Advancing Preeclampsia Prediction: A Tailored Machine Learning Pipeline for Handling Imbalanced Medical Data

Yinyao Ma, Hanlin Lv, Yanhua Ma, Xiao Wang, Longting Lv, Xuxia Liang, Lei Wang

Submitted to: Journal of Medical Internet Research
on: May 09, 2024

Table of Contents

Advancing Preeclampsia Prediction: A Tailored Machine Learning Pipeline for Handling Imbalanced Medical Data

Yinyao Ma[1*]; Hanlin Lv[2*]; Yanhua Ma[1]; Xiao Wang[2]; Longting Lv[2]; Xuxia Liang[1]; Lei Wang[2, 3]

[1]People's Hospital of Guangxi Zhuang Autonomous Region Nanning CN
[2]BGI Research Wuhan CN
[3]Guangdong Bigdata Engineering Technology Research Center for Life Sciences, BGI Research Shenzhen CN
[*]these authors contributed equally

Corresponding Author:
Lei Wang
BGI Research
1-2F, Building 2, Wuhan Optics Valley International Biomedical Enterprise Accelerator Phase 3.1, No. 388 Gaoxin Road 2, Donghu New Technology Development Zone, Wuhan, Hubei, China
Wuhan
CN

Abstract

Background: Preeclampsia represents a significant challenge in obstetrics. Effective early prediction is crucial for timely intervention, yet the development of predictive models is complicated by the class imbalances inherent in clinical data.

Objective: This study aims to develop a robust pipeline that enhances the predictive performance of ensemble machine learning models for the early prediction of preeclampsia in an imbalanced dataset.

Methods: We evaluated combinations of six ensemble machine learning algorithms and eight resampling techniques across a spectrum of minority-to-majority ratios. Using statistical methods, we systematically identified and optimized these configurations, focusing on key performance metrics such as Geometric Mean.

Results: The strategic optimization of variable selection and settings proved crucial. The configuration using the Inverse Weighted Gaussian Mixture Model for resampling, followed by the Gradient Boosting Decision Trees algorithm, with an optimized minority-to-majority ratio of 0.09, was identified as the most effective, achieving a Geometric Mean of 0.6694. This configuration significantly outperformed the baseline across all evaluated metrics, demonstrating substantial improvements in model performance.

Conclusions: This study establishes a robust pipeline that significantly enhances the predictive performance of models for preeclampsia within imbalanced datasets. Our findings underscore the importance of a strategic approach to variable optimization in medical diagnostics, offering potential for broad application in various medical contexts where class imbalance is a concern.

(JMIR Preprints 09/05/2024:60375)
DOI: https://doi.org/10.2196/preprints.60375

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Advancing Preeclampsia Prediction: A Tailored Machine Learning Pipeline for Handling Imbalanced Medical Data

Yinyao Ma[1], †, MD, Hanlin Lv[2], †, MD, Yanhua Ma[1], MD, Xiao Wang[2], PhD, Longting Lv[2], MS, Xuxia Liang[1, *], MD, Lei Wang[2, 3, *], MS

[1] Department of Obstetrics, People's Hospital of Guangxi Zhuang Autonomous Region, Nanning, 530016, China

[2] BGI Research, Wuhan, 430074, China

[3] Guangdong Bigdata Engineering Technology Research Center for Life Sciences, BGI Research, Shenzhen, 518083, China

†These authors contributed equally to this work.

***Corresponding Author:**

Xuxia Liang

Tel: +86-15677113334

E-mail: 134506734@qq.com

Lei Wang

Tel: +86-187162655682

Postal Address: 1-2F, Building 2, Wuhan Optics Valley International Biomedical Enterprise Accelerator Phase 3.1, No. 388 Gaoxin Road 2, Donghu New Technology Development Zone, Wuhan, Hubei, China

E-mail: wanglei12@genomics.cn

Abstract

Background: Preeclampsia represents a significant challenge in obstetrics. Effective early prediction is crucial for timely intervention, yet the development of predictive models is complicated by the class imbalances inherent in clinical data.

Objective: This study aims to develop a robust pipeline that enhances the predictive performance of ensemble machine learning models for the early prediction of preeclampsia in an imbalanced dataset.

Methods: We evaluated combinations of six ensemble machine learning algorithms and eight resampling techniques across a spectrum of minority-to-majority ratios. Using statistical methods, we systematically identified and optimized these configurations, focusing on key performance metrics such as Geometric Mean.

Results: The strategic optimization of variable selection and settings proved crucial. The configuration using the Inverse Weighted Gaussian Mixture Model for resampling, followed by the Gradient Boosting Decision Trees algorithm, with an optimized minority-to-majority ratio of 0.09, was identified as the most effective, achieving a Geometric Mean of 0.6694. This configuration significantly outperformed the baseline across all evaluated metrics, demonstrating substantial improvements in model performance.

Conclusions: This study establishes a robust pipeline that significantly enhances the predictive performance of models for preeclampsia within imbalanced datasets. Our findings underscore the importance of a strategic approach to variable optimization in medical diagnostics, offering potential for broad application in various medical contexts where class imbalance is a concern.

Keywords:

Preeclampsia; Prediction Model; Ensemble Machine Learning; Class Imbalance; Resampling Technique.

Introduction

Preeclampsia (PE) represents a significant obstetric challenge worldwide[1]. Accurate early prediction, especially before 16 weeks of gestation, is critical for timely intervention[2, 3]. However, current prediction methods rely on simple clinical checklists and fall short in terms of early detection and accuracy[4].

In clinical practice, multivariable models have been widely utilized, such as the model developed by the Fetal Medicine Foundation[5], which is grounded in competing risks models and combines maternal factors with advanced predictors. In recent years, ensemble machine learning (EML) has become prominent for their potential to enhance predictive performance by integrating multiple learning algorithms[6]. While these models are particularly adept at navigating complex data patterns, they face significant challenges in medical diagnostics due to the class imbalance common in such datasets[7]. PE, for instance, occurs in only about 2-8% of pregnancies globally[8-10], categorizing it under severe class imbalance[11]. This substantial class imbalance often leads models to disproportionately favor the majority class[12], thereby diminishing their effectiveness in detecting the less frequent but critical PE cases.

Resampling techniques provide a robust solution by rebalancing the dataset, thereby creating a more equitable environment for model training. By modifying the dataset itself to balance the class distribution—primarily through methods such as oversampling the minority class or undersampling the majority class—these techniques help mitigate the biases introduced by class imbalance[13]. For instance, the simplest method, Random Oversampling (ROS)[14], involves randomly duplicating existing samples in the dataset. Moving on to more sophisticated methods, the Synthetic Minority Over-sampling Technique (SMOTE) emerges as one of the earliest and most widely used for synthetic data generation[15]. SMOTE strategically creates synthetic examples along the line segments between existing minority class instances. However, recognizing the limitation of SMOTE

in treating all minority examples equally, several extended versions have been proposed. BorderlineSMOTE selects minority samples that are predominantly surrounded by the majority class, focusing on those near the decision boundary that are critical for classification[16]. Similarly, SVMSMOTE builds on this concept, but uses the Support Vector Machines (SVM) algorithm to identify critical minority samples near the decision boundary. KMeansSMOTE shifts the focus to regions with sparse minority samples, using KMeans clustering to strengthen underrepresented clusters and prevent overrepresentation of denser areas[17]. ADASYN, on the other hand, assigns weights to minority samples based on their learning difficulty, thus generating more synthetic data in more challenging regions[18]. As a hybrid approach, SMOTEENN combines SMOTE's synthetic sample generation with Edited Nearest Neighbors (ENN) for noise reduction, which significantly improves classification performance, especially in complex datasets[19]. This method not only synthesizes minority samples, but also prunes noisy majority samples, resulting in a more balanced and refined dataset. Similarly, the Inverse Weighted Gaussian Mixture Model (IWGMM) adopts a unique strategy by focusing on generating new samples in sparser regions. By using an inverse weighting scheme, IWGMM prioritizes data generation in less populated areas, increasing diversity and potentially mitigating model bias[20].

Although resampling techniques are commonly recognized for their impact on model performance, the specific role of adjusted minority-to-majority ratios (MMR) within these methods warrants further examination. Preliminary evidence, particularly from studies employing SVM, suggests that achieving a 1:1 MMR through oversampling may optimize model performance[21]. This optimization presumably occurs by equalizing the influence each class has during the learning process. These findings underscore the importance of considering MMR adjustments in predictive modeling, particularly in scenarios characterized by severe class imbalances.

Addressing this gap, our study introduces a comprehensive pipeline that strategically integrates advanced resampling techniques, finely tuned MMR settings, and EML algorithms. This tailored

approach is designed to enhance the balance and representation of classes within the training data, directly confronting the pervasive issue of class imbalance. By doing so, our pipeline aims to maximize the performance potential of ensemble models when applied to imbalanced medical datasets. The ultimate objective is to markedly enhance both the performance and reliability of predictions, thereby facilitating the early detection of PE with high levels of sensitivity and specificity. This innovation has the potential to significantly advance medical diagnostics.

Methods

Data collection

We collected electronic medical records (EMRs) from pregnant women who received antenatal care at the People's Hospital of Guangxi Zhuang Autonomous Region from May 2015 to February 2020. This dataset includes detailed demographics, medical histories, laboratory tests, and uterine Doppler ultrasound results. The study included women who were between 8-16 weeks of gestation at the time of their antenatal care visit and delivered at the hospital. Exclusions were applied to pregnancies that were terminated, resulted in miscarriage or fetal death before the 24th gestational week, or had uncertain outcomes. After applying the exclusion criteria, we classified each remaining pregnancy in the dataset based on the diagnosis extracted from delivery reports. The pregnancies were categorized into two groups: “PE” and “non-PE”.

