frontiers | Frontiers in Global Women's Health

TYPE Original Research
PUBLISHED 10 February 2025
DOI 10.3389/fgwh.2025.1456238

EDITED BY
Stephen Kennedy,
University of Oxford, United Kingdom

REVIEWED BY
Igor Victorovich Lakhno,
Kharkiv National Medical University, Ukraine
Mesfin Abebe,
Dilla University, Ethiopia
Tsion Mulat Tebeje,
Dilla University, Ethiopia

*CORRESPONDENCE
Tirualem Zeleke Yehuala
✉ sarazeleke3@gmail.com

RECEIVED 12 September 2024
ACCEPTED 16 January 2025
PUBLISHED 10 February 2025

CITATION
Yehuala TZ, Mengesha SB and Baykemagn ND (2025) Predicting pregnancy loss and its determinants among reproductive-aged women using supervised machine learning algorithms in Sub-Saharan Africa.
Front. Glob. Womens Health 6:1456238.
doi: 10.3389/fgwh.2025.1456238

Predicting pregnancy loss and its determinants among reproductive-aged women using supervised machine learning algorithms in Sub-Saharan Africa

Tirualem Zeleke Yehuala*, Sara Beyene Mengesha and Nebebe Demis Baykemagn

Department Health Informatics, Institute of Public Health, College of Medicine and Health Sciences, University of Gondar, Gondar, Ethiopia

Background: Pregnancy loss is a significant public health issue globally, particularly in Sub-Saharan Africa (SSA), where maternal health outcomes continue to be a major concern. Despite notable progress in improving maternal health, pregnancy-related complications, including s due to miscarriages, stillbirths, and induced abortions, continue to impact women's health, social wellbeing, and economic stability in the region. This study aims to identify the key predictors of pregnancy loss and develop effective predictive models for pregnancy loss among reproductive-aged women in SSA.
Methods: We derived the data for this cross-sectional study from the most recent Demographic and Health Survey of Sub-Saharan African countries. Python software was used to process the data, and machine learning techniques such as Random Forest, Decision Tree, Logistic Regression, Extreme Gradient Boosting, and Gaussian were applied. The performance of the predictive models was evaluated using several standard metrics, including the ROC curve, accuracy score, precision, recall, and F-measure.
Result: The final experimental results indicated that the Random Forest model performed the best in predicting pregnancy loss, achieving an accuracy of 98%, precision of 98%, F-measure of 83%, ROC curve of 94%, and recall of 77%. The Gaussian model had the lowest classification accuracy, with an accuracy of 92.64% compared to the others. Based on SHPY values, unmarried women may be more likely to experience pregnancy loss, particularly in contexts where premarital pregnancies are stigmatized. The use of antenatal care and family planning services can significantly impact the risk of pregnancy loss. Women from lower-income backgrounds may face challenges in accessing prenatal care or safe reproductive health services, leading to higher risks of loss. Additionally, higher levels of education are often correlated with increased awareness of family planning methods and better access to healthcare, which can reduce the likelihood of unintended pregnancy loss.

Conclusion: The Random Forest machine learning model demonstrates greater predictive power in estimating pregnancy loss risk factors. Machine learning can help facilitate early prediction and intervention for women at high risk of pregnancy loss. Based on these findings, we recommend policy measures aimed at reducing pregnancy loss Sub-Saharan African countries.

KEYWORDS

prediction, reproductive-aged women, pregnancy loss, machine learning, Sub-Saharan Africa

Introduction

Pregnancy loss refers to the end of a pregnancy before the fetus can survive outside the womb. It can occur spontaneously or be induced for medical or personal reasons. The main types of pregnancy loss include miscarriage, abortion, and stillbirth (1, 2). Pregnancy loss is a significant public health issue globally, particularly in Sub-Saharan Africa (SSA) (3–5). Worldwide, there are thought to be 73 million induced abortions performed annually. Approximately 45% of abortions are unsafe, and 97% of those occur in developing nations. In 2022, the global late adolescent birth rate for girls was estimated to be 1.5 per 1,000 women, with higher rates in SSA (6).

In Sub-Saharan Africa, the rate of unwanted pregnancies was 91 per 1,000 women, compared to 35 per 1,000 in Europe and North America for women aged 15–49 (7). Between 1990 and 1994 and 2015–2019, the proportion of unintended pregnancies resolved through abortion increased by 26% in Middle Africa, 44% in Eastern and Western Africa, and 72% in Southern Africa (8).

One of the major reasons for high fertility rates in most SSA countries is that a significant number of women in SSA do not use contraception (9). One of the main causes of maternal deaths and morbidities is unsafe abortion (10). It may result in issues with women's physical and mental health as well by way of financial and social hardships for them, their communities, and the impact on families and healthcare systems (11).

