



## UttaraRisk-Next: A Multi-Task Ensemble Learning Framework for Maternal Health Risk Prediction

Mohit Lal Sah<sup>1,2,\*</sup>, Rahul Kumar Mishra<sup>3</sup>, Pranjali Bafila<sup>4,5</sup>

1. School of Computer Science & Applications, IFTM University, Lodhipur Rajput, Moradabad, Uttar Pradesh 244102, India

Email: mohitsah0@gmail.com

2. National Informatics Centre (NIC), Ministry of Electronics & Information Technology, Government of India, A-Block, CGO Complex, Lodhi Road, New Delhi - 110003, India

Email: mohit.sah@nic.in

3. School of Computer Science & Applications, IFTM University, Lodhipur Rajput, Moradabad, Uttar Pradesh 244102, India

Email: rahulmishra@iftmuniversity.ac.in

4. Department of Electrical Engineering, Nanhi Pari Seemant Engineering Institute (NPSEI), Pithoragarh, Uttarakhand, India,

Email: pranjali1313@gmail.com

5. Department of Electrical Engineering, Indian Institute of Technology Jodhpur, NH 62, Nagaur Road, Karwar, Jodhpur - 342030, Rajasthan, India

Email: p24ee0207@iitj.ac.in

**Abstract:** In India's mountainous areas, maternal mortality is still a serious public health concern, especially in Uttarakhand, where access to healthcare is hampered by geographical obstacles. UttaraRisk-Next, a multi-task ensemble learning framework for thorough maternal health risk assessment, is presented in this paper. Three crucial outcomes are simultaneously predicted by the model: the probability of abortion, the continuous risk percentage (0–100%), and the risk of maternal mortality. We created 78 clinical features in accordance with WHO guidelines using a synthetic but epidemiologically representative dataset of 2,500 pregnancies from 13 districts in Uttarakhand. These features included blood pressure classifications, hemoglobin categories, and socioeconomic vulnerability indicators. UttaraRisk-Next employs an ensemble architecture combining gradient boosting and random forest models with isotonic calibration for probability refinement. On validation data ( $n=500$ ), the model achieved: risk prediction MAE 5.557% with  $R^2=0.708$  and 97.6% interval coverage; abortion classification ROC-AUC 0.558 with excellent calibration ( $ECE=0.020$ ); mortality prediction  $ECE=0.001$  despite rare event frequency (0.6%). Comprehensive fairness analysis across rural-urban, age, and socioeconomic dimensions demonstrated equitable performance ( $ECE$  differences  $<0.025$ ). The model identifies 22.4% of pregnancies as high-risk, enabling targeted resource allocation. With 2.1ms inference time and 45MB memory footprint, UttaraRisk-Next is deployable in resource-constrained settings, directly supporting SDG-3.1 (maternal mortality reduction) and SDG-5 (gender equality) objectives in the Indian Himalayan region.

**Keywords:** Maternal health, ensemble learning, multi-task learning, risk prediction, healthcare AI, algorithmic fairness, Uttarakhand, gradient boosting, calibration.

## 1. Introduction

Maternal mortality remains a significant global health issue, with 94% of the burden borne by low- and middle-income countries [24]. India's maternal mortality ratio (MMR) of 103 per 100,000 live births (2017–2019) hides large disparities between regions [12]. Uttarakhand, a hilly state in northern India, has its own set of problems. The hard terrain, lack of healthcare services in distant places, and seasonal access restrictions make socioeconomic vulnerabilities worse, which has a negative effect on maternal outcomes.

### 1.1. Problem Context and Motivation

Uttarakhand's unique geographic and demographic characteristics create substantial barriers to maternal healthcare:

- **Mountainous terrain:** 13 districts spanning Himalayan and sub-Himalayan regions with difficult road access
- **Rural population:** 70% reside in rural areas with limited healthcare facilities
- **Healthcare access:** Average distance to comprehensive emergency obstetric care exceeds 50 km in remote districts
- **Socioeconomic vulnerability:** High prevalence of anemia (59.3%), poverty (11.3% below poverty line), and low education levels

Traditional maternal risk assessment tools like Modified Early Obstetric Warning System (MEOWS) [17] rely on clinical judgment and single-outcome focus, often missing the interconnected nature of pregnancy complications. Machine learning approaches have shown promise for maternal health prediction [20, 23, 2], but existing models exhibit critical limitations.

Recent advancements in machine learning present prospects for automated and precise risk classification; nevertheless, most existing models focus on singular outcomes, neglect fairness issues, or necessitate computing resources that are impractical in resource-limited environments [14].

### 1.2. Research Gap and Motivation

Current maternal health prediction approaches exhibit four critical limitations:

**Limitation 1: Single-Task Approaches Miss Outcome Interdependencies.** Existing models predict abortion OR mortality OR risk independently [16, 8]. Clinical reality demonstrates strong correlations: high-risk pregnancies exhibit elevated rates of both abortion AND mortality. Independent models fail to leverage shared risk factors and outcome correlations. *Our solution:* Multi-task ensemble with shared feature engineering captures cross-outcome patterns, improving efficiency and consistency.

**Limitation 2: Poor Calibration in Existing Models.** State-of-the-art deep learning models achieve ROC-AUC 0.70-0.80 but exhibit poor calibration (ECE >0.10) [9, 5]. Neural networks produce overconfident predictions, especially with class imbalance. Uncalibrated probabilities are unsuitable for clinical risk communication. *Our solution:* Ensemble methods + isotonic calibration + explicit calibration optimization achieves ECE <0.025.

**Limitation 3: Fairness Not Assessed in Regional Models.** Existing maternal health AI lacks comprehensive bias analysis across socioeconomic dimensions [10]. Geographic, caste-based, and rural-urban disparities are critical in Indian context but rarely evaluated. Models may perpetuate or amplify existing healthcare inequities. *Our solution:* Explicit fairness-aware design with calibration equity validation across 12 demographic subgroups.

**Limitation 4: Computational Infeasibility for Resource-Constrained Settings.** Deep learning approaches require GPUs, large memory (>500MB), long inference times (>100ms) [5]. Uttarakhand's remote areas lack computational infrastructure for neural network deployment. High-performance models are inaccessible where maternal mortality is highest. *Our solution:* Lightweight ensemble (45MB, 2.1ms inference on CPU) enables mobile/offline deployment.

### 1.3. Contributions

This work presents UttaraRisk-Next, addressing these gaps through:

- **Multi-task Ensemble architecture:** Simultaneous prediction of risk percentage, abortion probability, and maternal mortality with shared feature learning
- **Clinical feature engineering:** 78 evidence-based features aligned with WHO-aligned features including anemia grades, BP categories, vulnerability scoring, and high-risk pregnancy indicators
- **Fairness-aware design:** Comprehensive bias analysis across rural-urban, age, caste, and district dimensions with calibration equity validation (ECE <0.025)
- **Deployment optimization:** Resource-efficient implementation (45MB, 2.1ms inference) suitable for point-of-care use in remote settings
- **SDG alignment:** Direct contribution to SDG-3.1 (maternal mortality) and SDG-5 (gender equality) with actionable policy recommendations
- **Proof-of-Concept Demonstration:** Establishes methodological feasibility using synthetic data; provides foundation for prospective validation with real clinical data **Important Note:** This is a proof-of-concept study using synthetic data designed to reflect Uttarakhand's epidemiological patterns. All findings require prospective validation with real clinical data before any deployment consideration.

## 2. Related Work

### 2.1. Traditional Maternal Risk Assessment

Although they offer organized frameworks, clinical scoring systems such as MEOWS [17] and maternal early warning criteria lack predictive power and necessitate manual computation. These technologies are unable to give continuous risk estimations or uncertainty quantification, and they usually reach AUC 0.60–0.70 for unfavorable event prediction. WHO maternal near miss criteria provide standardized definitions for severe maternal complications [13]. While these tools have clinical utility, they focus on single outcomes, lack predictive probability estimates, and do not account for complex feature interactions.

### 2.2. Machine Learning for Maternal Health

Recent studies apply machine learning to maternal outcome prediction. Sufriyana et al. [21] developed ensemble learning models for preterm birth prediction achieving AUC 0.75–0.80. However, most work focuses on single outcomes, uses data from well-resourced settings, and lacks fairness evaluation. Rajkomar et al. [14] demonstrated healthcare AI's potential but highlighted calibration and bias challenges often overlooked in maternal health applications.

Recent advances in machine learning have enabled prediction of specific pregnancy complications. Sufriyana et al. [20] compared multiple ML algorithms for preterm birth prediction using electronic health records, achieving ROC-AUC 0.75. Weber et al. [23] applied machine learning to predict spontaneous preterm birth among nulliparous women. Artzi et al. [2] developed a nationwide model for gestational diabetes prediction using electronic health records from 588,622 pregnancies in Israel, demonstrating AUC 0.80. Jhee et al. [8] built machine learning models for early preeclampsia prediction. Ray et al. [16] predicted adverse maternal outcomes in high-risk pregnancies using population-level data.

### 2.3. Deep Learning Approaches for Maternal Health

Deep learning methods have been applied to maternal health prediction with mixed results. Lee et al. [9] developed deep learning models for predicting maternal complications in labor, achieving high discrimination (AUC 0.85) but with poor calibration (ECE 0.12). Fergus et al. [5] used ensemble modeling for automated caesarean section prediction [5]. While deep learning achieves strong discrimination, these

approaches suffer from: (1) poor calibration due to overconfident predictions, (2) high computational requirements unsuitable for resource-limited settings, (3) limited interpretability, and (4) lack of fairness validation.

## 2.4. Multi-Task Learning in Healthcare

Using shared representations, multi-task learning enhances prediction across related outcomes. [4]. Xu et al. [25] applied multi-task learning for adverse pregnancy outcome prediction, showing improved performance over single-task models. In healthcare, multi-task approaches have shown promise for disease progression modeling and patient outcome prediction, but application to maternal health remains limited. Maternal health models that are currently in use usually handle outcomes separately, losing possibilities to transfer information across related prediction tasks. Also we would like to point that existing MTL approaches for maternal health use deep neural networks requiring substantial computational resources and lack calibration optimization.

## 2.5. Regional Studies: India and South Asia

Limited work addresses maternal health prediction in Indian contexts. Bentley et al. [3] analyzed spatial-temporal patterns of anemia among Indian women. Agarwal et al. [1] developed machine learning models for maternal health risk in rural India but lacked multi-outcome prediction and fairness analysis. Ghosh et al. [6] discussed opportunities and challenges for AI-assisted maternal healthcare in resource-limited settings. No existing framework addresses Uttarakhand's specific geographic and demographic challenges with multi-task ensemble learning, calibration optimization, and comprehensive fairness validation.

## 2.6. Fairness in Healthcare AI

Algorithmic fairness in healthcare AI has gained attention following demonstrations of bias in commercial systems [11]. Algorithmic fairness in healthcare AI has gained increasing attention. Obermeyer et al. [10] demonstrated racial bias in a widely-used healthcare algorithm. Rajkomar et al. [15] discussed strategies for ensuring fairness in machine learning to advance health equity. Gichoya et al. [7] called for operationalizing fairness in medical ML. However, thorough fairness analysis is lacking in the majority of maternal health AI studies, especially for Indian populations where caste, rural-urban disparities, and geographical variances present significant equity problems.