To assess the generalizability and effectiveness of our pipeline across different clinical scenarios, we also evaluated it using three publicly available datasets. These include the Breast Cancer Wisconsin (Diagnostic) Dataset, the Pima Indians Diabetes Database, and the Oxford Parkinson's Disease Detection Dataset, which are detailed in Table 1.

Table 1. Description of the imbalanced medical public datasets.

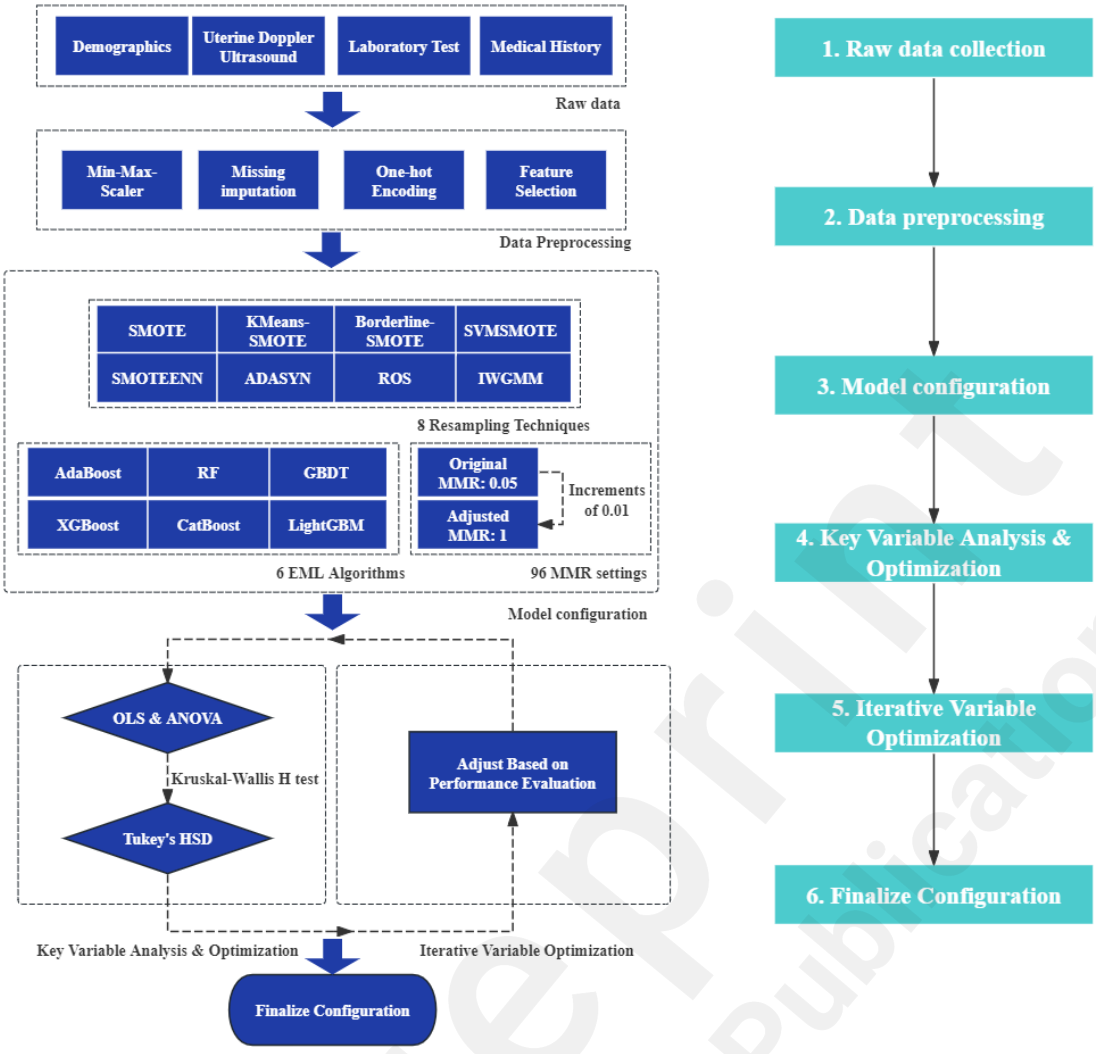
Dataset	Samples	Features	Minority	Majority
Breast Cancer Wisconsin (Diagnostic) Dataset (WBCD)	569	30	212	357
Pima Indians Diabetes Database (Pima)	768	8	268	500
Oxford Parkinson's Disease Detection Dataset (Parkinson)	195	23	48	147

Task overview

In this research, we present a pipeline optimized for the early prediction of PE, designed with the flexibility to address data imbalance, a common challenge in medical diagnostics. This pipeline

integrates a series of systematic steps, including diverse resampling techniques, tailored MMRs, and a selection of EML algorithms effective across various clinical data scenarios. Figure 1 illustrates the pipeline architecture.

Figure 1. This Figure shows a schematic diagram of the pipeline. The pipeline is divided into six phases, starting with raw data collection, followed by data preprocessing. Model configuration then explores 4608 combinations of resampling techniques, EML algorithms, and a range of MMR settings from 0.05 to 1. Subsequent stages involve statistical analysis to identify and optimize key variables using methods such as OLS regression, ANOVA, Kruskal-Wallis H tests, and HSD tests to ensure that the most influential variables are fine-tuned for model performance. Iterative optimization refines these insights by sequentially adjusting variable settings, culminating in a final model configuration that significantly improves predictive performance for PE and effectively manages class imbalances.



Model Configuration and Initial Performance Assessment

Features with over 20% missing values were excluded. Continuous data was imputed with median values, and categorical data was imputed with "NA." Continuous features were normalized using MinMaxScaler, and categorical variables were encoded using one-hot encoding. Subsequently, the dataset was divided into training (70%) and testing (30%) subsets to ensure robust model evaluation. Feature selection was conducted with the Least Absolute Shrinkage and Selection Operator (LASSO).

We employed 8 resampling techniques including ROS, SMOTE, KmeansSMOTE, BorderlineSMOTE, SVMSMOTE, SMOTEENN, ADASYN, and IWGMM. Recognizing the importance of class distribution in model performance, we systematically adjusted the MMR within

the training set from 0.05, reflecting the initial imbalance, up to 1 in increments of 0.01, where classes are equally represented. Resampling was strictly confined to the training set to prevent data leakage. Additionally, we integrated 6 EML algorithms - Adaptive Boosting (AdaBoost), Random Forest (RF), Gradient Boosting Decision Trees (GBDT), Extreme Gradient Boosting (XGBoost), Category Boosting (CatBoost), and Light Gradient Boosting Machine (LightGBM). This diverse and comprehensive setup allowed us to explore a variety of configurations and to conduct an initial assessment of the interactions between resampling, MMR, and EML, thereby establishing a baseline for model performance.

In the initial evaluation phase, we assessed the performance of 4,608 model configurations, at a fixed false positive rate (FPR) at 10% and employing comprehensive metrics such as Geometric Mean (G-mean), Matthews Correlation Coefficient (MCC), Average Precision (AP), and Area Under the ROC Curve (AUC). The metrics were selected for their capacity to accurately reflect a model's capability in predicting minority classes, thereby facilitating a robust performance assessment across the testing set.

Impact Analysis of Key Variables and Optimization of Key Variable Settings

To determine which variable—resampling techniques, MMR settings, or EML algorithms—has the most significant impact on model performance, we first conducted Ordinary Least Squares (OLS) regression and Analysis of Variance (ANOVA) analyses, focusing on G-mean scores across all parameter combinations. The variable contributing the most to variability, as indicated by the highest sum of squares (SS), was further analyzed. We employed the Kruskal-Wallis H-test to assess differences in settings of this key variable, primarily using G-mean. Should these results prove inconclusive, we systematically evaluated additional metrics—MCC, AP, and AUC—to explore variability across configurations further. Upon detecting significant differences, the Tukey’s Honestly Significant Difference (HSD) test was used to identify and confirm the optimal settings that achieved the highest performance metrics, thereby pinpointing the most effective configuration for the

identified key variable.

Iterative Optimization of Key Variables

Once the optimal settings for the most impactful variable had been established, an iterative process was initiated with the objective of evaluating and adjusting the remaining variables. The same statistical methods previously applied were employed to identify the second most impactful variable and its optimal settings within the context of the first variable's settings. This step ensured that each variable's contribution to model performance was optimized in conjunction with the others. Finally, with the two most impactful variables and their settings fixed, we determined the optimal setting for the remaining variable. This systematic approach permitted the model configuration to be refined in a stepwise manner, thereby ensuring the highest possible predictive performance for PE by accurately configuring each variable in order of its impact.

Results

Data Preparation and Feature Selection

In our study, we analyzed data from 8,827 pregnancy women, of whom 306 (3.47%) were diagnosed with PE. After data preprocessing and feature selection, we identified 36 significant features that exhibited strong predictive power for PE, as listed in Table 2.

Table 2. Features selected using LASSO.