Numerous studies had revealed barriers to pregnancy loss, such as insufficient understanding of family planning techniques and their accessibility, poor quality and restricted family planning service availability, and exorbitant expenses associated with family planning techniques, services, travel, and time (12–16).

Furthermore, previous research has found that the most important predictors of lower odds of pregnancy loss were having knowledge of modern contraceptive methods, marriage, and women with higher education (7, 17, 18). Moreover, studies showed that women from lower-income backgrounds had faced challenges in accessing prenatal care or safe reproductive health services, leading to higher risks of (19).

Currently, there is a gap in the literature based on recent data from East African countries. Classical statistical methods Although these models can provide valuable insights, their reliance on predefined rules and assumptions limits them (19–21), which can be limited in capturing complex, nonlinear interactions between variables, lack of automated methods for model selection and optimization, lack of flexibility to efficiently handle high-dimensional data, and patterns.

In contrast, machine learning models have the potential to improve predictive accuracy by learning from larger, more diverse datasets and identifying hidden patterns that traditional models may overlook (22, 23). Additionally, Machine learning (ML) models, on the other hand, have gained attention for their ability to automatically learn from data, capture these complex patterns, and make more accurate predictions (24).

The ultimate goal of this study is to contribute to improved reproductive health outcomes in Sub-Saharan Africa by providing actionable insights into the predictors of pregnancy. Through the use of machine learning, we aim to create a data-driven tool that can help health professionals predict and address the needs of at-risk populations while also offering evidence to support more effective and equitable care. Therefore, by overcoming the above limitations', this study aimed to predict pregnancy loss and identify its determinants among women of reproductive age in SSA, using a machine learning algorithm using the most recent demographic and health survey dataset.

Methods and materials

Study design and study period

This study adopted a design science approach for further analysis of the DHS, which was conducted from 2012 to 2023. The design science approach focuses on creating and evaluating artifacts (such as models, methods, constructs, or systems) to solve complex, real-world problems. Unlike traditional research, which may focus solely on theory-building or hypothesis testing, the design science approach emphasizes the design and innovation of solutions to address specific issues or needs (25) show in Figure 1.

Study setting

This study used the most recent dataset from the Demographic and Health Survey (DHS) for SSA. The dataset spans from 2012 to 2023, and the study covers 36 selected SSA. sub-Saharan Africa is regionally classified into West Africa, Southern Africa, East Africa,

Abbreviations

AUC, area under the curve; DHS, demographic and health surveys; DT, decision trees; XGB, extreme gradient boosting; ML, machine learning; RF, random forests; RFE, recursive feature elimination; SMOTE, synthetic minority oversampling techniques; SSA, Sub-Saharan Africa; IR, women's record.

