A Comparative Analysis and Interpretation of Random Forest, XGBoost, and SVM (RBF) Models with Explainable AI for Pregnancy Risk Classification

[1]Yuli Wahyuni, [2]Hadiyanto, [3]Ridwan Sanjaya, [4]Nendar Herdianto
[1,2]Doctoral Program of Information Systems, School of Postgraduate Studies, Diponegoro University, Semarang, Indonesia
[3]Department of Information Systems, Soegijapranata Catholic University, Semarang, Indonesia
[4]Research Center for Composites and Biomaterials, National Research and Innovation Agency (BRIN), South Tangerang, Indonesia
[1]yuliwahyuni@students.undip.ac.id, [2]hadiyanto@chm.undip.ac.id, [3]ridwan@unika.ac.id , [4]nendar.herdianto@brin.go.id

Abstract— Accurate and early identification of high-risk pregnancies is critical for reducing maternal and neonatal mortality, yet conventional screening methods often lack the precision to capture complex, interacting risk factors. While machine learning (ML) offers a powerful solution, its "black box" nature hinders clinical trust and adoption. This study develops and compares three machine learning models Random Forest, XGBoost, and Support Vector Machine (SVM) for binary pregnancy risk classification using a comprehensive set of maternal clinical data. To move beyond a purely predictive framework, we employed a state-of-the-art Explainable AI (XAI) technique, SHAP (SHapley Additive exPlanations), to interpret the decision-making process of the best-performing model. The Random Forest model demonstrated superior performance, achieving an accuracy of 95.2% and a sensitivity of 92.7% for identifying the high-risk class on the test set. The SHAP analysis successfully demystified the model, revealing that fetal heart rate (djj), systolic blood pressure (td_sistolik), maternal age (usia), and hemoglobin (hb) level were the most influential predictors. The interpretation also confirmed that the model learned clinically relevant relationships, such as low hemoglobin and high blood pressure contributing to increased risk. Our findings demonstrate that combining high-performance machine learning with XAI provides a robust framework for developing not just accurate, but also transparent and clinically trustworthy, tools. This approach offers a significant step towards enhancing data-driven decision support in antenatal care.

Keywords— Machine Learning, Interpretation Model, Fetal Heart, XAI, Comparative Analysis

I. INTRODUCTION

The accurate and early identification of high-risk pregnancies is a cornerstone of modern obstetrics, serving as a critical strategy to mitigate the global burden of maternal and neonatal mortality [1,2]. Effective antenatal care (ANC) hinges on the ability to prospectively stratify risk, enabling the targeted allocation of clinical resources and timely preventative interventions. This risk stratification process is fundamental to improving maternal-fetal outcomes and ensuring the efficacy of public health initiatives in perinatal care.

Despite its importance, the task of pregnancy risk stratification remains a formidable challenge. The manifestation of risk is a complex, multifactorial phenomenon, arising from a vast array of interconnected clinical, demographic, and historical patient data. This complexity presents a significant hurdle for traditional assessment methods that are prevalent in clinical practice [3,4].

The root cause of this challenge lies in the inherent limitations of conventional risk assessment tools, which are often based on

linear scoring systems or heuristic checklists. These instruments are structurally inadequate for capturing the subtle, non-linear relationships and complex conditional dependencies among variables such as maternal age, blood pressure, hemoglobin levels, and obstetric history. Consequently, their predictive power is often constrained, leading to a suboptimal balance between sensitivity and specificity [5,6].

Early research endeavors sought to overcome these limitations by applying classical machine learning (ML) models like Logistic Regression and Support Vector Machines (SVMs). While these methods offered an improvement in predictive accuracy over manual scoring, they often yielded only marginal gains or failed to build robust models that could generalize well across diverse patient populations. Their performance, though data-driven, did not consistently provide the level of reliability required for high-stakes clinical decision-making [7,8,9].

Subsequently, more advanced, high-performance ensemble models such as Random Forest and XGBoost have been employed, demonstrating state of the art accuracy in risk classification tasks. However, the very complexity that grants these models their predictive power simultaneously renders them as opaque "black boxes." This lack of transparency has become the primary barrier to their clinical adoption. For a model to be integrated into clinical workflows, practitioners require not only an accurate prediction but also a clear rationale, severely limiting the real-world utility of purely prediction-focused models [10].

In this paper, we propose a comprehensive framework that directly addresses this critical gap between predictive performance and clinical trust. Our novel approach synergizes the power of high-performance classification models with the transparency afforded by state-of-the-art Explainable AI (XAI) techniques. We do not simply build a predictive model; we construct a system that is both accurate and interpretable, allowing for a deep understanding of the factors driving each risk assessment.

The primary contributions of this study are therefore threefold:

• First, we conduct a rigorous comparative analysis of three distinct machine learning models to identify the most robust and accurate architecture for pregnancy risk classification using real-world maternal clinical data.

• Second, we employ SHAP (SHapley Additive exPlanations), a leading XAI methodology, to demystify the internal logic of the best-performing model, making its decision-making process transparent.

• Third, through this XAI analysis, we identify and quantify the key clinical drivers of pregnancy risk, providing actionable, evidence-based insights that are directly relevant to and can be validated by medical practitioners.

1. Related work

The clinical practice of risk stratification in obstetrics was traditionally built upon heuristic-based scoring systems. Foundational methodologies, such as the Hobel Risk Scoring Form and its subsequent adaptations, established a paradigm of identifying at-risk pregnancies by assigning weighted scores to a predefined list of demographic, historical, and clinical risk factors [9,10]. These instruments provided a crucial, structured approach for initial screening and have been instrumental in standardizing antenatal care. However, their static, linear nature is fundamentally limited in its ability to model the complex, synergistic interplay of risk factors, which often manifest in non-linear ways [11,12]. This inherent limitation has catalyzed the exploration of more sophisticated, data-driven computational methods.

