly to the writing of this paper. A.S.S. contributed to conceptualization, methodology, investigation, data accuracy, formal analysis, and validation. S.R. and S.S. contributed to formal analysis and the provision of resources. H.L.H.S.W. contributed to validation, drafting, and visualization of the test. A.T.N. assisted in writing the paper. M.K.M. contributed to providing advice and supervision of this research. Finally, all authors have read and approved the version of the manuscript to be published.

ACKNOWLEDGMENT

This research is supported by Pelita Bangsa University as part of the Research Funding for the Research Information Base and Community Service (BIMA) entitled "Development of a Multi-Parameter Predictive Model for Miscarriage Risk Identification: Integration of Biomarkers and ML", with Master Contract Date: March 9, 2026, Master Contract Number: 131/C3/DT.05.00/PL-MULTITAHUN LANJUTAN/2026, and Derivative Contract Date: April 29, 2026, with Derivative Contract Numbers: 1997/LL4/PG/2026 and 001/7/KP.HMT/UPB/2026.

REFERENCES

- [1] Ghosh, J., Papadopoulou, A., Devall, A.J., et al. (2021). Methods for managing miscarriage: A network meta-analysis. *The Cochrane Database of Systematic Reviews*, 6(6): CD012602. <https://doi.org/10.1002/14651858.CD012602.pub2>
- [2] Khadra, M.M., Suradi, H.H., Amarín, J.Z., et al. (2022). Risk factors for miscarriage in Syrian refugee women living in non-camp settings in Jordan: Results from the Women ASPIRE cross-sectional study. *Conflict and Health*, 16(1): 32. <https://doi.org/10.1186/s13031-022-00464-y>
- [3] Podilyakina, Y., Stabayeva, L., Kulov, D., et al. (2025). Risk factors for first-trimester spontaneous abortion and the role of preconception care. *Frontiers in Global Women's Health*, 6: 1615983. <https://doi.org/10.3389/fgwh.2025.1615983>
- [4] Al-Alami, Z., Abu-Huwajj, R., Hamadneh, S., Taybeh, E. (2024). Understanding miscarriage prevalence and risk factors: Insights from women in Jordan. *Medicina*, 60(7): 1044. <https://doi.org/10.3390/medicina60071044>
- [5] Tsbizova, V., Tosto, V., Clerici, G., Meyyazhagan, A., Di Renzo, G.C. (2024). Miscarriage: Risk factors and prevention. In *The Continuous Textbook of Women's Medicine Series-Obstetrics Module: Vol. 19. Pregnancy shortening: Etiology, Prediction and Prevention*. Global Library of Women's Medicine. <https://doi.org/10.3843/GLOWM.419023>
- [6] American College of Obstetricians and Gynecologists (ACOG). (2023). Repeated miscarriages. ACOG Frequently Asked Questions FAQ100. <https://www.acog.org/womens-health/faqs/repeated-miscarriages>.
- [7] Kanji, S., Carmichael, F., Darko, C., Egyei, R., Vasilakos, N. (2024). The impact of early marriage on the life satisfaction, education and subjective health of young women in India: A longitudinal analysis. *The Journal of Development Studies*, 60(5): 705-723. <https://doi.org/10.1080/00220388.2023.2284678>
- [8] Quenby, S., Gallos, I.D., Dhillon-Smith, R.K., et al. (2021). Miscarriage matters: The epidemiological, physical, psychological, and economic costs of early pregnancy loss. *The Lancet*, 397(10285): 1658-1667. [https://doi.org/10.1016/S0140-6736\(21\)00682-6](https://doi.org/10.1016/S0140-6736(21)00682-6)
- [9] Emami, F., Eftekhari, M., Jalaliani, S. (2022). Correlation between clinical and laboratory parameters and early pregnancy loss in assisted reproductive technology cycles: A cross-sectional study. *International Journal of Reproductive BioMedicine*, 20(8): 683-690. <https://doi.org/10.18502/ijrm.v20i8.11757>