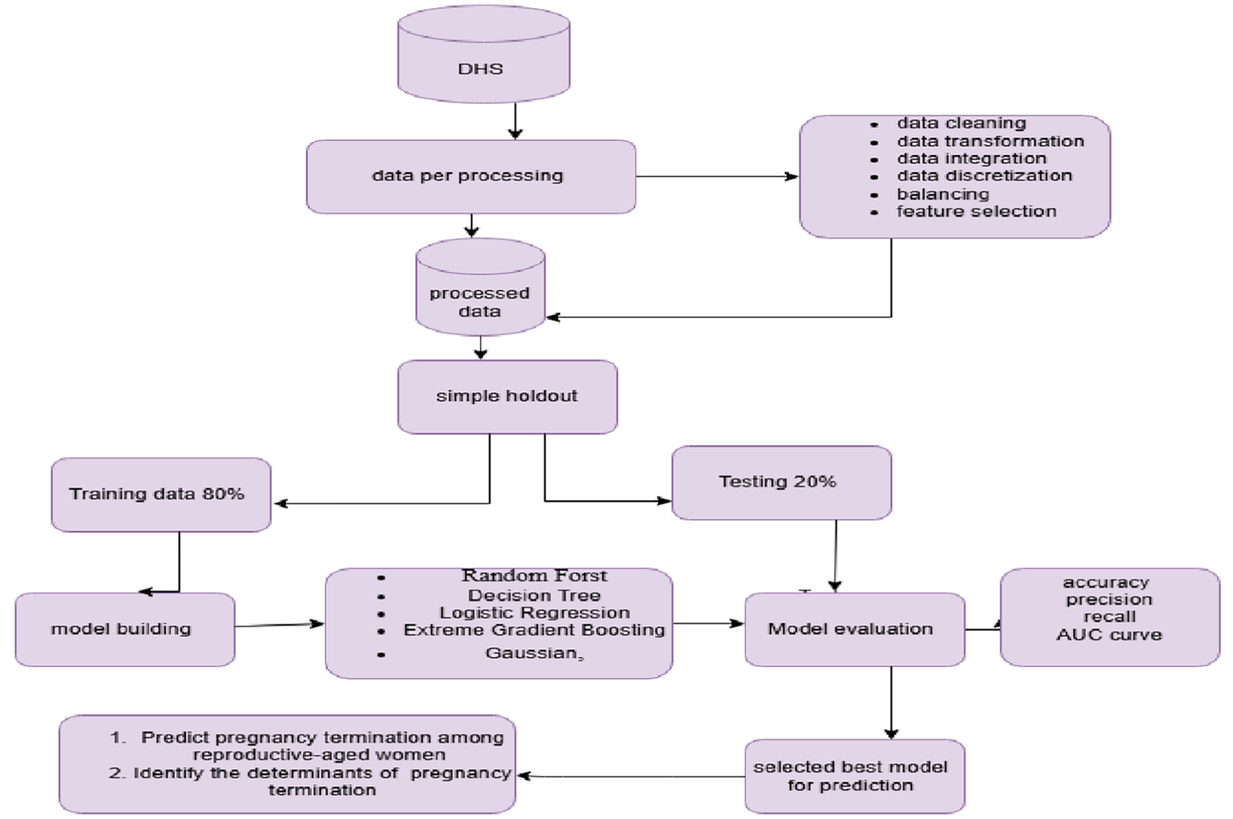

FIGURE 1
Flow diagram for data analysis and model building to predict pregnancy loss.

and Central Africa: East Africa (Burundi, Comoros, Ethiopia, Kenya, Madagascar, Malawi, Mozambique, Rwanda, Tanzania, Uganda, Zambia, and Zimbabwe), West Africa (Burkina Faso, Benin, Côte d'Ivoire, Ghana, Guinea, Liberia, Mali, Nigeria, Niger, Sierra Leone, Senegal, and Togo), Southern Africa (Lesotho, Namibia, Eswatini, and South Africa), and Central Africa (Angola, Democratic Republic of the Congo, Cameroon, Chad, Gabon, São Tomé and Príncipe, and Zambia).

Population, and eligibility criteria

This study included reproductive-age (15–49 years) women who had experienced pregnancy loss in the selected enumeration areas at the time of DHS data collection.

Data source

This study used the most recent DHS datasets obtained from the DHS Program website (http://www.dhsprogram.com) after permission was granted, following the submission of the study's justification and project title. The Women's Record (IR) dataset was used for this investigation. A two-stage probability sampling method, stratified by geographic region and urban/rural areas within each region, was applied to select study participants, ensuring the sample fully represents the target population of Sub-Saharan African countries. After data preparation, a weighted sample of 425,810 reproductive-age women was included in the study.

Sample size and sampling procedure

This study used a weighted sample of 425,810 reproductive-age women. Due to the non-proportional distribution of the sample size across different regions, variations between urban and rural areas, and potential differences in response rates, sampling weights were applied to maintain representativeness. Participants were selected using a two-stage stratified cluster sampling procedure.

In the first stage, Enumeration Areas (EAs) were randomly selected based on their clusters. A stratified sample of census EAs from both urban and rural areas was chosen with complete household listings, using systematic probability sampling. This sampling was based on the sampling frame containing population and household information from the Population and Housing Census (PHC). In the second stage, households within the selected EAs were chosen using equal probability systematic sampling. In each selected household, reproductive-aged women were interviewed using an individual questionnaire (26).

Study variables and measurements

Dependent variable

The dependent variable in this study was pregnancy loss among reproductive-aged women who experienced either miscarriage, abortion, or stillbirth that was dichotomized into two: "Yes" = 1 (if a woman encountered at least one of the three within 5 years preceding each survey, she is considered to have terminated pregnancy) and "No" = 0 if none of these events occurred.

Independent variables

In this study, we employed the fast recursive feature selection method to select the independent features.

Operational definition

Pregnancy loss

Pregnancy loss is generally used to refer to the ending of a pregnancy, whether intentional or accidental. It encompasses a broad range of situations, including miscarriage, abortion, and stillbirth.

Unsafe abortion

It refers to the of a pregnancy performed under conditions that do not meet established medical standards of care, leading to an increased risk of complications such as infection, hemorrhage, and injury to the reproductive organs.

Stillbirth

It refers to the death of a fetus at or after 20 weeks of gestation, but before or during birth. This includes any fetal death occurring before or during the labor and delivery process.

