



Development of an Explainable AI-Driven Early Warning System for Predicting Postpartum Haemorrhage in Low-Resource Environments

Bamikole Joseph OLOJIDO¹, AWORETAN Fayowole², OJAJUNI Oluwatosin James³

^{1,2,3}Department of Computer Science, Rufus Giwa Polytechnic Owo, Ondo State, Nigeria, West Africa and Faculty of Nursing, Achievers University Owo, Ondo State, Nigeria, West Africa

ARTICLE INFO	ABSTRACT
Published Online: 03 September 2026 Corresponding Author: Bamikole Joseph OLOJIDO	Postpartum haemorrhage (PPH) is a leading cause of maternal morbidity and mortality worldwide, with a disproportionate burden in low-resource settings where timely diagnosis is often difficult. This study presents an interpretable AI-based early warning framework for predicting PPH using machine learning. A retrospective dataset of 223 anonymized obstetric records from Mpilo Central Hospital, Zimbabwe, was used for model development. Preprocessing involved missing-value imputation, outlier capping, feature engineering, and normalization, generating clinically relevant features such as labour duration in minutes, binary obstetric risk indicators, and a composite maternal risk score. Random Forest and Multilayer Perceptron (MLP) models were trained and evaluated on an 80:20 split. Random Forest achieved superior performance, with 86.67% accuracy, 89.47% precision, 80.95% recall, an F1-score of 85.00%, and ROC-AUC of 0.8730 for the PPH class. Labour duration and delivery method were the most influential predictors. SHAP (Shapley Additive Explanations) was used to enhance interpretability and clinical transparency. The findings show that interpretable machine learning can deliver reliable predictive performance even on small clinical datasets from resource-constrained settings, supporting proactive obstetric intervention and improved decision-making.
KEYWORDS: Postpartum Haemorrhage, Artificial Intelligence in Healthcare, Explainable AI, Early Warning Systems, Maternal Risk Prediction	

1. INTRODUCTION

Postpartum haemorrhage (PPH) is one of the leading causes of maternal morbidity and mortality worldwide. It involves excessive bleeding after delivery that is typically sudden and rapid in progression. Despite advances in obstetric care, PPH remains a major cause of preventable maternal mortality, particularly in low- and middle-income countries. Its unpredictability underscores the need for early identification and timely treatment to reduce serious complications. Conventional risk assessment relies on clinical judgment and familiar risk factors, but these approaches often fail to capture complex interactions among demographic, obstetric, and intrapartum variables, and scoring systems may not adapt well to real-time changes during labour. Artificial intelligence (AI) and machine learning (ML) can process multidimensional data and detect patterns that conventional

statistics miss enabling data-driven, proactive early warning systems in obstetric care. Recent studies highlight the growing role of ML in PPH prediction. Raman et al. (2025) compared logistic regression, random forests, gradient boosting, and neural networks using demographic, historical, and dynamic labour data; neural networks achieved 93% accuracy and an AUC-ROC of 0.96, with real-time labour data improving prediction flexibility. Yao et al. (2025) developed five ML models on 1,225 vaginal births in Shanghai, finding XGBoost the strongest performer, with SHAP analysis identifying midwife experience, fear of childbirth, parity, second-stage labour time, episiotomy, and companionship as key predictors. Xiao et al. (2025) combined convolutional neural networks with T2-weighted MRI on 280 pregnant women, where a deep-learning-plus-clinical model achieved an AUC of 0.868, outperforming deep-learning-only (0.834) and clinical-only (0.784) models.

Lérias Cambeiro et al. (2025) applied K-means clustering for feature selection followed by logistic, ridge, and random forest regression, with ridge logistic regression performing best (accuracy 0.871, AUC 0.907).

The specific objectives of this study are to: (i) design an AI-based early warning system for postpartum haemorrhage; (ii) implement and optimize machine learning models within the proposed framework; and (iii) assess the clinical relevance and predictive performance of the developed system. Despite recent progress, few studies offer an integrated, interpretable, and clinically ready early warning system focused on prevention, motivating the development of more accurate and transparent AI tools to improve maternal outcomes.

