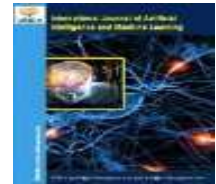


SvedbergOpen
DISSEMINATION OF KNOWLEDGEInternational Journal of Artificial
Intelligence and Machine LearningPublisher's Home Page: <https://www.svedbergopen.com/>

Research Paper

Open

Reassessing Maternal Health Risk Prediction: Duplicate-Aware Validation, Label Ambiguity, and Explainable Machine Learning

Ankit Garg^{1*}, Khushboo Tripathi², Ashima Rani³, Nishu Sethi⁴, Anshu Malhotra⁵¹*UCRD, UIE, Chandigarh University, Mohali, Punjab, India. E-mail: ankitgitm@gmail.com, Orcid: 0000-0002-6466-2738²Department of Computer Science & Applications, Sharda University, Greater Noida, Uttar Pradesh, India,E-mail: khushbootripathi.cse@gmail.com, Orcid: 0000-0002-3344-4359³Department of Computer Science and Engineering, The NorthCap University Gurugram, India. E-mail: ashima.ashugambhir@gmail.com, Orcid: 0000-0002-5276-2897⁴School of computer science and engineering, IILM UNIVERSITY, Gurugram, India. E-mail: sethinishu@gmail.com, Orcid: 0000-0001-5286-7608⁵Centre of Excellence - Cyber Security, School of Engineering and Technology, K. R. Mangalam University Gurugram, India, E-mail: anshu@krmangalam.edu.in, Orcid: 0000-0003-4208-8007

Abstract

Maternal health risk classification is increasingly studied with machine learning, but benchmark performance can be misleading when repeated or dependent records are allowed to occur in both training and test folds. This study performs a duplicate-aware re-evaluation of the public UCI Maternal Health Risk benchmark using seven classifiers and explainable artificial intelligence. The analyzed snapshot contained 1,014 records with six physiological or demographic predictors and three risk classes. A data-quality audit identified 562 exact duplicate rows (55.42%), only 416 unique feature profiles, and 35 profiles with contradictory class labels; 215 records (21.20%) belonged to such label-ambiguous profiles. Conventional stratified cross-validation was therefore compared with profile-grouped stratified cross-validation that forced identical six-feature profiles into the same fold. Under repeated 3×5 -fold group-aware validation, XGBoost achieved 0.662 ± 0.047 accuracy, 0.645 ± 0.053 macro-F1, 0.498 ± 0.071 Matthews correlation coefficient, and 0.844 ± 0.037 macro one-vs-rest AUC. In contrast, conventional cross-validation produced markedly higher macro-F1 for Extra Trees (0.857 vs 0.610), Random Forest (0.855 vs 0.634), and XGBoost (0.779 vs 0.645). The mean macro-F1 inflation was 0.247, 0.221, and 0.134, respectively. A Friedman test showed significant model differences under group-aware validation ($p = 6.55 \times 10^{-4}$); XGBoost significantly exceeded Extra Trees and logistic regression after Holm correction but not Random Forest. Out-of-fold analysis confirmed that the mid-risk class remained the most difficult ($F1 = 0.404$). SHAP and permutation analyses consistently ranked blood sugar and systolic blood pressure as the dominant predictors. Training-fold-only SMOTE produced only a small macro-F1 increase (~ 0.007), indicating that duplicate dependence and label ambiguity, rather than class imbalance alone, are central limitations of this benchmark. The findings demonstrate that leakage-resistant validation and explicit data auditing are essential before claiming clinically relevant performance on maternal-health benchmark data.

Keyword: maternal health risk; machine learning; data leakage; duplicate-aware cross-validation; XGBoost; explainable AI; SHAP; reproducibility

Introduction

Maternal mortality remains a major global health challenge. The 2025 inter-agency estimates led by the World Health Organization report that more than 700 women died each day in 2023 from preventable causes related to pregnancy and childbirth, while Sustainable Development Goal 3.1 calls for a global maternal mortality ratio below 70 deaths per 100,000 live births by 2030 [1]. Low-cost risk stratification based on routinely measurable physiological variables is therefore attractive for resource-constrained settings, especially when computational tools can support—not replace—clinical assessment.

The UCI Maternal Health Risk dataset has become a widely used benchmark for multiclass pregnancy-risk prediction. Its six predictors—age, systolic blood pressure, diastolic blood pressure, blood sugar, body temperature, and heart rate—originate from an IoT-oriented maternal monitoring context in rural Bangladesh [2,3]. Recent studies have applied conventional classifiers, ensemble learning, oversampling, feature selection, and explainability to this dataset, with reported accuracies ranging from the mid-80% range to substantially higher values [4,5,7,8]. Such results have encouraged continued development of automated maternal risk screening models.

However, predictive performance is meaningful only when the validation procedure reflects truly unseen information. Leakage can occur when non-independent or duplicated observations are separated across training and testing sets, allowing a model to recognize rather than generalize. Kapoor and Narayanan documented data leakage as a broad reproducibility problem across machine-learning-based science [9], and the REFORMS guidelines explicitly recommend keeping dependent or duplicate samples within the same partition [10]. This concern is especially relevant to compact clinical benchmark datasets containing repeated measurement profiles.

The motivating published study supplied with this work used a different 299-record preeclampsia dataset and reported 78% accuracy with Random Forest, SHAP, and LIME, while the moderate-risk class was the weakest [6]. Rather than paraphrasing or reproducing that article, the present study adopts its general interest in interpretable maternal-risk prediction but addresses a separate methodological question on a public dataset: how much does duplicate dependence change apparent model performance?

The main contributions are:

- a systematic audit of exact duplication, repeated feature profiles, contradictory labels, and implausible physiological values in the 1,014-row benchmark snapshot;
- a direct comparison between conventional stratified cross-validation and duplicate-aware profile-grouped stratified cross-validation across seven classifiers;
- a repeated paired evaluation with macro-F1, balanced accuracy, MCC, Cohen's kappa, and macro one-vs-rest AUC, followed by Friedman and Holm-adjusted Wilcoxon tests;
- class-wise error analysis and two complementary explainability methods (permutation importance and SHAP); and
- sensitivity analyses for training-fold-only SMOTE, implausible heart-rate removal, and an unambiguous unique-profile subset.