Miscarriage, which is referred to as spontaneous abortion**,** is the early loss of a fetus before the 20th week of gestation.

Data analysis procedure

This study used data from the DHS of SSA to predict pregnancy loss and identify its factors among women of reproductive age. We utilized Python and several key libraries, including Pandas, scikit-learn, imbalance-learn (lmblearn), numPy, and matplotlib, for data preparation, model development, model construction, model evaluation, and model deployment. In conclusion, we developed a predictive model that forecasts pregnancy loss and identifies its determinants.

Data pre-processing

Data processing is a machine-learning technique that transforms raw data into an understandable format (27). In this study, we employed the major data preprocessing steps, which included dimensionality reduction, data transformation, data discretization, data integration, and data cleaning. The advantages of data preprocessing include improving model accuracy through tasks such as data cleaning, exploratory data analysis, normalization, dimensionality reduction, data transformation, and data integration, all of which can positively impact the model's performance (28).

Data cleaning

In this study, we employed data cleaning to address outlier values, imbalanced outcome variables, noise, and missing values. Raw data cannot be directly used in the model testing and training process, as it often contains missing values, which can lead to biased or inaccurate results (29). A large fraction of the 425,810 records had fewer than 643 missing data points for the critical features, comprising nearly 6.6% of the dataset. Our dataset contains missing values for both categorical and continuous variables. Features such as maternal age: 3.2%, birth order: 2.3%, and wealth index: 1.1% have missing values. We used mean imputation for continuous variables and mode imputation for categorical variables to address these missing values.

Outliers are data points that differ significantly from other observations in the dataset. In this study, we identified outliers using visualization techniques such as box plots and Z-scores. To remove outliers, we use methods for identifying outliers by using the Z-score. The threshold is a Z-score greater than 3 or less than −3. This indicates that the data point is more than 3 standard deviations away from the mean and is therefore an outlier.

Feature selection

Feature selection is the process of removing irrelevant or redundant features during the development of a predictive model (30). Feature selection methods were applied during data preprocessing to achieve more efficient data. Furthermore, this study includes 48 features, making feature selection essential. Excessive features are time-consuming and resource-intensive, so it is important to speed up model building and improve the model's performance (31). As a result, identifying the most important features associated with pregnancy loss is a fundamental step. In this study, we employed Recursive Feature Elimination (RFE) to identify the most relevant variables for predicting pregnancy loss. Since RFE infers the relevance of features by estimating their importance through the algorithm, it selects the most important features.

Data transformation

Data transformation involves converting the data into a format suitable for analysis; this may include changing the data's types, scaling, normalizing, and renaming (32). In this study, we used the one-hot encoding technique to convert string data into integers, ensuring a uniform data type for machine learning classifiers. Before building the model, we also scaled the dataset to standardize it, making it suitable for analysis and improving model training and evaluation.

Data discretization

Data discretization is the method of converting continuous data (numerical values) into discrete categories or intervals (33). In this study, we employ binning as a technique to enhance the interpretability and performance of classification algorithms by transforming continuous input into categorical input. Data discretization was used, which limits the impact of outliers, makes it easier to analyze, and reduces noise by transforming continuous variables into categorical features to make the data easier to understand and analyze. For example, reproductive mother's age is continuous; attributes were discretized into 15–24, 25–34, and 35–49 according to DHS guidelines (34, 35).

Data standardization and data integration

This study uses the "standard Scaler ()" library of "sklearn" to normalize the data in order to avoid scaling issues for distance-based learning approaches like logistic regression and gaussian. In this study, we used 36 sub-Saharan African countries DHS datasets, integrated them based on identification variables, sorted both data files by the identification variables, determined the base (primary) file, and finally merged them using Python software.

Class balancing

Before training the prediction model, an unbalanced dataset was balancing (resampled); this might be viewed as a data preparation step (36). Synthetic Minority Oversampling (SMOTE) was used to balance the training data in order to prevent machine learning models from being biased toward the majority class (37). We utilized SMOTE oversampling by creating synthetic examples (new observations) that resemble the minority class by interpolating between minority classes samples in the feature space rather than creating exact copies of existing examples. SMOTE has been shown to almost continually increase classification model performance for resampling imbalanced datasets (38).

Model selection

The predicted variable in this study was binary classification, since pregnancy loss was divided into two "yes" or "no." For model building, four classifiers: fandom forests, Boost, Gaussian, and decision trees were used. The algorithms were chosen in accordance with previous research that used machine-learning methods to classify tasks (39, 40). Machine learning has been widely used to predict childbirth mode and assess potential maternal risks during pregnancy (24). The rationale behind the selected algorithm in this study is its ease of implementation, interpretability, training efficiency, reduction of over fitting, and speed in predicting unknown records (41, 42).