2. MATERIALS AND METHODS

2.1 Materials

All computation was performed on an MSI laptop with an Intel Core i7-8750H, 16GB RAM, and a 4GB NVIDIA GeForce GTX 1060, sufficient for neural network training with optional GPU acceleration. The software environment was managed with Anaconda for a reproducible Python 3.11 setup, with development in Jupyter Notebook. Key libraries included Pandas and NumPy for data handling; Scikit-learn for preprocessing, the Random Forest classifier, and evaluation metrics; PyTorch for building and training the MLP; Matplotlib and Seaborn for visualization; SHAP for model interpretability; and Joblib for saving trained models and preprocessing objects. The dataset was a publicly

available retrospective set of 223 anonymized patient records from Mpilo Central Hospital, Zimbabwe (Mendeley Data: data.mendeley.com/datasets/k7z2yywdn5/1), provided in Microsoft Excel format.

2.2 Methods

2.2.1 Data Collection

The dataset comprised 223 patient records and 16 variables covering demographic, obstetric, and neonatal information, including maternal age, gestational age, gravida, parity, number of children alive, previous obstetric complications, abnormalities, labour duration, delivery method, estimated blood loss, perineum outcome, birth weight, sex of newborn, and maternal HIV status. Records were anonymized standard obstetric data. PPH was defined as blood loss of 1,000 mL or more within 24 hours of delivery, following standard clinical definitions.

2.2.2 Data Loading and Inspection

Data manipulation used Pandas and NumPy, machine learning used PyTorch and Scikit-learn, and visualization used Matplotlib and Seaborn. Random seeds were fixed at 42 for reproducibility, and the dataset was loaded with Pandas' read_excel function. Inspection confirmed 223 rows and 16 columns (6 integer, 10 object variables, ~28.0 KB). Descriptive statistics showed a mean maternal age of 28.14 years (SD 6.98, range 15–44), mean gravida of 2.74, mean parity of 1.63, and mean estimated blood loss of 971.62 mL (range 500–3000 mL), indicating some extreme values.

```
Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 223 entries, 0 to 222
Data columns (total 16 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Patient ID                            223 non-null    int64
1   Age                                    223 non-null    int64
2   GA                                    223 non-null    object
3   B/U                                   223 non-null    object
4   Gravida                               223 non-null    int64
5   Parity                                223 non-null    int64
6   Children Alive                        223 non-null    int64
7   Past Obst Complications               223 non-null    object
8   Abnormalities                         223 non-null    object
9   Duration of Labour                    220 non-null    object
10  Delivery method                       223 non-null    object
11  Est Blood Loss                         223 non-null    int64
12  Perineum                              221 non-null    object
13  Birth weight                          221 non-null    object
14  Sex of Baby                           218 non-null    object
15  HIV Status                            223 non-null    object
dtypes: int64(6), object(10)
memory usage: 28.0+ KB
None
```

Figure 1: Dataset structure and variable summary

Categorical review found 199 booked and 24 unbooked patients, with inconsistent text formatting in obstetric-complication categories requiring standardization. Twelve missing values were identified: three in labour duration, two in perineum outcome, two in birth weight, and five in sex of baby, with all other variables fully recorded.

2.2.3 Data Preprocessing

A working copy preserved the original dataset. Missing values were addressed in two stages: forward fill for consecutive gaps, followed by K-Nearest Neighbors imputation (k=5) for remaining numerical gaps, successfully resolving all 12 missing entries. Outliers in numerical variables were capped at the 1st and 99th percentiles to limit the influence of extreme values (e.g., the maximum estimated

blood loss of 3000 mL) without discarding the target variable; patient ID and HIV status were excluded from outlier treatment.

2.2.4 Feature Engineering and Selection

Labour duration, originally text (e.g., "8h 40min"), was converted to total minutes (e.g., 520 min) via a custom function. Categorical obstetric variables were converted to binary indicators, such as a Has_Previous_C_S flag, alongside indicators for obstetric complications and abnormalities. A composite Risk_Score combined previous caesarean section, obstetric complications, abnormalities, gravida greater than one, and maternal age below 18 or above 35, reflecting differing maternal risk levels. Birth weight text (e.g., "2800g") was cleaned into numeric grams, and label encoding was applied to delivery method (Abortion, Forceps, LSCS, NVD), perineum outcome, and sex of baby.