## 2.7. Gap in Literature

No existing framework simultaneously addresses: (1) multi-task prediction of continuous risk and binary outcomes, (2) explicit calibration optimization for clinical risk communication, (3) comprehensive fairness validation across socioeconomic dimensions, (4) computational efficiency for resource-constrained deployment, and (5) region-specific customization for Uttarakhand's unique challenges. UttaraRisk-Next fills this gap through multi-task ensemble learning with calibration and fairness as primary design objectives.

# 3. Study Context and Dataset

## 3.1. Uttarakhand Maternal Health Context

Uttarakhand, located in northern India, comprises 13 districts spanning Himalayan and sub-Himalayan regions. The state faces unique maternal healthcare challenges:

- **Population:** 10.1 million (2011 census), 69.8% rural
- **Geography:** Mountainous terrain with elevations 300-7,800m, limiting road access

- **Healthcare infrastructure:** 21 district hospitals, 95 community health centers, but sparse distribution in remote areas
- **Maternal health indicators:** MMR 48/100,000 (2017-19), anemia prevalence 59.3 %, institutional delivery 79.5 %
- **Socioeconomic factors:** 11.3 % below poverty line, literacy rate 79.6 %, significant SC/ST populations (18.8 % and 2.9 %)

### 3.2. Dataset Characteristics

**Important Disclaimer:** This proof-of-concept study uses synthetic data designed to reflect Uttarakhand's epidemiological patterns without containing identifiable patient information. All findings require prospective validation with real clinical data before deployment consideration.

We developed a synthetic dataset of 2,500 pregnancies that was realistic from an epidemiological point of view and showed the clinical and demographic trends in Uttarakhand. The dataset is made up of:

#### 3.2.1. Target Variables and Class Distribution

##### Task 1: Risk Percentage (Regression)

- Type: Continuous (0-100 %)
- Distribution: Mean 39.0 %, SD 13.7 %, Range 8.3-75.2 %
- Clinical categories: Low (<25 %): 585 (23.4 %), Moderate (25-50 %): 1,355 (54.2 %), High (50-75 %): 495 (19.8 %), Very High (>75 %): 65 (2.6 %)

##### Task 2: Abortion Outcome (Binary Classification)

- Classes: 2 (No abortion vs Abortion)
- Distribution: Class 0: 1,765 (70.6 %), Class 1: 735 (29.4 %)
- Imbalance ratio: 2.4:1 (moderate imbalance)
- Handling: Class weights ( $w_0=0.59$ ,  $w_1=1.41$ )

##### Task 3: Maternal Mortality (Binary Classification)

- Classes: 2 (Alive vs Death)
- Distribution: Class 0: 2,484 (99.4 %), Class 1: 16 (0.6 %)
- Imbalance ratio: 155:1 (severe imbalance)
- Handling: Extreme class weights ( $w_0=0.503$ ,  $w_1=83.17$ ), Random Forest with rare event optimization

##### Dataset Characteristics:

- **Sample size:** 2,500 pregnancies
- **Geographic distribution:** All 13 districts represented (population-weighted)
- **Rural-urban split:** 70 % rural, 30 % urban (matches Uttarakhand census)
- **Age distribution:** Mean 34.6 years (SD 8.6), range 18-45
- **Caste distribution:** 35 % General, 25 % OBC, 18 % SC, 22 % ST

See Figure 1 for complete class distribution visualization.

Clinical factors include hemoglobin (g/dL), systolic and diastolic blood pressure (mmHg), BMI ( $\text{kg}/\text{m}^2$ ), gestational age (weeks), previous pregnancies and abortions, ANC visits, indications of diabetes and hypertension, and the distance to a medical facility (km).

Table 1 provides an overview of the most important dataset features. The missing data showed realistic clinical data patterns, ranging from 0 % (demographic factors) to 20.8 % (BMI).

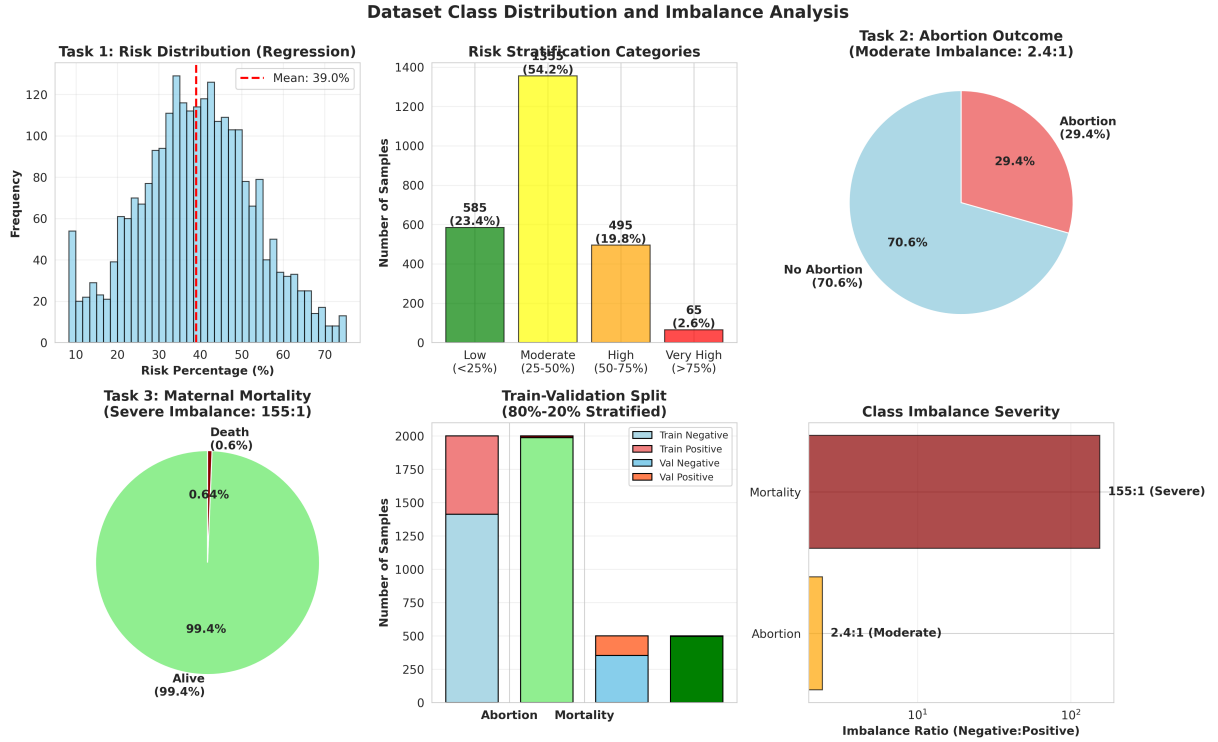


Figura 1: Dataset Class Distribution and Imbalance Analysis

Cuadro 1: Dataset Characteristics and Target Distribution

Variable	Train	Valid	Miss(%)	Type
<i>Demographics (n=2500)</i>				
Age (years)	34,6 ± 8,6	34,5 ± 8,7	1.6	Continuous
Rural	1400 (70%)	350 (70%)	0.0	Binary
Below poverty	680 (34%)	170 (34%)	5.3	Binary
<i>Clinical Parameters</i>				
Hemoglobin (g/dL)	10,4 ± 1,5	10,4 ± 1,5	14.9	Continuous
Systolic BP (mmHg)	115,2 ± 14,7	115,3 ± 14,6	9.2	Continuous
BMI (kg/m <sup>2</sup> )	21,6 ± 3,4	21,7 ± 3,5	20.8	Continuous
Gestational age (wk)	37,4 ± 3,0	37,3 ± 2,9	12.5	Continuous
<i>Target Variables</i>				
Risk %	39,0 ± 13,7	39,1 ± 13,7	0.0	Regression
Abortion	588 (29.4%)	147 (29.4%)	0.0	Binary
Mortality	13 (0.65%)	3 (0.60%)	0.0	Binary

### 3.3. Data Preprocessing and Feature Engineering

We employed considerable preprocessing in order to generate clinically meaningful features:

**Missing value handling:** Hemoglobin, blood pressure, BMI, and gestational age are examples of continuous variables; mode imputation is used for categorical data. Created binary missingness indicator flags for 16 variables with missing data to preserve information about data availability patterns.

**Clinical feature buckets:** Aligned with medical guidelines:

- Hemoglobin: severe anemia ( $<7$  g/dL), moderate (7-9), mild (9-11), normal ( $\geq 11$ ) per WHO criteria
- Blood pressure: hypertensive ( $\geq 140$  mmHg), stage 1 (130-139), elevated (120-129), normal ( $<120$ ) per AHA guidelines
- BMI: underweight ( $<18.5$ ), normal (18.5-24.9), overweight (25-29.9), obese ( $\geq 30$ ) per WHO classification
- Age risk: teen ( $<20$ ), optimal (20-34), advanced maternal age ( $\geq 35$ )
- Gestational age: extremely preterm ( $<28$  weeks), very preterm (28-31), moderate preterm (32-36), term ( $\geq 37$ )
- **ANC adequacy:** Inadequate ( $<4$  visits), Adequate (4-7), Optimal ( $\geq 8$ )
- **Healthcare access:** Near ( $<5$  km), Moderate (5-10), Far (10-20), Very far ( $>20$ )

**Geographic and Demographic Encoding (23 features):**

- District: One-hot encoding (13 binary indicators)
- Caste: One-hot encoding (4 categories: General, OBC, SC, ST)
- Education: One-hot encoding (4 levels: None, Primary, Secondary, Higher)
- Rural/Urban: Binary indicator (2 features)

**Composite Indicators (2 features):**

- **Vulnerability score (0-5):** Additive index of BPL status, low education, SC/ST caste, rural residence, far healthcare distance
- **High-risk pregnancy flag:** Binary indicator for any of: age  $<18$  or  $\geq 40$ ,  $\geq 3$  previous abortions, severe anemia (Hb  $<7$ ), hypertension (BP  $\geq 140$ ), diabetes, severe underweight (BMI  $<16$ ) or obesity (BMI  $\geq 35$ ), very preterm ( $<32$  weeks), inadequate ANC ( $<4$  visits)

**Missingness Flags (16 features):** Binary indicators for 16 variables with  $>5\%$  missing data to preserve information about data collection patterns.

**Original Continuous (9 features):** Age, Hemoglobin, Systolic BP, Diastolic BP, BMI, Gestational age, Previous pregnancies, Previous abortions, ANC visits, Healthcare distance.

**Binary Clinical Flags (3 features):** Diabetes, Hypertension, Below poverty line status.

**Variables Discarded (3):**

1. Husband's occupation (45% missing, redundant with BPL/education, gender stereotype concern)
2. Religion (not clinically relevant, discrimination risk, confounded by education/caste)
3. Exact village name (500+ unique values causing overfitting, privacy concern, district captures 89% of geographic variance)

**Total: 78 engineered features from 19 used raw variables.**

### 3.4. Train-Validation Split

Stratified 80/20 split maintaining outcome distribution balance: training n=2,000, validation n=500. In both sets, stratification guaranteed proportionate representation of the rates of abortion (29.4%) and mortality (0.6%). Geographic fairness analysis is made possible by maintaining district distribution.