Our aim for this study is to apply ML for pregnancy loss among reproductive-aged women and to provide insight for the government and policymakers. Random forests, XGBoost, decision trees, and Gaussian were used to identify the predictors due to their favorable prediction performance in prior research (24, 43, 44).

Decision tree

We're using an entirely novel integrated supervised learning algorithm developed for this study to effectively handle enormous volumes of survey data. Because it is easy to use and predictable, the study's methodology combines theory and practice in a novel and innovative way. One of the most widely used methods for representing predictions is the decision tree (45). Because decision trees are easily interpretable and can efficiently deal with large, complicated datasets without imposing a complicated parametric structure, they can be quite powerful when used in ensemble algorithms and are resistant to outliers.

Random forest

Random forest is a machine learning algorithm that ensembles multiple decision trees to make predictions for classification and regression problems (46). The concept of multiple random tree generation is used in each split decision, along with a voting system, sample bagging, training bootstrapping, and randomly selected features. Random Forest overcomes these limitations of decision trees by using an ensemble of decision trees (46). Due to an ensemble of models, bootstrap aggregating, sometimes known as bagging, increases prediction accuracy and stability.

Gradient boosting

Extreme Gradient Boosting (XGBoost) is a decision-tree-based ensemble machine learning algorithm that uses a gradient boosting framework to improve the speed and efficiency of boosted tree algorithms (47). It can manage the issue of over fitting, but scalability on larger datasets is a concern due to its sensitivity to outliers.

Data splitting

In this study, we utilized a simple holdout method, with 80% (340,648 samples) for training and 20% (85,162 samples) for testing, to ensure robust model evaluation. The dataset was divided into training and testing sets, with the training set used to build and train the model and the testing set used to evaluate its performance on previously unseen data.

Model evaluation

The performance of the trained model is evaluated using the testing dataset. Then, the performance of the trained models was evaluated using the test set based on the criteria of accuracy score, AUC curve, precision (P), recall (R), and F-measure as follows: The confusion matrix is a matrix of N * N, where N is the number of predicted classes, and it displays the number of correct (48) and incorrect predictions made by the classification model relative to the target value. Subsequently, the test set was used to assess the trained models' performance using the accuracy score, AUC curve, precision (P), recall (R), and

TABLE 1 Socio demographic characteristics of the study participant, evidences from DHS (*N* = 425,810).

Variable	Frequency (%)	Pregnancy loss(PT)	
		YES	NO
Marital status			
Never married	116,903 (64%)	2,784 (0.6%)	114,119 (27%)
Married	273,548 (28%)	23,855 (6%)	249,693 (58%)
Widowed, divorced and separated	35,359 (8)	2,424 (0.5)	32,935 (8)
Contraceptive use			
Yes	342,402 (80%)	23,178 (5%)	319,224 (75%)
No	83,408 (20%)	5,885 (12%)	319,224 (18%)
Watching TV			
Yes	240,766 (56%)	14,827 (4%)	225,939 (53%)
No	185,044 (44%)	14,236 (10%)	170,808 (34%)
Parity			
Null parity	114,951 (27%)	4,194 (1%)	110,757 (26%)
Prim parity	59,203 (14%)	5,445 (2%)	53,758 (12%)
Multi parity	143,454 (34%)	11,945 (3%)	131,509 (31%)
Grand parity	108,202 (25)	7,479 (2%)	100,723 (23%)
Maternal age			
15–24 years	169,095 (40%)	8,239 (1%)	160,856 (39%)
25–34 years	133,148 (31%)	12,191 (3%)	120,957 (27%)
35–49 years	123,567 (29%)	8,633 (2%)	114,934 (27%)
Independence			
Living Alone	43,076 (10)	3,830 (1%)	39,246 (9%)
Living with jointly with husband	100,939 (24%)	8,863 (2%)	92,075 (21%)
Living with others	281,796(66%)	1,670(4%)	265,426(62%)

F-measure criteria.

$$\text{Precision} = (\text{TP})/(\text{TP} + \text{FP}) \tag{1}$$

$$\text{Recall} = (\text{TP})/(\text{TP} + \text{FN}) \tag{2}$$

$$\text{F–Measure} = (2{*}\text{Precision}{*}\text{Recall})/(\text{Precision} + \text{Recall}) \tag{3}$$

$$\text{Accuracy} = ((\text{TP} + \text{TN})/\text{TP} + \text{TN} + \text{FP} + \text{FN})) \times 100 \tag{4}$$