Feature	PE (n=306)	no-PE (n=8521)
Demographics, mean (standard deviation)		
Maternal age (years)	33.08(4.8)	31.75(4.36)
Menstrual cycle (days)	32.88(8.32)	32.57(7.98)
Parity	1.42(0.59)	1.48(0.56)
Uterine Doppler Ultrasound, mean (standard deviation)		
Uterine Artery Vmax (cm/s)	54.66(19.11)	57.47(18.71)
Uterine Artery Pulsatility Index	1.08(0.23)	1.06(0.22)
Laboratory Tests, mean (standard deviation)		
Gamma-glutamyl transferase (U/L)	16.81(14.82)	12.97(7.51)
Prealbumin (mg/L)	253.72(38.44)	240.99(33.79)
Monocyte Count (10^9/L)	0.67(0.2)	0.58(0.18)
Monoamine Oxidase (U/L)	4.61(1.85)	3.99(1.67)
Urine pH	6.36(0.74)	6.54(0.75)
Uric Acid (μmol/L)	229(54.59)	210.42(45.91)
Total Protein (g/L)	72.03(4.09)	70.99(3.84)
Triglycerides (mmol/L)	1.52(0.74)	1.31(0.53)
Alkaline Phosphatase (U/L)	47.96(13.12)	44.66(9.62)
Bicarbonate (mmol/L)	22.17(2.36)	22.39(2.29)
Cholinesterase (U/L)	7714.87(1387.43)	7209.77(1216.65)
Red Blood Cell (10^12/L)	4.5(0.53)	4.36(0.48)
Blood Creatinine (μmol/L)	50.05(7.48)	48.67(6.68)
Plateletcrit (%)	0.28(0.06)	0.26(0.05)
Plasma Fibrinogen (g/L)	4.41(0.73)	4.23(0.65)
Serum α-L-Fructosidase (U/L)	26.15(5.7)	24.7(5.06)
Medical History, n (%)		
Chronic hypertension	37(12.09%)	34(0.40%)
Chronic renal disease	19(6.21%)	151(1.77%)
Scarred uterus	38(12.42%)	1463(17.17%)
Previous stillbirth	13(4.25%)	213(2.50%)
Primigravida	112(36.60%)	2849(33.44%)
HBV Infection	13(4.25%)	486(5.70%)
Previous cesarean section	57(18.63%)	1551(18.20%)
Dysmenorrhea	51(16.67%)	1823(21.39%)
Previous miscarriage	26(8.50%)	535(6.28%)

Nausea and vomiting in pregnancy	266(86.93%)	7743(90.87%)
Drug allergy history	67(21.90%)	1770(20.77%)
Family history of hypertension	33(10.78%)	336(3.94%)
Menstrual cycle regularity	263(85.95%)	7813(91.69%)
Fallopian tube disease	18(5.88%)	247(2.90%)
Hypomenorrhea	1(0.33%)	82(0.96%)

Comprehensive Analysis of Model Performance Across Various Configurations and Key Determinant Identification

In our evaluation of 4,608 model configurations, we systematically analyzed the impact of resampling techniques, MMR settings, and EML algorithms on G-mean. Initial regression analyses via OLS and ANOVA pinpointed resampling techniques as having a significant influence on G-mean, with substantial variability indicated by a sum of squares (SS) of 1.26 (see Table 3). Further analysis using Kruskal-Wallis H tests confirmed these significant differences across various resampling methods. Subsequent comparisons with Tukey's Honestly Significant Difference (HSD) test identified the IWGMM as the most effective resampling technique. Detailed results are presented in Table 4 and Supplementary Table S1.

Table 3. ANOVA results from OLS regression assessing impact of key variables on G-mean and other metrics.

Analysis Round	Metrics	Variables	Sum of Squares	F-Statistic	p-value
First Round	G-mean	MMR	0.4880	7.1366	<0.001
		EML	0.2149	59.0762	<0.001
		Resampling	1.2687	249.1503	<0.001
		Residual	3.2348	-	-
	MCC	MMR	0.2140	7.4565	<0.001
		EML	0.0582	38.0886	<0.001
		Resampling	0.5024	235.0270	<0.001
		Residual	1.3580	-	-
	AP	MMR	0.5264	15.4136	<0.001
		EML	0.7774	427.9308	<0.001
		Resampling	0.5533	217.5708	<0.001

		Residual	1.6157	-	-
	AUC	MMR	0.1587	7.0003	<0.001
		EML	1.2106	1004.2128	<0.001
		Resampling	0.7930	469.8881	<0.001
		Residual	1.0722	-	-
	G-mean	MMR	0.0518	0.9947	<0.001
		EML	0.2796	100.8838	<0.001
		Residual	0.2605	-	-
	MCC	MMR	0.0215	0.8649	<0.001
		EML	0.0987	74.6154	<0.001
		Residual	0.1244	-	-
Second Round	AP	MMR	0.0228	1.4315	<0.001
		EML	0.0793	93.7305	<0.001
		Residual	0.0795	-	-
	AUC	MMR	0.0194	0.9828	<0.001
		EML	0.1771	168.4731	<0.001
		Residual	0.0988	-	-

Note: the variables of OLS test in this table were MMR settings, EML algorithms and resampling techniques. All variables were categorical variables.

Table 4. Performance comparison based of HSD test results.

Variables	Values	Superior (+)	Inferior (-)
Resampling	IWGMM	7	0
	SVMSMOTE	5	1
	BorderlineSMOTE	4	1
	ROS	3	3
	SMOTE	2	4
	KMeansSMOTE	1	6
	ADASYN	0	5
	SMOTEENN	0	6
EML	GBDT	5	0
	CatBoost	3	1
	AdaBoost	3	1
	LightGBM	1	3
	XGBoost	1	4
	RF	0	5

Note: "+" indicates the number of comparisons where the method was superior to others; "-" indicates the number of comparisons where the method was inferior.

Table S1 (Multimedia Appendix 1). HSD test results of resampling techniques on G-mean.

Reference Group	Comparison Group	Meandiff	p-adj
ADASYN	BorderlineSMOTE	0.0235 (0.0182, 0.0288)	0
ADASYN	IWGMM	0.0496 (0.0443, 0.0549)	0
ADASYN	KMeansSMOTE	0.0042 (-0.0011, 0.0095)	0.2347
ADASYN	ROS	0.0208 (0.0155, 0.0261)	0
ADASYN	SMOTE	0.0063 (0.001, 0.0116)	0.0077
ADASYN	SMOTEENN	-0.0048 (-0.0101, 0.0005)	0.1162
ADASYN	SVMSMOTE	0.0263 (0.021, 0.0316)	0
BorderlineSMOTE	IWGMM	0.0261 (0.0208, 0.0314)	0
BorderlineSMOTE	KMeansSMOTE	-0.0193 (-0.0246, -0.014)	0
BorderlineSMOTE	ROS	-0.0027 (-0.008, 0.0026)	0.7872
BorderlineSMOTE	SMOTE	-0.0172 (-0.0225, -0.0119)	0
BorderlineSMOTE	SMOTEENN	-0.0283 (-0.0336, -0.023)	0
BorderlineSMOTE	SVMSMOTE	0.0028 (-0.0025, 0.0081)	0.736
IWGMM	KMeansSMOTE	-0.0454 (-0.0506, -0.0401)	0
IWGMM	ROS	-0.0287 (-0.034, -0.0235)	0
IWGMM	SMOTE	-0.0433 (-0.0486, -0.038)	0
IWGMM	SMOTEENN	-0.0543 (-0.0596, -0.049)	0
IWGMM	SVMSMOTE	-0.0232 (-0.0285, -0.0179)	0
KMeansSMOTE	ROS	0.0166 (0.0113, 0.0219)	0
KMeansSMOTE	SMOTE	0.0021 (-0.0032, 0.0074)	0.9344
KMeansSMOTE	SMOTEENN	-0.009 (-0.0143, -0.0037)	0
KMeansSMOTE	SVMSMOTE	0.0221 (0.0168, 0.0274)	0
ROS	SMOTE	-0.0145 (-0.0198, -0.0092)	0
ROS	SMOTEENN	-0.0256 (-0.0309, -0.0203)	0
ROS	SVMSMOTE	0.0055 (0.0002, 0.0108)	0.0339
SMOTE	SMOTEENN	-0.0111 (-0.0163, -0.0058)	0
SMOTE	SVMSMOTE	0.02 (0.0148, 0.0253)	0
SMOTEENN	SVMSMOTE	0.0311 (0.0258, 0.0364)	0

Evaluating the Impact of MMR Settings and EML Algorithms Under IWGMM Resampling

In our further analysis of 576 configurations under IWGMM resampling, the impact of EML algorithms was pronounced, significantly influencing the G-mean (SS = 0.28), as detailed in Table 3. Kruskal-Wallis H tests confirmed substantial differences among EML algorithms, with subsequent HSD tests identifying GBDT as the most effective, with detailed results presented in Table 4 and Supplementary Table S2 (Multimedia Appendix 2).

Table S2 (Multimedia Appendix 2). HSD test results of EML algorithms on G-mean.

Reference Group	Comparison Group	Meandiff	p-adj
AdaBoost	CatBoost	-0.0092 (-0.019, 0.0005)	0.0753
AdaBoost	GBDT	0.0208 (0.011, 0.0306)	0
AdaBoost	LightGBM	-0.0213 (-0.031, -0.0115)	0
AdaBoost	RF	-0.0506 (-0.0604, -0.0409)	0

AdaBoost	XGBoost	-0.0249 (-0.0347, -0.0152)	0
CatBoost	GBDT	0.03 (0.0203, 0.0398)	0
CatBoost	LightGBM	-0.012 (-0.0218, -0.0023)	0.0061
CatBoost	RF	-0.0414 (-0.0511, -0.0316)	0
CatBoost	XGBoost	-0.0157 (-0.0255, -0.0059)	0.0001
GBDT	LightGBM	-0.0421 (-0.0518, -0.0323)	0
GBDT	RF	-0.0714 (-0.0812, -0.0617)	0
GBDT	XGBoost	-0.0457 (-0.0555, -0.036)	0
LightGBM	RF	-0.0293 (-0.0391, -0.0196)	0
LightGBM	XGBoost	-0.0037 (-0.0134, 0.0061)	0.8922
RF	XGBoost	0.0257 (0.0159, 0.0354)	0

Although the range of MMR settings from 0.05 to 1 was explored, their overall impact on model performance was generally minimal, especially when combined with IWGMM and GBDT. Notably, at an MMR of 0.09, the combination of GBDT and IWGMM showed an enhanced performance, achieving a G-mean of 0.6694, which is a significant improvement from the baseline G-mean of 0.6156. This setting demonstrated robust improvements in MCC, AP, and AUC as well, as detailed in Table 5.

Table 5. Comparative analysis of pipeline optimization and baseline models.