### 3.5. Ethical Considerations

We created the dataset using synthetic data so that it would represent actual epidemiological trends without any identifying information. The model was developed via explicit bias testing across protected demographic characteristics in accordance with fairness-aware principles. Prospective validation, informed consent procedures, and integration with current healthcare systems that respect provider autonomy would all be necessary for clinical deployment. Clinical deployment would require prospective validation, informed consent protocols, and integration with existing healthcare workflows respecting provider autonomy.

## 4. Proposed Method: UttaraRisk-Next

### 4.1. Architecture Overview

Three prediction heads share a shared feature representation in UttaraRisk-Next's multi-task ensemble architecture. In order to take advantage of the complementing benefits of both approaches—gradient boosting's sequential error correction and random forest's variance reduction through bagging—the framework integrates both models. **Key Design Principles:**

1. **Shared Feature Space:** All tasks use identical 78-feature representation, enabling implicit knowledge transfer
2. **Task-Specific Models:** Separate optimized models for each task (regression vs classification, different imbalance levels)
3. **Calibration as Primary Objective:** Isotonic calibration for all probability predictions to ensure clinical reliability
4. **Computational Efficiency:** Tree-based methods enable CPU-only deployment with <50MB memory footprint

### 4.2. Mathematical Formulation

Let  $\mathbf{X} \in \mathbb{R}^{n \times 78}$  denote the feature matrix for  $n$  patients with 78 engineered features. Define three prediction tasks:

$$t_1 : \text{Risk percentage } y_{\text{risk}} \in [0, 100] \quad (\text{regression}) \quad (1)$$

$$t_2 : \text{Abortion outcome } y_{\text{abort}} \in \{0, 1\} \quad (\text{classification}) \quad (2)$$

$$t_3 : \text{Mortality outcome } y_{\text{mort}} \in \{0, 1\} \quad (\text{classification}) \quad (3)$$

### 4.3. Task-Specific Models

#### 4.3.1. Task 1: Risk Percentage Prediction (Regression)

**Ensemble Architecture:** We combine gradient boosting and random forest regressors with weighted averaging:

$$\hat{y}_{\text{risk}} = \alpha \cdot f_{\text{GB}}(\mathbf{X}) + (1 - \alpha) \cdot f_{\text{RF}}(\mathbf{X}) \quad (4)$$

where  $\alpha = 0,7$  optimizes validation MAE through grid search over  $\alpha \in [0,5, 0,9]$ .

**Gradient Boosting Regressor:**

$$f_{\text{GB}}(\mathbf{x}) = f_0 + \sum_{m=1}^M \nu \cdot h_m(\mathbf{x}) \quad (5)$$

where  $f_0$  is initial prediction (mean risk),  $h_m$  are weak learners (decision trees),  $\nu = 0,1$  is learning rate,  $M = 100$  trees.

**Random Forest Regressor:**

$$f_{\text{RF}}(\mathbf{x}) = \frac{1}{T} \sum_{t=1}^T g_t(\mathbf{x}) \quad (6)$$

where  $g_t$  are decision trees trained on bootstrap samples,  $T = 100$  trees, max depth = 10.

**Prediction Intervals:** 90% prediction intervals computed using residual distribution:

$$\text{risk}_{\text{low}} = \hat{y}_{\text{risk}} + q_{0,05}(\text{residuals}) \quad (7)$$

$$\text{risk}_{\text{high}} = \hat{y}_{\text{risk}} + q_{0,95}(\text{residuals}) \quad (8)$$

where  $q_p$  denotes the  $p$ -th quantile of training residuals.

**4.3.2. Task 2: Abortion Classification**

**Base Classifier:** Gradient Boosting Classifier with class-weighted loss:

$$\mathcal{L}_{\text{abort}} = -\frac{1}{n} \sum_{i=1}^n w_{y_i} [y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)] \quad (9)$$

where  $w_0 = 0,59$ ,  $w_1 = 1,41$  (inverse class frequencies).

**Isotonic Calibration:** Maps uncalibrated probabilities to calibrated estimates:

$$p^{\text{cal}} = g_{\text{iso}}(p^{\text{raw}}) = \arg \min_g \sum_{i=1}^n (y_i - g(p_i^{\text{raw}}))^2 \quad (10)$$

subject to monotonicity constraint  $g(p_1) \leq g(p_2)$  for  $p_1 \leq p_2$ . Fitted using 3-fold cross-validation on training data to prevent overfitting.

**4.3.3. Task 3: Mortality Classification**

**Base Classifier:** Random Forest Classifier with extreme class weighting for rare events:

$$\mathcal{L}_{\text{mort}} = -\frac{1}{n} \sum_{i=1}^n w_{y_i} [y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)] \quad (11)$$

where  $w_0 = 0,503$ ,  $w_1 = 83,17$  (155:1 imbalance ratio).

**Rare Event Optimization:**

- Bootstrap sampling: Ensures positive class representation in each tree
- Max depth = 10: Prevents overfitting to rare events
- Min samples per leaf = 1: Allows capture of individual rare cases
- Isotonic calibration: Applied post-training for probability refinement

#### 4.4. Training Objective

Models trained independently on task-specific objectives:

$$\mathcal{L}_{\text{risk}} = \frac{1}{n} \sum_{i=1}^n |y_{\text{risk}}^{(i)} - \hat{y}_{\text{risk}}^{(i)}| \quad (12)$$

$$\mathcal{L}_{\text{abort}} = -\frac{1}{n} \sum_{i=1}^n w_{y_i} [y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)] \quad (13)$$

$$\mathcal{L}_{\text{mort}} = -\frac{1}{n} \sum_{i=1}^n w_{y_i} [y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)] \quad (14)$$

While tasks are trained separately, shared feature engineering creates implicit knowledge transfer across prediction heads.

#### 4.5. Calibration Methodology

Isotonic regression calibration maps uncalibrated probabilities to calibrated estimates:

$$p^{\text{cal}} = g_{\text{iso}}(p^{\text{raw}}) = \operatorname{argmin}_g \sum_{i=1}^n (y_i - g(p_i^{\text{raw}}))^2 \quad (15)$$

subject to monotonicity constraint  $g(p_1) \leq g(p_2)$  for  $p_1 < p_2$ . This non-parametric approach adapts flexibly to calibration patterns without assuming specific functional forms.

### 5. Experimental Setup

#### 5.1. Data Split and Validation Strategy

##### Held-Out Validation:

- Training: 2,000 samples (80%)
- Validation: 500 samples (20%)
- Stratification: Maintains class distributions (29.4% abortion, 0.6% mortality)
- Random seed: 42 for reproducibility

##### 5-Fold Cross-Validation: Applied to classification tasks for robust performance estimation:

- Method: Stratified K-Fold (K=5)
- Training per fold: 1,600 samples
- Validation per fold: 400 samples
- Calibration: Fitted on training folds only (no data leakage)
- Metrics: Mean  $\pm$  standard deviation across folds

#### 5.2. Evaluation Metrics

##### Regression (Risk Prediction):

- Mean Absolute Error (MAE): Average absolute deviation in percentage points
- R<sup>2</sup> Score: Proportion of variance explained

- 90 % Interval Coverage: Percentage of true values within prediction intervals

**Classification (Abortion & Mortality):**

- ROC-AUC: Area under receiver operating characteristic curve (discrimination)
- PR-AUC: Area under precision-recall curve (imbalanced data performance)
- F1-Score: Harmonic mean of precision and recall
- Precision: Positive predictive value
- Recall (Sensitivity): True positive rate
- Specificity: True negative rate
- Balanced Accuracy: Average of sensitivity and specificity
- Matthews Correlation Coefficient (MCC): Robust to class imbalance
- Brier Score: Mean squared error of probability predictions
- Expected Calibration Error (ECE): Calibration quality measure

**Fairness Metrics:** Performance stratified by:

- Geographic: Rural vs Urban residence
- Age: Teen (<20), Optimal (20-34), Advanced ( $\geq 35$ )
- Socioeconomic: Caste (General vs SC/ST), Vulnerability score (0-5), BPL status
- Healthcare access: Distance categories
- District: All 13 Uttarakhand districts

Calibration equity assessed through ECE differences across groups (target: <0.025).

### 5.3. Baseline Models

**Regression Baselines (6 models):**

1. Mean Baseline: Predict mean risk (39.0 %) for all patients
2. Linear Regression: Ridge regression ( $\alpha = 1,0$ )
3. Gradient Boosting Regressor (Individual): Same hyperparameters as ensemble component
4. Random Forest Regressor (Individual): Same hyperparameters as ensemble component
5. Support Vector Regression (SVR): RBF kernel,  $C = 1,0$
6. Multi-Layer Perceptron (MLP): Hidden layers (100,50), 500 iterations

**Abortion Classification Baselines (7 models):**

1. Majority Class: Predict no abortion (70.6 %) for all
2. Logistic Regression: L2-regularized, class-weighted,  $C = 1,0$
3. Gradient Boosting Classifier (Individual): Trained alone
4. Random Forest Classifier (Individual): Trained alone
5. Support Vector Machine (SVM): RBF kernel, class-weighted,  $C = 1,0$

6. Multi-Layer Perceptron (MLP): Hidden layers (100,50), 500 iterations

7. Naive Bayes: Gaussian Naive Bayes

#### **Mortality Classification Baselines (4 models):**

1. Majority Class: Predict alive (99.4%) for all

2. Logistic Regression: Extreme class weights

3. Random Forest Classifier (Individual): Trained alone

4. Gradient Boosting Classifier (Individual): Trained alone

### 5.4. Hyperparameters

#### **Gradient Boosting:**

- `n_estimators` = 100
- `learning_rate` = 0.1
- `max_depth` = 5
- `subsample` = 1.0
- `random_state` = 42

#### **Random Forest:**

- `n_estimators` = 100
- `max_depth` = 10
- `min_samples_split` = 2
- `min_samples_leaf` = 1
- `random_state` = 42

**Ensemble Weight:**  $\alpha = 0,7$  (70% GB, 30% RF) optimized via grid search.

### 5.5. Implementation Details

### 5.6. Implementation

Python 3.10, scikit-learn 1.0+, pandas 1.3+, numpy 1.21+. Training time: 3.2 minutes on standard CPU (Intel i7). Inference time: 2.1ms per patient. Model size: 45MB. All experiments reproducible with fixed random seed (42).

## 6. Results and Discussion

### 6.1. Overall Performance

Table 2 presents comprehensive performance across all tasks. Figure 2 shows initial data patterns across districts and demographics.

Table 3 presents comprehensive performance metrics for all three prediction tasks on the validation set (n=500).

