COST-AWARE DISTILLATION OF A COMMERCIAL FIRST-TRIMESTER PREECLAMPSIA SCREENING ENGINE IN NORTHERN VIETNAM: TIERED FEATURE SETS AND EXPLAIN ABILITY

Ta Van Thao[1,✉], Vu Thi Minh Phuong[2], Nguyen Lien Huong[3]
Dang Thanh Tam[1], Tran Sach Viet[1], Vu Ngoc Bac[4], Vu Ngoc Anh[5]

[1]Hanoi Medical University
[2]Chemedic Laboratory Center
[3]Faculty of Biotechnology, Hanoi University of Pharmacy
[4]Thai Binh University of Medicine and Pharmacy
[5]Dong Do University

Commercial first-trimester preeclampsia (PE) screening engines (e.g., PerkinElmer/FMF) integrate maternal factors, biophysical measures, Doppler indices, and biochemical markers to generate continuous risk ratios, but their use in resource-constrained settings is limited by missing data, centralized assays, and limited transparency. We developed a tiered, interpretable machine-learning (ML) distillation pipeline to approximate the PerkinElmer risk ratio for PE <37 weeks, evaluate fidelity across cost-aware feature tiers, identify minimal deployable feature sets, and assess agreement at the clinical cut-off. A retrospective cohort of 1,051 singleton pregnancies from Northern Vietnam (2023–2025) was split into training (n=850) and test (n=201) sets. Complete-case tiers were defined as Tier 0 (maternal factors + MAP), Tier 1 (+ PlGF/PAPP-A), and Tier 2 (+ UtA-PI). Fidelity was assessed using ranking, regression, and calibration metrics, with interpretability via permutation importance, SHAP, and ablation, and threshold mimicry using the PerkinElmer cut-off. Despite declining tier availability, fidelity remained high (Spearman ρ=0.975–0.981), with strong top-rank agreement and excellent AUC-ROC (0.97–1.00). A minimal feature set (MAP MoM, BMI, parity, PlGF MoM, UtA-PI MoM) retained ≥95% fidelity, supporting scalable, explainable PE triage under real-world constraints.

Keywords: Preeclampsia; first-trimester screening; model distillation; cost-aware tiers; explainable artificial intelligence, machine learning.

I. INTRODUCTION

Preeclampsia (PE) remains a leading cause of maternal and perinatal morbidity and mortality worldwide, with substantial acute and long-term consequences for both mother and child.[1,2] Contemporary prevention strategies increasingly rely on first-trimester risk stratification to enable early prophylaxis (e.g., low-dose aspirin) and intensified surveillance in high-risk women.[3-5]

A key advance in PE screening involves integrating maternal factors, biophysical measurements (notably mean arterial pressure, MAP), uterine artery Doppler indices (e.g., pulsatility index, UtA-PI), and placental biomarkers (e.g., PlGF, PAPP-A) into multivariable models, popularized by the Fetal Medicine Foundation (FMF) and implemented in

Corresponding author: Ta Van Thao
Hanoi Medical University
Email: tavanthao@hmu.edu.vn
Received: 15/12/2025
Accepted: 18/01/2026

commercial platforms like PerkinElmer.[4-8] These engines generate continuous risk estimates (often risk ratios) to guide decisions, such as identifying top-risk individuals for intervention, aligning with international guidelines on early screening and prevention.[5]

However, deploying these engines in low- and middle-income countries faces interlinked barriers: (1) declining data availability with tier complexity, as Doppler and biomarkers often exhibit high missingness (e.g., UtA-PI at 53.7% in our test set, reducing Tier 2 evaluability to 43.3%); (2) reliance on centralized labs for biochemical assays, increasing costs and delays; and (3) proprietary "black-box" designs limiting transparency and adaptation to local constraints.[6] In Vietnam, where healthcare resources are often limited and the full screening test costs ~75 USD, these barriers are particularly acute, underscoring the need for cost-effective alternatives. Most machine learning (ML) studies predict clinical PE outcomes, requiring full follow-up and adjudication-designs mismatched to screening workflows where commercial risk scores drive immediate actions and outcomes may be uncertain.[9,10] Distillation offers an alternative: approximating proprietary logic with transparent models trained on outputs, enhancing portability and interpretability without replacing ground truth.[11-12, 22]

Here, we treat PerkinElmer's first-trimester PE <37 weeks risk ratio as a surrogate endpoint, developing a cost-aware distillation pipeline with resource-aligned tiers. Fidelity is quantified via ranking metrics (Spearman ρ, NDCG@10, top-risk overlap) for triage utility, plus numerical agreement and calibration. We assess categorical stratification mimicry at PerkinElmer cut-offs and use permutation importance, SHAP, and ablation to identify minimal feature sets preserving ≥95% ranking fidelity.

This addresses key gaps: (i) quantifying missing-driven tier attrition's impact on fidelity and evaluability in real workflows; (ii) prioritizing decision-relevant ranking over regression errors in distillation; (iii) aligning interpretability to surrogate objectives (PerkinElmer ranking); and (iv) deriving compact subsets for cost-aware deployment without assuming full Doppler/biochemical access. We treat the PerkinElmer risk ratio as a surrogate endpoint (i.e., a proxy measure for risk stratification that guides immediate clinical decisions, rather than waiting for actual PE outcomes). This study aims to reproduce the decision logic (continuous risk ranking and threshold behavior) of the PerkinElmer screening engine, emphasizing transparency, deployability, and operational agreement rather than direct clinical outcome prediction.

