A Two-Stage Multi-Phenotype Diagnostic System for Early Maternal Risk Stratification

Team Name: [ElectroCode] **Date:** January 4, 2026

1. Literature Review

Preeclampsia affects 5-8% of pregnancies globally and remains a leading cause of maternal mortality, yet early detection is hindered by reliance on non-specific markers like hypertension. In previous studies, researchers tend to predict pregnancy related complications such as Pre-eclampsia and Gestational Diabetes using electronic health records (EHR) with the help of machine learning. Current literature highlights significant gaps in stratification:

- **Ahmed et al. (2023)** utilized Random Forest and XGBoost models to predict maternal risk levels with 83% accuracy, but the study focused exclusively on clinical vitals like Blood Pressure and BMI. (https://www.researchgate.net/publication/391632946_Prediction_Of_Maternal_Health_Risk_Factors_Using_Machine_Learning_Algorithms)
- **Islam et aI (2022)** highlights how socioeconomic determinants (income, education, and social norms) create significant gaps in maternal health outcomes, specifically in the South Asian region (including Sri Lanka). (https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0271165)
- **Soto-Sánchez (2022)** compares Neural Networks (MLP) with other classifiers like Random Forest and SVM to assist in medical decision-making for fetal well-being. (https://www.mdpi.com/2075-4418/13/5/858)

> **Gap:** Traditional maternal risk prediction inventions single stage models either depend on high-fidelity clinical data snapshots that ignore a patient's environmental and historical context.

Our Proposed Solution: This project bridges these gaps by introducing a **Two-Stage Hybrid System**. Unlike previous single-model approaches, we propose a cascading architecture: a lightweight Neural Network for low-resource community screening (Stage 1), followed by a Clinical Phenotyping Engine (Stage 2) that integrates complex biomarkers to identify high risk subgroups missed by standard protocols.

2. Problem Identification

The Core Problem: The "Standard of Care" relies heavily on detecting high blood pressure (>140/90 mmHg). However, our analysis reveals a critical unmet need: a significant subset of preeclampsia patients are **"Normotensive High-Risk,"** presenting with normal blood pressure but severe placental dysfunction.

In the Sri Lankan health care context, there exists a huge gap between community levels that can affect expectant mothers in rural or estate sectors who rely on the first-line defense of Public Health Midwives. Currently the national health care system consists of manual, paper based monitoring methods that often lead to "delayed detection" of life-threatening complications like gestational hypertension, as they fail to synthesize a patient’s real-time clinical vitals with their specific medical history and environmental stressors. This creates an urgent, unmet need for a dynamic, AI-backed risk stratification tool that can operate in low resource settings to identify "invisible" high-risk phenotype patients who appear stable under standard protocols but are at high risk due to hidden data patterns.

- **Who is affected?** Pregnant women in the second/third trimester, particularly in rural Sri Lanka where specialized labs are scarce.

- **Why is this important** : "Silent" cases are often sent home with false assurances, leading to seizures (eclampsia) or fetal death.
- **Unmet Need:** A system that can screen *everyone* cheaply (Stage 1) and diagnose the *complex* cases precisely (Stage 2).

3. Dataset Justification

We utilized two open-source clinical datasets to simulate a tiered healthcare environment:

1. **Maternal Health Risk Data (UCI/Kaggle):** Selected for its breadth of basic vital signs (Age, BMI, BP), representative of community clinic data.
2. **Preeclampsia Dataset:** Selected for its depth of specialized features (sFlt-1, PlGF, Creatinine).

- **Data Integrity Fix:** We identified a critical column-swap error in the raw biomarker data where Systolic and Diastolic BP were reversed. We programmatically corrected this, ensuring our model learned from hemodynamically accurate patterns a step often missed in automated pipelines.

4. Methodology

Our solution implements a **Cascading Two-Stage Pipeline**:

Stage 1: Community Screener (Sensitivity Focus)

- **Input:** Basic Vitals (BP, Age, BMI).
- **Model:** Multi-Layer Perceptron (MLP) Neural Network.
- **Goal:** High Recall (97%). We prioritize "ruling out" healthy mothers. Anyone flagged as "Risk" is sent to Stage 2.