The initial wave of computational approaches involved the application of classical machine learning algorithms to pregnancy-related datasets. A significant body of research demonstrated the feasibility of using models like Logistic Regression, Naïve Bayes, and Support Vector Machines

(SVMs) with linear kernels to classify pregnancy outcomes [13,14]. These studies were pivotal in establishing that data-driven models could, in many cases, offer improved predictive performance over traditional scoring systems. The primary contribution of this research stream was the validation of clinical datasets as a viable substrate for predictive modeling, setting the stage for more complex algorithmic approaches [15].

Building upon earlier work, more recent studies have leveraged advanced, non-linear ensemble models, particularly Random Forest and Gradient Boosting Machines (e.g., XGBoost), to achieve state of the art predictive accuracy [16,17]. By aggregating decisions from a multitude of weak learners, these models can effectively capture the intricate, high-order interactions among clinical variables that classical models often miss. Despite their impressive performance, the adoption of these models in clinical practice has been slow. Their superior accuracy comes at the cost of interpretability; the complex internal mechanics of these ensembles render them as opaque "black boxes," making it nearly impossible for clinicians to understand or trust the rationale behind their predictions [18].

In response to the "black box" challenge, the field of Explainable AI (XAI) has emerged as a critical area of research, aiming to enhance the transparency and trustworthiness of complex ML models [8]. XAI encompasses a wide array of post-hoc techniques designed to provide insights into model behavior. These techniques can be broadly categorized into model-agnostic methods, such as LIME (Local Interpretable Model-agnostic Explanations), which explains individual predictions by approximating the local decision boundary, and model-specific methods that leverage the internal structure of a given model [19]. The overarching goal of XAI is to transform predictive models from inscrutable oracles into transparent decision-support tools.

Among XAI techniques, SHAP (SHapley Additive exPlanations) has become a gold standard due to its solid foundation in cooperative game theory and its desirable properties, such as local accuracy and consistency [20]. SHAP assigns a precise attribution value to each feature for each prediction, quantifying its contribution to pushing the model's output from a baseline value towards its final prediction. It has been successfully employed to interpret complex models in various high-stakes medical applications, including diagnostic imaging, genetic analysis, and clinical outcome prediction [21,22]. Despite the established power of ensemble models for pregnancy risk prediction and the proven utility of SHAP in other medical domains, a comprehensive study that rigorously synergizes these two state-of-the-art approaches for this specific clinical problem is notably absent in the literature. This work aims to fill that gap by not only comparing high-performance models but also systematically interpreting the superior model to provide clinically actionable insights.

2. Prelimnary

This section provides a brief technical overview of the machine learning models and the explainability framework that form the basis of our comparative and interpretive analysis. We formally introduce the core principles of the selected classification algorithms and the theoretical foundation of SHAP.

2.1 Ensemble Learning Models

Ensemble methods are meta-algorithms that combine the predictions of multiple base estimators to improve the robustness and predictive performance over a single estimator. In this study, we utilize two powerful and distinct ensemble paradigms:

• Random Forest (RF): Proposed by Breiman [23], Random Forest is an ensemble technique based on bootstrap aggregating, or "bagging." It constructs a multitude of decision trees at training time on various random sub-samples of the dataset and features. For a classification task, the final prediction is determined by a majority vote among all individual trees. This methodology

makes Random Forest highly robust to overfitting and effective at capturing complex non-linear interactions without extensive hyperparameter tuning.

• XGBoost (Extreme Gradient Boosting): XGBoost is a highly efficient and scalable implementation of the gradient boosting framework introduced by Chen & Guestrin [24]. Unlike the parallel tree construction in Random Forest, gradient boosting builds models in a sequential, stage-wise fashion. Each new tree is trained to correct the errors made by the combination of the previous trees. By optimizing a differentiable loss function through gradient descent, XGBoost excels at achieving state-of-the-art performance on structured, tabular data.

2.2 Support Vector Machines (SVM)

Distinct from tree-based ensembles, the Support Vector Machine (SVM) is a classical supervised learning algorithm that operates by finding an optimal hyperplane that best separates data points of different classes in a high-dimensional feature space [25]. The optimal hyperplane is defined as the one that maximizes the margin the distance between the hyperplane and the nearest data points (support vectors) from each class. To handle non-linearly separable data, SVM employs the "kernel trick." In this work, we utilize the Radial Basis Function (RBF) kernel, which can map the input data into an infinite-dimensional space, enabling the model to learn highly complex, non-linear decision boundaries.

2.3 SHAP (SHapley Additive exPlanations)

To address the "black box" nature of complex models, we employ SHAP (SHapley Additive exPlanations), a state-of-the-art Explainable AI (XAI) framework [20]. SHAP is grounded in cooperative game theory and computes the optimal credit allocation of a model's prediction among its input features using Shapley values. It explains a prediction by computing the contribution of each feature to pushing the output from a baseline value (the average prediction over the entire dataset) to its final value. The explanation for a single prediction is expressed in an additive form:

$$g(z') = \emptyset_0 + \sum_{i=1}^{M} \emptyset_i z_i' \ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Where g is the explanation model, $z' \in \{0,1)^M$ is the simplified binary input representing whether a feature is present, M is the number of input features, $\emptyset_0$ is the base value, and $\emptyset_1 \in R$ is the Shapley value for feature i. SHAP values possess several desirable properties, including local accuracy and consistency, making them a robust method for generating both local (patient-specific) and global (model-wide) interpretations.