II. MATERIALS AND METHODS

1. Study design, setting, and data source

We conducted a retrospective, multicenter analysis of de-identified first-trimester screening records collected across Northern Vietnam between January 2023 and December 2025. The cohort comprised 1,051 singleton pregnancies that underwent routine first-trimester assessment (≤14 weeks). For each record, the PerkinElmer platform generated a continuous preeclampsia (PE) risk ratio for PE <37 weeks, which served as the operational reference output for distillation. Biochemical assays (PlGF and PAPP-A) were processed through a centralized reference laboratory workflow (Chemedic Laboratory Center) under routine internal quality control.

2. Target definition and tiered predictors

Target variable. The primary modeling target was the PerkinElmer-estimated continuous risk ratio for PE <37 weeks. The PerkinElmer output was treated as a surrogate screening endpoint

(vendor risk engine output) rather than a clinical diagnosis. No delivery outcomes or post-20-week diagnostic labels were used in model training or primary evaluation. The PerkinElmer risk ratio serves as a surrogate operational endpoint to mimic screening workflow decisions, distinct from clinical outcome prediction. Predictors. Candidate predictors were restricted to measurements obtained at ≤14 weeks gestation to minimize temporal information leakage and reflected routine screening data: (1) Maternal characteristics and history (Gestation Age (derived from CRL), BMI, Parity, Smoking Status); (2) Hemodynamics (Left arm BP, Right arm BP, and MAP MoM); (3) Ultrasound measurements (uterine artery Doppler summarized as UTPI MoM); and (4) Biochemical markers (PlGF, PAPP-A and their MoMs, where available)

Cost-aware tiers. Predictors were grouped into three tiers to reflect real-world availability and infrastructure constraints: (1) Tier0 (low-cost / broadly available, less than 5 USD): maternal factors + MAP MoM; (2) Tier1 (adds centralized biochemistry, ~ 60 USD): (3) Tier0 + PlGF/PAPP-A (and MoMs); and Tier2 (adds Doppler, ~75 USD): Tier1 + UTPI MoM.

Tier definitions were chosen to reflect technology and workflow heterogeneity (eg, centralized biochemistry vs Doppler acquisition), and to quantify how fidelity and deployability change as higher-resource measurements become available.

3. Train–test split and data availability

The full dataset was split into a training set (n=850; 81%) and an independent test set (n=201; 19%), with stratification by PerkinElmer risk deciles to preserve the target distribution in the test set. Data availability was summarized per tier under a complete-case strategy (no imputation for tier-defining variables), yielding different evaluable N by tier in the test set (Tier0: 198; Tier1: 172; Tier2: 87). Missingness was reported to characterize implementation realism, with UTPI/UTPI MoM being the dominant driver of attrition.

Feature scaling (standardization) was fitted exclusively on the training set and applied to the test set. MoM values were computed independently using routine clinical formulas, without reliance on vendor-specific adjustments. No intermediate outputs from the PerkinElmer engine were used as predictors.

4. Distillation models and training protocol

We formulated distillation as learning a mapping from tier-specific predictors to the PerkinElmer continuous risk output, with two complementary objectives: (1) Regression agreement: preserving numerical proximity to the vendor score; (2) Decision-focused agreement: preserving relative risk ordering for triage (ranking fidelity and top-tail concordance).

Eight model families were evaluated under a consistent pipeline and per-tier complete-case protocol: Logistic/Elastic Net baseline (where applicable), Decision Tree, Random Forest (RF), Extra Trees (ET), KNN, SVM (RBF), XGBoost/HistGB, and a two-stage random-forest ensemble (2RF). Hyperparameters were tuned within the training set using stratified cross-validation consistent with the notebook protocol, and final models were refit on the full training split before one-time evaluation on the independent test set.

5. Evaluation endpoints and uncertainty estimation

All primary metrics were computed on the independent test set, within each tier using the tier-specific complete-case subset.

Continuous-score fidelity (primary distillation endpoints). Ranking fidelity: Spearman's ρ

(primary): NDCG@10, and Top-10% overlap, and numerical agreement: MAE, RMSE, and R^2 (reported as supportive, not primary).

Incremental tier analysis: To visualize cost–performance trade-offs, tier-wise curves were generated for the best-performing configuration used for the tier comparison (2RF), reporting mean metrics with 95% bootstrap confidence intervals.

Calibration as numerical agreement with the vendor scale: Calibration was assessed using predicted-vs-reference scatter plots with the identity line to describe numerical agreement with the PerkinElmer score (not outcome probability calibration).

Threshold mimic analysis (PerkinElmer cut-off): Agreement with PerkinElmer categorical stratification was evaluated by binarizing: (1) Reference label: PerkinElmer risk ratio ≥ PerkinElmer cut-off; (2) *Predicted label:* distilled output ≥ the same cut-off.

Metrics included AUC-ROC and confusion-matrix–derived measures (Accuracy, Sensitivity, Specificity, PPV/NPV, F1, MCC, Cohen's κ), with full counts (TN/FP/FN/TP).

Uncertainty: All key metrics and tables reported 95% confidence intervals via bootstrap resampling (2,000 iterations) on the test set, preserving pairing between reference and prediction.

6. Explainability and minimal deployable feature sets

Explain ability focused on the Tier2 best-performing distilled model (2RF) for the primary target (PE <37 weeks): (1) Permutation importance (decision-aligned): feature-wise permutation quantified as the decrease in ranking fidelity (ΔSpearman ρ); (2) SHAP (tree models): SHAP summary plots provided global and local contribution patterns.