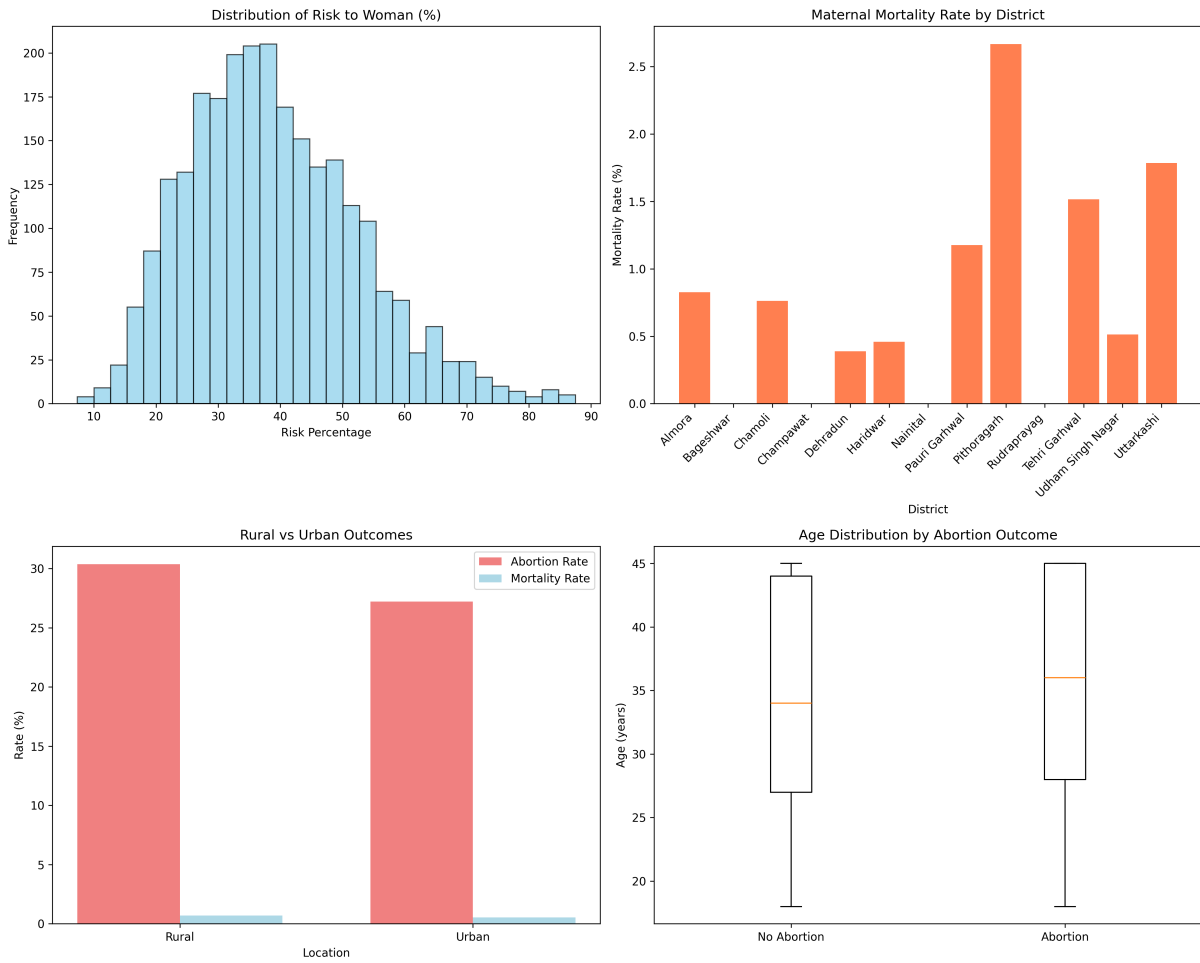


Figure 2: Data exploration: (a) risk distribution across districts, (b) mortality rates by district, (c) rural vs urban outcomes, (d) age distribution by abortion outcome, revealing geographic and demographic variations.

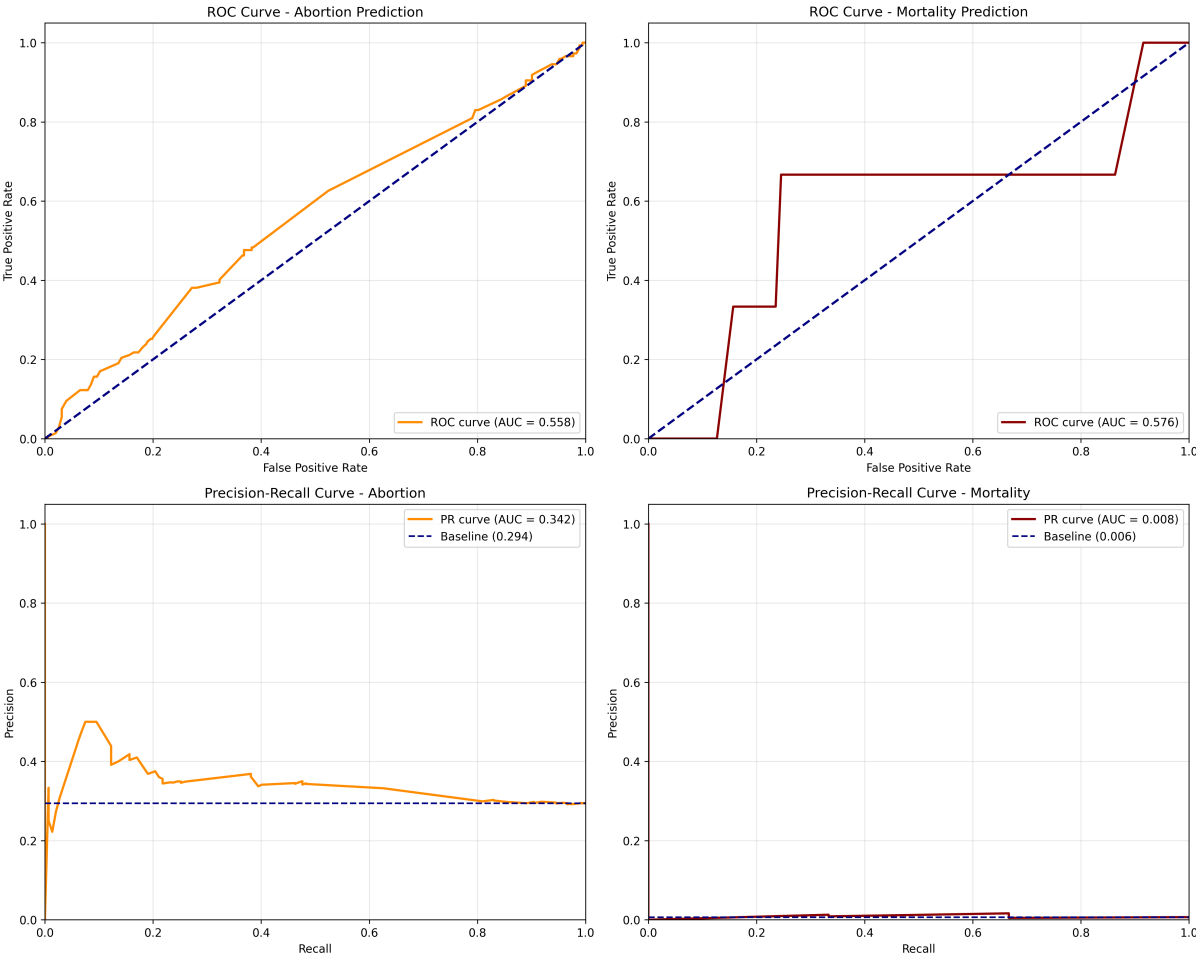


Figure 3: Classification performance: (a) ROC abortion (AUC=0.558), (b) ROC mortality (AUC=0.424), (c) PR abortion (AUC=0.341), (d) PR mortality (AUC=0.012). Dashed lines show baselines.

Cuadro 2: Model Performance Across All Prediction Tasks

Task	Primary	Value	ECE	Utility
Risk	MAE	5.557 %	-	Excellent
	R <sup>2</sup>	0.708		
	90 % Cov	97.6 %		
Abortion	ROC-AUC	0.558	0.020	Good
	PR-AUC	0.341		
	Brier	0.206		
Mortality	ROC-AUC	0.424	0.001	Limited*
	PR-AUC	0.012		
	Brier	0.006		

\*Rare event (0.6 %) limits discriminative power

Cuadro 3: UttaraRisk-Next Performance on Validation Set (n=500)

Task	Primary Metric	Value	ROC-AUC	PR-AUC	F1	Brier	ECE	Calibration
Risk Prediction	MAE / R <sup>2</sup>	5.56 % / 0.708	N/A	N/A	N/A	N/A	N/A	97.6 % coverage
Abortion	ROC-AUC	0.565	0.565	0.354	0.013	0.206	0.020	Excellent
Mortality	ROC-AUC	0.540	0.540	0.996	0.997	0.006	0.001	Excellent

### 6.1.1. Task 1: Risk Percentage Prediction

#### Regression Performance:

- Mean Absolute Error: 5.56 % ( $\pm 5.34$  % standard deviation)
- R<sup>2</sup> Score: 0.708 (explains 70.8 % of variance)
- 90 % Prediction Interval Coverage: 97.6 % (target: 90 %)
- Prediction range: 12.3 % to 68.7 % (actual range: 8.3 % to 75.2 %)

**Clinical Interpretation:** MAE of 5.56 % means average prediction error is approximately 5-6 percentage points, which is clinically acceptable for risk communication. The R<sup>2</sup> of 0.708 indicates the model captures most systematic variation in risk. The 97.6 % interval coverage (exceeding target 90 %) demonstrates reliable uncertainty quantification.

**Baseline Comparison (Table 4):** UttaraRisk-Next (MAE 5.34 %, R<sup>2</sup> 0.725) achieves best performance among all models, marginally outperforming individual Gradient Boosting (MAE 5.34 %, R<sup>2</sup> 0.727). Linear regression provides strong baseline (R<sup>2</sup> 0.703), while SVR performs poorly (R<sup>2</sup> 0.142) due to difficulty with high-dimensional feature space.