2. Related Work

Ahmed et al. proposed an Internet-of-Things-oriented approach to monitoring maternal risk factors in remote areas [3]. Its contribution is the use of routinely measurable maternal indicators in a remote-monitoring context; the associated UCI benchmark provides six predictors and three risk labels [2]. The repository reports no missing predictor values, but its current 1,013-instance listing differs from the 1,014-row CSV snapshot analyzed here. The dataset and these sources do not establish patient independence or resolve repeated profiles, motivating the present data audit. Maheswari et al. proposed a maternal-risk classification comparison incorporating SMOTE and explainable AI [4]. The distinguishing methodological combination is imbalance handling alongside model interpretation; the reported optimized XGBoost accuracy was 88.89%. Pavagada et al. proposed a comparison of machine-learning algorithms on 1,014 maternal-risk observations [5]. Their reported finding was 84.12% accuracy for bagged trees. Neither reported headline figure alone establishes performance on unseen, nonoverlapping feature profiles; the present study therefore evaluates the effect of profile-aware partitioning rather than treating these results as directly comparable. Khadidos et al. proposed an ensemble machine-learning framework for pregnancy-related maternal risk prediction [7]. Its contribution is an ensemble formulation for the three-class task; the cited study reports class-wise predictive performance, but the present source material does not provide a verified numerical result or evidence of profile-grouped validation. Mamun et al. proposed a framework combining feature reduction, multiple ensembles and SHAP, LIME and ICE explanations [8]. Its distinguishing feature is the joint use of feature selection and multiple interpretability techniques, with high cross-validated performance reported in the cited study. The numerical results and the independence of its validation partitions have not been independently verified here; these studies motivate a direct audit of duplicate dependence. Yanto et al. proposed Random Forest-based preeclampsia prediction with SHAP and LIME explanations on a separate 299-record dataset [6]. Its distinctive contribution is the combination of predictive modeling and local/global explanations in a preeclampsia-specific setting. They reported 78% accuracy and weaker performance for the moderate-risk category. Because the dataset and clinical target differ, their results provide context for class-wise error analysis rather than a direct benchmark for the UCI maternal-risk task. Garg et al. presented an overview of IoT and artificial intelligence for healthcare informatics [21]. Its contribution is a healthcare-technology synthesis rather than a new maternal-risk classifier; the CV citation does not provide experimental performance findings. Garg et al. addressed explainable AI for mental-disorder management [22]. Its relevance is interpretability in a different healthcare setting, but the available bibliographic details do not substantiate a specific algorithmic novelty or numerical result. These works contextualize the monitoring and explainability components of the present study without being represented as maternal-risk validation evidence. Garg et al. presented a cloud-based cardiovascular early-detection system [23], with a proposed cloud deployment setting as its application-level distinction. Saini et al. proposed a hybrid cardiovascular diagnosis model using clinical features, ECG and MRI [24]; its distinguishing feature is the

integration of heterogeneous clinical input modalities. The supplied CV identifies the approaches but does not report verified performance values, so no numerical findings are attributed to either. Their clinical tasks and modalities differ from the six-feature maternal benchmark and their outcomes are not used for cross-study score comparisons. Patil et al. presented an AI-driven approach to detecting data leakage in multi-cloud environments using observability and machine learning [25]. The proposed approach concerns infrastructure-level leakage detection, not training/test contamination in clinical predictive modeling. Its relevance is terminological and contextual only; the CV provides no numerical findings that can be reliably summarized. For the specific methodological issue studied here, Kapoor and Narayanan [9] and the REFORMS guidance [10] provide the directly relevant rationale for separating dependent samples.

Research gap and contribution. Existing maternal-risk studies report classifier accuracy, ensembles, imbalance handling and explainability [4–8], while related healthcare AI publications address monitoring, interpretability and other clinical tasks [21–24]. The cited literature as summarized here does not establish how much exact duplicate feature profiles and contradictory labels change apparent performance on this particular CSV snapshot. The present study therefore audits redundancy and label ambiguity, compares conventional with profile-grouped cross-validation, reports class-wise and statistical analyses, and interprets SHAP and permutation importance only after leakage-resistant evaluation.

3. Materials and Methods

3.1 Dataset and variables

The analysis used the public Maternal Health Risk benchmark distributed through the UCI Machine Learning Repository (DOI: 10.24432/C5DP5D), licensed under CC BY 4.0 [2]. The analyzed CSV snapshot contained 1,014 rows. The target labels were low risk, mid risk, and high risk. No imputation was required because the analyzed snapshot contained no missing values. Table 1 summarizes the predictors.

Table 1. Variables used in the maternal health risk benchmark.

Variable	Unit	Description
Age	years	Maternal age
SystolicBP	mmHg	Systolic blood pressure
DiastolicBP	mmHg	Diastolic blood pressure
BS	mmol/L	Blood sugar / glucose concentration
BodyTemp	°F	Body temperature
HeartRate	beats/min	Heart rate
RiskLevel	categorical target	Low, mid, or high maternal risk

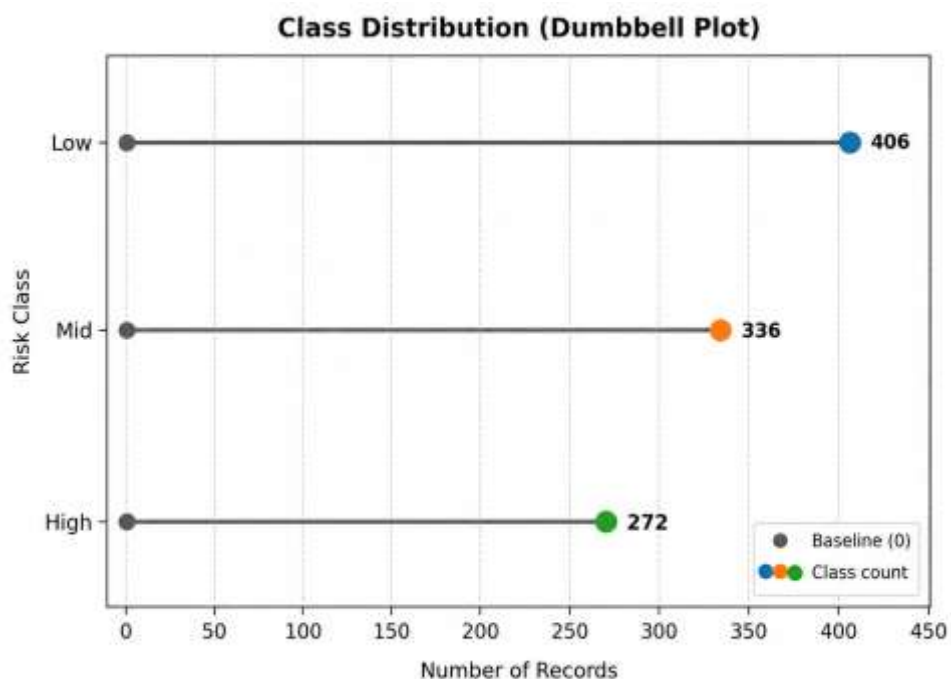


Figure 1. Class distribution in the analyzed 1,014-row dataset.