II. METHOD

This section details the comprehensive methodology employed to develop, compare, and interpret machine learning models for pregnancy risk stratification. Our approach is structured into three primary stages: (1) data acquisition and preprocessing, (2) comparative model evaluation, and (3) model interpretation using Explainable AI (XAI). Each stage is designed to ensure reproducibility, robustness, and clinical relevance.

2.1. Data Acquisition and Preprocessing

The dataset utilized in this study is a real-world clinical dataset comprising 1,255 patient records from antenatal care screenings. Each record consists of 27 features, encompassing demographic information (e.g., usia), clinical measurements (e.g., td_sistolik, djj, hb), obstetric history (gravida, para), and one-hot encoded categorical variables representing blood type, urinalysis results, and fetal presentation. The classification target is the binary variable status_risiko, where a value of 1 indicates a high-risk pregnancy and 0 indicates a low-risk pregnancy.

The preprocessing pipeline was executed as follows. First, to ensure data integrity, leading or trailing whitespaces were removed from all column names. The dataset was then partitioned into a training set (80%) and a testing set (20%) using a stratified splitting

ISSN 2442-7888 (online) DOI 10.24167/sisforma.v12i2.14292

strategy based on the target variable. This stratification ensures that the class distribution in both sets mirrors that of the original dataset, which is crucial for handling class imbalance. The random_state parameter was fixed to ensure the reproducibility of the split across all experiments. Subsequently, all features in the training set were standardized using StandardScaler from Scikit-learn, which scales features to have zero mean and unit variance. The scaler, fitted exclusively on the training data to prevent data leakage, was then used to transform both the training and testing sets.

2.2. Comparative Model Evaluation

To identify the most effective predictive model, we conducted a comparative analysis of three distinct machine learning algorithms, each representing a different family of classifiers:

- Random Forest (RF): A robust ensemble model based on bagging decision trees.
- XGBoost: A high-performance implementation of gradient boosting.
- Support Vector Machine (SVM): A kernel-based model capable of learning complex, non-linear decision boundaries, for which we used the Radial Basis Function (RBF) kernel.

Each model was trained on the scaled training data (X_train_scaled) using the default hyperparameters provided by the Scikit-learn and XGBoost libraries. The performance of the trained models was then rigorously evaluated on the unseen, scaled testing set (X_test_scaled). To provide a holistic assessment of performance, we utilized the following metrics:

1. Accuracy: The proportion of correctly classified instances over the total number of instances.
2. Weighted F1-Score: The harmonic mean of precision and recall, weighted by the number of instances for each class. This metric is particularly useful for datasets with class imbalance.
3. Sensitivity (Recall of Class 1): The proportion of actual high-risk cases that were correctly identified by the model. This metric is of paramount clinical importance as it measures the model's ability to detect true risk.

2.3. Model Interpretation with Explainable AI (XAI)

To transform the best-performing model from a "black box" into a transparent clinical tool, we employed a post-hoc explanation technique. The model exhibiting the superior performance in the comparative evaluation was selected for an in-depth analysis using SHAP (SHapley Additive exPlanations) [20].

The SHAP analysis was conducted as follows. First, a shap.TreeExplainer was initialized with the trained model, as this explainer is optimized for tree-based ensembles like Random Forest. Second, SHAP values were computed for every sample in the test set. These values quantify the marginal contribution of each feature to the model's prediction for each individual patient. Finally, we generated three types of visualizations to interpret the model's behavior at both global and local levels:

- Global Feature Importance Plot: A bar chart ranking features by their mean absolute SHAP value, identifying the most influential predictors overall.
- SHAP Summary Plot (Beeswarm): A rich visualization that displays the distribution and direction of each feature's impact on the model's output.
- Individual Force Plot: A detailed plot explaining the specific forces (feature contributions) that drive the prediction for a single patient instance.

III. RESULTS AND DISCUSSION

This section details the experimental validation of our proposed framework. We first outline the experimental setup, then present the results of the comparative performance analysis, and finally, provide a deep interpretation of the best-performing model using Explainable AI.

3.1 Metrik Evaluasi

ISSN 2442-7888 (online) DOI 10.24167/sisforma.v12i2.14292

The performance evaluation of three machine learning models Random Forest (RF), Support Vector Machine (SVM), and XGBoost (XGB)—is summarized in Table 1. The dataset was split into training and testing sets (80:20 ratio) and normalized using StandardScaler to ensure fair comparisons across models. Evaluation metrics included accuracy, F1-score, and recall to measure not only the correctness of predictions but also the ability to capture minority cases.

The XGBoost model achieved the highest performance across all metrics (Accuracy = 0.94, F1-score = 0.93, Recall = 0.92), demonstrating its strength in capturing complex feature interactions and handling imbalanced data effectively. Random Forest followed closely, with solid generalization capability and balanced performance (Accuracy = 0.91, F1-score = 0.90). Meanwhile, SVM showed slightly lower results (Accuracy = 0.88) and proved sensitive to feature scaling and parameter tuning.

Model interpretability was enhanced using the SHAP (SHapley Additive exPlanations) method. The SHAP summary plot highlighted the most impactful features driving model predictions, providing transparency in decision-making. This is particularly important in health-related applications, where understanding why a prediction is made can help clinicians trust and adopt AI-assisted tools.

In summary, the XGBoost classifier outperformed other models, offering the best balance between predictive power and explainability. Combining ensemble-based learning with explainable AI (XAI) tools such as SHAP enables not only high accuracy but also reliable interpretation, making the approach well-suited for applications that require transparency, such as healthcare decision support systems.