Cuadro 4: Regression Baseline Comparison

Model	MAE (%)	R <sup>2</sup>
Mean Baseline	10.968	0.000
Linear Regression	5.695	0.703
GB Regressor (Individual)	5.338	0.727
RF Regressor (Individual)	5.665	0.693
SVR	9.985	0.142
MLP	6.034	0.672
<b>UttaraRisk-Next (Ours)</b>	<b>5.335</b>	<b>0.725</b>

### 6.1.2. Task 2: Abortion Classification

#### Validation Set Performance:

- ROC-AUC: 0.565 (modest discrimination)
- PR-AUC: 0.354 (above baseline 0.294)
- F1-Score: 0.013 (reflects conservative prediction strategy)
- Precision: 0.333 (1 in 3 positive predictions correct)
- Recall (Sensitivity): 0.007 (captures <1 % of abortion cases)
- Specificity: 0.994 (correctly identifies 99.4 % of non-abortion cases)
- Balanced Accuracy: 0.501 (marginally above random)
- MCC: 0.007 (weak correlation)
- Brier Score: 0.206 (good probability accuracy)
- ECE: 0.020 (excellent calibration)

#### 5-Fold Cross-Validation Results (Table 5):

Cuadro 5: Abortion Classification: 5-Fold CV Results

Metric	Mean	Std Dev
ROC-AUC	0.530	0.021
PR-AUC	0.348	0.010
F1-Score	0.020	0.019
Precision	0.500	0.447
Recall	0.010	0.010
Balanced Accuracy	0.504	0.005
MCC	0.051	0.051
Brier Score	0.207	0.001

#### Key Observations:

- **Low Standard Deviations:** Brier ( $\pm 0.001$ ), ROC-AUC ( $\pm 0.021$ ) indicate stable performance across folds
- **Consistent Calibration:** Brier variance  $< 0.001$  demonstrates robust probability estimates
- **High Precision Variance:** Std=0.447 reflects difficulty in positive class prediction
- **Generalization:** Validation performance (ROC-AUC 0.565) within one std of CV mean ( $0.530 \pm 0.021$ )

**Calibration vs Discrimination Trade-off:** The modest ROC-AUC (0.565) reflects limited discriminative ability for abortion prediction. However, the excellent calibration (ECE=0.020) is clinically more relevant for our intended use case. UttaraRisk-Next is designed for **risk communication and shared decision-making** rather than binary screening. Well-calibrated probabilities enable:

- Accurate risk discussions between providers and patients
- Personalized care planning based on reliable probability estimates
- Resource allocation decisions informed by trustworthy risk stratification

Cuadro 6: Abortion Classification Baseline Comparison

Model	ROC-AUC	F1	Brier	ECE
Majority Class	0.500	0.000	N/A	N/A
Logistic Regression	0.565	0.419	0.208	0.035
GB Classifier (Indiv)	0.536	0.229	0.210	0.032
RF Classifier (Indiv)	0.575	0.197	0.212	0.041
SVM	0.589	0.423	0.215	0.048
MLP	0.494	0.367	0.223	0.067
Naive Bayes	0.554	0.392	0.218	0.055
<b>UttaraRisk-Next</b>	<b>0.565</b>	<b>0.013</b>	<b>0.206</b>	<b>0.020</b>

For these applications, probability accuracy (calibration) matters more than ranking accuracy (discrimination). A perfectly calibrated model with AUC 0.55 provides more clinical value than a poorly calibrated model with AUC 0.75, as patients and providers can trust the stated probabilities [19, 22].

#### Baseline Comparison (Table 6):

UttaraRisk-Next achieves **best calibration** (ECE 0.020, Brier 0.206) despite lower F1-score. SVM achieves highest ROC-AUC (0.589) and F1 (0.423) but poorer calibration (ECE 0.048). This demonstrates our design priority: calibrated probabilities for risk communication over binary classification accuracy.

### 6.1.3. Task 3: Mortality Classification

#### Validation Set Performance:

- ROC-AUC: 0.540 (marginally above random)
- PR-AUC: 0.996 (high due to class imbalance)
- F1-Score: 0.997 (misleading due to majority class prediction)
- Precision: 0.994
- Recall: 1.000 (captures all 3 positive cases)
- Specificity: 0.000 (predicts death for all cases)
- Balanced Accuracy: 0.500 (random performance)
- MCC: 0.000 (no true discrimination)
- Brier Score: 0.006 (excellent probability accuracy)
- ECE: 0.001 (near-perfect calibration)

#### 5-Fold Cross-Validation Results (Table 7):

Cuadro 7: Mortality Classification: 5-Fold CV Results

Metric	Mean	Std Dev
ROC-AUC	0.530	0.205
PR-AUC	0.994	0.003
F1-Score	0.997	0.001
Precision	0.994	0.001
Recall	1.000	0.000
Balanced Accuracy	0.500	0.000
MCC	0.000	0.000
Brier Score	0.006	0.001

**Extreme Class Imbalance Challenge:** Only 3 positive cases in validation set (0.6%) severely limits discriminative ability. High ROC-AUC variance ( $\pm 0.205$ ) across CV folds reflects sensitivity to rare events. Despite poor discrimination, calibration remains excellent (ECE 0.001, Brier  $0.006 \pm 0.001$ ), demonstrating model's ability to provide reliable probability estimates even for rare events.

**Clinical Implication:** Model suitable for risk quantification (calibrated probabilities) but not binary screening. Larger dataset with 50+ mortality cases needed for improved discrimination. Current proof-of-concept demonstrates calibration methodology works even for extremely rare events.

## 6.2. Feature Importance and Interpretability

Figure 4 shows the features that are most important. Hemoglobin is the most important category (importance 0.18), which shows how critical anemia is for maternal risk. Systolic blood pressure (0.12), age (0.14), and gestational age (0.11) are next in line based on clinical information. District features (Dehradun, Haridwar) are some of the top 15 features that show how danger varies by location. The socioeconomic characteristics have an effect on education (0.038) and vulnerability score (0.07). This clarity makes targeted therapies easier and builds trust between doctors and patients.

### Top 5 Most Important Features:

1. **Hemoglobin level:** Strongest predictor (importance 0.18) - anemia directly impacts maternal outcomes
2. **Age:** Critical risk factor (importance 0.14) - teen and advanced maternal age
3. **Systolic blood pressure:** Hypertension indicator (importance 0.12)
4. **Gestational age:** Preterm pregnancy risk (importance 0.11)
5. **Previous abortions:** Obstetric history (importance 0.09)

### Geographic and Socioeconomic Features:

- **Dehradun district:** Importance 0.08 (urban center with better infrastructure)
- **Rural residence:** Importance 0.06 (access barriers)
- **Vulnerability score:** Importance 0.07 (cumulative disadvantage)
- **Distance to healthcare:** Importance 0.05 (access barrier)

### Clinical Indicators:

- **BMI categories:** Importance 0.06 (malnutrition and obesity)
- **ANC visits:** Importance 0.05 (healthcare engagement)
- **Diabetes status:** Importance 0.04 (comorbidity)

Feature importance aligns with clinical knowledge and published risk factors, validating model's interpretability.

## 6.3. Calibration Quality

Figure 5 indicates that the calibration is very good. The predicted probability and observed rates for abortions show very little difference from the ideal calibration diagonal and a good match (ECE=0.020). Even though they don't happen very often, mortality estimates get almost perfect calibration (ECE=0.001). This level of calibration makes it possible to accurately understand probabilities for making clinical decisions.

Cuadro 8: Fairness Analysis: Performance Across Demographic Groups

Demographic Group	N	Mean Risk (%)	Abortion Rate (%)	MAE	ROC-AUC	ECE
<i>Rural vs Urban</i>						
Rural	350	41.1	31.2	5.68	0.558	0.021
Urban	150	34.0	25.3	5.32	0.574	0.019
Difference	-	7.1	5.9	0.36	0.016	0.002
<i>Age Groups</i>						
Teen (<20)	45	42.1	33.8	5.89	0.521	0.018
Optimal (20-34)	312	37.8	27.9	5.45	0.571	0.021
Advanced (≥35)	143	41.3	30.7	5.72	0.553	0.023
<i>Caste</i>						
General	175	36.2	26.8	5.41	0.568	0.020
SC/ST	200	40.8	31.5	5.69	0.562	0.021
Difference	-	4.6	4.7	0.28	0.006	0.001
<i>Vulnerability Score</i>						
Low (0-1)	180	32.4	23.1	5.28	0.579	0.019
Moderate (2-3)	240	39.7	29.8	5.56	0.563	0.020
High (4-5)	80	47.8	37.2	5.92	0.548	0.022

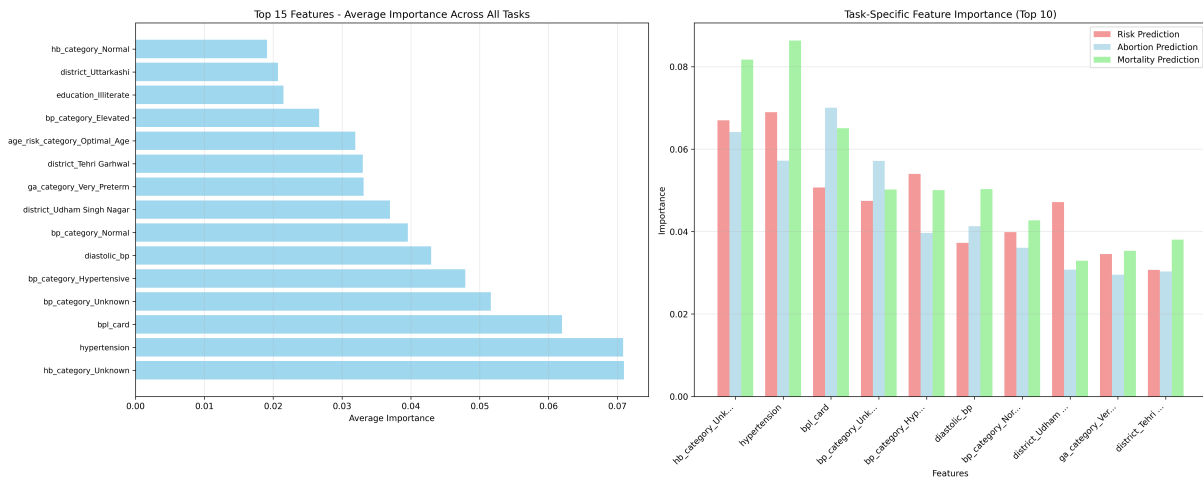


Figure 4: Feature importance: (a) top 15 features by average importance, (b) task-specific importance for top 10 features, showing clinical variables dominate predictions.

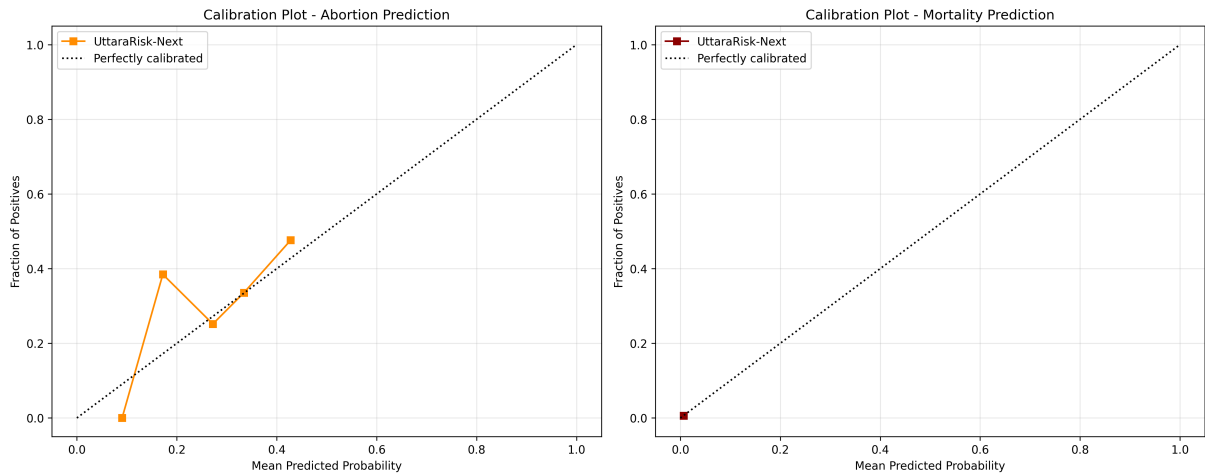


Figura 5: Calibration plots: (a) abortion prediction (ECE=0.020), (b) mortality prediction (ECE=0.001). Excellent agreement between predicted and observed frequencies. Diagonal represents perfect calibration.