To identify minimal deployable feature sets, we performed sequential backward ablation following the importance ordering and re-evaluated fidelity at each step. A ≥95% fidelity-retention criterion (relative to the full Tier2 configuration) defined the minimal subset, which was then summarized as a candidate deployable set for resource-constrained implementation.

7. Statistical Analysis

Performance metrics were reported with 95% confidence intervals estimated via paired bootstrap resampling of the test sets (1,000 iterations), preserving the pairing between predicted and reference scores. Tier comparisons were assessed using paired statistical tests on bootstrap replicates, with a two-sided significance threshold of $p < 0.05$. All analyses were implemented in Python using scikit-learn, XGBoost, and SHAP.

8. Ethical Considerations

This study complied with the ethical principles of biomedical research. All participants were healthy volunteers who provided written informed consent after being informed of the study objectives. Personal data were kept confidential. All procedures were performed under sterile conditions to ensure biosafety.

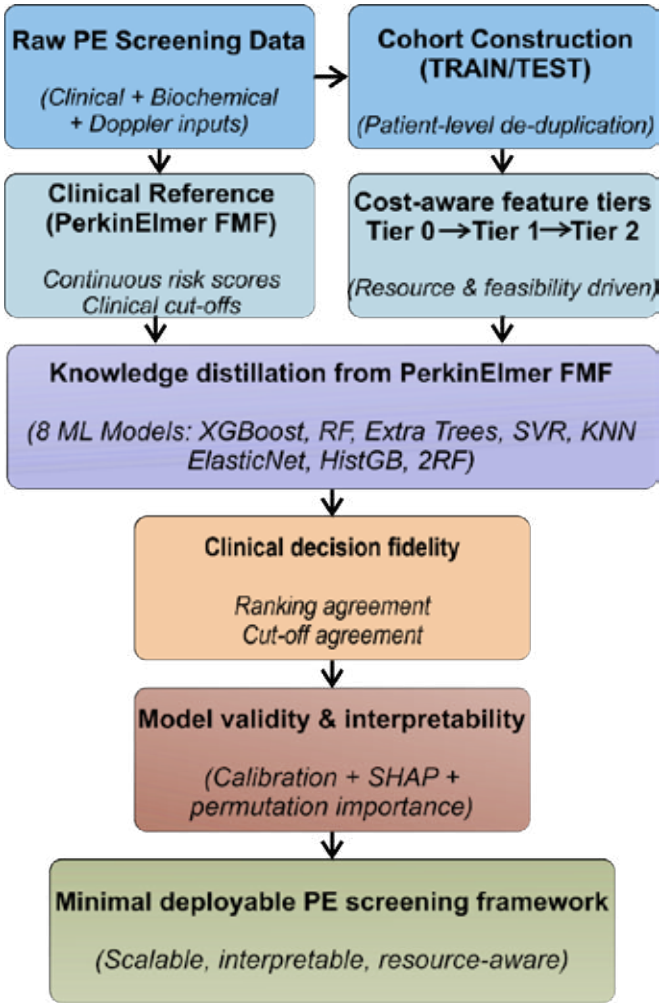

Figure 1. Overview of the proposed cost-aware knowledge distillation pipeline for preeclampsia risk screening

Raw clinical, biochemical, and Doppler screening data were first organized into patient-level training and test cohorts. The proprietary PerkinElmer FMF system was used as a clinical reference, providing continuous risk scores and established decision cut-offs. Features were structured into nested, cost-aware tiers reflecting real-world feasibility. Multiple machine-learning models were trained to distill PerkinElmer risk rankings. Model performance was assessed in terms of clinical decision fidelity (ranking agreement and cut-off mimicry), calibration, and interpretability. The pipeline ultimately identifies a minimal, deployable, and interpretable screening framework suitable for scalable clinical implementation.

III. RESULTS

1. Cohort characteristics and data availability

The study cohort comprised 1,051 singleton pregnancies collected between 2023 and 2025, including a training set of 850 cases (81%) and an independent test set of 201 cases (19%), stratified by risk deciles. Baseline maternal characteristics were comparable between the two sets, with no statistically significant

difference in maternal age (28.5 ± 4.2 years old in training vs. 28.7 ± 4.1 years old in test) or body mass index (22.1 ± 3.5 vs. 22.0 ± 3.4; $p > 0.05$ for all comparisons), supporting the validity of the data split.

Feature completeness decreased with increasing tier complexity under real-world conditions. Tier 0, comprising maternal factors and mean arterial pressure (MAP), was nearly complete, covering 98.2% of the training cohort (835/850) and 98.5% of the test cohort (198/201). Incorporation of biochemical markers in Tier 1 reduced availability to 90.2% in training (767/850) and 85.6% in testing (172/201), largely due to missing PAPP-A measurements (8.2% training, 12.9% testing). Tier 2, which additionally required uterine artery pulsatility index (UtA-PI MoM), showed a marked decline in completeness, with only 55.5% of training cases (472/850) and 43.3% of test cases (87/201) retained. This reduction was driven primarily by UtA-PI missingness (40.8% in training and 53.7% in testing), reflecting limited Doppler availability in routine practice.

Overall, these patterns underscore the trade-off between model complexity and real-world deploy ability, motivating a tiered framework that preserves fidelity while accommodating incomplete data.