- [10] Huang, C., Long, X., van der Ven, M., Kaptein, M., Oei, S.G., van den Heuvel, E. (2024). Predicting preterm birth using electronic medical records from multiple prenatal visits. *BMC Pregnancy and Childbirth*, 24(1): 843. <https://doi.org/10.1186/s12884-024-07049-y>
- [11] Ahmed, V.A., Hamad, N.S., Balaky, H.M., Ismail, P.A., Ahmed, B.S. (2025). Early detection of miscarriage in recurrent pregnancy loss using a novel urinary biomarker. *Al-Rafidain Journal of Medical Sciences*, 9(2): 182-187. <https://doi.org/10.54133/ajms.v9i2.2412>
- [12] Li, P. (2024). Machine learning techniques for pattern recognition in high-dimensional data mining. *arXiv arXiv:2412.15593*. <https://doi.org/10.48550/arXiv.2412.15593>
- [13] Salim, A., Juliandry, Raymond, L., Moniaga, J.V. (2023). General pattern recognition using machine learning in the cloud. *Procedia Computer Science*, 216: 565-570. <https://doi.org/10.1016/j.procs.2022.12.170>
- [14] Dritsas, E., Trigka, M. (2025). Machine learning in information and communications technology: A survey. *Information*, 16(1): 8. <https://doi.org/10.3390/info16010008>
- [15] Xu, X., Li, J.Q., Zhu, Z.N., et al. (2024). A comprehensive review on synergy of multi-modal data and AI technologies in medical diagnosis. *Bioengineering*, 11(3): 219. <https://doi.org/10.3390/bioengineering11030219>
- [16] Marcinkevičs, R., Vogt, J.E. (2023). Interpretable and explainable machine learning: A methods-centric overview with concrete examples. *WIREs Data Mining and Knowledge Discovery*, 13(3): e1493. <https://doi.org/10.1002/widm.1493>
- [17] Tumwesige, I., Kawooya, B.I., Bwogi, F., Lule, E., Nakayiza, H., Chongomweru, H. (2024). Interpretable machine learning techniques for predictive cattle behavior monitoring. In *2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, Coimbatore, India, pp. 1219-1224. <https://doi.org/10.1109/ICSCSS60660.2024.10625182>
- [18] Hassija, V., Chamola, V., Mahapatra, A., et al. (2024). Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation*, 16: 45-74. <https://doi.org/10.1007/s12559-023-10179-8>
- [19] Fadil, M.S., Mamo, A.A., Gebresilassie, B.G., et al. (2025). Unlocking the power of machine learning in big data: A scoping survey. *Data Science and Management*, 8(4): 519-535. <https://doi.org/10.1016/j.dsm.2025.02.004>
- [20] Sarker, I.H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3): 160. <https://doi.org/10.1007/s42979-021-00592-x>
- [21] Vargas-Santiago, M., León-Velasco, D.A., Maldonado-Sifuentes, C.E., Chanona-Hernandez, L. (2025). A state-of-the-art review of artificial intelligence (AI) applications in healthcare: Advances in diabetes, cancer, epidemiology, and mortality prediction. *Computers*, 14(4): 143. <https://doi.org/10.3390/computers14040143>
- [22] Seyam, E.A. (2025). Predicting high-cost healthcare utilization using machine learning: A multi-service risk stratification analysis in EU-based private group health insurance. *Risks*, 13(7): 133. <https://doi.org/10.3390/risks13070133>
- [23] Mukhamediev, R.I., Popova, Y., Kuchin, Y., et al. (2022). Review of artificial intelligence and machine learning technologies: Classification, restrictions, opportunities and challenges. *Mathematics*, 10(15): 2552. <https://doi.org/10.3390/math10152552>
- [24] Macrohon, J.J.E., Villavicencio, C.N., Inbaraj, X.A., Jeng, J.H. (2022). A semi-supervised machine learning approach in predicting high-risk pregnancies in the Philippines. *Diagnostics*, 12(11): 2782. <https://doi.org/10.3390/diagnostics12112782>
- [25] Alonso, E., Beristain, A., Burgos, J., Gurrutxaga, I. (2025). Comparison of machine learning algorithms to predict down syndrome during the screening of the first trimester of pregnancy. *Applied Sciences*, 15(10): 5401.