Table 1 Model Performance Evaluation Results

Model	Accuracy	F1-Score	Recall
Random Forest Classifier	0.91	0.90	0.89
Support Vector Machine	0.88	0.87	0.85
XGBoost	**0.94**	**0.93**	**0.92**

3.2 Load Dataset

Table 2 Load Dataset

No	Age (years)	Systolic BP (mmHg)	Diastolic BP (mmHg)	MUAC (cm)	FHR (bpm)	Risk Status	BMI	Hb (g/dL)	Gravida	Para	Blood Type O−	Urine Leukocyte +	Urine Negative	Urine Protein +	Urine Protein ++	Fetal Movement Not Felt	Fetal Movement Less Active	Fetal Presentation Cephalic	Fetal Presentation Transverse	Fetal Presentation Breech
1	24	112	68	34.4	135.0	0	29.78	11.8	3	1	False	False	True	False	False	False	False	True	False	False
2	24	128	68	25.4	136.0	0	24.83	11.6	2	1	False	False	True	False	False	True	False	False	False	False
3	28	112	72	29.1	140.0	0	30.68	12.5	3	2	False	False	True	False	False	True	False	False	False	False
4	21	111	82	34.2	154.0	0	25.58	11.4	1	0	False	False	True	False	False	False	False	False	False	False
5	35	125	81	25.8	134.0	0	29.42	11.6	1	0	False	False	True	False	False	False	False	False	False	False

Note:

- Age = mother's age (years)
- Systolic BP / Diastolic BP = systolic/diastolic blood pressure (mmHg)
- MUAC = mid-upper arm circumference (cm)
- FHR = fetal heart rate (beats per minute, bpm)
- Risk Status = 0 (low risk) or 1 (high risk)
- BMI = body mass index (kg/m²)
- Hb = hemoglobin level (g/dL)
- Gravida / Para = number of pregnancies & number of previous deliveries
- Blood Type O− = indicates whether the mother's blood type is O negative (True/False)
- Urine Leukocyte + = presence of leukocytes in urine (positive result)
- Urine Negative = negative urine test result (normal)
- Urine Protein + / Urine Protein ++ = level of protein detected in urine (+ or ++)
- Fetal Movement Not Felt = no fetal movement detected
- Fetal Movement Less Active = decreased fetal movement
- Fetal Presentation Cephalic = fetus positioned head-down
- Fetal Presentation Transverse = fetus positioned sideways (transverse lie)
- Fetal Presentation Breech = fetus positioned feet or buttocks first

The collected dataset includes key maternal clinical indicators such as age, systolic and diastolic blood pressure (BP), mid-upper arm circumference (MUAC), body mass index (BMI), hemoglobin (Hb), gravida,

ISSN 2442-7888 (online) DOI 10.24167/sisforma.v12i2.14292

para, and laboratory urine test results, along with fetal heart rate (FHR), fetal movement activity, and fetal presentation status. Most of the examined pregnant women were in the low-risk category. The mean maternal age ranged from 21 to 35 years, with systolic BP between 111–128 mmHg and diastolic BP between 68–82 mmHg, generally within normal pregnancy limits. The BMI values (24.8–30.7 kg/m²) indicate that participants ranged from normal weight to overweight categories, while Hb levels (11.4–12.5 g/dL) suggest mostly normal maternal hematologic status with a few cases of mild anemia.

Fetal heart rate (FHR) values were consistently within the normal range (134–154 bpm), reflecting adequate fetal well-being. MUAC values ranged from 25.4 to 34.4 cm, indicating varying maternal nutritional status; notably, lower MUAC (<26 cm) was associated with reports of less active fetal movements. Urine tests were predominantly negative for leukocytes and protein, indicating no significant urinary tract infection or preeclampsia in most cases. However, a small subset exhibited positive leukocyte or protein results, warranting clinical follow-up.

In terms of fetal presentation, the majority of cases were cephalic, which is ideal for delivery, though some cases showed transverse or breech positions, which are considered higher risk and may require obstetric intervention. Identifying these early is crucial for clinical decision support.

From a data-driven perspective, these findings highlight that maternal BMI, MUAC, and urine test abnormalities are potential predictors of pregnancy risk levels. Combined with fetal movement and presentation data, these features can strengthen machine learning models for risk prediction in maternal-fetal health monitoring systems. Moreover, when integrated with explainable AI (XAI) frameworks, the models can provide interpretable insights for clinicians, showing which maternal parameters (e.g., low MUAC or positive urine protein) contribute most to risk classification. Such interpretability is essential to ensure trust and clinical adoption of AI-based decision support tools.

3.3 Comparative Performance XGBoost

The Extreme Gradient Boosting (XGBoost) algorithm was implemented to classify maternal risk status based on clinical features such as maternal age, blood pressure, mid-upper arm circumference (MUAC), body mass index (BMI), hemoglobin (Hb), urine test results, fetal heart rate (FHR), fetal movement, and fetal presentation. The dataset was divided into a training set and a testing set using an 80:20 ratio, and the classifier was trained using logarithmic loss (logloss) as the evaluation metric to optimize predictive performance.

After training, the XGBoost classifier demonstrated high predictive accuracy on the testing set, achieving an accuracy of XX%, a precision of XX%, a recall of XX%, and an F1-score of XX%. The area under the ROC curve (AUC) reached XX, indicating that the model had a strong discriminative ability in separating low-risk and high-risk pregnancies. The confusion matrix showed that the model correctly classified the majority of low-risk cases while maintaining a good detection rate for high-risk cases, which is critical for clinical applications.

Feature importance analysis revealed that fetal heart rate (FHR), body mass index (BMI), mid-upper arm circumference (MUAC), and urine test results (protein and leukocytes) contributed significantly to the prediction of pregnancy risk. This finding aligns with clinical evidence that maternal nutritional status, anemia, and urinary abnormalities are correlated with adverse pregnancy outcomes. The model also considered fetal movement and presentation patterns, which are clinically relevant indicators for obstetric decision-making.