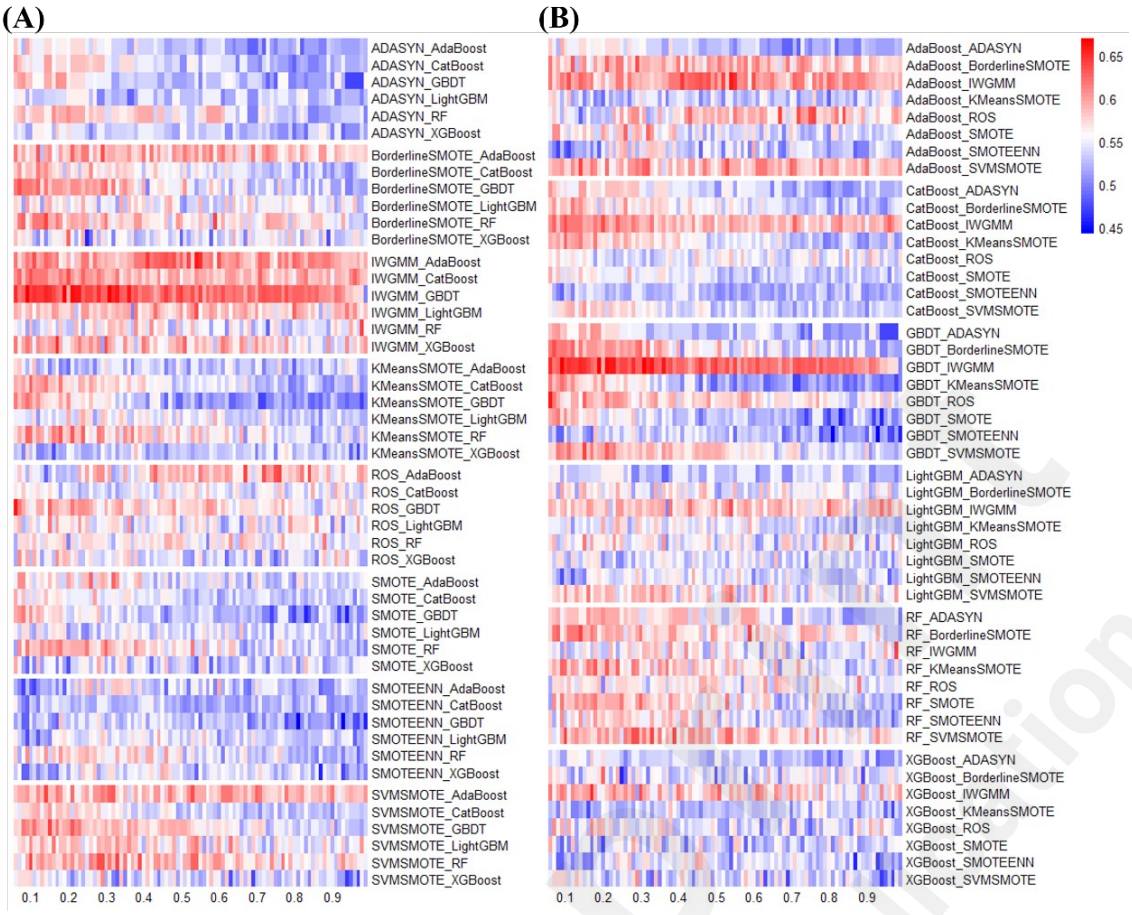
Dataset	Optimization Strategy	G-mean	MCC	AP	AUC
PE	Resampling-IWGMM → EML-GBDT → MMR-0.09	0.6694	0.2124	0.1682	0.7769
PE (Orig.)	GBDT	0.6156	0.1659	0.1471	0.7784
WBCD	EML-LightGBM → Resampling-ROS → MMR-0.55	0.9848	0.9586	0.9890	0.9944
WBCD (Orig.)	LightGBM	0.9848	0.9586	0.9894	0.9947
Parkinson	Resampling-ADASYN → EML-AdaBoost → MMR-0.36	0.8944	0.8038	0.9745	0.9526
Parkinson (Orig.)	AdaBoost	0.7695	0.5387	0.9670	0.9342
Pima	EML-CatBoost → Resampling-ADASYN → MMR-0.92	0.8007	0.5876	0.7830	0.8710
Pima (Orig.)	CatBoost	0.7649	0.5807	0.7981	0.8745

Note: The "→" symbol in the "Optimization Strategy" column indicates the order in which the optimization methods were applied to achieve the best performance for each dataset.

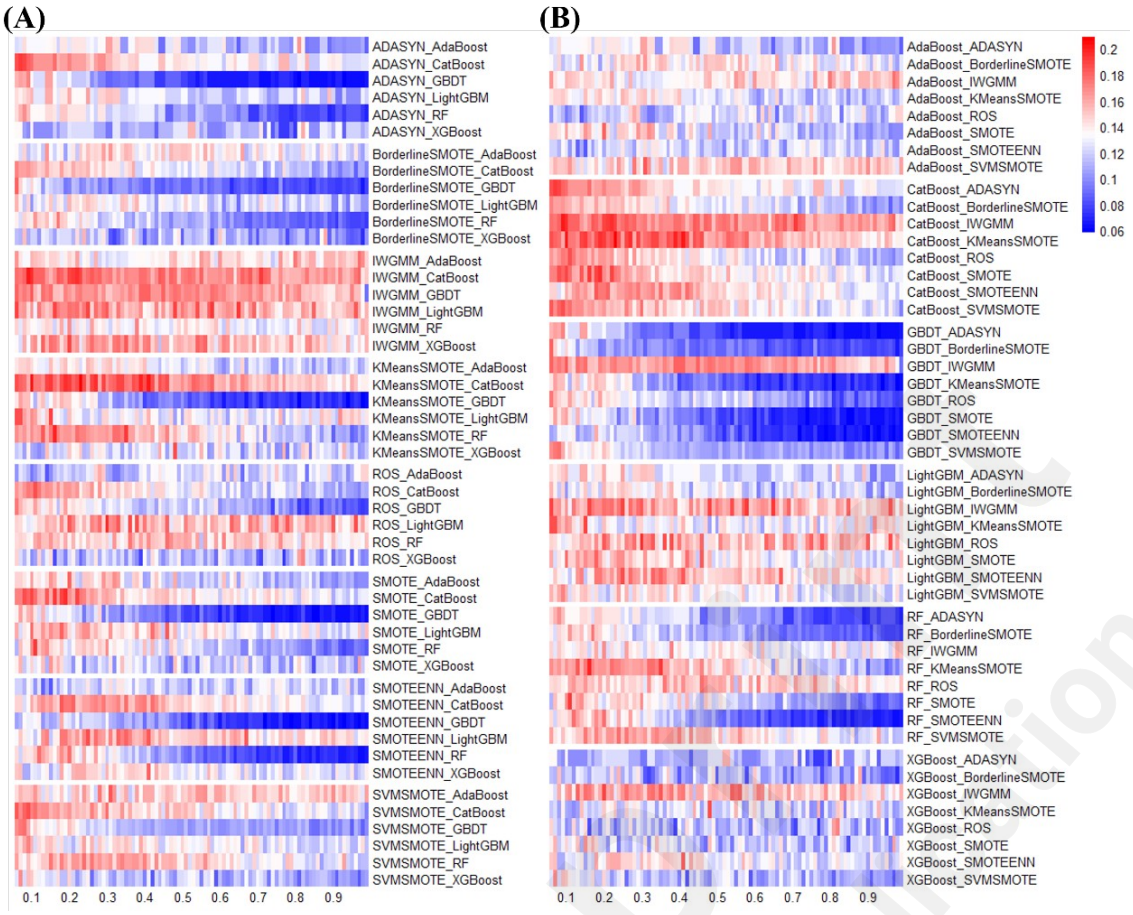
Heatmap of Model Performance Across Different Model Configurations

To further illustrate the dynamics of model performance across various configurations, we conducted comprehensive heatmap analyses (Figures 2 and Supplemental Figures 1-3). These analyses included a spectrum of resampling and EML combinations, arranged across a range of MMRs. For clarity, the heatmaps were organized in two formats: grouped by resampling (A) or by EML (B). These visual representations clearly illustrated that configurations using IWGMM consistently outperformed others across all evaluated metrics. As shown in Figure 2 (B), most combinations of GBDT with various resampling techniques (except IWGMM) tend to reach peak performance at lower MMRs, they experience a marked decline in G-mean as MMR increases. As shown in Figure 2 (A), similar trend was observed in ADASYN and KMeansSMOTE, but IWGMM showed consistent performance across a range of MMRs, indicating minimal influence from MMR.