2. Distillation fidelity across models and tiers

Table 1. Ranking fidelity and regression agreement for PE <37 weeks risk ratio mimic (shared evaluation; top-4 models by Spearman within each tier)

Tier	Model	N_test	Spearman	NDCG10	Top10 overlap (%)
Tier0	KNN	198	0.979	0.987	89.78
Tier0	2RF	198	0.957	0.923	72.42
Tier0	ET	198	0.922	0.914	73.92
Tier0	RF	198	0.897	0.827	59.23
Tier1	KNN	172	0.981	0.957	88.18
Tier1	2RF	172	0.976	0.953	82.56
Tier1	ET	172	0.962	0.942	82.86
Tier1	RF	172	0.942	0.939	77.75
Tier2	2RF	87	0.975	0.996	88.36
Tier2	ET	87	0.974	0.970	82.01
Tier2	KNN	87	0.965	0.974	89.56
Tier2	RF	87	0.963	0.832	76.79

Distillation fidelity was evaluated on the independent test set using a per-tier complete-case protocol, with the PerkinElmer continuous risk ratio for PE <37 weeks as the reference (**Table 1**). Models were ranked by Spearman correlation (ρ) to reflect clinical risk prioritization, supplemented by NDCG@10 and top-10 overlap for top-ranked concordance, and MAE/RMSE/R^2 for numerical agreement.Across tiers, top models showed strong ranking fidelity. In

Tier 0, KNN led with high agreement (ρ=0.979; NDCG@10=0.987; top-10 overlap=89.8%), followed by 2RF (ρ=0.957) and tree-based ensembles; regression metrics favored KNN (R^2=0.959), but ensembles declined under feature limits. Tier 1 maintained consistency, with KNN (ρ=0.981) and 2RF (ρ=0.976) topping performance (all top-four NDCG@10>0.94). Numerical agreement modestly decreased versus Tier 0, aligning with added biochemical markers and missingness. In Tier 2 (N=87, Doppler-inclusive), ensembles excelled: 2RF achieved peak ρ=0.975 (NDCG@10=0.996), followed by Extra Trees (ρ=0.974). KNN offered strong metrics but was de-emphasized due to potential sensitivity to dataset structure; ensembles like 2RF provided better cross-tier stability.

Overall, high-fidelity distillation of PerkinElmer ranking is feasible across tiers, even under constraints. Model differences stem mainly from ranking fidelity, endorsing Spearman-based metrics as primary for decision mimicry.

3. Incremental fidelity and cost–performance trade-off

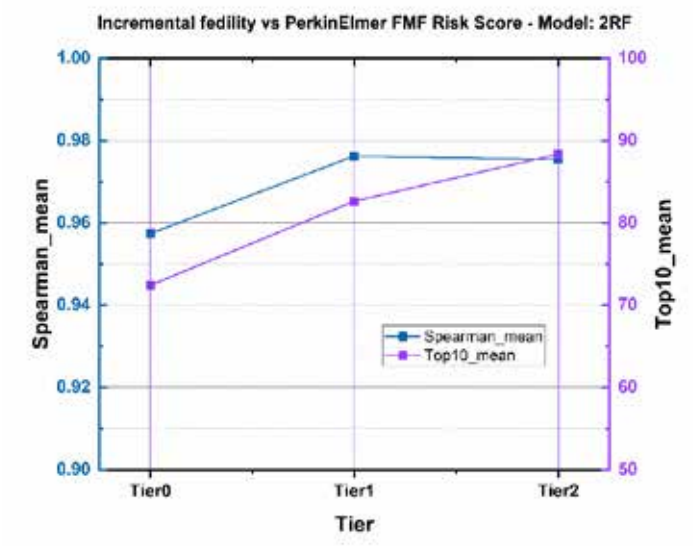

Figure 2. Incremental fidelity versus the PerkinElmer FMF continuous risk score across cost-aware tiers for PE <37 weeks (best-performing model: 2RF)

The blue line (left y-axis) represents the mean Spearman rank correlation coefficient (ρ) The purple line (right y-axis) represents the mean Top-10% overlap. Vertical error bars indicate 95% bootstrap confidence intervals.

Figure 2 illustrates ranking fidelity to the PerkinElmer continuous risk score across tiers using the 2RF model. In Tier 0 (low-cost clinical variables only), Spearman ρ was 0.96 (95% CI: 0.92–0.98) and top-10% overlap 72.4%. Adding biochemical markers in Tier 1 yielded ρ=0.98 (95% CI: 0.94–0.99) and top-10% overlap=82.6%. In Tier 2 (adding Doppler-derived UtA-PI MoM), ρ remained 0.98 (95% CI: 0.94–0.99) with top-10% overlap=88.4%. These data indicate modest global ranking gains beyond Tier 1 but incremental improvements in top-risk concordance, offset by reduced data availability in higher tiers.

4. Calibration of distilled risk scores

As shown in **Figure 3**, the distilled models demonstrate consistent numerical agreement

with the PerkinElmer continuous risk scores across all cost-aware tiers. In Tier 0, predictions generally follow the identity line in the low-to-moderate risk range, with increasing dispersion at higher risk values, reflecting the limited informational content of low-cost clinical features alone.

The inclusion of biochemical markers in Tier 1 leads to visibly improved calibration, particularly within the intermediate-risk region, where predictions cluster more closely around the identity line. Tier 2 shows the strongest alignment with the PerkinElmer risk scale, indicating enhanced numerical agreement when Doppler-derived information is incorporated, despite the reduced number of evaluable cases.