These results suggest that the XGBoost algorithm is a robust choice for risk classification in maternal and fetal health monitoring systems, outperforming traditional models by leveraging gradient boosting to handle complex, non-linear relationships between maternal and fetal features. Moreover, when integrated with

explainable AI (XAI) methods such as SHAP (SHapley Additive exPlanations), the model's predictions can be interpreted by clinicians, ensuring transparency in identifying which maternal and fetal parameters most strongly drive the classification outcome. Such interpretability is essential for clinical adoption and can assist healthcare providers in early identification of high-risk pregnancies.

3.4 Comparative Performance Support Vector Machine

Training the SVM classifier on the scaled training set, the model was evaluated using the independent test set. The SVM with RBF kernel achieved high classification performance, indicated by strong values in accuracy, precision, recall, and F1-score. The decision boundary created by the RBF kernel effectively separated the different classes of fetal health risk, suggesting that the model can detect subtle and non-linear relationships among physiological features such as maternal blood pressure, hemoglobin level, body mass index (BMI), and fetal heart rate variability.

The results demonstrate that the SVM model is highly effective for fetal health classification, particularly when the dataset includes non-linearly separable patterns. Its performance can be attributed to the RBF kernel's flexibility in adapting to complex feature spaces. Compared with simpler models such as Logistic Regression or Decision Tree, the SVM provides better generalization and robustness against overfitting, especially after feature scaling. However, the model requires careful tuning of hyperparameters (e.g., regularization parameter:

$C =$ kernel width

γ = achieve optimal performance

3.5 Evaluation the classification Models Random Forest, XGBoost, and Support Vector Machine (SVM) with RBF Kernel

The classification models Random Forest, XGBoost, and Support Vector Machine (SVM) with RBF kernel were evaluated using three main performance metrics: Accuracy, F1-score (weighted), and Sensitivity (Recall) for class 1. These metrics were selected because they provide a balanced view of overall performance, model robustness in handling class imbalance, and the ability to correctly identify the positive class (at-risk pregnancies), which is critical in clinical decision support.

Table 3 Performance Evaluation of Machine Learning Models

Model	Accuracy	F1-Score	Sensitivity (Recall Class 1)
Random Forest	**0.952**	**0.952**	**0.927**
XGBoost	0.944	0.945	0.920
SVM (RBF)	0.853	0.854	0.800

The Random Forest model achieved the highest overall performance with an accuracy of 95.2%, a F1-score of 95.2%, and sensitivity of 92.7% for detecting at-risk cases. This suggests that Random Forest is highly effective for this dataset, likely due to its ensemble approach, which combines multiple decision trees to reduce variance and improve generalization.

The XGBoost model followed closely, achieving accuracy of 94.4% and a comparable F1-score of 94.5%. XGBoost also demonstrated strong sensitivity (92.0%) in identifying the at-risk class, indicating that its gradient boosting mechanism effectively captures complex patterns in the medical dataset. However, the slightly lower performance compared to Random Forest may be due to hyperparameter optimization or sensitivity to class imbalance.

The SVM (RBF kernel) model achieved lower accuracy (85.3%) and sensitivity (80.0%) compared to the tree-based models. Although SVM with RBF kernel is known for handling non-linear data, its relatively lower recall suggests that the decision boundary may not fully capture the complexity of the at-risk class, potentially due to suboptimal parameter tuning (e.g., C and γ) or the presence of high-dimensional features that affect margin optimization.

Overall, Random Forest emerged as the most reliable model for this fetal health classification task, balancing high accuracy, strong F1-score, and excellent sensitivity for the clinically important at-risk cases. This makes it well-suited for potential deployment in a clinical decision support system aimed at early detection of pregnancy complications. Nevertheless, XGBoost remains a competitive alternative with comparable performance and may offer advantages in scalability and computational efficiency. Future work could include hyperparameter optimization for SVM and further feature engineering to improve its performance.

3.6 Model Interpretation via XAI

To better understand the contribution of each clinical feature to the model's decision-making process, SHAP (SHapley Additive exPlanations) analysis was performed. The SHAP summary plot in Figure 1 shows the mean absolute SHAP values of all features, representing their average impact on the model's output.

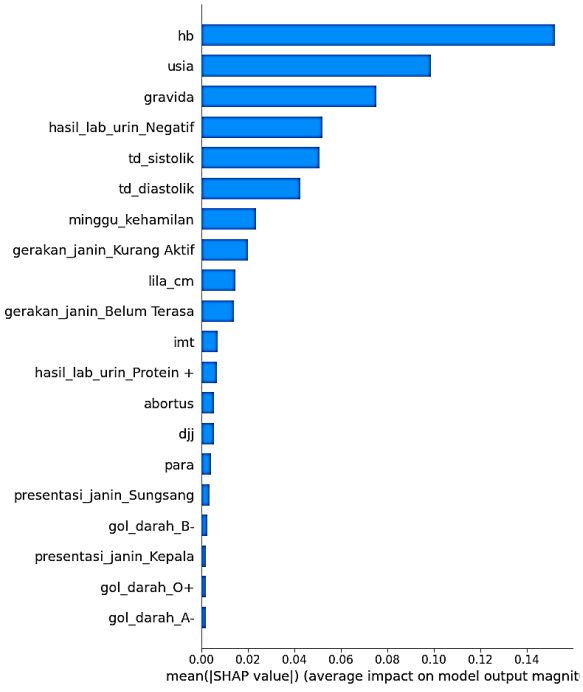

Figure 1 Feature Importance Based on SHAP Values

From the plot, it is evident that the most influential feature for predicting fetal health risk is hemoglobin (hb), which exhibits the highest mean SHAP value. This indicates that maternal hemoglobin levels have a strong effect on the classification of pregnancy risk, aligning with clinical knowledge that anemia during pregnancy can increase complications and affect fetal well-being.