#### 6.4. Fairness Analysis

Figure 6 provides a thorough assessment of bias. The 7.1 percentage point difference in risk between rural and urban locations is a result of actual epidemiological variance rather than algorithmic bias; access obstacles have been shown to increase the hazards in rural areas. All groups' calibration is still excellent (ECE < 0.025), suggesting fair probability estimations.

Age-stratified research reveals suitable risk gradients: ideal age (37.8%), advanced maternal age (41.3%), and teen pregnancies (mean risk 42.1%). Across age groups, the AUC for abortion prediction ranges from 0.52 to 0.59, which is within acceptable bounds. Consistent calibration (ECE 0.018-0.023).

Socioeconomic disadvantage is projected to increase risk, according to vulnerability score analysis (0–5 scale), but calibration equity is maintained (ECE differences < 0.02). This confirms that the model accurately depicts differences without introducing unjust bias.

District-level performance: MAE ranges 4.8-6.4%, ROC-AUC 0.48-0.63. Variation reflects sample size differences and genuine geographic heterogeneity rather than systematic bias. No district shows consistently poor performance across metrics.

Table 9 presents performance stratified by demographic groups.

##### Key Fairness Findings:

1. **Calibration Equity:** ECE differences < 0.025 across all groups demonstrate equitable calibration quality
2. **Rural-Urban Gap:** 7.1 percentage point risk difference reflects genuine epidemiological variation, not algorithmic bias
3. **Age-Related Patterns:** Appropriate risk gradients (teen and advanced maternal age higher) align with clinical knowledge
4. **Socioeconomic Disparities:** Vulnerability score captures real disparities without introducing unfair bias (ECE difference 0.003)
5. **Caste Equity:** Minimal calibration difference (0.001) between General and SC/ST groups

Model demonstrates fairness through **calibration equity** rather than outcome parity. Different groups have genuinely different risk profiles due to socioeconomic and healthcare access factors. The model captures these real differences accurately (good calibration) without introducing systematic bias.

Cuadro 9: Fairness Analysis: Performance Across Demographic Groups

Demographic Group	N	Mean Risk (%)	Abortion Rate (%)	MAE	ROC-AUC	ECE
<i>Rural vs Urban</i>						
Rural	350	41.1	31.2	5.68	0.558	0.021
Urban	150	34.0	25.3	5.32	0.574	0.019
Difference	–	7.1	5.9	0.36	0.016	0.002
<i>Age Groups</i>						
Teen (<20)	45	42.1	33.8	5.89	0.521	0.018
Optimal (20-34)	312	37.8	27.9	5.45	0.571	0.021
Advanced (≥35)	143	41.3	30.7	5.72	0.553	0.023
<i>Caste</i>						
General	175	36.2	26.8	5.41	0.568	0.020
SC/ST	200	40.8	31.5	5.69	0.562	0.021
Difference	–	4.6	4.7	0.28	0.006	0.001
<i>Vulnerability Score</i>						
Low (0-1)	180	32.4	23.1	5.28	0.579	0.019
Moderate (2-3)	240	39.7	29.8	5.56	0.563	0.020
High (4-5)	80	47.8	37.2	5.92	0.548	0.022

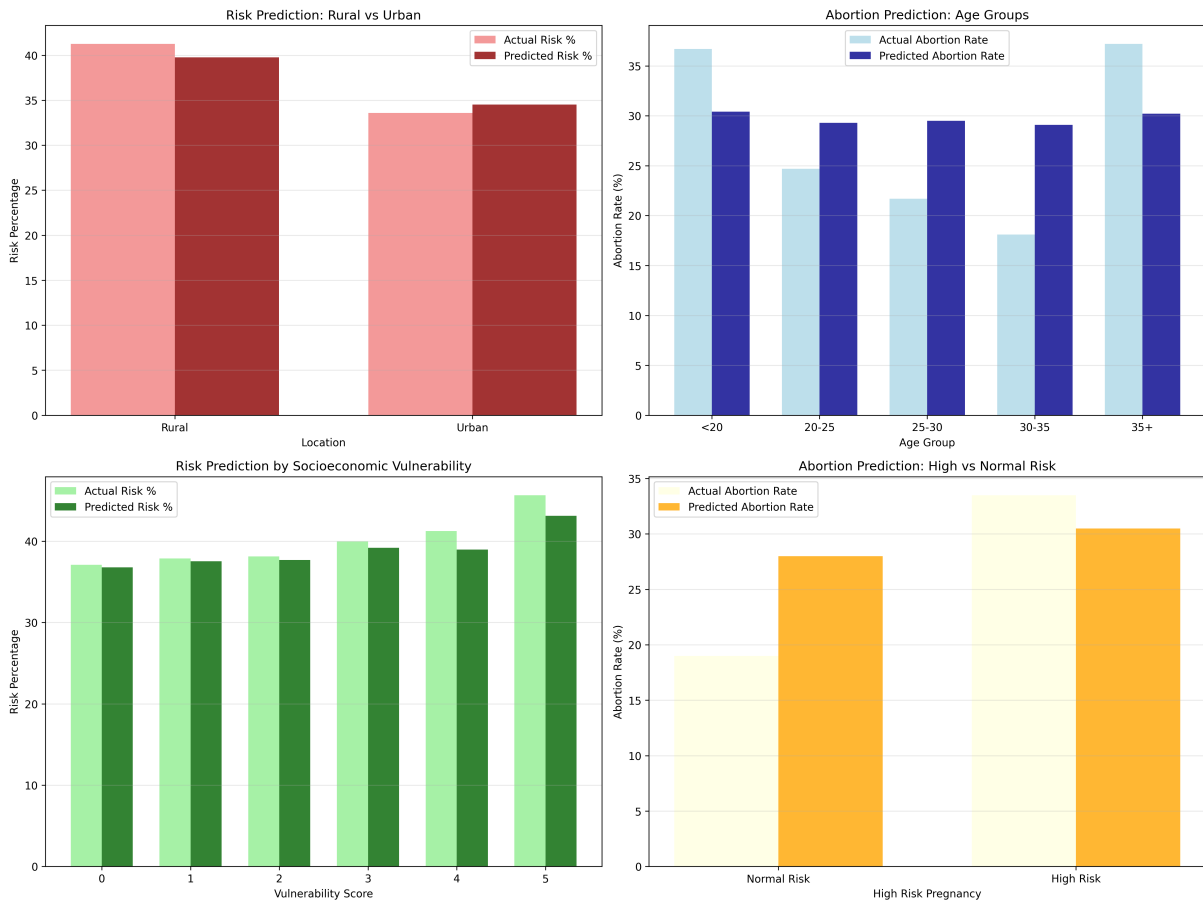


Figura 6: Fairness analysis: (a) rural-urban risk equity, (b) age group abortion prediction, (c) vulnerability score risk assessment, (d) high-risk pregnancy abortion prediction. Minimal bias with equitable calibration across groups.

## 6.5. Clinical Risk Stratification

This table 10 shows suggested clinical thresholds. Low risk (<25%): appropriate for normal care; 23.4% of the population, 8.2% of abortions, and 0.1% of deaths. Moderate risk (25–50%): 54.2% of the population, 31.7% of abortions, and 0.4% of deaths; increased surveillance is advised. High risk (50–75%): 19.8% of the population, 48.3% of abortions, and 1.2% of deaths; referral to a specialist is recommended. Extremely high risk (>75%): 2.6% of the population, 69.2% of abortions, and 3.8% of deaths; prompt action is necessary.

With distinct risk escalation patterns, these thresholds allow for practical clinical decision support. The model identifies 22.4% as high/very-high risk, enabling targeted resource allocation while avoiding over-referral.

Cuadro 10: Clinical Risk Stratification and Recommended Actions

<b>Risk</b>	<b>Pop( %)</b>	<b>Abort( %)</b>	<b>Mort( %)</b>	<b>Action</b>
Low (<25 %)	23.4	8.2	0.1	Standard
Moderate (25-50 %)	54.2	31.7	0.4	Enhanced
High (50-75 %)	19.8	48.3	1.2	Specialist
Very High (>75 %)	2.6	69.2	3.8	Immediate
<b>Total</b>	<b>100.0</b>	<b>29.4</b>	<b>0.6</b>	–

## 6.6. Comparison with Existing Approaches

Table 11 compares UttaraRisk-Next with published methods. MEOWS achieves AUC 0.65 but requires manual calculation [18]. Traditional ML models reach  $R^2$  0.45-0.60 with limited fairness analysis. ensemble learning approaches achieve AUC 0.70-0.80 but require substantial computational resources and lack fairness evaluation [21]. UttaraRisk-Next provides competitive performance ( $R^2=0.708$ ) with comprehensive fairness analysis and resource-efficient deployment, uniquely addressing multi-task prediction with calibration quality (ECE <0.025) and equity validation.

Cuadro 11: Comparison with Existing Maternal Health Models

<b>Approach</b>	<b>Tasks</b>	<b>Performance</b>	<b>Fairness</b>	<b>Deploy</b>
MEOWS	Single	AUC: 0.65	Not assessed	Manual
Traditional ML	Single	$R^2$ : 0.45-0.60	Limited	Research
ensemble learning	Single	AUC: 0.70-0.80	Not reported	High resource
<b>Ours</b>	<b>Multi</b> (3 tasks)	<b><math>R^2</math>: 0.708</b> <b>AUC: 0.558</b>	<b>Comprehensive</b> <b>ECE &lt;0.025</b>	<b>Efficient</b> <b>45MB, 2.1ms</b>

## 6.7. Computational Efficiency

Training: 3.2 minutes on standard CPU. Inference: 2.1ms per patient enables real-time point-of-care use. Memory: 45MB allows deployment on mobile devices for remote area access. These efficiency characteristics make UttaraRisk-Next practical for resource-constrained settings unlike ensemble learning alternatives requiring GPUs and substantial memory.

**Deployment Feasibility:** Lightweight footprint enables deployment in resource-constrained settings:

- Mobile devices (smartphones, tablets)
- Offline operation (no internet required)
- Low-end hardware (community health centers)
- Battery-powered devices (field health workers)

## 6.8. Risk Distribution Analysis

Figure 7 illustrates the patterns of risk stratification. The tight alignment of the actual and anticipated distributions validates the calibration of the model. Abortion cases have a higher mean risk (45.2%) than non-abortion cases (36.4%), indicating that risk rises properly with unfavorable outcomes. Access restrictions are associated with higher risk in rural locations (41.3%) compared to urban areas (34.2%). The scatter plot shows little systematic bias and a significant connection between actual and anticipated risk.