Results

Socio demographic characteristics of the study participant

This study investigated a sample of 425,810 reproductive-aged women from 36 countries in Sub-Saharan Africa, all of which were part of a Demographic and Health Survey. The results showed that 7% of women had a pregnancy loss, while 93% had not. As shown in Table 1 women who had never married (0.6%) had a lower prevalence of pregnancy loss compared to women who were married, who had a higher pregnancy loss rate (6%). Among women who had experienced a pregnancy loss, approximately 5% (22, 49) of women who had used contraceptives experienced pregnancy loss, compared to 12% (5,885) of women who had not used contraceptives. Furthermore, women who were exposed to watching TV (53%) had a lower rate of pregnancy loss, whereas 34% of women who did not watch TV had a higher prevalence of pregnancy loss (10%).

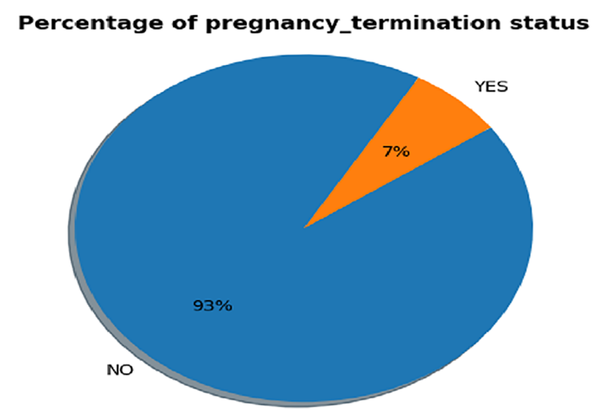

FIGURE 2
Prevalence of pregnancy loss status among reproductive-aged women.

Class balancing

In this study, 29,806 (7%) women had a pregnancy loss, while 396,004 (93%) did not, as shown in Figure 2. To prevent machine learning models from being biased toward the majority class, the training data was balanced using the SMOTE. The synthetic minority oversampling technique was applied to overcome

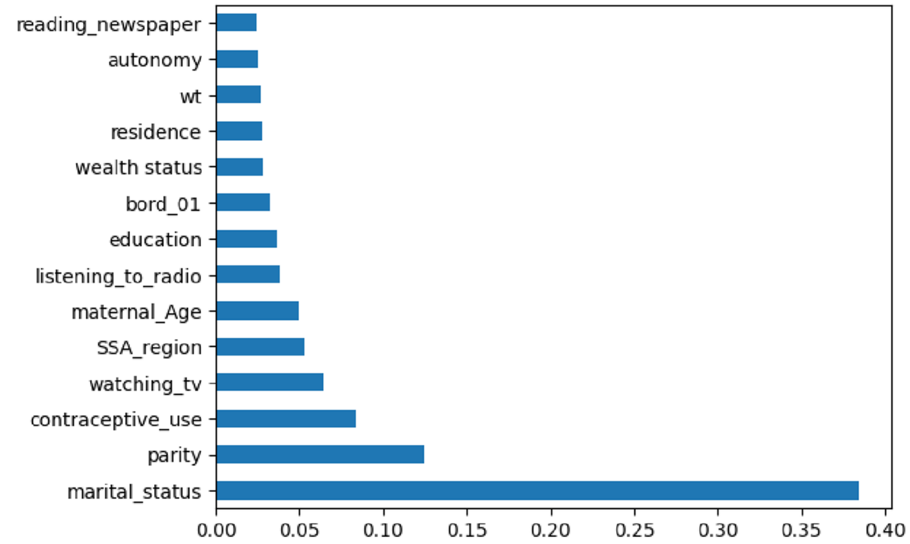

FIGURE 3
Important features selected by recursive feature elimination (RFE).

sampling bias by generating new, synthetic observations that approximate the minority class. This is done by interpolating between minority-class samples in the feature space rather than simply replicating existing examples. As a result, the distribution of the classes was balanced, creating an equal and symmetric distribution in each category, which helps build more reliable predictive models. We obtained a balanced sample of women who had a pregnancy loss with counts 396,004 and not with count 396,004.

Importance feature selection

Feature selection and variable importance ranking are techniques used to identify a subset of relevant features by removing irrelevant or redundant ones. The importance of feature selection lies in its ability to reduce the cost of learning by decreasing the number of features, improving model performance, and minimizing storage and computation time. The dataset contains 48 features. To select the most important ones, we employed RFE. This approach provides flexibility in controlling the number of features retained and can effectively handle correlated features. As shown in Figure 3, the selected features included marital status, parity, contraceptive use, wealth status, residence, birth order, maternal age, reading newspapers, listening to the radio, country, watching TV, weight, and independence. These determinants were used for model building.