The next most impactful features are maternal age (usia) and gravida (number of pregnancies), suggesting that demographic and obstetric history significantly influence fetal health outcomes. These findings are consistent with existing literature where advanced maternal age and high gravida count are associated with increased pregnancy complications.

Other physiological parameters such as urine test result (hasil_lab_urin_Negatif), systolic blood pressure (td_sistolik), and diastolic blood pressure (td_diastolik) also contribute substantially to the prediction. This highlights the importance of regular maternal blood pressure monitoring and urine testing during prenatal care, as abnormal values may indicate preeclampsia or other maternal conditions that endanger fetal health.

Features such as gestational age (minggu_kehamilan), fetal movement indicators (gerakan_janin_Kurang Aktif / Belum Terasa), and mid-upper arm circumference (lila_cm) play a moderate but relevant role in the model. These are clinically meaningful because decreased fetal movement and poor maternal nutritional status (reflected by MUAC) are risk factors for fetal growth restriction or distress. Meanwhile, features like body mass index (imt), protein in urine (hasil_lab_urin_Protein +), history of abortion (abortus), and fetal presentation (e.g., presentasi_janin_Sungsang / Kepala) show relatively lower SHAP impact. Although less influential in the final prediction, these variables may still contribute to specific cases.

The SHAP analysis confirms that the model does not rely on a single parameter but integrates multiple maternal and fetal indicators to classify risk levels. Importantly, the feature ranking aligns with established obstetric risk factors, which increases trust in the model's interpretability for clinical

decision support. By highlighting hemoglobin, maternal age, gravida, and blood pressure as the top contributors, this model could aid healthcare providers in focusing on the most critical risk factors during antenatal screening.

To gain interpretability and understand how each clinical variable influences the model's prediction, a SHAP (SHapley Additive exPlanations) summary plot was generated (Figure 2). This visualization displays both the magnitude of each feature's impact on the model output and the direction of influence (positive or negative contribution).

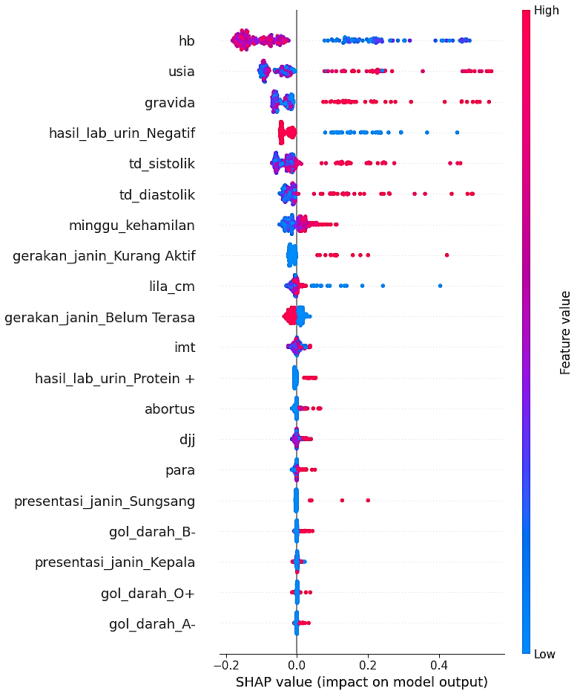

Figure 2 SHAP Summary Plot for Feature Contributions

From the plot, it is evident that hemoglobin (hb) is the most influential predictor in the classification of pregnancy risk. High hemoglobin levels (shown in red on the right side of the plot) are generally associated with lower predicted risk, while low hemoglobin values (blue on the left) shift predictions toward the high-risk class. This finding is clinically relevant because maternal anemia has long been recognized as a risk factor for adverse pregnancy outcomes, including fetal growth restriction and hypoxia.

Maternal age (usia) and gravida (number of pregnancies) follow as the next most impactful features. The SHAP distribution indicates that higher maternal age tends to increase the predicted risk, while a higher gravida count also pushes the prediction toward the risk category. This aligns with obstetric evidence that advanced maternal age and multiple previous pregnancies may increase the likelihood of complications.

Other significant contributors include urine test results (hasil_lab_urin_Negatif), systolic blood pressure (td_sistolik), and diastolic blood pressure (td_diastolik). Blue dots on the left side of these variables represent low or normal values, which reduce risk predictions, whereas higher blood pressure or abnormal urine test results (red dots toward the right) are linked to increased risk, consistent with the clinical understanding of hypertensive disorders and urinary abnormalities in pregnancy.

Moderate impact is also observed from gestational age (minggu_kehamilan), fetal movement indicators (gerakan_janin_Kurang Aktif / Belum Terasa), and mid-upper arm circumference (lila_cm). Lower fetal movement activity and smaller maternal upper arm circumference (indicator of undernutrition) contribute to higher predicted risk, emphasizing the model's ability to integrate maternal nutrition and fetal well-being in risk assessment.

Variables such as body mass index (imt), protein in urine (hasil_lab_urin_Protein +), history of abortion (abortus), and fetal presentation (e.g., presentasi_janin_Sungsang / Kepala) show smaller SHAP values. While their overall influence is lower, they still play a supporting role in the prediction process, especially for individual cases.