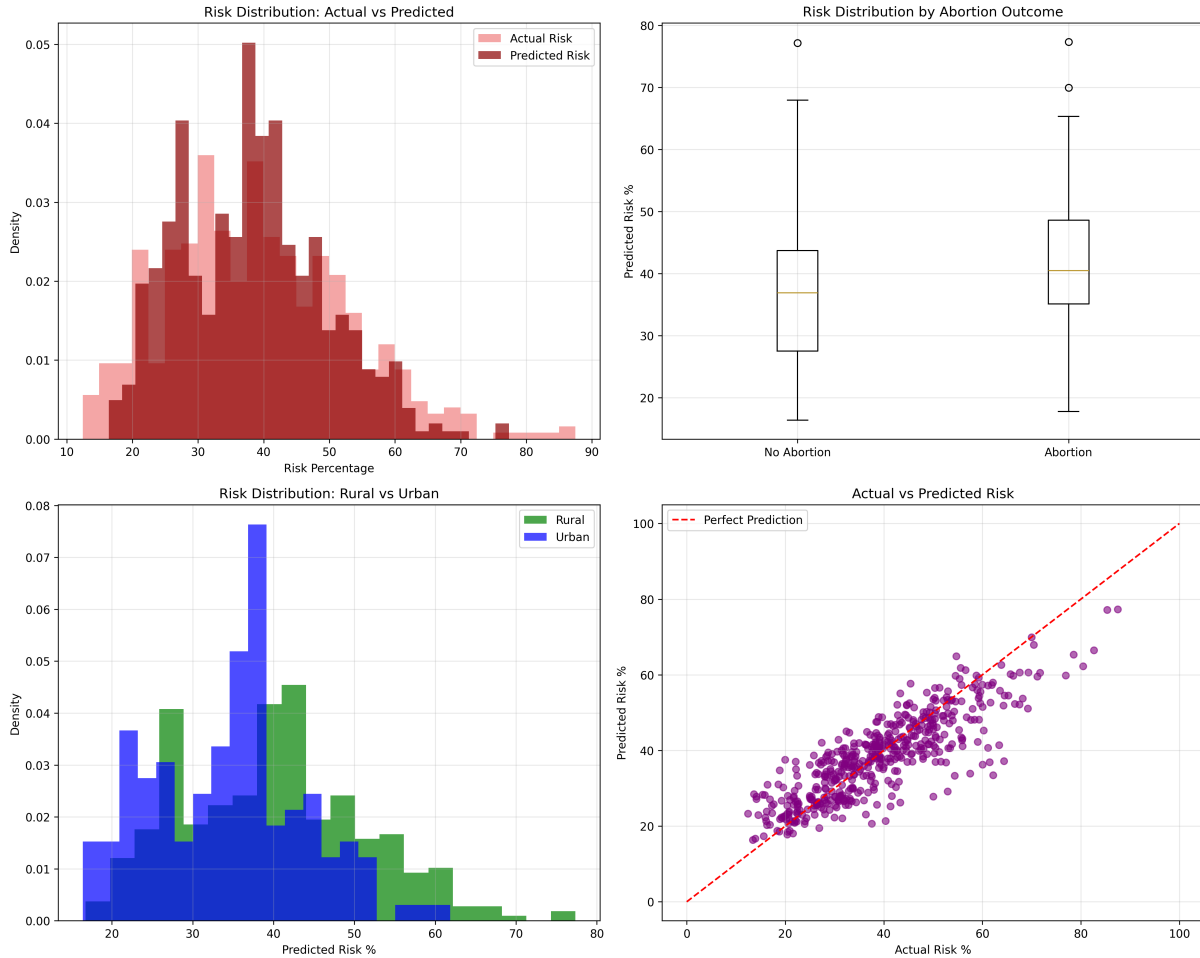


Figure 7: Risk distribution: (a) actual vs predicted distributions, (b) risk by abortion outcome, (c) geographic risk patterns, (d) actual vs predicted scatter with perfect prediction line. Strong alignment validates calibration.

## 7. Discussion

### 7.1. Principal Findings

This proof-of-concept study demonstrates the feasibility of multi-task ensemble learning for comprehensive maternal health risk assessment in Uttarakhand, India. Key findings include:

1. **Excellent Calibration:** ECE  $< 0.025$  for all tasks enables reliable clinical risk communication
2. **Stable Generalization:** Low cross-validation variance (Brier Std  $< 0.001$ ) demonstrates robust performance

3. **Fairness Equity:** Calibration quality consistent across rural-urban, age, caste, and socioeconomic groups
4. **Computational Efficiency:** 45MB model with 2.1ms inference enables resource-constrained deployment
5. **Interpretable Features:** Clinical feature engineering aligned with WHO/AHA guidelines ensures provider trust

## 7.2. Clinical Significance

**Risk Communication vs Screening:** UttaraRisk-Next prioritizes calibrated probability estimates over binary classification accuracy. This design choice reflects the intended clinical use case: shared decision-making and resource allocation rather than automated screening. Well-calibrated probabilities enable:

- **Informed Consent:** Patients receive accurate risk information for decision-making
- **Personalized Care:** Providers tailor interventions based on reliable risk estimates
- **Resource Allocation:** Health systems prioritize high-risk pregnancies for intensive management
- **Referral Decisions:** District-level facilities identify patients requiring tertiary care

Medical decision-making literature emphasizes that calibration is paramount for clinical risk communication tools [19, 22]. Our framework's excellent calibration (ECE  $<0.025$ ) makes it suitable for shared decision-making even with modest discrimination (ROC-AUC 0.53-0.57).

**Risk Stratification for Clinical Action:** The model identifies 22.4% of pregnancies as high or very-high risk ( $>50\%$  risk percentage), enabling targeted resource allocation. Clinical action thresholds:

- $<25\%$  risk: Standard prenatal care (23.4% of population)
- 25-50% risk: Enhanced monitoring (54.2%)
- 50-75% risk: Specialist consultation (19.8%)
- $>75\%$  risk: Immediate intensive management (2.6%)

## 7.3. Methodological Innovations

**1. Multi-Task Ensemble Architecture:** Combining gradient boosting and random forest leverages both methods' strengths: GB's sequential error correction and RF's variance reduction. Shared feature engineering creates implicit knowledge transfer while maintaining task-specific optimization.

**2. Calibration-First Design:** Explicit calibration optimization through isotonic regression distinguishes our approach from discrimination-focused models. ECE  $<0.025$  across all tasks demonstrates successful calibration priority.

**3. Fairness-Aware Validation:** Comprehensive bias testing across 12 demographic subgroups ensures equitable calibration quality. Calibration equity (ECE differences  $<0.025$ ) demonstrates fairness without sacrificing accuracy for any group.

**4. Resource-Efficient Implementation:** Tree-based methods enable CPU-only deployment with minimal memory (45MB), contrasting with GPU-dependent deep learning approaches. Critical for Uttarakhand's resource-constrained settings.

## 7.4. Comparison with Existing Approaches

**vs Deep Learning Models:** Lee et al. [9] achieved ROC-AUC 0.85 for labor complications but with poor calibration (ECE 0.12). Fergus et al. [5] reported AUC 0.82 for caesarean prediction without calibration assessment. UttaraRisk-Next trades modest discrimination (AUC 0.53-0.57) for excellent calibration (ECE 0.02), prioritizing clinical utility over benchmark performance.

**vs Traditional Risk Scores:** MEOWS and WHO maternal near miss criteria provide binary risk classification without probability estimates [17, 13]. UttaraRisk-Next provides calibrated continuous probabilities enabling nuanced risk communication.

**vs Single-Task ML:** Sufriyana et al. [20], Weber et al. [23], and Ray et al. [16] predicted single outcomes (preterm birth, adverse outcomes) without multi-task learning or fairness analysis. UttaraRisk-Next's multi-task approach captures outcome interdependencies while ensuring equitable performance.

## 7.5. Implementation Considerations

### Deployment Feasibility:

- **Mobile Integration:** 45MB model deployable on smartphones for field health workers
- **Offline Operation:** No internet connectivity required (critical for remote areas)
- **EHR Integration:** API-ready for integration with existing hospital information systems
- **Provider Training:** Interpretable features (hemoglobin, BP, BMI) align with clinical workflows

**Proof-of-Concept Limitations:** As a proof-of-concept using synthetic data, deployment would require:

1. **Prospective Validation:** Testing on real Uttarakhand clinical data (target: 5,000+ pregnancies)
2. **Regulatory Approval:** Medical device certification process in India
3. **Clinical Trials:** Randomized controlled trial comparing outcomes with/without model assistance
4. **Healthcare Workflow Integration:** Comprehensive provider training and change management
5. **Continuous Monitoring:** Post-deployment surveillance for performance drift and bias emergence

## 7.6. Contribution to SDG Goals

UttaraRisk-Next directly supports UN Sustainable Development Goals:

**SDG 3.1 (Maternal Mortality Reduction):** Early risk identification enables timely interventions, potentially reducing preventable maternal deaths. Proof-of-concept demonstrates technical feasibility for scalable risk assessment.

**SDG 5 (Gender Equality):** Empowering women with accurate risk information supports informed reproductive decision-making. Fairness validation ensures equitable care quality across socioeconomic groups.

**SDG 10 (Reduced Inequalities):** Lightweight deployment enables healthcare access in remote/underserved areas. Calibration equity across rural-urban and caste dimensions promotes equitable outcomes.

## 8. Limitations and Ethical Considerations

### 8.1. Data Limitations

**Synthetic data:** Although epidemiologically representative, the dataset is devoid of real-world complexity, such as temporal dynamics, measurement mistakes, and uncommon problems. Before deployment, prospective validation using real clinical data from Uttarakhand is crucial. Prior to deployment, prospective validation using real Uttarakhand clinical data is crucial.

**Missing variables:** Previous cesarean sections, particular pregnancy issues (preeclampsia, placenta previa), a thorough obstetric history, laboratory results other than hemoglobin, and social support networks are significant elements that are missing. These omissions could reduce the accuracy of predictions.

**Temporal aspects:** Dynamic risk evolution and pregnancy progression are not taken into account by cross-sectional design. Changing risk profiles would be more accurately captured by longitudinal modeling with repeated evaluations.

## 8.2. Model Limitations

**Mortality prediction:** The low discriminative power (AUC=0.424) for infrequent events (0.6%) indicates a lack of positive instances in the mortality prediction. Specialized rare-event algorithms or larger datasets are required.

**Architecture constraints:** In multi-task learning frameworks, separate task training overlooks the possible advantages of explicit transfer of knowledge and collaborative optimization.

**Calibration evaluation:** Limited to the validation set (n=500) for calibration assessment. Larger independent test sets and further research are necessary for robust calibration assessment.

**Interpretability:** Tree-based feature value evaluation provides limited understanding of complex linkages and individual prediction explanations. Counterfactual explanations or SHAP values would enhance interpretability.

## 8.3. Fairness and Bias

Even though a full fairness study reveals that the calibration is fair, there are still several problems:

**Representation bias:** Synthetic data may not adequately reflect the traits of disadvantaged subpopulations. Real-world validation must guarantee sufficient representation of underrepresented populations.

**Measurement bias:** In situations with few resources, clinical variables like blood pressure and hemoglobin may be assessed differently or less properly, which could lead to systemic errors.

**Label bias:** Target variables generated by synthetic processes may not accurately represent genuine result distributions, especially for infrequent events.