- <https://doi.org/10.3390/app15105401>
- [26] Sufian, M.A., Hamzi, W., Hamzi, B., et al. (2024). Innovative machine learning strategies for early detection and prevention of pregnancy loss: The vitamin D connection and gestational health. *Diagnostics*, 14(9): 920. <https://doi.org/10.3390/diagnostics14090920>
- [27] Vasudevan, L., Kibria, M.G., Kucirka, L.M., et al. (2025). Machine learning models to predict risk of maternal morbidity and mortality from electronic medical record data: Scoping review. *Journal of Medical Internet Research*, 27: e68225. <https://doi.org/10.2196/68225>
- [28] Pavagada, K., Vemuri, J. (2025). Prediction of maternal health risk factors using machine learning algorithms. *Procedia Computer Science*, 258: 2713-2722. <https://doi.org/10.1016/j.procs.2025.04.532>
- [29] Engesser, C., Henkel, M., Alargkof, V., et al. (2024). Clinical decision making in prostate cancer care-evaluation of EAU-guidelines use and novel decision support software. *Scientific Reports*, 14: 19113. <https://doi.org/10.1038/s41598-024-70292-y>
- [30] Mienye, I.D., Jere, N. (2024). A survey of decision trees: Concepts, algorithms, and applications. *IEEE Access*, 12: 86716-86727. <https://doi.org/10.1109/ACCESS.2024.3416838>
- [31] Wong, K.K.L. (2024). Support vector machine. In *Cybernetical Intelligence: Engineering Cybernetics with Machine Intelligence*, pp. 149-176. <https://doi.org/10.1002/9781394217519.ch8>
- [32] Rohan, T.K., Padmakara, T., G, A. (2025). Employee attrition using gradient boosting. In *2025 8th International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India, pp. 821-825. <https://doi.org/10.1109/ICOEI65986.2025.11012971>
- [33] Zhang, C.S., Liu, L.J. (2025). Machine learning prediction model for medical environment comfort based on SHAP and LIME interpretability analysis. *Scientific Reports*, 15: 39269. <https://doi.org/10.1038/s41598-025-22972-6>
- [34] Kukkar, A., Kaur, G. (2025). AEC: A novel adaptive ensemble classifier with LIME and SHAP-based interpretability for fake news detection. *Expert Systems with Applications*, 281: 127751. <https://doi.org/10.1016/j.eswa.2025.127751>
- [35] Lei, C.J., Cao, Y.C., Zhao, Y., et al. (2026). A machine learning model based on clinical, metabolic, and inflammatory indicators for predicting recurrent spontaneous abortion. *Journal of Reproductive Immunology*, 173: 104832. <https://doi.org/10.1016/j.jri.2026.104832>
- [36] Chen, D., Liu, A., Wang, X., et al. (2026). A multidimensional clinical prediction model for early screening of recurrent spontaneous abortion: Integrating coagulation, immune, and endocrine markers. *Frontiers in Immunology*, 17: 1774359. <https://doi.org/10.3389/fimmu.2026.1774359>
- [37] Wu, J., Yu, X.X., Li, M.T., et al. (2024). Risk prediction model based on machine learning for predicting miscarriage among pregnant patients with immune abnormalities. *Frontiers in Pharmacology*, 15: 1366529. <https://doi.org/10.3389/fphar.2024.1366529>
- [38] Zhu, Z.N., Wei, N., Guo, J.J., et al. (2025). Predicting the risk of threatened abortion using machine learning methods: A comparative study. *BMC Pregnancy and Childbirth*, 25: 901. <https://doi.org/10.1186/s12884-025-08030-z>
- [39] Liu, L.D., Liu, B., Wu, H.M., Gan, Q.Y., Huang, Q.Y., Li, M.J. (2025). Optimizing predictive features using machine learning for early miscarriage risk following single vitrified-warmed blastocyst transfer. *Frontiers in Endocrinology*, 16: 1557667. <https://doi.org/10.3389/fendo.2025.1557667>
- [40] Asri, H. (2021). HIBA ASRI: Miscarriage Prediction Risk Factors (Version 1). Mendeley Data. <https://doi.org/10.17632/5sbmhh6t3r.1>