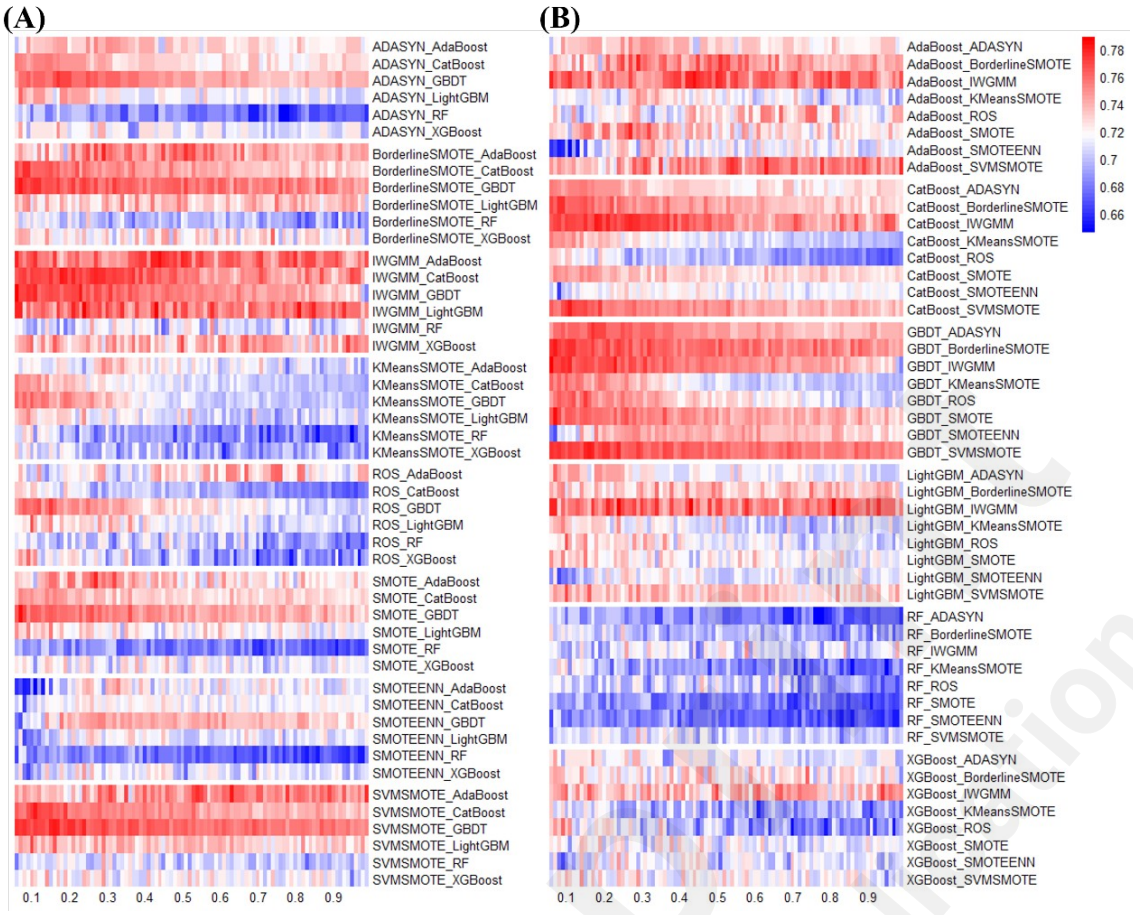
Figure 2. Heatmaps of G-mean Performance Across Resampling and EML Combinations at Various MMR Settings. This figure presents two heatmaps detailing the G-mean performance variations across a range of MMRs. Panel (A) is organized by resampling techniques, highlighting that combinations using IWGMM generally yield higher G-means. SVMSMOTE also shows robust performance, whereas SMOTEENN often results in lower G-means. Panel (B) is organized by EML algorithms and reveals a distinct pattern: while most combinations, particularly GBDT paired with various resampling techniques except IWGMM, achieve optimal performance at lower MMRs, a significant decline in G-mean is observed as MMR increases.



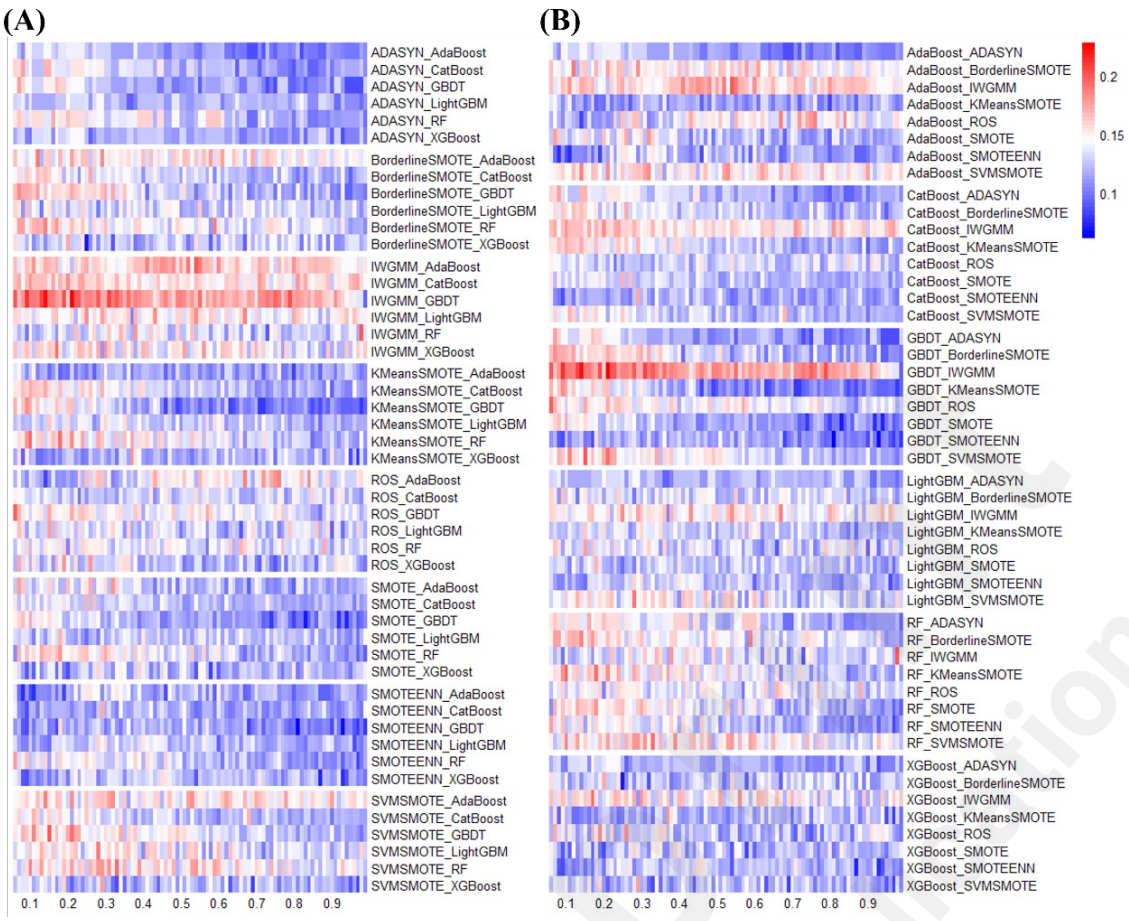
Supplemental Figure 1 (Multimedia Appendix 3). Heatmaps of AP Performance Across Resampling and EML Combinations at Various MMR Settings. Panel (A) is grouped by resampling techniques and Panel (B) by EML algorithms.



Supplemental Figure 2 (Multimedia Appendix 4). Heatmaps of AUC Performance Across Resampling and EML Combinations at Various MMR Settings. Panel (A) is grouped by resampling techniques and Panel (B) by EML algorithms.



Supplemental Figure 3 (Multimedia Appendix 5). Heatmaps of MCC Performance Across Resampling and EML Combinations at Various MMR Settings. Panel (A) is grouped by resampling techniques and Panel (B) by EML algorithms.



Validation on Public Datasets

Our method was further validated across three publicly available medical datasets, affirming the robustness and general applicability of our approach. This validation confirms our pipeline’s capacity to handle diverse imbalanced datasets and improve overall predictive performance, as summarized in Table 6.

Discussion

Our analysis confirmed that the strategic order of variable optimization was crucial; the most effective sequence in our data involved prioritizing resampling, followed by EML, and finally optimizing the MMR. Deviations from this sequence, such as prioritizing EML like RF over resampling like SVMSMOTE with an MMR of 0.33, resulted in marked decrease in effectiveness, evidenced by the G-mean of 0.6393, MCC of 0.1909, AUC of 0.7101, and AP of 0.1645. External validations further confirmed the robustness and adaptability of our pipeline across various medical contexts, demonstrating its potential to significantly enhance predictive performance in diverse imbalanced medical datasets.

In evaluating model performance in imbalanced datasets, selecting metrics that effectively capture the performance nuances becomes crucial. The G-mean, serving as our primary metric, was indispensable for integrating sensitivity and specificity[22-24]. By maintaining a constant false positive rate of 10%, indicates that 10% of non-PE cases might be incorrectly identified as positive, we ensured that variations in G-mean accurately mirrored shifts in sensitivity. This method not only facilitates consistent comparisons across studies but also aligns with established practices in PE research[25-28], thereby underscoring the utility of G-mean in robust model performance assessment. Further exploration of metrics such as MCC, AP, and AUC in our supplementary analyses revealed that while G-mean offers a comprehensive overview, each metric contributes uniquely to understanding model performance, suggesting that a multifaceted approach to metric selection and calibration is essential for advancing predictive accuracy in clinical diagnostics.