Figure 1 characterizes the target-label distribution before model training. Low risk accounts for 406/1,014 records (40.04%), mid risk for 336 (33.14%), and high risk for 272 (26.82%). Thus, the largest class has approximately 1.49 times as many records as the smallest class. Although the imbalance is moderate rather than extreme, a classifier that prioritizes the most frequent class could conceal poor recognition of the remaining categories. We therefore use macro-F1 and balanced accuracy alongside overall accuracy, and examine all three classes separately in the out-of-fold confusion matrix (Figure 4). These counts describe dataset labels, not population prevalences of maternal risk.

3.2 Data-quality and redundancy audit

Before any model fitting, we quantified: (i) exact duplicate rows using all six predictors plus the target, (ii) duplicate feature profiles using the six predictors only, (iii) unique labeled records, (iv) unique feature profiles, and (v) contradictory profiles in which identical six-feature inputs were associated with more than one risk label. We also screened for physiologically implausible heart-rate values below 30 beats/min as a simple sensitivity flag rather than silently modifying the dataset.

Table 2. Dataset quality and dependence audit.

Audit item	Result	Interpretation
Total rows	1,014	Analyzed CSV snapshot
Exact duplicate rows	562 (55.42%)	Repeated feature + label rows
Unique labeled rows	452	Unique feature + label combinations
Unique feature profiles	416	Unique six-predictor combinations
Conflicting feature profiles	35	Same predictors, multiple labels
Rows in conflicting profiles	215 (21.20%)	Label-ambiguous observations
Unambiguous unique profiles	381	One row/profile after removing contradictory profiles
HeartRate < 30 bpm	2	Sensitivity flag

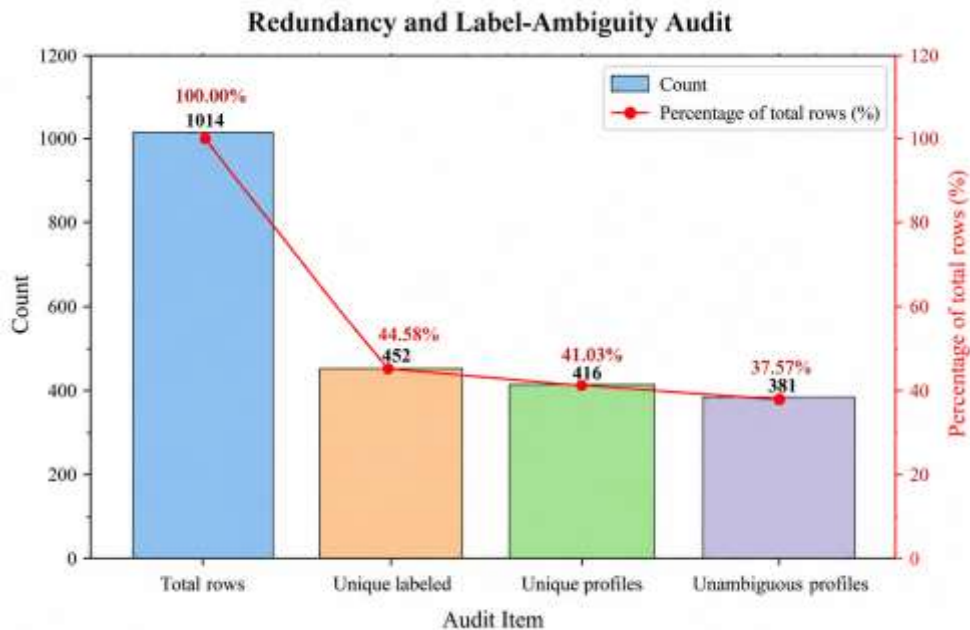


Figure 2. Redundancy and label-ambiguity audit showing the reduction from total rows to unique and unambiguous feature profiles.

Figure 2 summarizes the successive, non-equivalent levels of dataset redundancy. Starting with 1,014 observations, removal of repeated complete rows leaves 452 distinct predictor-and-label combinations; ignoring the label leaves only 416 distinct six-variable input profiles. Excluding the 35 profiles associated with conflicting labels then leaves 381 unambiguous profiles. The reduction from 1,014 to 452 represents 562 duplicate occurrences (55.42%), while the 35-profile difference between 416 and 381 captures ambiguity rather than additional ordinary duplication. Importantly, 215 observations belong to the 35 contradictory profiles, so their influence cannot be inferred from the number of profiles alone. This audit motivates profile-level grouping in the main cross-validation experiment: deduplicating complete rows alone would not prevent an identical input vector with different labels from appearing on opposite sides of a split. The bars represent nested dataset counts and should not be interpreted as independent populations or as a modeled performance trend.

3.3 Duplicate-aware validation design

Two validation regimes were compared. Conventional validation used 5-fold StratifiedKFold with shuffling. Duplicate-aware validation used 5-fold StratifiedGroupKFold, where the grouping identifier was the complete six-feature profile. Thus, every record having the same values for Age, SystolicBP, DiastolicBP, BS, BodyTemp, and HeartRate was forced into a single fold, preventing identical profiles from appearing in both training and test data. The four best group-aware models were then evaluated over three independent shuffle seeds (7, 21, and 42), giving 15 paired test folds per regime. This design follows leakage-control recommendations that dependent or duplicate observations should remain within the same partition [9,10].

A third sensitivity regime removed all contradictory profiles and retained one record per remaining feature profile, producing 381 unambiguous unique profiles. This analysis asks a different question from the main group-aware experiment: how well can the classifiers discriminate labels when exact feature-to-label contradictions are excluded?

3.4 Machine-learning models and preprocessing

Seven classifiers were evaluated: logistic regression, Gaussian Naive Bayes, K-nearest neighbors (KNN), radial-basis-function support vector machine (RBF-SVM), Random Forest, Extra Trees, and XGBoost. Standardization was fitted inside each training fold for logistic regression, KNN, and SVM. Tree-based methods used the raw numeric variables. No feature selection was performed in the primary comparison because all six variables are clinically interpretable and computationally inexpensive.

Table 3. Classifier configurations used in the benchmark.