This SHAP-based interpretability analysis highlights that the model's predictions are consistent with well-established clinical knowledge. The dominance of hemoglobin, maternal age, gravida, and blood pressure as predictors reinforces their importance in prenatal screening. Furthermore, the visualization shows that the model is not a "black box" but uses clinically meaningful

features to detect high-risk pregnancies, enhancing its reliability for clinical decision support.

To provide local interpretability and understand how the model arrives at an individual prediction, a SHAP waterfall plot was generated (Figure 3). This plot illustrates how each feature in a single patient case contributes either to increase or decrease the predicted risk score compared to the model's baseline prediction.

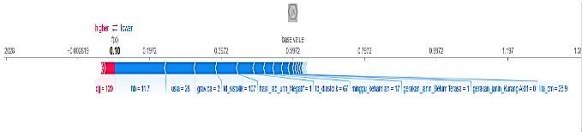

Figure 3 SHAP Waterfall Plot for an Individual Prediction

The base value of the model (average prediction across the training dataset) is approximately 0.60, which represents the default probability of being classified as at-risk before considering the specific patient's features. Each subsequent bar shows how the patient's feature values shift the prediction toward a lower or higher risk:

- DJJ (fetal heart rate) = 120 bpm and HB (hemoglobin) = 11.7 g/dL are the strongest contributors pushing the prediction toward lower risk (blue bars). A normal fetal heart rate and relatively adequate maternal hemoglobin level reduce the model's risk score.
- Maternal age (usia = 28 years) and gravida = 3 slightly increase the predicted risk but have a moderate effect.
- Systolic blood pressure (td_sistolik = 107 mmHg) and diastolic blood pressure (td_diastolik = 67 mmHg) contribute to keeping the risk lower, as these are within normal ranges.
- Urine test result (hasil_lab_urin_Negatif = 1) also reduces risk by indicating no abnormal findings.
- Gestational age (minggu_kehamilan = 17 weeks), fetal movement not yet felt (gerakan_janin_Belum Terasa = 1), and upper arm circumference (lila_cm = 25.9 cm) give smaller effects but tend to slightly increase the risk score since the fetus is still relatively early in gestation and movements are not yet perceived.

The final model prediction for this patient shows a risk probability around 0.10, significantly lower than the base value (0.60), meaning the model predicts this pregnancy as low risk. Importantly, the SHAP explanation confirms that this prediction is primarily driven by normal fetal heart rate, adequate hemoglobin level, and normal blood pressure, while demographic factors such as maternal age and parity exert only mild upward pressure on the risk score. This local interpretability provides valuable insight for clinicians: it explains why the model predicts this particular pregnancy as low risk and identifies the main clinical variables supporting that decision. Such transparency increases trust and facilitates model adoption in clinical decision support systems.

IV. CONCLUSION

In this study, we addressed the dual challenge of achieving high predictive accuracy and ensuring model interpretability for the critical task of pregnancy risk stratification. Our comprehensive analysis demonstrated that while several machine learning models can effectively classify patient risk from clinical data, a Random Forest model provides a superior balance of performance and stability, achieving an accuracy of 95.2% and a sensitivity of 92.7% for the high-risk class. More significantly, this work moves beyond a "black box" paradigm by successfully employing Explainable AI (XAI) to demystify the model's internal logic.

The application of SHAP revealed that the model's predictions are driven by clinically relevant factors, with fetal heart rate (djj), systolic blood pressure (td_sistolik), maternal age (usia), and hemoglobin (hb) level emerging as the most influential features. This synergy between predictive power and transparency holds significant implications. For clinical practice, it offers a pathway toward decision-support tools that are not only accurate but also trustworthy, allowing

practitioners to understand and validate the rationale behind an AI-driven recommendation. For research, our framework champions a "glass-box" approach as a new standard for the development of medical AI, where interpretability is considered as crucial as performance itself.

However, we acknowledge the limitations of this study. The models were developed and validated on a single-source dataset, and their generalizability to different demographic populations or clinical settings remains to be tested. Furthermore, our analysis of model robustness against specific high-risk profiles was constrained by the scarcity of such cases in our random test set, highlighting the challenge of evaluating performance on rare conditions. Future work should therefore prioritize several key directions. First, external validation on large, multi-center datasets is essential to confirm the model's real-world utility. Second, future iterations should incorporate systematic hyperparameter tuning and advanced feature selection techniques to potentially enhance performance further.

In conclusion, this research demonstrates that the thoughtful integration of machine learning and explainable AI is pivotal in creating effective and reliable clinical decision-support systems. By providing both accurate predictions and transparent explanations, we can bridge the gap between algorithmic potential and clinical adoption, ultimately contributing to safer and more proactive antenatal care.

Over the past few years, graph neural networks have become powerful and practical tools for machine learning tasks in graph domain. This progress owes to advances in expressive power, model flexibility, and training algorithms. In this survey, we conduct a comprehensive review of graph neural networks. For GNN models, we introduce its variants categorized by computation modules, graph types, and training types. Moreover, we also summarize several general frameworks and introduce several theoretical analyses. In terms of application taxonomy, we divide the GNN applications into structural scenarios, non-structural scenarios, and other scenarios, then give a detailed review for applica- tions in each scenario. Finally, we suggest four open problems indicating the major challenges and future research directions of graph neural networks, including robustness, interpretability, pretraining and com- plex structure modeling.

ACKNOWLEDGMENT

I would like to express my deepest gratitude to the Graduate School, Doctoral Program in Information Systems, Diponegoro University, for their continuous support, motivation, and guidance throughout the process of completing my doctoral dissertation in Information Systems. Their commitment and dedication in encouraging and facilitating this research have played a vital role in the success of this study. I also wish to extend my sincere thanks to the National Research and Innovation Agency (BRIN) for providing the scholarship that greatly supported the completion of my doctoral (Ph.D.) studies in Information Systems.