Fifteen modelling features were selected, spanning age, gestational age, gravida, parity, children alive, labour duration in minutes, estimated blood loss, birth weight, engineered binary variables, risk score, and encoded categorical variables, yielding a 223-by-15 feature matrix with no remaining non-numeric columns. Min-Max scaling normalized all features to a 0–1 range using $X_{scaled} = (X - X_{min}) / (X_{max} - X_{min})$, ensuring no single feature dominated model training due to differing magnitudes.

2.2.5 Model Development

The final feature table comprised 223 observations and 13 predictor variables, with a binary PPH target (blood loss ≥ 1000 mL): 118 No-PPH cases and 105 PPH cases (47.09% PPH), a relatively balanced distribution. Stratified splitting (train_test_split) produced 178 training and 45 testing samples, preserving class proportions, and features/labels were converted to PyTorch tensors for neural network training.

The Random Forest classifier used 500 trees, a maximum depth of 6, minimum samples of 5 to split and 2 per leaf, balanced-subsample class weighting, and a fixed random state of 42. The MLP was a fully connected network with an input layer matching the 13 selected features, two hidden layers (32 and 16 neurons, ReLU activation), 0.3 dropout after each hidden layer, and a single-neuron output for binary

classification. It was trained with the Adam optimizer (learning rate 0.0005, weight decay $1e-4$), BCEWithLogitsLoss with a positive class weight of 1.1, 80 epochs, batch size 16, and gradient clipping (max norm 1.0) on a CUDA-enabled GPU where available.

2.2.6 Evaluation and Interpretability

Both models were tested on the same 45-sample hold-out set using accuracy, precision, recall, F1-score, and ROC-AUC, with confusion matrices summarizing true/false positives and negatives. Random Forest feature importances were extracted to identify the most influential predictors. SHAP DeepExplainer was applied to the MLP using 50 test samples and 100 background training samples, producing summary plots for global feature effects and force plots identifying the top three contributing features for individual predictions, enhancing clinical transparency. Trained model weights were saved via PyTorch state_dict, and preprocessing components (scaler, encoders, feature list) were serialized with Joblib for future deployment.

3. RESULTS AND DISCUSSION

3.1 Model Performance

Both models were evaluated on 45 test observations (24 No-PPH, 21 PPH). The Random Forest achieved an overall accuracy of 0.8667: for No-PPH, precision 0.8462, recall 0.9167, F1 0.8800; for PPH, precision 0.8947, recall 0.8095, F1 0.8500; macro-average F1 0.8650, weighted-average F1 0.8660, and ROC-AUC 0.8730, indicating strong discriminative power. Its confusion matrix showed 22 true negatives, 17 true positives, 2 false positives, and 4 false negatives.

The MLP achieved an overall accuracy of 0.8222: for No-PPH, precision 0.8077, recall 0.8750, F1 0.8400; for PPH, precision 0.8421, recall 0.7619, F1 0.8000; macro-average F1 0.8200, weighted-average F1 0.8213, and ROC-AUC 0.8115. Its confusion matrix showed 21 true negatives, 16 true positives, 3 false positives, and 5 false negatives — one more of each error type than Random Forest, reflecting its slightly lower performance.

Table 1: Model comparison summary

Metric	RF No PPH	RF PPH	RF Overall	MLP No PPH	MLP PPH	MLP Overall
Precision	0.8462	0.8947	—	0.8077	0.8421	—
Recall	0.9167	0.8095	—	0.8750	0.7619	—
F1 Score	0.8800	0.8500	—	0.8400	0.8000	—
Accuracy	—	—	0.8667	—	—	0.8222
ROC AUC	—	—	0.8730	—	—	0.8115