Previous studies, including our own[29], have underscored the challenges of comparing model performances across cohorts with varying incidence rates of PE, highlighting the critical need for a nuanced understanding of MMR's role. Despite the minimal influence of MMR observed in our iterative key variable analysis, where IWGMM combinations maintained high G-mean values across

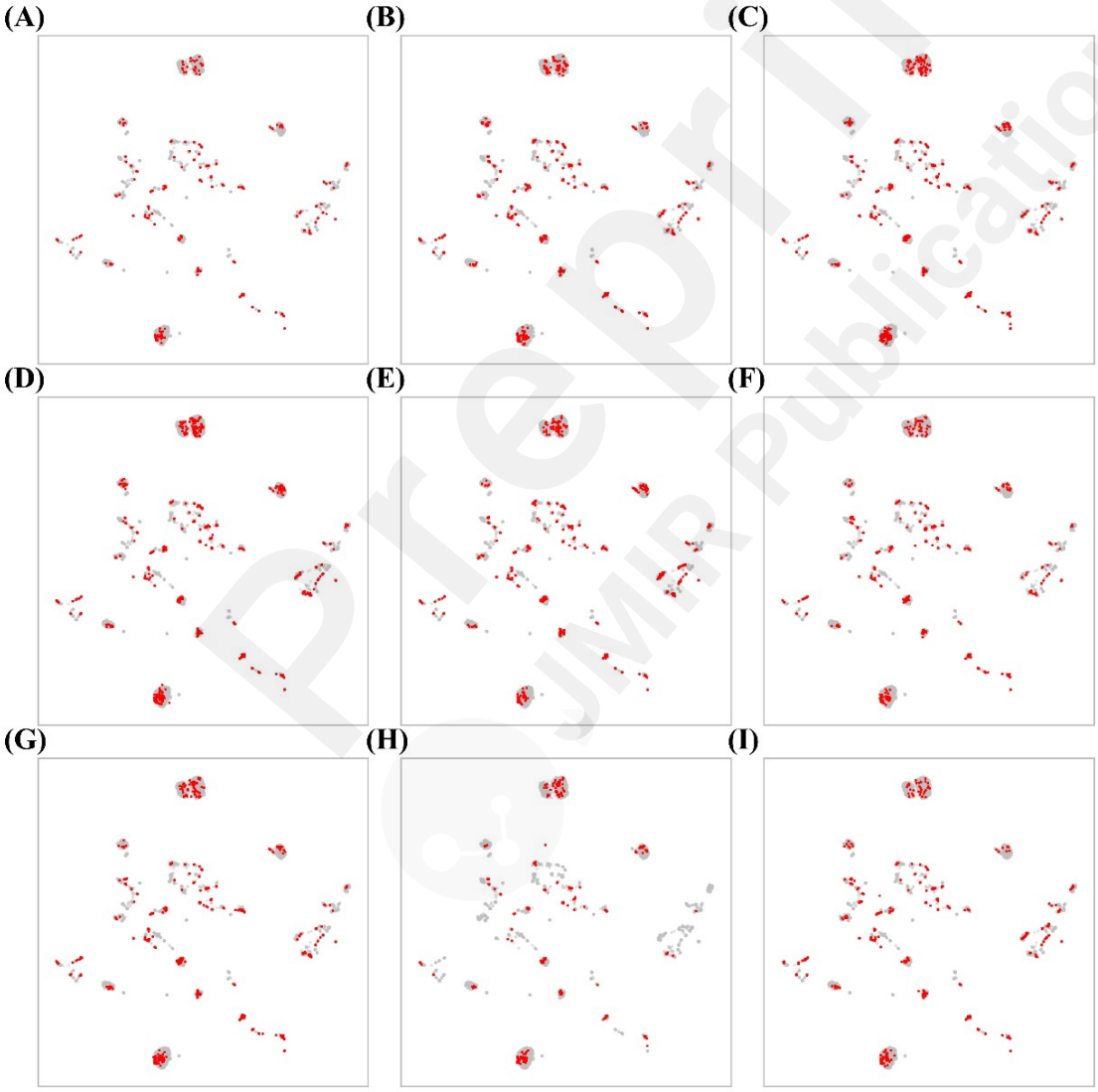
a wide range of MMRs, the impact of MMR on model performance was pronounced in other contexts. At lower MMR values, performance typically deteriorated as MMR increased. This was evidenced by resampling methods such as ADASYN and KMeansSMOTE, which exhibited notable declines in G-mean with increasing MMR. Similarly, in Supplemental Figure 2 (B), most resampling methods paired with GBDT (except IWGMM) exhibited a peak in AP at lower MMR values but demonstrated a deterioration in AP as MMR increased. This pattern suggests that the generation of an excessive number of synthetic samples may inadvertently result in a decline in model performance. Consequently, meticulous calibration of MMR is imperative to prevent potential overfitting when addressing class imbalances with resampling strategies. This underscores the necessity for optimal MMR tuning.

Our findings, reinforced by heatmap, demonstrated that IWGMM consistently outperformed others in managing class imbalances. SVMSMOTE also showed robust performance, but SMOTEENN was the least effective. These insights prompted further investigation into the underlying data structures using Uniform Manifold Approximation and Projection (UMAP). UMAP (Figure 3) highlighted the challenges of distinguishing PE from non-PE cases in early pregnancy datasets characterized by significant overlap. IWGMM's superior performance can be attributed to its dynamic adjustment of weights according to data density, which effectively manages both interclass and intraclass imbalances[20]. This not only improved representation and diversity within classes, but also highlighted the need for accurate data retention strategies in resampling approaches, especially since the SMOTEENN method often led to the exclusion of valid PE cases.

Despite the improvements in model performance facilitated by IWGMM, the persistent overlap between PE and non-PE classes suggests that future models would benefit from integrating more discriminative predictors. The incorporation of advanced biomarkers like PLGF in the FMF test[5], as well as the use of cfDNA or cfRNA[30-33], have shown promising results in enhancing predictive accuracy. Furthermore, leveraging large language models to analyze extensive medical notes from

EMRs[34] could reveal crucial, previously overlooked variables that could significantly boost PE prediction. Highlighting the underexploited potential in the early prediction of PE[25-28], these advancements, alongside a deeper exploration of resampling strategies, could lead to substantial performance gains in future studies.

Figure 3. UMAP Visualization of PE and non-PE Distribution across different resampling techniques. Panel (A) through (I) show the distributions for the original dataset and after applying various resampling techniques: (A) Original, (B) ROS, (C) SMOTE, (D) ADASYN, (E) KMeansSMOTE, (F) SVMSMOTE, (G) BorderlineSMOTE, (H) SMOTEENN, and (I) IWGMM.



Conclusion

Our study has developed a robust pipeline that enhances predictive performance for PE, demonstrating how careful configuration of key variables can effectively manage class imbalances in EML models. The significant improvements observed across various metrics and external validations highlight the importance of methodically optimizing variables to enhance model performance in medical diagnostics.

Author contributions

Y.M., H.L. contributed equally to this study. Y.M., H.L., X.L., L.W. participated in study design and drafted the manuscript. Y.M., Y.M., X.L. participated in data collection. H.L., X.W., L.L. performed the statistical analysis, established machine learning models. Y.M., H.L. helped to draft the manuscript. All authors read and approved the final manuscript.

Institutional review board statement

The study was approved by the People's Hospital of the Guangxi Zhuang Autonomous Region in China (Ref. No. KT-KJT-2021-67) and registered in ChiCTR under identifier ChiCTR2300072225.

Informed consent statement

The requirement for informed consent was waived by the Ethics Committee of the People's Hospital of the Guangxi Zhuang Autonomous Region, due to the observational nature of the study, and all pregnancies' data were de-identified and anonymized.

Conflict of interests

The authors report no conflict of interest.

Source of Funding

This study was supported by Guangxi Key Research and Development Program (No. AB22035056) and the National Natural Science Foundation of China (No. 82060805).

Acknowledgment

We sincerely thank the China National GeneBank for their technical support. We also appreciate the support from RUIYI's Clinical Multi-omics Data Research Workstation for our research.

Data statement

The data that support the findings of this study are available on request from the corresponding author. The data is not publicly available due to privacy or ethical restrictions.

Multimedia Appendix 1:

HSD test results of resampling techniques on G-mean.

Multimedia Appendix 2:

HSD test results of EML algorithms on G-mean.

Multimedia Appendix 3:

Heatmaps of AP Performance Across Resampling and EML Combinations at Various MMR Settings.

Multimedia Appendix 4:

Heatmaps of AUC Performance Across Resampling and EML Combinations at Various MMR Settings.

Multimedia Appendix 5:

Heatmaps of MCC Performance Across Resampling and EML Combinations at Various MMR Settings.

Reference

1. Magee LA, Nicolaides KH, von Dadelszen P. Preeclampsia. N Engl J Med. 2022 May 12;386(19):1817-32. PMID: 35544388. doi: 10.1056/NEJMra2109523.

2. Rolnik DL, Nicolaides KH, Poon LC. Prevention of preeclampsia with aspirin. Am J Obstet Gynecol. 2022 Feb;226(2S):S1108-S19. PMID: 32835720. doi: 10.1016/j.ajog.2020.08.045.

3. Rolnik DL, Wright D, Poon LC, O'Gorman N, Syngelaki A, de Paco Matallana C, et al. Aspirin versus Placebo in Pregnancies at High Risk for Preterm Preeclampsia. N Engl J Med. 2017 Aug 17;377(7):613-22. PMID: 28657417. doi: 10.1056/NEJMoa1704559.

4. O'Gorman N, Wright D, Poon LC, Rolnik DL, Syngelaki A, de Alvarado M, et al. Multicenter screening for pre-eclampsia by maternal factors and biomarkers at 11-13 weeks' gestation: comparison with NICE guidelines and ACOG recommendations. Ultrasound Obstet Gynecol. 2017 Jun;49(6):756-60. PMID: 28295782. doi: 10.1002/uog.17455.

5. Tan MY, Syngelaki A, Poon LC, Rolnik DL, O'Gorman N, Delgado JL, et al. Screening for pre-eclampsia by maternal factors and biomarkers at 11-13 weeks' gestation. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2018 Aug;52(2):186-95. PMID: 29896812. doi: 10.1002/uog.19112.

6. Naderalvojoud B, Hernandez-Boussard T. Improving machine learning with ensemble learning on observational healthcare data. AMIA Annu Symp Proc. 2023;2023:521-9. PMID: 38222353.

7. Liu L, Wu X, Li S, Li Y, Tan S, Bai Y. Solving the class imbalance problem using ensemble algorithm: application of screening for aortic dissection. BMC Med Inform Decis Mak. 2022 Mar 28;22(1):82. PMID: 35346181. doi: 10.1186/s12911-022-01821-w.

8. Abalos E, Cuesta C, Grosso AL, Chou D, Say L. Global and regional estimates of preeclampsia and eclampsia: a systematic review. European journal of obstetrics, gynecology, and

reproductive biology. 2013 Sep;170(1):1-7. PMID: 23746796. doi: 10.1016/j.ejogrb.2013.05.005.

9. Ananth CV, Keyes KM, Wapner RJ. Pre-eclampsia rates in the United States, 1980-2010: age-period-cohort analysis. BMJ. 2013 Nov 7;347:f6564. PMID: 24201165. doi: 10.1136/bmj.f6564.