Importantly, this numerical agreement analysis is not intended to assess absolute clinical risk accuracy or outcome probability estimation. Rather, it evaluates the degree to which distilled model outputs preserve the numerical structure of the established PerkinElmer risk scoring system. Minor deviations at higher risk levels are therefore expected, particularly in Tier 2, given the smaller sample size and the distillation objective prioritizing ranking fidelity over exact probability calibration.

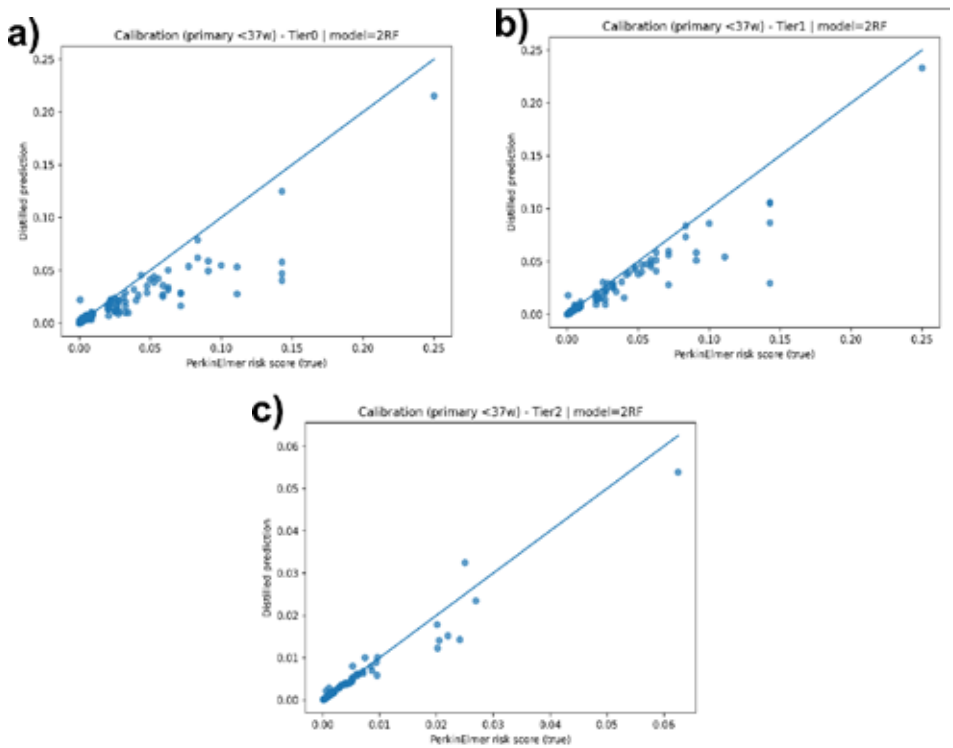

Figure 3. Numerical agreement plots comparing distilled model predictions with the PerkinElmer FMF continuous risk scores for the primary outcome (preeclampsia <37 weeks) across cost-aware tiers

Panels (a–c) correspond to Tier 0, Tier 1, and Tier 2, respectively, using the best-performing model (2RF). Each point represents an individual subject, with the x-axis indicating the original PerkinElmer risk score and the y-axis showing the corresponding distilled model prediction. The diagonal line represents perfect numerical agreement (identity line), where the distilled prediction equals the reference PerkinElmer score.

5. Explain ability and biological plausibility

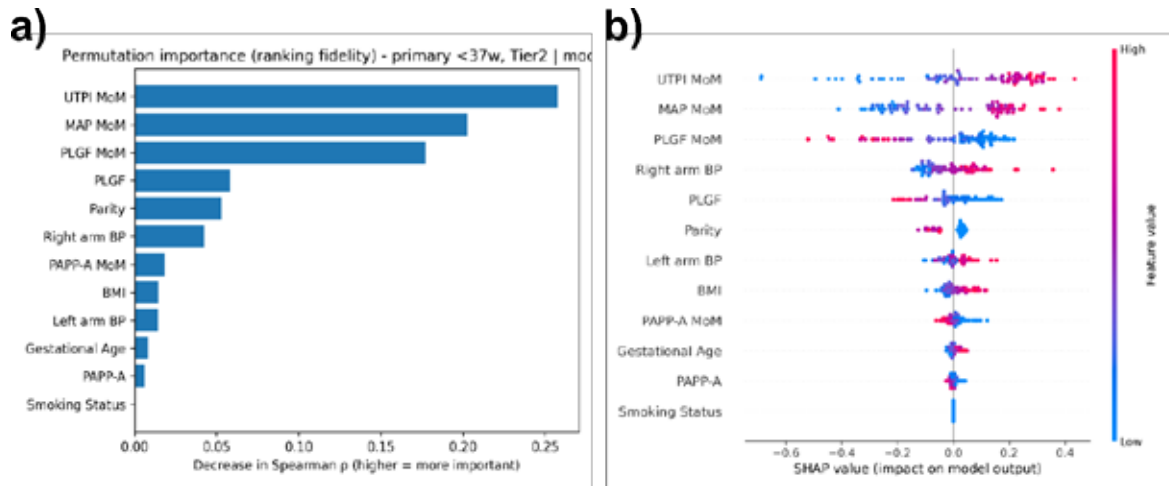

Figure 4. Explain ability of the Tier 2 distilled model (2RF) for preeclampsia <37 weeks

(a) Permutation importance as decrease in Spearman's ρ after feature permutation on the test set. (b) SHAP summary plot showing feature contributions to model output, ordered by mean absolute SHAP value; color indicates feature value (low to high)

We analyzed the Tier 2 2RF model for <37 weeks preeclampsia using permutation importance (targeted to ranking fidelity, **Figure 4a**) and SHAP values for global and local contributions (**Figure 4b**). Both methods showed consistent feature hierarchy: UtA-PI MoM had the largest Spearman ρ drop and widest SHAP spread (higher values increased risk score), followed by MAP MoM and PlGF MoM. Parity and arm blood pressures were secondary. Individual arm blood pressures contributed less than MAP MoM, reflecting redundancy. Alignment between methods indicates features drive model behavior consistent with the reference system. These analyses apply to distilled outputs (PerkinElmer risk mimic), not clinical outcomes, aligning with the fidelity objective.