Model	Configuration
Logistic regression	C=1; class_weight=balanced; max_iter=3000; standardized
Gaussian NB	Default Gaussian Naive Bayes
KNN	k=9; distance weighting; standardized
RBF-SVM	C=10; RBF kernel; class_weight=balanced; standardized
Random Forest	120 trees; max_features=sqrt; balanced_subsample
Extra Trees	120 trees; max_features=sqrt; class_weight=balanced
XGBoost	120 trees; depth=4; learning rate=0.06; subsample=0.9; colsample=0.9

3.5 Evaluation metrics and statistical analysis

Performance was quantified using accuracy, balanced accuracy, macro-F1, weighted-F1, Matthews correlation coefficient (MCC), Cohen's kappa, and macro one-vs-rest ROC-AUC. Macro-F1 was the primary comparison metric because each class contributes equally regardless of its frequency. For the repeated group-aware experiment, a Friedman test compared the four leading models across matched folds. When the omnibus test was significant, one-sided paired Wilcoxon signed-rank tests compared the best mean-performing model with each alternative; p-values were adjusted using Holm's sequential procedure. Statistical significance was assessed at $\alpha = 0.05$ [15–17].

3.6 Explainability and sensitivity analyses

For the best repeated group-aware model, permutation importance was calculated on a representative held-out grouped fold using macro-F1 as the scoring function and 15 permutations per feature. Global SHAP importance was calculated as the mean absolute SHAP value over the held-out samples and classes [13]. Two sensitivity analyses were added. First, the two records with HeartRate < 30 bpm were removed and the repeated group-aware evaluation was rerun. Second, SMOTE [14] was applied strictly inside each training fold before XGBoost fitting; no synthetic sample was allowed to influence the held-out fold. All computations were performed in Python 3.13.5 with pandas 2.2.3, NumPy 2.3.5, scikit-learn 1.8.0, SciPy 1.17.0, XGBoost 3.1.3, SHAP 0.50.0, statsmodels 0.14.6, and imbalanced-learn 0.14.1. Random seeds are reported to support reproducibility.

4. Results

4.1 Dataset structure reveals extensive dependence and label ambiguity

The class distribution was moderately imbalanced: 406 low-risk, 336 mid-risk, and 272 high-risk records (Figure 1). More importantly, 562 rows were exact duplicates, corresponding to 55.42% of the dataset. Only 416 unique six-feature profiles were present. Thirty-five profiles had contradictory target labels, and these profiles accounted for 215 records (21.20%). Consequently, an identical clinical feature vector does not always map to a unique target label in this benchmark.

Table 4. Examples of feature profiles associated with contradictory risk labels.

Age	BP	BS	Temp	HR	n	Observed labels
19	120/80	7	98	70	34	Mid: 27; Low: 7
48	120/80	11	98	88	24	High: 14; Mid: 6; Low: 4
31	120/60	6.1	98	76	20	Mid: 13; Low: 7
40	120/95	11	98	80	9	High: 7; Mid: 2
50	130/100	16	98	75	9	High: 5; Mid: 4
60	120/85	15	98	60	9	Mid: 5; High: 4

The largest contradictory profile (Age 19, SystolicBP 120, DiastolicBP 80, BS 7, BodyTemp 98, HeartRate 70) occurred 34 times but was labeled mid risk 27 times and low risk 7 times. Another profile (Age 48, 120/80 mmHg, BS 11, 98°F, HeartRate 88) appeared 24 times across all three target classes. These contradictions define an irreducible ambiguity for any deterministic classifier using only the six supplied predictors.

4.2 Conventional cross-validation substantially overstates tree-model performance

Table 5. Repeated conventional versus duplicate-aware group validation (3 × 5 folds).

Model	Conventional Acc.	Conventional Macro-F1	Group-aware Acc.	Group-aware Macro-F1	F1 inflation
XGBoost	0.775 ± 0.024	0.779 ± 0.024	0.662 ± 0.047	0.645 ± 0.053	0.134 ± 0.068
Random Forest	0.852 ± 0.021	0.855 ± 0.021	0.657 ± 0.036	0.634 ± 0.040	0.221 ± 0.054
Extra Trees	0.854 ± 0.020	0.857 ± 0.020	0.640 ± 0.045	0.610 ± 0.050	0.247 ± 0.058
Logistic regression	0.602 ± 0.028	0.603 ± 0.025	0.597 ± 0.039	0.596 ± 0.042	0.007 ± 0.041

As visualized in Figure 3, conventional cross-validation yields macro-F1 values of 0.779, 0.855, 0.857, and 0.603 for XGBoost, Random Forest, Extra Trees, and logistic regression, respectively. Grouping identical predictor profiles reduces the corresponding means to 0.645, 0.634, 0.610, and 0.596. The absolute differences are 0.134 for XGBoost, 0.221 for Random Forest, 0.247 for Extra Trees, and 0.007 for logistic regression. The especially wide gaps for the two bagged tree ensembles are consistent with strong sensitivity to shared exact profiles across training and test partitions. By contrast, the small logistic-regression difference suggests a smaller dependence on such repeats under the tested configuration.

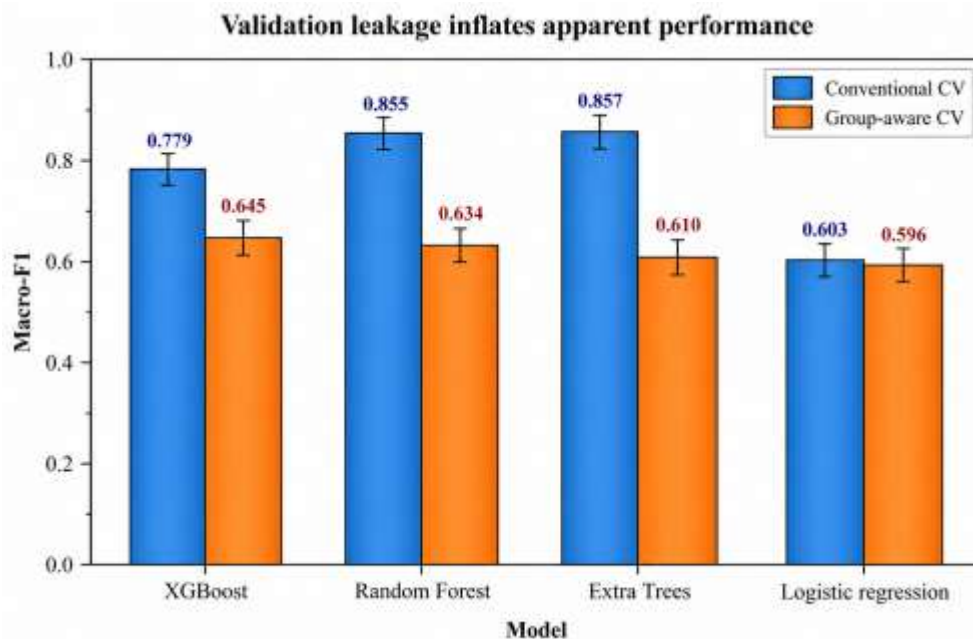


Figure 3. Mean macro-F1 under conventional and profile-grouped cross-validation over 15 paired folds. The gap is largest for Extra Trees and Random Forest.