Comparisons of selected machine learning model

In order to identify determinants and predict pregnancy loss among reproductive-aged women, the study used supervised machine learning techniques such as logistic regression, decision tree (DT), XGB (Extreme Gradient Boosting), and Gaussian Naive Bayes for analysis. The study compared these machine learning methods using the same testing parameters to build predictive models. The evaluation metrics included accuracy, Area under the Curve (AUC), precision, recall, and F-measure. Since Random Forest (RF) performed the best overall, it was chosen as the top model. The results are displayed in Figure 4 and Table 2. Random Forest achieved an accuracy of 98%, precision of 98%, F-measure of 83%, ROC curve of 94%, and recall of 77%. Additionally, Random Forest had a high true positive rate (99.3%), a low false positive rate (2%), a true negative rate of 94.6%, and an extremely low false negative rate (0.006%). The AUC for Random Forest was 94%. The Decision Tree model had an accuracy of 97.19%, precision of 82%, recall of 78%, F-measure of 80%, and an ROC curve of 90%. Among the proposed models, the Gaussian Naive Bayes classifier

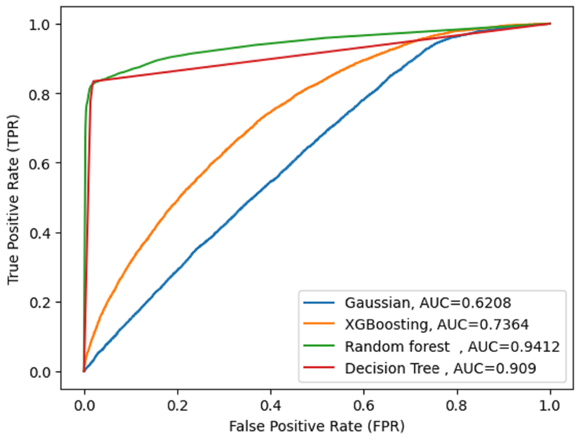

FIGURE 4
AUC curve score of machine learning model.

performed the least well, with an accuracy of 92.9%, precision of 12.3%, recall of 50%, F-measure of 10%, and an AUC curve of 62%.

Model explanation using SHAP values

Interpreting the results of machine learning algorithms can be significantly more challenging than traditional statistical analysis methods. It is often difficult to understand how predictions are made. However, techniques like SHAP (Shapley Additive Explanations), proposed by Lundberg and Lee, provide a unified framework for interpreting the outputs of a wide range of machine learning models. SHAP values are used to gain insights into the contributions of individual features to the model's predictions, helping to explain how each feature influences the final decision (20).

In this study, we employed the Random Forest classifier in combination with the model-agnostic SHAP technique to identify the most significant predictors of pregnancy loss. By evaluating the mean absolute SHAP values across the dataset, we were able to identify the primary predictors for women who had experienced a miscarriage, abortion, or stillbirth.

The positive contributions (in red) features for women had none of a miscarriage, abortion or stillbirth

As shown in Figure 5, where the SHAP values are positive, the features contribute to women who have not experienced a miscarriage, abortion, or stillbirth. This is represented by the red line, indicating the category coded as ‘1’ or a high value. These features tend to increase the predicted likelihood of not terminating a pregnancy. Examples of these features include being a married woman, being in the 25–34 maternal age range, having a second birth order, having secondary or higher education, watching TV, and using modern contraception methods. In total, five features have a positive impact on reproductive-aged women who have not experienced miscarriage, abortion, or stillbirth.

The negative contributions (in blue) features for women had experienced a miscarriage, abortion or stillbirth

As shown in Figure 6, where the SHAP values are negative, this is represented by the blue line, indicating the category coded as ‘0’ (no) or a low value. Increasing the value of these features tends to increase the likelihood of pregnancy loss (such as a miscarriage, abortion, or stillbirth). For example, women who are uneducated, have no experience with watching TV, have given birth to two children, are in the 15–24 maternal age range, do not use modern contraception methods, and have no media exposure appear to have a negative impact on pregnancy outcomes.

Discussion

This study used a classification machine learning method to compare, identify, and recognize specific risk factors related to

TABLE 2 Accuracy, precision, recall and F-measure for the machine learning algorithms.