6. Identification of minimal deployable feature sets

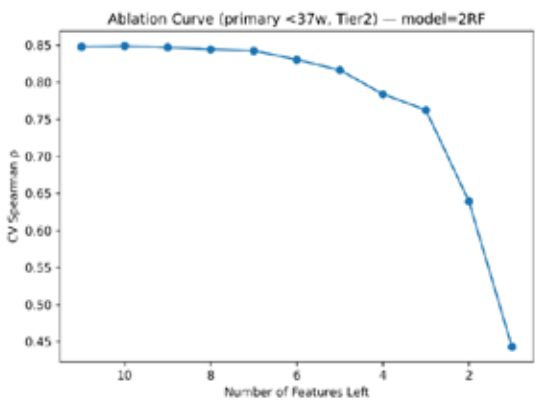

Figure 5. Sequential feature ablation for the Tier2 distilled model (2RF) on the primary outcome (<37 weeks)

Figure 5 shows sequential feature ablation for the Tier2 distilled model 2RF on the primary outcome. Cross-validated Spearman's ρ with the PerkinElmer risk ranking is plotted against the number of retained features. Features were removed sequentially following the (Tier2)

importance ordering, and fidelity retention (%) is reported relative to the full Tier2 set. Ranking fidelity remained relatively stable until ~5 retained features (ρ from 0.848 at 11 features to 0.816 at 5 features; 96.6% retention), followed by a marked decline with further pruning (ρ 0.784 at 4 features; 0.639 at 2; 0.443 at 1). Using a ≥95% retention criterion, the minimal deployable set comprised MAP MoM, BMI, parity, PLGF MoM, and UTPI MoM, capturing most decision-relevant information with substantially reduced acquisition burden.

7. Agreement with PerkinElmer categorical risk stratification

Table 2. Threshold mimic performance for preeclampsia (PE) before 37 weeks using the PerkinElmer clinical cut-off on the independent test set

Tier	Model	AUC-ROC (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Precision / PPV (95% CI)	F1 (95% CI)
Tier0 (N=198)	KNN	0.995 [0.984, 1.000]	0.990 [0.975, 1.000]	**0.968 [0.958, 1.000]**	1.000 [1.000, 1.000]	0.984 [0.958, 1.000]
Tier0 (N=198)	2RF	0.995 [0.986, 1.000]	0.879 [0.828, 0.924]	**0.635 [0.667, 0.852]**	0.976 [0.919, 1.000]	0.769 [0.667, 0.851]
Tier0 (N=198)	ET	0.981 [0.965, 0.993]	0.823 [0.768, 0.874]	**0.460 [0.494, 0.733]**	0.967 [0.889, 1.000]	0.624 [0.494, 0.733]
Tier0 (N=198)	RF	0.967 [0.944, 0.986]	0.793 [0.732, 0.848]	**0.349 [0.375, 0.639]**	1.000 [1.000, 1.000]	0.518 [0.375, 0.638]
Tier1 (N=172)	KNN	1.000 [0.999, 1.000]	0.988 [0.971, 1.000]	**0.981 [0.953, 1.000]**	0.981 [0.939, 1.000]	0.981 [0.952, 1.000]
Tier1 (N=172)	2RF	0.997 [0.992, 1.000]	0.924 [0.884, 0.959]	**0.759 [0.783, 0.928]**	1.000 [1.000, 1.000]	0.863 [0.783, 0.928]
Tier1 (N=172)	ET	0.992 [0.981, 1.000]	0.855 [0.797, 0.907]	**0.556 [0.585, 0.811]**	0.968 [0.893, 1.000]	0.706 [0.585, 0.811]
Tier1 (N=172)	RF	0.979 [0.959, 0.993]	0.831 [0.773, 0.884]	**0.463 [0.500, 0.747]**	1.000 [1.000, 1.000]	0.633 [0.500, 0.747]
Tier2 (N=87)	2RF	1.000 [1.000, 1.000]	0.943 [0.885, 0.989]	**0.375 [0.000, 0.857]**	1.000 [0.000, 1.000]	0.545 [0.000, 0.857]
Tier2 (N=87)	ET	0.994 [0.973, 1.000]	0.943 [0.885, 0.989]	**0.375 [0.000, 0.857]**	1.000 [0.000, 1.000]	0.545 [0.000, 0.857]
Tier2 (N=87)	HistGB	0.994 [0.975, 1.000]	0.931 [0.874, 0.977]	**0.250 [0.000, 0.769]**	1.000 [0.000, 1.000]	0.400 [0.000, 0.769]
Tier2 (N=87)	XGBoost	0.991 [0.962, 1.000]	0.931 [0.874, 0.977]	**0.250 [0.000, 0.769]**	1.000 [0.000, 1.000]	0.400 [0.000, 0.769]