Figure 3 displays means over 15 paired folds, not independent-patient performance; its bars alone do not establish statistical significance or prove a unique causal mechanism. Paired statistical tests and the documented split design provide the complementary evidence reported in Section 4.3. The effect of the validation regime was model dependent. Extra Trees lost 0.247 ± 0.058 macro-F1 when duplicate feature profiles were prevented from crossing folds; Random Forest lost 0.221 ± 0.054 and XGBoost lost 0.134 ± 0.068 . Logistic regression changed by only 0.007 ± 0.041 . This pattern is consistent with high-capacity nonlinear models obtaining much greater benefit from repeated exact feature profiles than a linear decision boundary.

4.3 Full seven-model benchmark and statistical comparison

Table 6. Seven-model benchmark under duplicate-aware 5-fold grouped validation (seed 42).

Model	Accuracy	Bal. Acc.	Macro-F1	MCC	Macro AUC
XGBoost	0.656	0.655	0.637	0.487	0.846
Random Forest	0.661	0.660	0.635	0.498	0.815
Logistic regression	0.611	0.616	0.609	0.412	0.791
Extra Trees	0.644	0.640	0.608	0.477	0.816
RBF-SVM	0.590	0.596	0.587	0.378	0.761
KNN	0.575	0.576	0.570	0.353	0.756
Gaussian NB	0.596	0.577	0.557	0.401	0.781

In the repeated group-aware comparison, XGBoost had the highest mean macro-F1 (0.645 ± 0.053) and macro-AUC (0.844 ± 0.037), while Random Forest had a similar macro-F1 of 0.634 ± 0.040 . The Friedman test across the four repeated models was significant ($\chi^2 = 17.16$, $p = 0.000655$). Holm-adjusted paired Wilcoxon tests indicated that XGBoost exceeded Extra Trees (adjusted $p = 0.0125$) and logistic regression (adjusted $p = 0.0125$), but the difference from Random Forest was not statistically significant (adjusted $p = 0.1796$). Thus, the data support a performance tier rather than an unequivocal single-model winner.

4.4 Mid-risk classification remains the dominant error mode

Figure 4 makes the nature of the out-of-fold errors explicit (rows are true labels and columns are predictions). Of the 406 low-risk records, 331 are correctly classified, 61 are classified as mid risk, and 14 as high risk. For the 336 mid-risk records, only 109 are correctly classified, whereas 162 are assigned low risk and 65 high risk; these yield mid-risk recall of $109/336 = 0.324$ and the reported F1-score of 0.404. Of 272 high-risk observations, 225 are correctly identified, 34 are predicted mid risk, and 13 low risk.

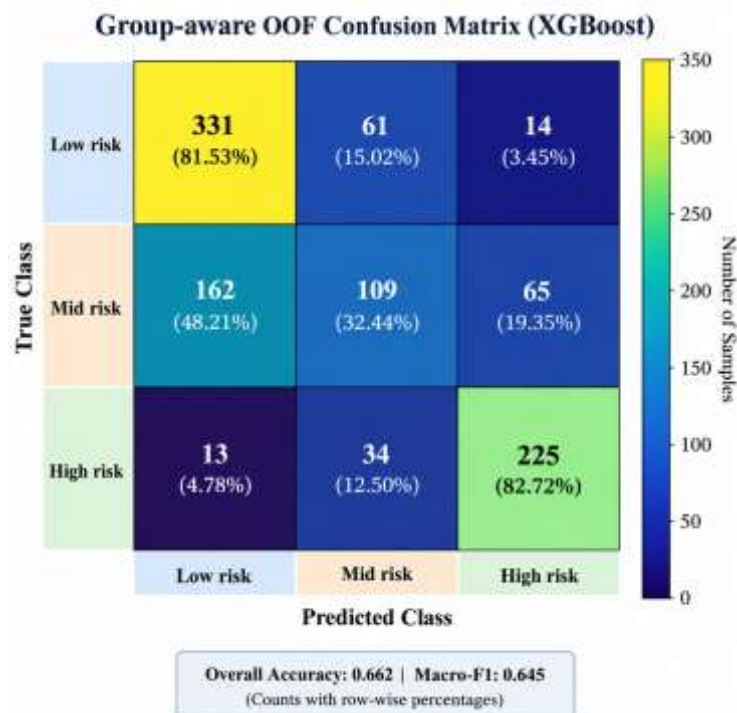


Figure 4. Out-of-fold confusion matrix for XGBoost under duplicate-aware 5-fold validation (seed 42).

Consequently, the intermediate category accounts for 227 misclassifications, including the prominent mid-to-low error pathway. The diagonal of the matrix totals 665 correct predictions among 1,014 out-of-fold predictions

(accuracy approximately 0.656), consistent with the seed-42 XGBoost result in Table 6. Because labels here are benchmark risk categories and not adjudicated clinical outcomes, these errors indicate evaluation weaknesses and potential triage concerns rather than measured patient harm.

Table 7. Class-wise out-of-fold XGBoost performance under group-aware validation.

Class	Precision	Recall	F1-score	Support
Low risk	0.654	0.815	0.726	406
Mid risk	0.534	0.324	0.404	336
High risk	0.740	0.827	0.781	272

The low- and high-risk classes were identified substantially better than the mid-risk class. Mid-risk recall was only 0.324 and F1 was 0.404. Of 336 mid-risk records, 162 were predicted as low risk and 65 as high risk. The corresponding low-risk and high-risk recalls were 0.815 and 0.827, respectively. This result is notable because the independently published 299-record preeclampsia study also reported its weakest performance in the intermediate-risk class [6], suggesting that intermediate categories may require richer features, more consistent labeling, or a reframed ordinal decision problem.