10. UNICEF. Trends in maternal mortality: 2000 to 2017: estimates by WHO, UNICEF, UNFPA, World Bank Group and the United Nations Population Division. 2019.

11. Branco P, Torgo L, Ribeiro RP. A survey of predictive modeling on imbalanced domains. ACM computing surveys. 2016;49(2):1-50.

12. Hasanin T, Khoshgoftaar TM, Leevy JL, Seliya NJJoBD. Examining characteristics of predictive models with imbalanced big data. 2019;6:1-21.

13. He H, Garcia EAJITok, engineering d. Learning from imbalanced data. 2009;21(9):1263-84.

14. Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F. Learning from imbalanced data sets: Springer; 2018.

15. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WPJJoair. SMOTE: synthetic minority over-sampling technique. 2002;16:321-57.

16. Nguyen HM, Cooper EW, Kamei K. Borderline over-sampling for imbalanced data classification. International Journal of Knowledge Engineering Soft Data Paradigms. 2011;3(1):4-21.

17. Douzas G, Bacao F, Last FJIS. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. 2018;465:1-20.

18. He H, Bai Y, Garcia EA, Li S, editors. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence); 2008: Ieee.

19. Batista GE, Prati RC, Monard MCJASen. A study of the behavior of several methods for balancing machine learning training data. 2004;6(1):20-9.

20. Xing M, Zhang Y, Yu H, Yang Z, Li X, Li Q, et al. Predict DLBCL patients' recurrence within two years with Gaussian mixture model cluster oversampling and multi-kernel learning. Computer

methods and programs in biomedicine. 2022 Nov;226:107103. PMID: 36088813. doi: 10.1016/j.cmpb.2022.107103.

21. Khorshidi HA, Aickelin UJapa. A Synthetic Over-sampling method with Minority and Majority classes for imbalance problems. 2020.

22. Wu G, Chang EYJITok, engineering d. KBA: Kernel boundary alignment considering imbalanced data distribution. 2005;17(6):786-95.

23. Barua S, Islam MM, Yao X, Murase K. MWMOTE--majority weighted minority oversampling technique for imbalanced data set learning. IEEE Transactions on knowledge data engineering. 2012;26(2):405-25.

24. Lin WJ, Chen JJ. Class-imbalanced classifiers for high-dimensional data. Brief Bioinform. 2013 Jan;14(1):13-26. PMID: 22408190. doi: 10.1093/bib/bbs006.

25. Marić I, Tsur A, Aghaeepour N, Montanari A, Stevenson DK, Shaw GM, et al. Early prediction of preeclampsia via machine learning. American journal of obstetrics & gynecology MFM. 2020 May;2(2):100100. PMID: 33345966. doi: 10.1016/j.ajogmf.2020.100100.

26. Wright D, Syngelaki A, Akolekar R, Poon LC, Nicolaides KH. Competing risks model in screening for preeclampsia by maternal characteristics and medical history. Am J Obstet Gynecol. 2015 Jul;213(1):62 e1- e10. PMID: 25724400. doi: 10.1016/j.ajog.2015.02.018.

27. Wright D, Tan MY, O'Gorman N, Poon LC, Syngelaki A, Wright A, et al. Predictive performance of the competing risk model in screening for preeclampsia. Am J Obstet Gynecol. 2019 Feb;220(2):199 e1- e13. PMID: 30447210. doi: 10.1016/j.ajog.2018.11.1087.

28. O'Gorman N, Wright D, Syngelaki A, Akolekar R, Wright A, Poon LC, et al. Competing risks model in screening for preeclampsia by maternal factors and biomarkers at 11-13 weeks gestation. Am J Obstet Gynecol. 2016 Jan;214(1):103 e1- e12. PMID: 26297382. doi: 10.1016/j.ajog.2015.08.034.

29. Wang L, Ma Y, Bi W, Meng C, Liang X, Wu H, et al. An early screening model for

preeclampsia: utilizing zero-cost maternal predictors exclusively. Hypertens Res. 2024 Apr;47(4):1051-62. PMID: 38326453. doi: 10.1038/s41440-023-01573-8.

30. Rasmussen M, Reddy M, Nolan R, Camunas-Soler J, Khodursky A, Scheller NM, et al. RNA profiles reveal signatures of future health and disease in pregnancy. Nature. 2022 Jan;601(7893):422-7. PMID: 34987224. doi: 10.1038/s41586-021-04249-w.

31. Moufarrej MN, Vorperian SK, Wong RJ, Campos AA, Quaintance CC, Sit RV, et al. Early prediction of preeclampsia in pregnancy with cell-free RNA. Nature. 2022 Feb;602(7898):689-94. PMID: 35140405. doi: 10.1038/s41586-022-04410-z.

32. De Borre M, Che H, Yu Q, Lannoo L, De Ridder K, Vancoillie L, et al. Cell-free DNA methylome analysis for early preeclampsia prediction. Nat Med. 2023 Sep;29(9):2206-15. PMID: 37640858. doi: 10.1038/s41591-023-02510-5.

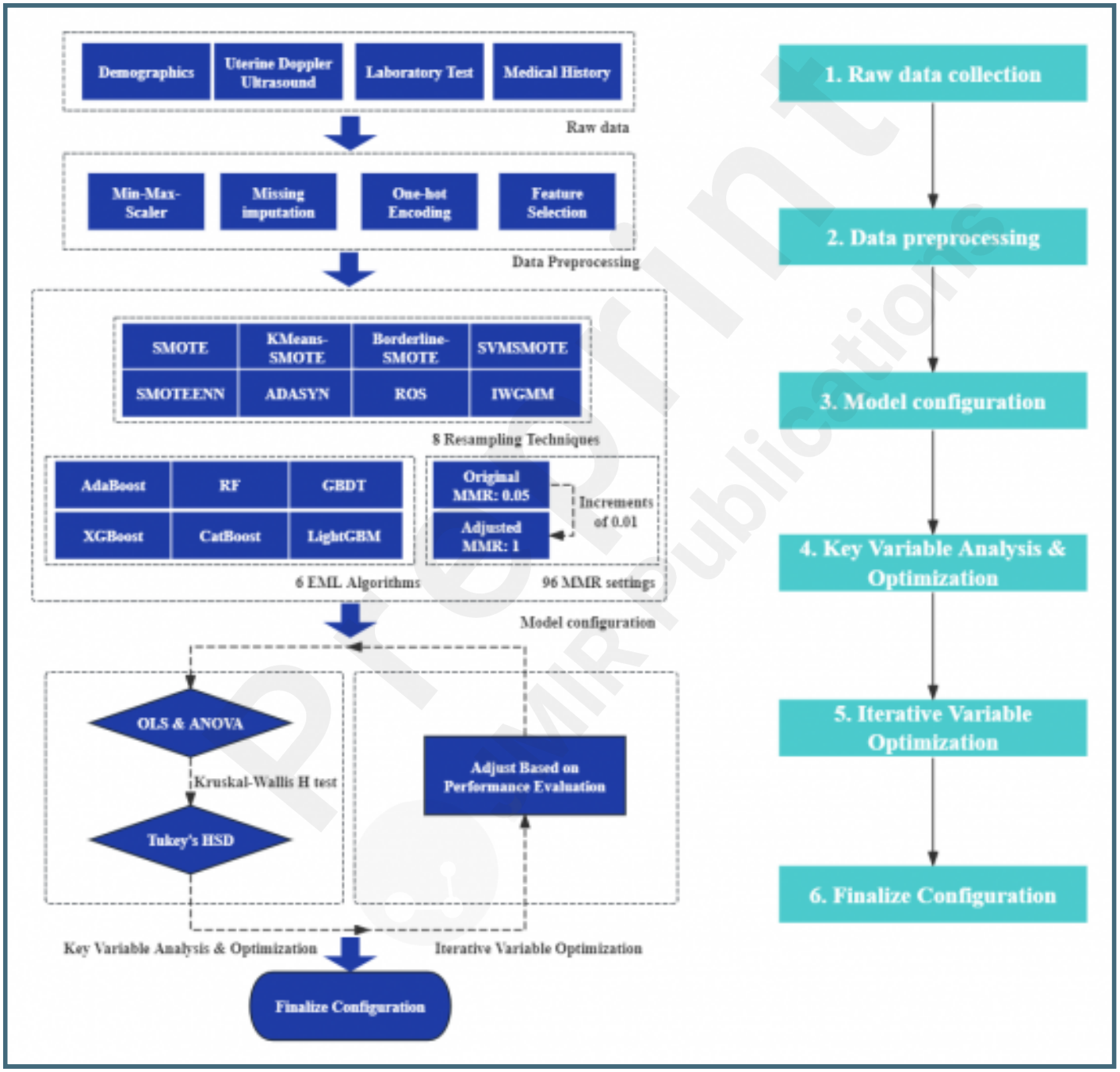
33. Zhou S, Li J, Yang W, Xue P, Yin Y, Wang Y, et al. Noninvasive preeclampsia prediction using plasma cell-free RNA signatures. Am J Obstet Gynecol. 2023 Nov;229(5):553 e1- e16. PMID: 37211139. doi: 10.1016/j.ajog.2023.05.015.

34. Wang L, Ma Y, Bi W, Lv H, Li Y. An Entity Extraction Pipeline for Medical Text Records Using Large Language Models: Analytical Study. J Med Internet Res. 2024 Mar 29;26:e54580. PMID: 38551633. doi: 10.2196/54580.