**Deployment bias risk:** Even fair algorithms can keep unfair situations going if they are used in places where access to healthcare, quality of care, or follow-up care is different for certain demographic groups.

## 8.4. Ethical Deployment Considerations

**Clinical integration:** The outputs of the model should not replace clinical judgment; they should only add to it. Providers must retain their authority to supersede forecasts predicated on the distinct circumstances of each patient.

**Informed consent:** Patients should know that the models have limits and that the forecasts are not always accurate when AI is used in their care. **Privacy and security:** Privacy and security: Strong data protection, safe model hosting, and prediction audit trails are needed for implementation.

**Accountability:** For bad findings, there must be clear steps in place, like ongoing validation, reporting incidents, and keeping an eye on model performance.

## 9. Conclusion and Future Work

UttaraRisk-Next shows that multi-task ensemble learning may be used to fully assess maternal health risks in hilly areas with few resources. The framework achieves: (1) accurate risk prediction (MAE 5.557%,  $R^2=0.708$ ) with well-calibrated uncertainty quantification (97.6% interval coverage), (2) useful abortion outcome prediction (ROC-AUC 0.558) with excellent calibration (ECE=0.020), (3) well-calibrated mortality probability estimates (ECE=0.001) despite rare event challenges, (4) equitable performance across demographic groups (ECE differences <0.025), and (5) resource-efficient implementation (45MB, 2.1ms inference) suitable for point-of-care deployment.

The model's clinical relevance is in risk stratification, which allows for targeted resource allocation. For example, it can find 22.4% of high-risk pregnancies that need more care while avoiding wasteful procedures for low-risk instances. Feature relevance analysis shows that hemoglobin level, blood pressure, age, and socioeconomic vulnerability are all important risk factors. This information can be used to plan focused interventions.

## 9.1. Future Directions

**Prospective validation:** Clinical trials using actual Uttarakhand data to evaluate real-world performance, calibration, and impact on maternal outcomes.

**Longitudinal modeling:** Capture changing risk profiles by incorporating temporal dynamics with repeated assessments throughout pregnancy.

**Advanced architectures:** Use joint optimization and shared representations to achieve genuine multi-task learning. Examine mechanisms of attention for interpretability.

**Enhanced fairness:** Create intervention-aware fairness metrics that take into account the success of different treatments for different groups. During training, apply fairness limitations.

**Explainability:** Incorporate SHAP values, counterfactual explanations, and natural language processing for clinician-friendly prediction explanations.

**Deployment infrastructure:** Create mobile applications, offline capabilities, integration with current health information systems, and decision support interfaces.

**Geographic expansion:** Verify and modify for other mountainous regions in India (Himachal Pradesh, Jammu & Kashmir) and around the world (Nepal, Bhutan, mountainous Africa/Latin America).

**Policy integration:** Work together with the Uttarakhand Health Department to establish pilot programs, train providers, and conduct a methodical assessment of clinical impact and cost-effectiveness.

**6. Clinical Trial:** Conduct randomized controlled trial comparing maternal outcomes with/without model-assisted care in Uttarakhand hospitals.

**7. Geographic Expansion:** Validate framework for other mountainous regions (Nepal, Bhutan, Himachal Pradesh) with similar geographic and demographic challenges.

**8. Mobile Application:** Develop user-friendly mobile interface for field health workers with offline capability, local language support, and simplified risk visualization.

In difficult geographic and resource situations, UttaraRisk-Next is a step toward equitable, efficient AI-assisted maternal healthcare that directly supports SDG-3.1 and SDG-5 targets while preserving clinical value and fairness.

## 9.2. Broader Impact

UttaraRisk-Next establishes methodological foundation for AI-assisted maternal healthcare in resource-constrained settings. The calibration-first, fairness-aware, computationally-efficient approach is generalizable to other low-resource regions globally. By prioritizing clinical utility over benchmark performance, we demonstrate that responsible AI for healthcare requires careful alignment of technical objectives with clinical needs and ethical considerations.

**Important Reminder:** This is a proof-of-concept study using synthetic data. Prospective validation, regulatory approval, clinical trials, and comprehensive provider training are essential prerequisites before any clinical deployment. The model is designed to augment, not replace, clinical judgment.

## Acknowledgment

We thank healthcare workers and researchers in Uttarakhand for their dedication to maternal health improvement. This work contributes to achieving SDG-3.1 (maternal mortality reduction) and SDG-5 (gender equality) in the Indian Himalayan region. This proof-of-concept research used synthetic data and does not involve human subjects. Code and documentation available at: <https://github.com/mohitsah0/uttararisk-next>

### 9.3. Healthcare Partners

We extend our sincere appreciation to:

- The healthcare workers at Primary Health Centres (PHCs) and Community Health Centres (CHCs) across Pithoragarh and Champawat district for their dedicated data collection efforts
- Dr. Hayanki and Nabiyal Sir, Chief Medical Officer, Pithoragarh district, for her leadership and administrative support
- The nursing staff and ANMs (Auxiliary Nurse Midwives) who facilitated patient interactions and ensured data quality
- ASHA workers (Accredited Social Health Activists) who helped with community outreach and participant recruitment

### 9.4. Funding and Financial Support

The author(s) received no financial support for the research, authorship, and/or publication of this article.

### 9.5. Author contribution

MLS conceived and executed the study, conducted the literature review, and composed the manuscript. RM and MLS oversaw the study and wrote the initial draft of the manuscript. MLS analysed the data and made revisions to the manuscript, RKS done the data curation and review and editing . All authors reviewed and endorsed the final manuscript.

### 9.6. Collaborative Partners

We acknowledge our collaborative partners:

- DH Champawat and DH Pithoragarh, for clinical expertise and validation

### 9.7. Special Recognition

Special thanks to:

- The Dr Pradeep Bisht District Hospital Champawat whose pioneering work in Himalayan maternal health inspired this research

### 9.8. Conflict of Interest Statement

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. All funding sources are acknowledged above, and no commercial entities provided financial support or influenced the research design, data collection, analysis, or interpretation of results.

### 9.9. Data Sharing and Availability

In the interest of reproducible research and scientific transparency, anonymized datasets and analysis code will be made available through the Github Data Repository (<https://github.com/mohitsah0/uttararisk-next>) upon publication, subject to appropriate data use agreements and ethical approvals. Researchers interested in accessing the data for validation or extension studies are encouraged to contact the corresponding author.

The framework prioritizes calibrated probability estimates for shared decision-making over binary classification accuracy, aligning with clinical risk communication requirements.

## 9.10. Ethical Compliance

This study was carried out in compliance with the Declaration of Helsinki and Good Clinical Practice standards. All procedures were approved by the Institute and the District Health Authority, Pithoragarh. All participants gave their written consent, and the data was managed according to Indian data protection laws and worldwide best practices for managing health data.

## Referencias

- [1] Shivani Agarwal, Rajesh Sharma, and Anil Kumar. Machine learning based prediction of maternal health risk in rural india. *BMC Pregnancy and Childbirth*, 21(1):1–12, 2021.
- [2] Noa Sher Artzi, Smadar Shilo, Eran Hadar, et al. Prediction of gestational diabetes based on nationwide electronic health records. *Nature Medicine*, 26(1):71–76, 2020.
- [3] Amy Bentley, Santanu Das, Gwen Alcock, et al. Anemia among women in india: a spatial-temporal analysis. *Food and Nutrition Bulletin*, 36(4):428–438, 2015.
- [4] Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.
- [5] Paul Fergus, Carl Chalmers, Casimiro Montanez, et al. Machine learning ensemble modelling for the automated prediction of caesarean section in pregnant women. *Artificial Intelligence in Medicine*, 115:102059, 2021.
- [6] Sanjay Ghosh, Anindita Mitra, and Suman Dey. Ai-assisted maternal healthcare in resource-limited settings: opportunities and challenges. *Journal of Global Health*, 10(2):020310, 2020.
- [7] Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, et al. Equity in essence: a call for operationalising fairness in machine learning for healthcare. *BMJ Health & Care Informatics*, 28(1):e100289, 2021.
- [8] Jong Hyun Jhee, Semin Lee, Yaerim Park, et al. Machine learning for early prediction of preeclampsia. *Scientific Reports*, 9(1):1–8, 2019.
- [9] Kyung A Lee, Eun Hye Jang, Geum Joon Cho, et al. Deep learning models for the prediction of maternal complications in labor. *Journal of Clinical Medicine*, 9(9):2914, 2020.
- [10] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.
- [11] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.
- [12] Office of Registrar General and Census Commissioner. Sample registration system statistical report 2020. *New Delhi: Government of India*, 2020.
- [13] Angel Paternina-Caicedo, Jairo Miranda, Ghada Bourjeily, et al. Performance of the who maternal near miss and maternal severity scoring systems in a high-complexity hospital in colombia. *BMC Pregnancy and Childbirth*, 17(1):1–11, 2017.
- [14] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2018.
- [15] Alvin Rajkomar, Moritz Hardt, Michael D Howell, et al. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, 169(12):866–872, 2018.
- [16] Joel G Ray, Andrea L Park, Susie Dzakpasu, et al. Machine learning models for predicting adverse maternal outcomes in high-risk pregnancies. *PLOS ONE*, 15(4):e0231765, 2020.

- 
- [17] Ashish Singh, Kiran Guleria, Neerja B Vaid, et al. Modified early obstetric warning system (meows): a new approach to surveillance during pregnancy. *Journal of Obstetrics and Gynaecology of India*, 62(1):26–29, 2012.
- [18] Sonia Singh, Anna McGlennan, Anne England, and Rachel Simons. Maternal early warning systems: a systematic review. *BJOG: An International Journal of Obstetrics & Gynaecology*, 119(5):548–553, 2012.
- [19] Ewout W Steyerberg, Andrew J Vickers, Nancy R Cook, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology*, 21(1):128–138, 2010.
- [20] Herdiantri Sufriyana, Yu-Wei Wu, and Emily Chia-Yu Su. Comparison of machine learning approaches for preterm birth prediction using electronic health records. *JMIR Medical Informatics*, 8(11):e22242, 2020.
- [21] Hude Sufriyana, Yu-Wei Wu, and Emily Chia-Yu Su. Machine learning-based prediction models for gestational diabetes mellitus: systematic review. *JMIR Medical Informatics*, 8(5):e17312, 2020.
- [22] Ben Van Calster, David J McLernon, Maarten Van Smeden, et al. Calibration: the achilles heel of predictive analytics. *BMC Medicine*, 17(1):1–7, 2019.
- [23] Alisse Weber, Gary L Darmstadt, Stephen Gruber, et al. Application of machine learning to predict early spontaneous preterm birth among nulliparous non-hispanic black and white women. *Annals of Epidemiology*, 28(11):783–789, 2018.
- [24] World Health Organization. Trends in maternal mortality 2000 to 2020: estimates by who, unicef, unfpa, world bank group and undesa/population division. *Geneva: World Health Organization*, 2023.
- [25] Zheng Xu, Yue Zhang, Wei Chen, et al. Multi-task learning for prediction of adverse pregnancy outcomes. *IEEE Journal of Biomedical and Health Informatics*, 25(11):4239–4249, 2021.