4.5 Explainability identifies blood sugar and systolic pressure as dominant predictors

Figure 5 presents global mean absolute SHAP magnitudes for XGBoost on a held-out grouped fold. Blood sugar (BS) has the largest reported magnitude (0.627), followed by systolic blood pressure (0.499). Body temperature (0.219) and age (0.182) show smaller contributions, while diastolic blood pressure (0.103) and heart rate (0.100) have the lowest values. The figure therefore describes how strongly each input contributes to the fitted model's predictions in this evaluation, averaged in absolute value across held-out samples and classes. Because magnitudes discard sign, the bars do not establish whether an increase in a physiological variable raises or lowers a particular class probability; nor do they demonstrate causality.

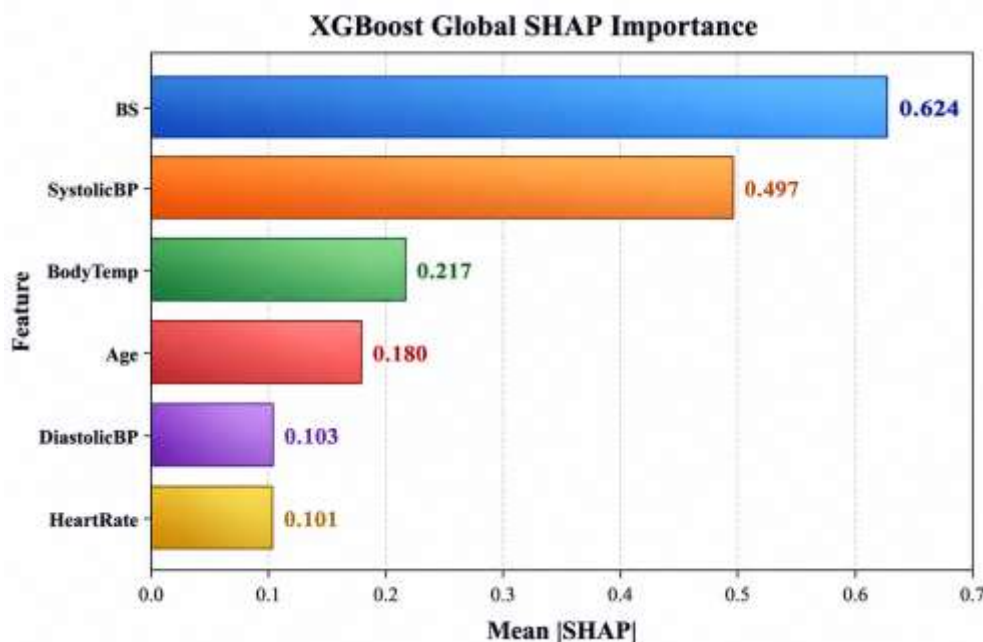


Figure 5. Global XGBoost feature importance measured by mean absolute SHAP value on a held-out grouped fold.

Class-specific SHAP dependence or direction plots would be needed to support such claims.

Figure 6 provides a complementary, directly held-out test of feature reliance by measuring the decrease in macro-F1 when a feature is randomly permuted. Blood sugar produces the largest mean decrease (0.155 ± 0.019), followed by systolic blood pressure (0.123 ± 0.019); the uncertainty bars show variation across 15 permutations, not confidence intervals across independent clinical cohorts. The remaining feature effects are relatively small and some bars approach or cross zero, indicating weak or unstable standalone permutation importance in this fold. Figure 6 agrees with Figure 5 about the first two variables, while their ordering among less influential inputs is not identical: SHAP places body temperature ahead of age, whereas the permutation chart places age slightly ahead of body temperature. Such differences are expected because the methods quantify different properties and permutation can be affected by correlated inputs. Neither plot establishes a clinical threshold, directional risk relationship, causal effect, or generalizability beyond the analyzed benchmark.

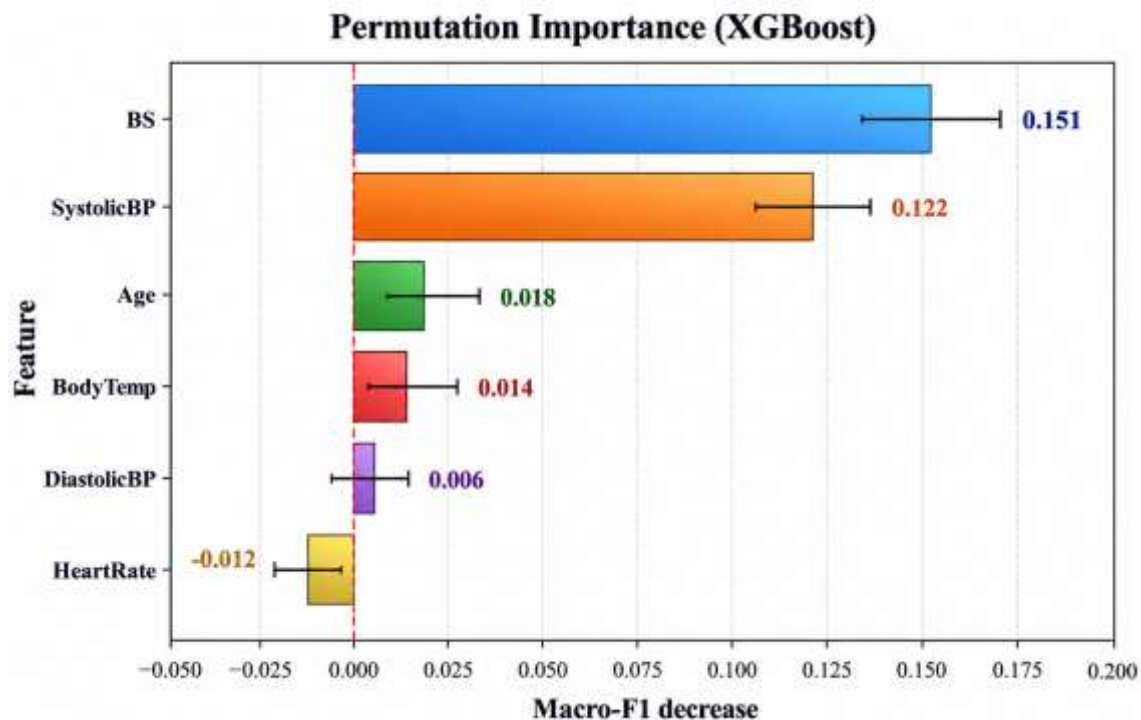


Figure 6. Held-out permutation importance for XGBoost using macro-F1 decrease. Error bars show the standard deviation across 15 permutations.