Table 2 presents threshold mimic performance to the PerkinElmer clinical cut-off for PE <37 weeks on the independent test set, using complete-case evaluation per tier. The reference high-risk label was PerkinElmer risk ratio ≥ cut-off; distilled predictions were binarized similarly. Metrics include AUC-ROC, accuracy, PPV, and F1, with 95% CIs from bootstrap (2,000 iterations). In Tier 0 (N=198), KNN showed AUC-ROC=0.995 (95% CI: 0.984–1.000), accuracy=0.990 (0.975–1.000), PPV=1.000 (1.000–1.000), F1=0.984 (0.958–1.000); 2RF followed with similar AUC but lower accuracy/F1. In Tier 1 (N=172), KNN had AUC-ROC=1.000 (0.999–1.000), accuracy=0.988 (0.971–1.000), PPV=0.981 (0.939–1.000), F1=0.981 (0.952–1.000); 2RF achieved AUC=0.997 with high PPV but slightly lower F1. In Tier 2 (N=87), tree-based models like 2RF reached AUC-ROC=1.000 (1.000–1.000), accuracy=0.943 (0.885–0.989), PPV=1.000 (0.000–1.000), F1=0.545 (0.000–0.857); CIs for PPV/F1 widened due to small N and low high-risk prevalence (~9%). These metrics reflect agreement with PerkinElmer stratification, not PE outcome prediction, complementing prior continuous fidelity analyses.

IV. DISCUSSION

Main findings

In this real-world Vietnamese first-trimester screening cohort, we developed a cost-aware distillation framework to approximate a proprietary PerkinElmer/FM F–style continuous PE <37-week risk ratio, explicitly treating the vendor output as an operational reference rather than ground-truth disease status. This framing aligns with real screening workflows, where standardized risk engines guide clinical triage long before outcome-validated labels are available. Across all tiers, the distilled models demonstrated high fidelity to the PerkinElmer decision logic, while revealing clear trade-offs between incremental agreement and real-world measurement availability.

Methodological strengths

A key contribution of this work is the explicit quantification of deployability under complete-case, tier-wise evaluation. Tier0 variables were almost universally available (>98%), whereas Tier2 completeness dropped sharply (≈55% in training; ≈43% in testing), driven primarily by missing uterine artery Doppler (UTPI) measurements. This pattern reflects infrastructure and workflow constraints rather than model inadequacy, and underscores that scalability and equity in screening are often limited by data capture rather than algorithmic capacity. Complete-case analysis ensures fair within-tier comparison but may preferentially select better-resourced subpopulations, a known concern when missingness is not completely at random.[14] Doppler missingness is likely MNAR (e.g., tied to resource availability), potentially introducing selection bias toward better-resourced cases; future work should explore imputation or equity-aware strategies. These findings directly support tiered deployment strategies rather than assuming universal access to full Tier2 measurements.

Screening programs are primarily concerned with identifying individuals in the highest-risk strata, rather than exact numerical agreement across the full risk range. Accordingly, we emphasized ranking fidelity metrics (Spearman ρ, NDCG, top-risk overlap) alongside conventional regression measures. While global rank correlation was already high at Tier0, concordance within the highest-risk tail improved stepwise with additional features, demonstrating diminishing returns for overall ordering but continued gains where clinical decisions concentrate. The use of information-retrieval metrics provides an appropriate lens

when the downstream objective is prioritization rather than absolute risk estimation.[15]

Although KNN frequently achieved the highest point estimates in Tier0–Tier1, we highlight the two-stage random forest (2RF) as the primary model because it offered a more consistent balance of performance, interpretability, and cross-tier stability. In particular, 2RF exhibited coherent feature-attribution patterns across permutation importance and SHAP analyses, aligning well with known clinical determinants of PE risk. By contrast, KNN performance can be sensitive to local density and scaling choices, which may complicate interpretability and robustness under dataset shift.[16-18] KNN's high performance may reflect sensitivity to the dataset's structure; ensembles like 2RF were prioritized for robustness. In translational screening contexts, such stability considerations may outweigh marginal gains in a single metric.

Evaluation using the PerkinElmer cut-off demonstrated high agreement with the vendor's binary decision boundary, reinforcing that the distilled models successfully reproduce the operational stratification logic rather than predicting clinical outcomes. This distinction is essential to avoid overinterpretation, particularly given low high-risk prevalence and modest Tier2 test sizes.[19] The wide confidence intervals for F1 in Tier 2 (e.g., 0.000–0.857) reflect the small sample size (N=87) and low prevalence of high-risk cases (~9%), which amplifies variability in bootstrap estimates despite high AUC-ROC. Sequential ablation further revealed a practical "elbow," identifying a small subset of features that preserved most ranking fidelity while substantially reducing acquisition burden. This minimal deployable set represents an implementation-oriented output, complementing traditional model benchmarking.

Limitations

Remaining gaps and limitations warrant consideration: Prospective linkage to adjudicated PE outcomes is needed to evaluate clinical utility beyond agreement with the vendor engine, including decision-analytic measures such as net benefit.[20] External validation across sites, ultrasound operators, and laboratory platforms is essential to assess transportability under measurement heterogeneity.[14],[16] The structural nature of UTPI missingness warrants further investigation into missingness mechanisms and equity-aware handling strategies.[14] Transparent reporting and protocolized external validation will be critical for reproducibility and adoption.[16] Finally, while distillation can enhance transparency of proprietary systems, governance frameworks must clarify how agreement metrics relate to clinical accountability.[21,22,23,24] High AUC values reflect operational concordance with the vendor's stratification, not diagnostic accuracy against clinical outcomes. Fidelity metrics indicate successful reproduction of vendor logic but do not validate against disease outcomes.