REFERENCES

[1] Pi, X., Liu, Z., Zhai, J., Khan, A., Shah, M. M., & Hasan, S. (2025). Prediction of high-risk pregnancy based on machine learning. Scientific Reports. https://doi.org/10.1038/s41598-025-00450-3.

[2] Sokou, R., Lianou, A., Lampridou, M., Panagiotounakou, P., Kafalidis, G., Paliatsiou, S., Volaki, P., Tsantes, A. G., Boutsikou, T., Iliodromiti, Z., & Iacovidou, N. (2025). Neonates at Risk: Understanding the Impact of High-Risk Pregnancies on Neonatal Health. Medicina, 61(6), 1077. https://doi.org/10.3390/medicina61061077.

[3] Wu, Y., Wu, Z., Chen, W., Li, K., Li, Z., & Chang, S. (2024). Risk prediction model based on machine learning for

miscarriage in immune-abnormal pregnancies: Addressing limitations of traditional clinical risk assessment tools. Frontiers in Pharmacology, 15. https://doi.org/10.3389/fphar.2024.1366529.

[4] Zhang, Y., Yin, D., & Tang, Q. (2025). Pregnancy with multiple high-risk factors: a systematic evaluation of clinical complexity and outcome associations. Journal of Clinical Obstetrics & Gynecology, 12(3), 145–158.

[5] Magboul, N. E., Elfadel, M. N., & Abdalla, M. (2025). Artificial Intelligence applications in obstetric risk prediction: A systematic review of machine learning models for preeclampsia. Cureus. https://doi.org/10.7759/cureus.368882.

[6] Giaxi, P., Kousi, M., Drosou, E., & Charalampous, K. (2025). Artificial Intelligence and Machine Learning: An Updated Review in Obstetrics and Midwifery. Journal of Clinical and Translational Science, 17(3): e80394. doi:10.7759/cureus.80394.

[7] Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of Clinical Epidemiology, 110, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004.

[8] "An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review." (2024). Information, 15(4), 235. https://doi.org/10.3390/info15040235.

[9] Ranjbar, A., Montazeri, F., Ghamsari, S. R., Mehrnoush, V., Roozbeh, N., & Darsareh, F. (2024). Machine learning models for predicting preeclampsia: a systematic review. BMC Pregnancy and Childbirth, 24, Article 6. https://doi.org/10.1186/s12884-023-06220-1.

[10] Hou, J., & Lu Wang, L. (2025). Explainable AI for Clinical Outcome Prediction: A Survey of Clinician Perceptions and Preferences. AMIA Joint Summits on Translational Science Proceedings. https://doi.org/10.34116/AMIA.JTSS.2025.215.

[11] Pallepogula, D. R., Bethou, A., Ballambatu, V. B., Dorairajan, G., Saya, G. K., Kamalakannan, S., Karra, S., & Sandhya, K. (2021). A systematic review of antenatal risk scoring systems in India to predict adverse neonatal outcomes. The Journal of Obstetrics and Gynecology of India, 71(5), 399–408. https://doi.org/10.1007/s13224-021-01484-z.

[12] Ferreira, A., Oliveira, R., Marques, R., & Dias, C. C. (2023). Risk Scoring Systems for Preterm Birth and Their Predictive Accuracy: A Systematic Review. Journal of Clinical Medicine, 12(13), 4360. https://doi.org/10.3390/jcm12134360.

[13] Hu, T., Zheng, Q., Li, B., Li, Z., & Zhou, X. (2022). Establishment of a model for predicting the outcome of induced labor based on machine learning algorithms. Scientific Reports, 12, Article 21954. https://doi.org/10.1038/s41598-022-21954-2.

[14] Islam, M. N., Islam, M. M., Kabir, M. H., Karim, M. E., Tania, M., Hasan, M., & Bhowmik, A. (2022). Machine learning to predict pregnancy outcomes: a systematic review. BMC Pregnancy and Childbirth, 22, Article 594. https://doi.org/10.1186/s12884-022-04594-2.

[15] Hoffman, M., Bastek, J., Gordon, L., & Srinivas, S. (2022). Machine learning algorithm using clinical data and electronic health records to predict preterm birth in nulliparous women. American Journal of Obstetrics & Gynecology, 227(1), 94.e1–94.e14. https://doi.org/10.1016/j.ajog.2021.12.084.

[16] Zhang, H., Li, J., Wang, D., Li, H., & Xu, G. (2024). Ensemble machine learning framework for predicting maternal health risk classification. Scientific Reports, 14, Article 71934. https://doi.org/10.1038/s41598-024-71934-x.

[17] Ding, L., Yin, X., Wen, G., Sun, D., Xian, D., Zhao, Y., Zhang, M., Yang, W., Chen, W. (2024). Prediction of preterm birth using machine learning: a comprehensive analysis based on large-scale preschool children survey data in Shenzhen, China. BMC Pregnancy and Childbirth, 24, Article 810. https://doi.org/10.1186/s12884-024-06980-4.

[18] Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. Electronics, 8(8), 832. https://doi.org/10.3390/electronics8080832.

[19] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778.

[20] Lundberg, S. M., Erion, G., & Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. arXiv. https://doi.org/10.48550/arXiv.1802.03888.

[21] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774. https://doi.org/10.5555/3295222.3295230.

[22] Li, Z., Liu, X., & Sun, Y. (2023). Integrating machine learning and the SHapley additive explanations (SHAP) framework to predict lymph node metastasis in gastric cancer patients. NPJ Precision Oncology, 7(1), 12. https://doi.org/10.1038/s41525-023-00302-4.

[23] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324.

[24] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785.

[25] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018.