RANDOM FOREST				
	precision	recall	f1-score	support
No PPH	0.8462	0.9167	0.8800	24
PPH	0.8947	0.8095	0.8500	21
accuracy			0.8667	45
macro avg	0.8704	0.8631	0.8650	45
weighted avg	0.8688	0.8667	0.8660	45
RF AUC: 0.8730				

Figure 2: Random Forest classification report

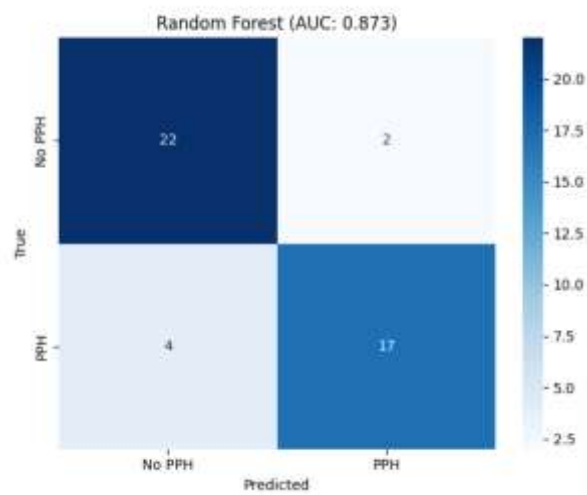


Figure 3: Random Forest confusion matrix

MLP				
	precision	recall	f1-score	support
No PPH	0.8077	0.8750	0.8400	24
PPH	0.8421	0.7619	0.8000	21
accuracy			0.8222	45
macro avg	0.8249	0.8185	0.8200	45
weighted avg	0.8238	0.8222	0.8213	45
MLP AUC: 0.8115				

Figure 4: MLP classification report

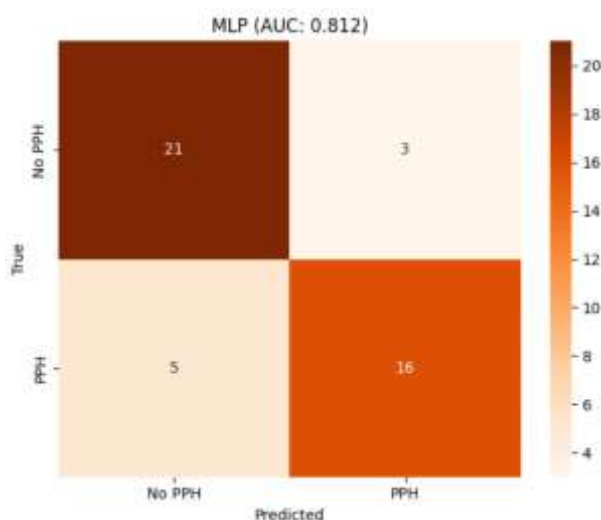


Figure 5: MLP confusion matrix

3.2 Feature Importance

Random Forest feature importance analysis identified Duration_of_Labour_minutes (≈ 0.33) and Delivery_Method_Encoded (≈ 0.30) as the dominant predictors, together accounting for over 0.60 of total importance. Age (≈ 0.09), Perineum_Encoded (≈ 0.08), and Birth_Weight_numeric (≈ 0.06) contributed moderately,

while Risk_Score, Gravida, Children_Alive, Parity, Has_Previous_C_S, and Sex_of_Baby_Encoded contributed marginally (≈ 0.02 – 0.05 each). Has_Obst_Complications and Has_Abnormalities contributed negligibly, indicating that procedural and labour-related variables were stronger predictors in this dataset than historical complication factors.