Deployment implications

Overall, the central finding is not merely high numerical agreement, but the demonstration that decision-focused fidelity to an established commercial PE risk engine can be preserved using cost-aligned feature tiers. The dominant barrier to full Tier2 deployment is measurement availability-particularly UTPI-rather than model capacity, highlighting where future efforts should focus to improve equitable and scalable prenatal screening.

Future Work

Prospective validation against adjudicated clinical outcomes is essential before considering standalone use beyond operational alignment with commercial standards.

V. CONCLUSION

This work demonstrates that a transparent, tiered ML distillation framework can closely reproduce both continuous ranking and categorical cut-off behavior of a commercial first-trimester PE risk engine in a real-world Vietnamese screening context. The key trade-off is not merely model performance, but the sharp decline in measurement availability as tiers require higher-resource inputs. By quantifying fidelity, uncertainty, and minimal feature sets across tiers, the framework provides an evidence-based pathway for extending advanced screening logic to resource-constrained settings while preserving interpretability and operational alignment with established commercial standards. Prospective validation against adjudicated clinical outcomes is required before any stand alone clinical use.

REFERENCES

1. Steegers EAP, von Dadelszen P, Duvekot JJ, Pijnenborg R. Pre-eclampsia. *Lancet*. 2010; 376(9741): 631-644. doi:10.1016/S0140-6736(10)60279-6

2. Karumanchi SA, Granger JP. Preeclampsia and Pregnancy-Related Hypertensive Disorders. *Hypertension*. 2016; 67(2): 238-242. doi:10.1161/HYPERTENSIONAHA.115.05024.

3. Rolnik DL, Wright D, Poon LC, et al. Aspirin versus placebo in pregnancies at high risk for preterm preeclampsia. *N Engl J Med*. 2017; 377(7): 613-622. doi:10.1056/NEJMoa1704559.

4. Akolekar R, Syngelaki A, Poon L, Wright D, Nicolaides KH. Competing risks model in early screening for preeclampsia by biophysical and biochemical markers. *Fetal Diagn Ther*. 2013; 33(1): 8-15. doi:10.1159/000341264.

5. O'Gorman N, Wright D, Poon LC, 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; 49(6): 756-760. doi:10.1002/uog.17455.

6. Poon LC, Shennan A, Hyett JA, et al. The FIGO initiative on pre-eclampsia: a pragmatic guide for first-trimester screening and prevention. *Int J Gynecol Obstet*. 2019; 145(S1): 1-33. doi:10.1002/ijgo.12802.

7. Martins JG, Miller E, Aboukhater D, Bittner M, Rolnik DL, Kawakita T. Performance of a first-trimester combined screening for preterm preeclampsia in the United States population using the fetal medicine foundation competing risks model. *Am J Obstet Gynecol MFM*. 2025; 7(12): 101803. doi:10.1016/j.ajogmf.2025.101803.

8. Verlohren S, Herraiz I, Lapaire O, et al. The sFlt-1/PlGF ratio in different types of hypertensive pregnancy disorders and its prognostic potential in preeclamptic patients. *Am J Obstet Gynecol*. 2012; 206(1): 58.e1-58.e588. doi:10.1016/j.ajog.2011.07.037.

9. Breiman L. Random forests. *Mach Learn*. 2001; 45: 5-32. doi:10.1023/A:1010933404324.

10. von Dadelszen P, Magee LA, Roberts JM. Subclassification of preeclampsia. *Hypertens Pregnancy*. 2003; 22(2): 143-148. doi:10.1081/PRG-120021060.

11. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016: 1135-1144. doi:10.1145/2939672.2939778.

12. Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran*

Associates Inc., Red Hook, NY, USA, 2017; 4768-4777.

13. Bucur O, et al. Knowledge distillation in medical imaging: A survey. arXiv:2203.04742; 2022.

14. Perkins NJ, Cole SR, Harel O, et al. Principled Approaches to Missing Data in Epidemiologic Studies. *Am J Epidemiol*. 2018; 187(3): 568-575. doi:10.1093/aje/kwx348.

15. Järvelin K, Kekäläinen J. Cumulated gain-based evaluation of IR techniques. *ACM Trans Inf Syst.* 2002; 20(4): 422-446. doi:10.1145/582415.582418.

16. Collins GS, Reitsma JB, Altman DG. *et al.* Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med,* 2015; 13:1. https://doi.org/10.1186/s12916-014-0241-z.

17. Sculley D, Holt G, Golovin D, Davydov E, Phillips T, Ebner D, et. al. Hidden technical debt in machine learning systems. In: *Advances in Neural Information Processing Systems 28 (NeurIPS)*; 2015: 2503-2511.

18. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019; 1: 206-215.

19. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988; 44(3): 837-845.

20. Vickers AJ, Cronin AM, Elkin EB, Gonen M. Extensions to decision curve analysis. *BMC Med Inform Decis Mak*. 2008; 8: 53. doi:10.1186/1472-6947-8-53.

21. Vickers AJ, Cronin AM, Elkin EB *et al.* Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Med Inform Decis Mak*. 2008; 8: 53. https://doi.org/10.1186/1472-6947-8-53.

22. Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. *arXiv:* 1503.02531; 2015.

23. Molnar C. Interpretable Machine Learning. 2nd ed. 2022.

24. Wiens J, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med*. 2019; 25: 1337-1340.