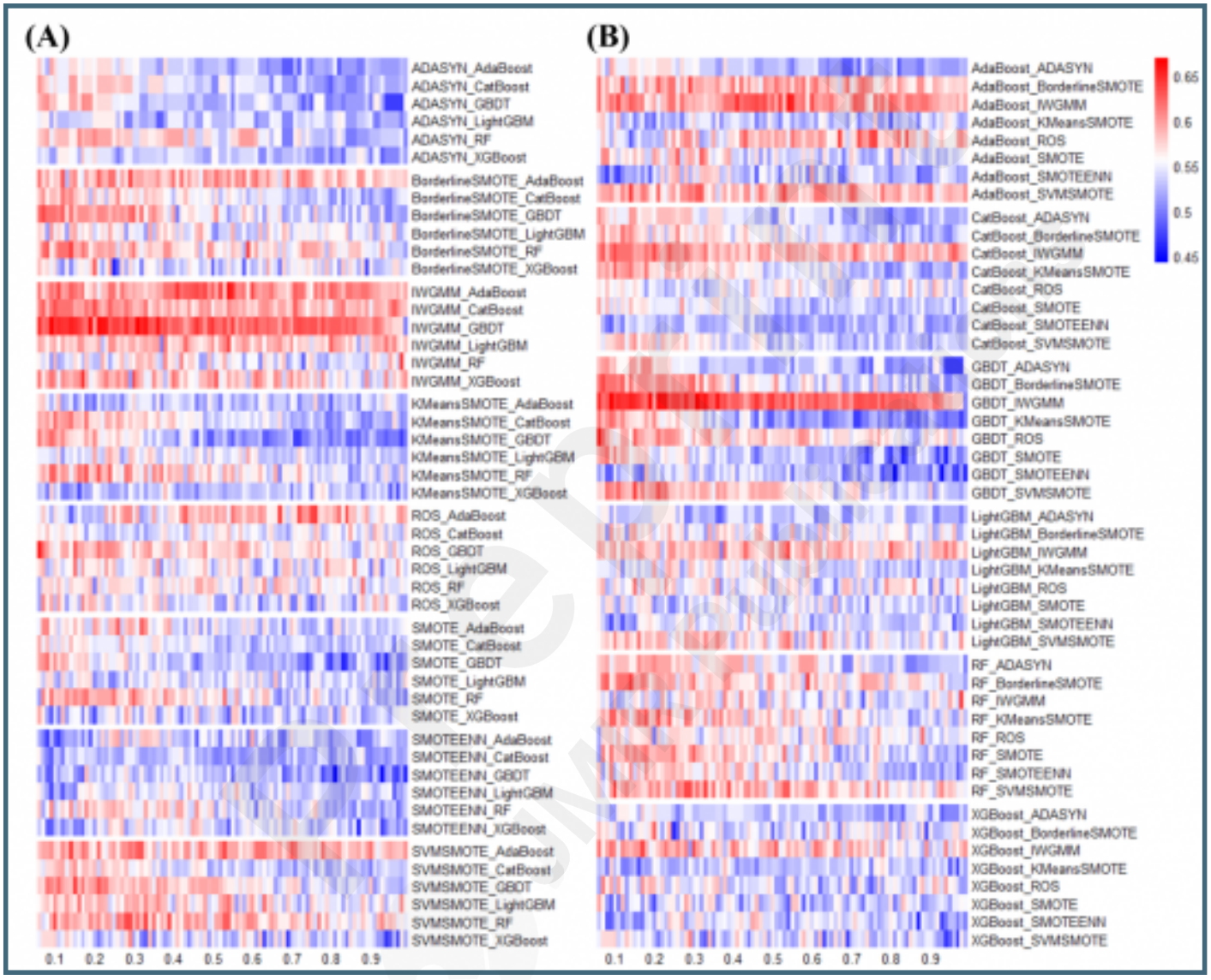
Supplementary Files

Figures

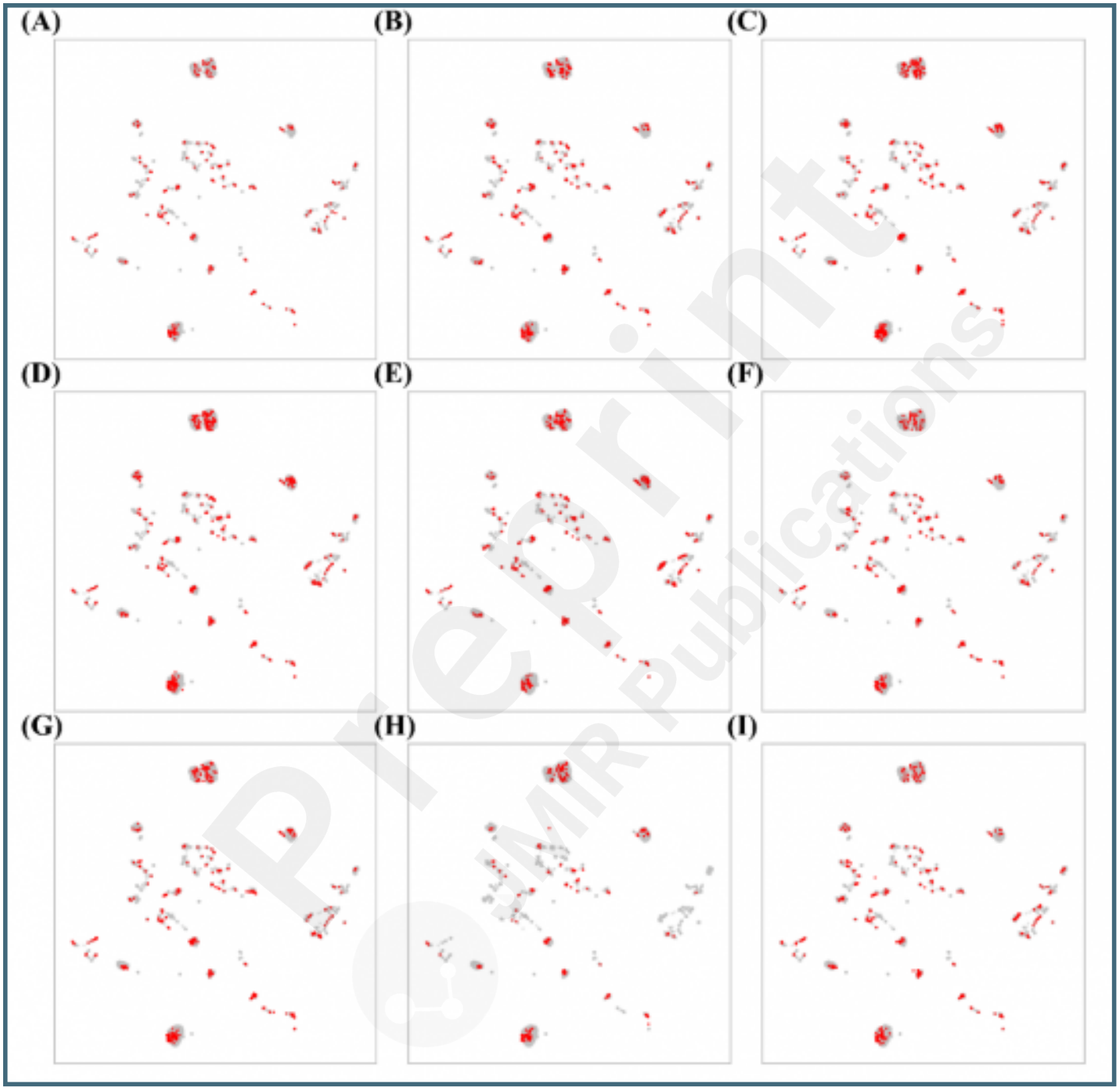
This Figure shows a schematic diagram of the pipeline. The pipeline is divided into six phases, starting with raw data collection, followed by data preprocessing. Model configuration then explores 4608 combinations of resampling techniques, EML algorithms, and a range of MMR settings from 0.05 to 1. Subsequent stages involve statistical analysis to identify and optimize key variables using methods such as OLS regression, ANOVA, Kruskal-Wallis H tests, and HSD tests to ensure that the most influential variables are fine-tuned for model performance. Iterative optimization refines these insights by sequentially adjusting variable settings, culminating in a final model configuration that significantly improves predictive performance for PE and effectively manages class imbalances.



Heatmaps of G-mean Performance Across Resampling and EML Combinations at Various MMR Settings. This figure presents two heatmaps detailing the G-mean performance variations across a range of MMRs. Panel (A) is organized by resampling techniques, highlighting that combinations using IWGMM generally yield higher G-means. SVMSMOTE also shows robust performance, whereas SMOTEENN often results in lower G-means. Panel (B) is organized by EML algorithms and reveals a distinct pattern: while most combinations, particularly GBDT paired with various resampling techniques except IWGMM, achieve optimal performance at lower MMRs, a significant decline in G-mean is observed as MMR increases.



UMAP Visualization of PE and non-PE Distribution across different resampling techniques. Panel (A) through (I) show the distributions for the original dataset and after applying various resampling techniques: (A) Original, (B) ROS, (C) SMOTE, (D) ADASYN, (E) KMeansSMOTE, (F) SVMSMOTE, (G) BorderlineSMOTE, (H) SMOTEENN, and (I) IWGMM.



Multimedia Appendixes

HSD test results of resampling techniques on G-mean.
URL: http://asset.jmir.pub/assets/c10685db8fc796b78eaf031d65f29de8.docx

HSD test results of EML algorithms on G-mean.
URL: http://asset.jmir.pub/assets/18320a29c55d226123fbefdbd4ac811e.docx

Heatmaps of AP Performance Across Resampling and EML Combinations at Various MMR Settings.
URL: http://asset.jmir.pub/assets/2a1554c258727a7c08018fc4a8137b78.png

Heatmaps of AUC Performance Across Resampling and EML Combinations at Various MMR Settings.
URL: http://asset.jmir.pub/assets/0757dfd661cd8cb099c9fd2e2d45f5fc.png

Heatmaps of MCC Performance Across Resampling and EML Combinations at Various MMR Settings.
URL: http://asset.jmir.pub/assets/0556e17e97f1f05e732ed20f5bdb2b80.png