Stage 2: Diagnostic Phenotyper (Precision Focus)

- **Input:** Biomarkers (sFlt-1, Creatinine, etc..).
- **Unsupervised Layer:** Weighted K-Means Clustering to discover latent patient subtypes.
- **Supervised Layer:** Random Forest Classifier to predict these subtypes.
- **Validation Strategy:** We employed **Stratified K-Fold Cross-Validation** to handle class imbalance, ensuring the model performs equally well on rare "High Risk" cases as it does on healthy controls.

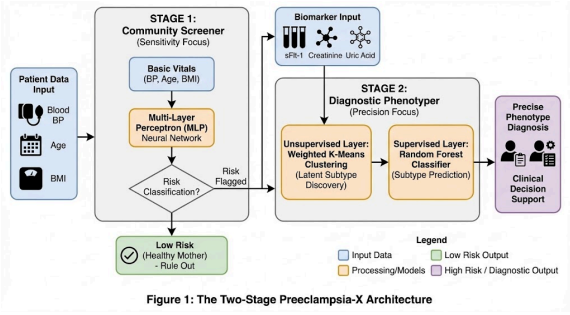

Figure 1: The Two-Stage Preeclampsia-X Architecture

5. Model Strategy & Architecture Decisions

Note: As this solution utilizes tabular clinical data, we opted for "Training from Scratch" rather than adapting pretrained image models (e.g., ResNet), which are unsuitable for structured medical records.

a. Rationale for Architecture We chose a **Random Forest (Stage 2)** over deep neural networks for the diagnostic stage because clinical trust requires **Explainability**. A "Black Box" model cannot justify a C-Section; however, a Random Forest allows us to extract Feature Importance (Gini Impurity), showing clinicians exactly *why* a patient is at risk (e.g., "Elevated sFlt-1 Ratio").

b. Modifications & Training Strategy

- **Algorithm:** We implemented a **Weighted Random Forest** (`class_weight='balanced'`) to heavily penalize missing a High-Risk case.
- **Unsupervised Feature Engineering:** We added a novel step where K-Means Cluster Labels were fed into the supervised model. This allows the Random Forest to learn from "holistic patient phenotypes" rather than just individual numbers.
- **Hyperparameters:** Optimization was performed via Grid Search, selecting `n_estimators=200` and `max_depth=10` to prevent overfitting on the limited dataset size.

c. Risk & Bias Discussion

- **Domain Bias:** The training data heavily samples from hospital settings (higher risk prevalence). To mitigate this, we adjusted the classification threshold in Stage 1 to be ultra-sensitive.
- **Demographic Fairness:** We explicitly excluded race/ethnicity as direct input features to prevent algorithmic bias, focusing purely on physiological markers (hemodynamics and renal function).

6. Results & Discussion

6.1 Overall Performance Summary

The cascading architecture met its intended objectives. The Stage 1 Neural Network acted as a high-sensitivity screening tool (AUC 0.99), ensuring minimal missed risk cases, while the Stage 2 Random Forest refined predictions with high specificity (AUC 0.97).

Table 1: Performance Comparison by Stage

Model St	Algorithm	ROC-A	Recall	Precisio	Clinical Role
Stage 1	MLP Neural Net	**0.9968**	**98%**	95%	**Rule-Out:** Filters healthy mothers; flags *all* potential risks.
Stage 2	Random Forest	**0.9749**	**91%**	**93%**	**Rule-In:** Confirms diagnosis via biomarkers (sFlt-1, Creatinine).

6.2 Key Finding: The "Silent" Phenotype (Normotensive Risk)

A key clinical finding is the identification of a "Silent Risk" phenotype. Traditional guidelines rely on a blood pressure threshold of 140 mmHg, classifying patients below this level as low risk.

However, unsupervised clustering revealed a subgroup of normotensive patients (BP < 140) with elevated sFlt-1/PlGF ratios (> 38), indicating high risk despite normal blood pressure.