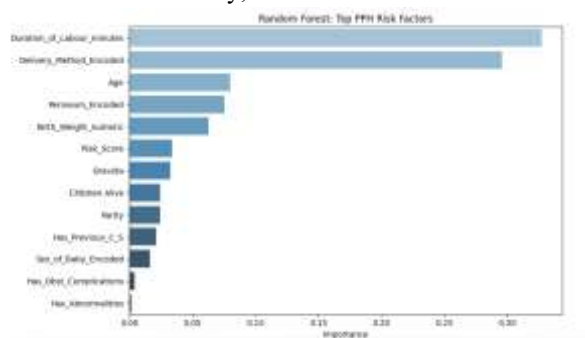


Figure 6: Random Forest feature importance

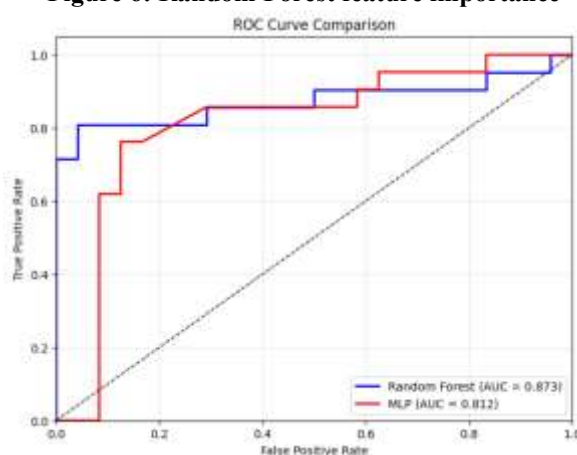


Figure 7: ROC-AUC curve comparison (Random Forest vs MLP)

3.3 Discussion

Random Forest outperformed the MLP across all key metrics, notably producing fewer false negatives — a clinically important distinction, as missed PPH cases can delay life-saving intervention. Compared with recent literature, Raman et al. reported an AUC of 0.96 using larger datasets and dynamic labour parameters; Yao et al. achieved strong results with XGBoost on 1,225 records [8]; Lérias Cambeiro et al. obtained an AUC of 0.907 with ridge logistic regression [10]; and Xiao et al. reached an AUC of 0.868 by fusing imaging with clinical data. While this study's AUC of 0.8730 does not exceed the highest reported values, it remains competitive given the smaller sample size (223 records) and reliance on routinely collected clinical variables alone, without imaging or multicenter data — demonstrating that reliable predictive performance is achievable in low-resource environments.

A key strength of this study is the integration of SHAP-based interpretability, addressing a common limitation of accuracy-focused models by clarifying which factors particularly labour duration and delivery method drive predictions,

thereby improving clinical transparency and trust. These results support the feasibility of AI-based early warning systems in resource-limited settings, though larger multicenter validation is needed to confirm generalizability.

4. CONCLUSION

This study designed, implemented, and evaluated an AI-based early warning system for predicting postpartum haemorrhage using routine obstetric data from Mpilo Central Hospital, Zimbabwe. Through systematic preprocessing, feature engineering, and model comparison, the Random Forest classifier achieved an accuracy of 0.8667, balanced precision and recall for the PPH class, and a ROC-AUC of 0.8730 on unseen test data, with labour duration and delivery method identified as the dominant risk predictors. The framework demonstrates that machine learning can deliver clinically valuable, interpretable predictions from small datasets typical of low- and middle-income settings, supporting proactive responses such as uterotonic management, fluid replacement, or surgical pre-treatment. The trained model and preprocessing pipeline provide a

foundation for integration into clinical decision-support systems, with potential to reduce avoidable maternal morbidity and mortality from PPH in resource-constrained environments.

5. RECOMMENDATIONS

- i. Perform external validation of the Random Forest model using independent datasets from other low-resource maternity units to assess generalizability.
- ii. Incorporate additional real-time intrapartum variables (e.g., vital-sign trends, uterine tone) and psychosocial factors (e.g., fear of childbirth) to improve predictive quality.
- iii. Explore advanced ensemble algorithms such as XGBoost or hybrid statistical-ML models to improve accuracy and calibration on similar small-scale datasets.
- iv. Develop and pilot the model as a clinical decision-support tool in maternity wards to assess its effect on timely intervention and outcomes.
- v. Incorporate additional interpretability methods (e.g., LIME, counterfactual explanations) alongside SHAP to strengthen clinician trust.
- vi. Investigate adapting the framework for mobile or low-compute deployment in rural or resource-poor healthcare facilities.
- vii. Partner with multicenter studies to build larger, harmonized African datasets, enabling more advanced deep learning models and reducing data scarcity.