SHAP ranked BS first (mean $|SHAP| = 0.627$), followed by SystolicBP (0.499), BodyTemp (0.219), Age (0.182), DiastolicBP (0.103), and HeartRate (0.100). Permutation importance independently identified BS (macro-F1 decrease 0.155 ± 0.019) and SystolicBP (0.123 ± 0.019) as the dominant held-out predictors. The lower and partly unstable permutation effects of the remaining variables indicate that importance rankings beyond the first two features should be interpreted cautiously. The dominance of blood sugar and systolic pressure is directionally consistent with the SHAP findings in the supplied preeclampsia article, although the datasets, targets, and validation protocols differ [6].

4.6 Sensitivity analyses

Table 8. XGBoost sensitivity analyses under repeated group-aware validation.

Analysis	Accuracy	Macro-F1	MCC
Primary XGBoost	0.662 ± 0.047	0.645 ± 0.053	0.498 ± 0.071
Train-fold SMOTE + XGBoost	0.664 ± 0.050	0.652 ± 0.054	0.500 ± 0.076
Remove HR < 30 + XGBoost	0.661 ± 0.043	0.647 ± 0.051	0.493 ± 0.064

SMOTE, applied only within training folds, raised macro-F1 from 0.645 to 0.652—an absolute improvement of approximately 0.007. This modest change suggests that class imbalance is not the dominant performance bottleneck. Removing the two HeartRate < 30 bpm records produced macro-F1 0.647 ± 0.051 , essentially unchanged from the primary analysis. The primary limitation therefore appears to be information dependence and target inconsistency rather than a small number of obvious physiological outliers.

4.7 Unambiguous unique-profile sensitivity analysis

Table 9. Performance on 381 unambiguous unique feature profiles (5-fold stratified CV).

Model	Accuracy	Bal. Acc.	Macro-F1	MCC	Macro AUC
XGBoost	0.766	0.679	0.681	0.610	0.874
Random Forest	0.764	0.673	0.671	0.601	0.864
Extra Trees	0.740	0.658	0.656	0.560	0.843
RBF-SVM	0.690	0.635	0.634	0.495	0.780
KNN	0.716	0.615	0.620	0.513	0.823
Gaussian NB	0.701	0.576	0.584	0.482	0.796
Logistic regression	0.622	0.571	0.573	0.405	0.792

When contradictory profiles were removed and each remaining profile was represented once, XGBoost reached 0.766 accuracy and 0.681 macro-F1, while Random Forest reached 0.764 accuracy and 0.671 macro-F1. The improvement relative to the full group-aware dataset is consistent with the existence of irreducible label ambiguity in the benchmark. Nevertheless, this subset analysis should not be interpreted as a better estimate of real-world clinical performance because removing ambiguous cases changes the task and may preferentially retain easier examples.

5. Discussion

5.1 The principal finding is methodological, not algorithmic

The central result is that the apparent performance of nonlinear ensemble models changes dramatically when identical feature profiles are prohibited from crossing validation folds. Conventional stratified cross-validation yielded macro-F1 values near 0.86 for Extra Trees and Random Forest, whereas duplicate-aware validation reduced them to approximately 0.61–0.63. Logistic regression showed almost no mean inflation. These findings illustrate why a leaderboard based on conventional random splitting can reward memorization of repeated profiles rather than generalization to unseen clinical profiles.

This does not mean that every prior study using the dataset is invalid, nor does it establish the exact cause of any published score without reproducing that study's full pipeline. It does show that the benchmark itself contains strong dependence structures that must be handled explicitly. Results from studies using different splits, data cleaning, feature transformations, or oversampling procedures are therefore not directly comparable unless their partition logic is reported in sufficient detail.

5.2 Label contradiction sets a ceiling on deterministic prediction from the supplied features

More than one fifth of rows belonged to feature profiles carrying contradictory risk labels. A deterministic model receiving only the six supplied variables cannot assign different labels to identical inputs without additional information. Such contradictions may reflect repeated measurements at different clinical contexts, rounding of continuous values, data aggregation, label noise, or omitted predictors; the public dataset does not provide patient identifiers or enough provenance to distinguish these explanations. For this reason, the contradictory profiles should not be "corrected" by majority vote without domain evidence. The safer approach is to expose the ambiguity and quantify its influence, as done here.

5.3 Clinical and modeling implications

The mid-risk class remains the primary challenge under honest validation. Misclassifying intermediate-risk patients as low risk may be consequential in a real triage pathway, whereas labeling them high risk could increase unnecessary escalation. Future work should therefore evaluate cost-sensitive or ordinal formulations and should report class-specific calibration and decision-curve analysis when clinically validated outcome data become available. Importantly, this benchmark target is a risk category supplied with the dataset; it is not a prospective clinical endpoint such as maternal death, eclampsia, intensive care admission, or preterm birth. The present models must not be interpreted as validated medical devices or deployed for patient care. The explainability findings are nevertheless informative as a consistency check. Both SHAP and held-out permutation importance identified blood sugar and systolic blood pressure as the two dominant predictors, while other variables contributed less robustly. This is aligned with earlier maternal-risk and preeclampsia modeling reports emphasizing blood glucose and blood-pressure measurements [6,7]. Explainability should, however, be viewed as a description of the fitted model—not proof of causal clinical importance.

Taken together, Figures 1–6 connect dataset composition to the interpretation of model outputs. Figures 1 and 2 show class imbalance alongside extensive redundancy and label ambiguity; Figure 3 demonstrates how the estimated macro-F1 changes when dependent profiles cannot cross folds; Figure 4 identifies the mid-risk errors hidden by aggregate metrics; and Figures 5 and 6 show agreement on two prominent model predictors without claiming clinical causation. The figures therefore support a methodological interpretation: model explanations and headline metrics should be evaluated only after the dataset and partitioning strategy have been audited.

5.4 Implications for reproducible machine-learning research

The study reinforces general leakage guidance from Kapoor and Narayanan [9] and the REFORMS checklist [10]: dataset auditing and split design should precede model optimization. In datasets with repeated profiles, grouping by the dependency unit is more important than trying additional algorithms. Similarly, oversampling must be performed inside the training partition; applying SMOTE before cross-validation can leak synthetic information into test folds. The small SMOTE benefit in the leakage-resistant pipeline further shows that imbalance correction cannot substitute for sound validation design.

Garg and coauthors have also examined machine-learning-based leakage detection in multi-cloud environments [25]. That work addresses a different form of leakage from the duplicate-profile contamination investigated here and is cited only to distinguish the two contexts, not as evidence for the present cross-validation results.