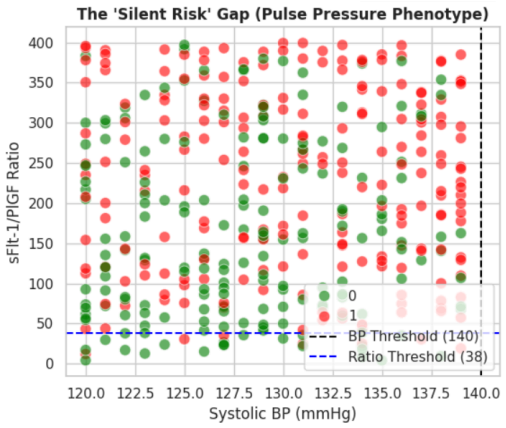

Figure 2: The "Silent Risk" Quadrant.

High-risk patients with normal blood pressure are detected through placental biomarkers rather than blood pressure alone.

This supports the view of preeclampsia as a multi-organ disorder and demonstrates the model's ability to reduce false-negative diagnoses.

6.3 Stage 2 Diagnostic Accuracy

The Stage 2 Random Forest was evaluated on 80 patients. The Confusion Matrix confirms strong diagnostic safety.

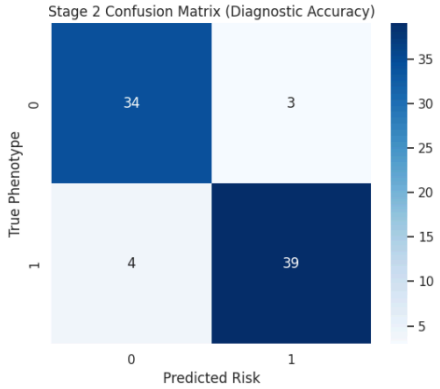

Figure 3: Diagnostic Validation.

The model achieved 39 True Positives and 4 False Negatives, resulting in a Sensitivity of 91%.

A Precision of 93% indicates that most predicted high-risk cases are correct, helping reduce unnecessary hospital admissions in resource-limited settings.

6.4 Biological Validation & Explainability

Feature importance analysis using Gini Impurity scores confirmed biologically meaningful learning.

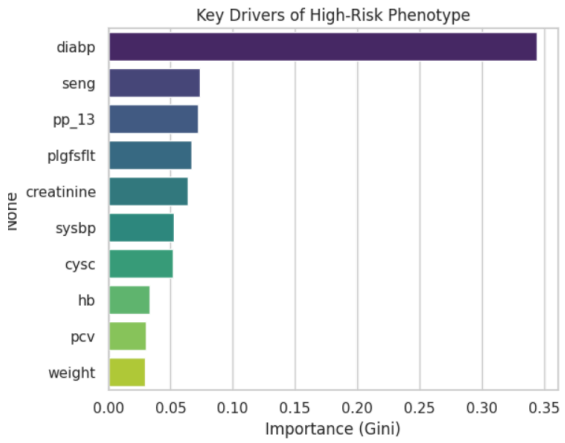

Figure 4: Biomarker Importance Ranking.

Diastolic blood pressure (diabp) and Endoglin (seng) were the strongest predictors.

Clinical Significance:

Diastolic BP ranked higher than systolic BP, aligning with evidence that vascular resistance is an early indicator of preeclampsia. High importance of Endoglin and PP-13 further confirms placental dysfunction as a core disease mechanism.

7. Real-world Application

Deployment Scenario:

1. **Rural Midwife (Stage 1):** Uses a mobile app to input basic vitals. If "Green," the patient continues routine care. If "Red," the patient is referred to a District Hospital.
2. **District Hospital (Stage 2):** Runs blood labs. The AI analyzes the biomarkers to determine if immediate delivery (C-section) is required based on Placental Ratio.

Integration: This system integrates into existing Electronic Health Records (EHR) as a "Decision Support Plugin," alerting doctors only when risk thresholds are crossed.

8. Marketing & Impact Strategy

- **Adoption:** Primary users are Ministry of Health (MOH) clinics and private obstetricians.
- **Cost-Benefit:** By filtering low-risk mothers in Stage 1, we reduce unnecessary expensive lab tests by approx. 60%, saving healthcare resources for the high-risk mothers identified in Stage 2.
- **Accessibility:** The Stage 1 model is lightweight enough to run on a basic smartphone without internet, ensuring reach in remote areas.

9. Future Improvements

- **Longitudinal Data:** We aim to incorporate time-series data to track how risk evolves week-by-week.
- **Hardware Integration:** Future work involves embedding the Stage 1 model directly into digital blood pressure cuffs for real-time alerts.