REFERENCES

1. Elser, H., Mumford, S., Grantz, K., Pollack, A., Mendola, P., Mills, J., Yeung, E., Zhang, C., Schisterman, E., & Hinkle, S. (2025). Postpartum Haemorrhage and Long-Term Mortality. *Paediatric and Perinatal Epidemiology*. <https://doi.org/10.1111/ppe.13166>.
2. Lee, S., & Kang, Y. (2022). Postpartum hemorrhage. *Taiwanese Journal of Obstetrics & Gynecology*, 61(1), 5-7. <https://doi.org/10.1016/j.tjog.2021.11.003>.
3. Riches, J., Jafali, J., Twabi, H., et al. (2025). Avoidable factors associated with maternal death from postpartum haemorrhage: a national Malawian surveillance study. *BMJ Global Health*, 10. <https://doi.org/10.1136/bmjgh-2024-015781>.
4. Sumikura, H. (2025). Postpartum Hemorrhage: Preventable Maternal Deaths and the Role of Coagulation Monitoring. *Journal of Obstetrics and Gynaecology Research*, 51. <https://doi.org/10.1111/jog.70134>.
5. Ende, H., & Butwick, A. (2021). Current State and Future Direction of Postpartum Hemorrhage Risk Assessment. *Obstetrics & Gynecology*, 138, 924-930. <https://doi.org/10.1097/aog.0000000000004579>.
6. Joshi, D. (2025). Why Traditional Statistical Methods Need to Evolve in the Age of Artificial Intelligence: A Biostatistical Perspective. *Nepal Journal of Public Health*. [https://doi.org/10.70280/njph\(2025\)v2i1.31](https://doi.org/10.70280/njph(2025)v2i1.31).
7. Raman, R., Kumar, V., Pillai, B., Verma, A., Rastogi, S., & Pandey, D. (2025). Predictive Machine Learning Models for Postpartum Hemorrhage Risk Assessment. 2025 IITCEE, 1-6. <https://doi.org/10.1109/iitcee64140.2025.10915212>.
8. Yao, X., Bao, Y., Wu, N., Shan, S., Xu, Y., Huo, K., Huang, R., & Ying, H. (2025). Interpretable machine-learning-based prediction of postpartum haemorrhage in normal vaginal births in Shanghai, China. *Frontiers in Medicine*, 12. <https://doi.org/10.3389/fmed.2025.1670987>.
9. Xiao, Y., Hu, Z., Yuan, B., & Sunv, Z. (2025). Postpartum Hemorrhage Prediction Based on Convolutional Neural Networks. 2025 DIPCA, 182-186. <https://doi.org/10.1109/dipca65051.2025.11042290>.
10. Lérias-Cambeiro, M., Mugeiro-Silva, R., Rodrigues, A., Dias-Domingues, T., Lança, F., & Carneiro, A. (2025). Enhancing Postpartum Haemorrhage Prediction Through the Integration of Classical Logistic Regression and Machine Learning Algorithms. *Mathematics*. <https://doi.org/10.3390/math13213376>.
11. Nelli, F. (2023). Python Data Analytics: With Pandas, NumPy, and Matplotlib. <https://doi.org/10.1007/978-1-4842-9532-8>.
12. Adugna, T., Xu, W., & Fan, J. (2022). Comparison of Random Forest and Support Vector Machine Classifiers for Regional Land Cover Mapping Using Coarse Resolution FY-3C Images. *Remote Sensing*, 14, 574. <https://doi.org/10.3390/rs14030574>.
13. Nohara, Y., Matsumoto, K., Soejima, H., & Nakashima, N. (2021). Explanation of Machine Learning Models Using Shapley Additive Explanation and Application for Real Data in Hospital. *Computer Methods and Programs in Biomedicine*, 214, 106584. <https://doi.org/10.1016/j.cmpb.2021.106584>.
14. Parida, S., Gerostathopoulos, I., & Bogner, J. (2025). How Do Model Export Formats Impact the Development of ML-Enabled Systems? A Case Study on Model Integration. 2025 IEEE/ACM CAIN, 48-59. <https://doi.org/10.1109/cain66642.2025.00014>.