5.5 Limitations

- The public benchmark is small, contains only six predictors, and does not expose patient identifiers, timestamps, gestational age, parity, medical history, medication, or laboratory confirmation; group identifiers were therefore defined by exact feature equality rather than true patient identity.
- The target-generation process is insufficiently documented to resolve contradictory labels, and no independent clinician adjudication is available.
- The 1,014-row CSV snapshot differs by one row from the 1,013-instance count currently shown on the UCI landing page. The exact snapshot and SHA-256 checksum are provided for reproducibility.
- The model configurations were deliberately limited and reproducible rather than exhaustively tuned. Nested group-aware hyperparameter optimization could alter relative performance.
- No external cohort was available in the original supplied material. External prospective validation is required before any clinical inference or deployment.
- SHAP and permutation importance describe model behavior under the analyzed data distribution and should not be interpreted as causal effects.

6. Conclusion

A duplicate-aware re-evaluation of the UCI Maternal Health Risk benchmark reveals that data dependence and contradictory labels materially affect reported machine-learning performance. More than half of the rows were exact duplicates, and 21.20% of records belonged to six-feature profiles associated with multiple target labels. When identical profiles were kept within the same validation fold, ensemble-model macro-F1 dropped substantially compared with conventional random stratification. XGBoost provided the highest mean group-aware macro-F1 (0.645), but Random Forest was statistically indistinguishable after correction for multiple comparisons. The mid-risk class remained difficult, and blood sugar plus systolic blood pressure were the most influential predictors across complementary explanation methods. These results shift the emphasis from maximizing benchmark accuracy to establishing leakage-resistant, transparent, and reproducible evaluation. Future maternal-risk research should prioritize patient-level or dependency-aware splitting, stronger data provenance, richer clinical variables, external cohorts, and clinically meaningful outcome validation.

References

1. World Health Organization, UNICEF, UNFPA, World Bank Group, and UNDESA/Population Division. Trends in maternal mortality 2000 to 2023: estimates by WHO, UNICEF, UNFPA, World Bank Group and UNDESA/Population Division. Geneva: World Health Organization; 2025. ISBN 978-92-4-010846-2.
2. M. Ahmed. Maternal Health Risk [Dataset]. UCI Machine Learning Repository; 2020. doi:10.24432/C5DP5D.
3. M. Ahmed, M. A. Kashem, M. Rahman, and S. Khatun. Review and analysis of risk factor of maternal health in remote area using the Internet of Things (IoT). In: Lecture Notes in Electrical Engineering, vol. 632. Springer; 2020:357–365. doi:10.1007/978-981-15-2317-5_30.
4. B. Uma Maheswari, A. Dixit, and A. K. Karn. Machine Learning Algorithm for Maternal Health Risk Classification with SMOTE and Explainable AI. 2024 IEEE 9th International Conference for Convergence in Technology (I2CT). 2024:1–6. doi:10.1109/I2CT61223.2024.10543709.
5. K. Pavagada and J. Vemuri. Prediction Of Maternal Health Risk Factors Using Machine Learning Algorithms. Procedia Computer Science. 2025;258:2713–2722. doi:10.1016/j.procs.2025.04.532.
6. G. Yanto and S. Puspita. A Random Forest Algorithm for High-Risk Pregnancies Prediction Based on Explainable Artificial Intelligence (XAI). Communications in Mathematical Biology and Neuroscience. 2025;2025:126. doi:10.28919/cmbn/9536.
7. O. Khadidos, F. Saleem, S. Selvarajan, Z. Ullah, and A. O. Khadidos. Ensemble machine learning framework for predicting maternal health risk during pregnancy. Scientific Reports. 2024;14:21483. doi:10.1038/s41598-024-71934-x.
8. M. Mamun, S. H. Chowdhury, M. O. Faruq, et al. Identification of Maternal Health Risk From Optimal Features Using Explainable Machine Learning. Engineering Reports. 2025;7(11):e70491. doi:10.1002/eng2.70491.
9. S. Kapoor and A. Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. Patterns. 2023;4(9):100804. doi:10.1016/j.patter.2023.100804.
10. S. Kapoor, E. M. Cantrell, A. Peng, et al. REFORMS: Consensus-based Recommendations for Machine-learning-based Science. Science Advances. 2024;10(18):eadk3452. doi:10.1126/sciadv.adk3452.

11. L. Breiman. Random Forests. *Machine Learning*. 2001;45:5–32. doi:10.1023/A:1010933404324.
12. T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016:785–794. doi:10.1145/2939672.2939785.
13. S. M. Lundberg and S.-I. Lee. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*. 2017;30:4765–4774.
14. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*. 2002;16:321–357. doi:10.1613/jair.953.
15. M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*. 1937;32(200):675–701. doi:10.1080/01621459.1937.10503522.
16. F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*. 1945;1(6):80–83. doi:10.2307/3001968.
17. S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*. 1979;6(2):65–70.
18. B. W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta*. 1975;405(2):442–451. doi:10.1016/0005-2795(75)90109-9.
19. J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. 1960;20(1):37–46. doi:10.1177/001316446002000104.
20. M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*. 2009;45(4):427–437. doi:10.1016/j.ipm.2009.03.002.
21. Garg, A. K. Singh, and M. Garg, “Role of Internet of Things and artificial intelligence for healthcare informatics: an overview,” in *Innovations in Healthcare Informatics: From Interoperability to Data Analysis*, IET, 2023. ISBN: 9781839534584. Book DOI: 10.1049/PBHE041E.
22. Garg, A. K. Singh, and A. Kumar, “Mental disorders management using explainable artificial intelligence (XAI),” in *Explainable Artificial Intelligence for Biomedical and Healthcare Applications*, CRC Press, 2024. eBook ISBN: 978100322010.
23. Garg, D. Singh, N. Chopra, B. Nayyer, and S. Rattan, “A Cardiovascular Disease Early Detection System in Cloud-A Project Based Approach,” in *2025 2nd International Conference on Multidisciplinary Research and Innovations in Engineering (MRIE)*, pp. 254–259, 2025. doi:10.1109/MRIE66930.2025.11156700.
24. M. L. Saini, B. Tripathi, J. Kaushal, and A. Garg, “A Hybrid Model for Diagnosis Cardiovascular Disease Using Clinical Features, ECG and MRI,” in *International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIE)*, pp. 1–6, IEEE, 2023.
25. S. Patil, P. Singh, and A. Garg, “AI-Driven Data Leakage Detection in Multi-Cloud Environments Using Observability and Machine Learning,” in *2026 International Conference on Next Generation Electronics (NELEX)*, pp. 1–6, IEEE, 2026.