ML model	Accuracy	Precision	Recall	F-measure	AUC
Random forest(RF)	98%	98%	77%	83%	94%
DecisionTree (DT)	97.19%	82%	78%	80%	90%
XGB(Extreme Gradient Boosting)	93.17%	92%	18%	3.7%	73%
Gaussian	92.92%	12.3%	0.5%	10%	62%

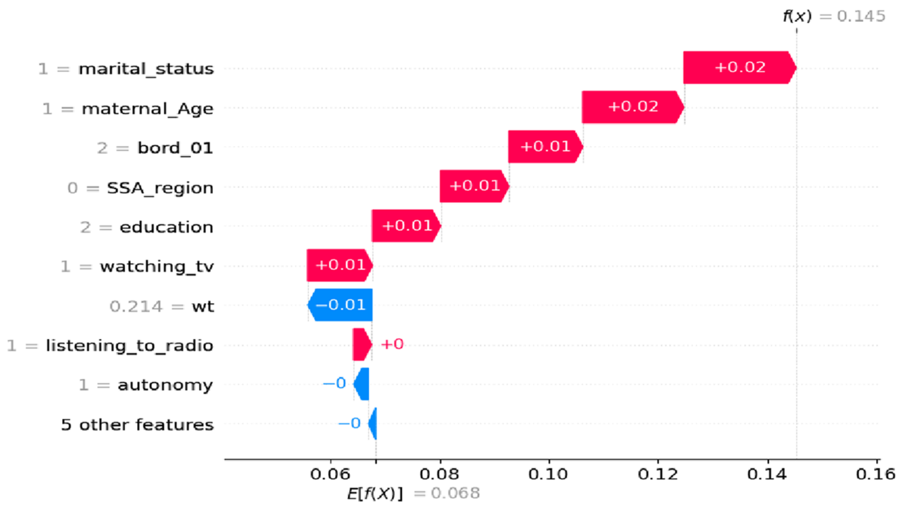

FIGURE 5
Waterfall plot displaying prediction of the 5th index observation high value.

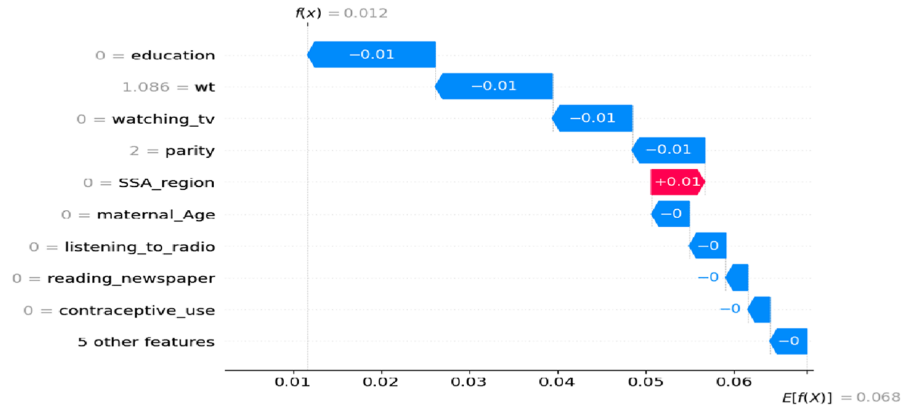

FIGURE 6
Waterfall plot displaying prediction of the zero index observation for low values.

pregnancy loss among reproductive-aged women in SSA, which could be targeted for intervention. When compared to other machine learning classifiers, such as Decision Tree (DT), Extreme Gradient Boosting (XGB), and Naïve Bayes Gaussian, the Random Forest (RF) model demonstrated the best predictive power for identifying risk factors for pregnancy loss. It achieved an accuracy of 98.0%, recall of 77.0%, precision of 98.0%, F1 score of 98.0%, and AUC of 94%. The feature importance analysis, performed using RFE, identified the following as critical predictors of pregnancy loss in sub-Saharan Africa countries: marital status, parity, contraceptive use, wealth status, place of residence, birth order, maternal age, reading newspapers, listening to the radio, watching TV, weight, and independence.

Some of these variables had already been proven to be predictors of pregnancy loss in previously published studies. The first important feature of SSA predictors of pregnancy loss was marital status. We found that never married women were more likely to report having a pregnancy loss compared with those who were married. This could be due to a tendency toward a delay in women's age at marriage, which has been suggested to result in increased sexual activity among never-married women and raise their risk of unintended pregnancy. These women may also obtain a pregnancy because they feel that having a child would interfere with future opportunities. Yet, fear of school dropout was mentioned as a primary reason for demanding an induced abortion service among women of reproductive age in SSA (15, 20). Younger women and those who are unmarried were found to have higher probabilities of pregnancy loss (50).

With regard to parity, multiparous women were less likely to experience pregnancy loss compared to primiparous women. This may be because women without children are more likely to be teenagers, and the likelihood of unwanted pregnancies, which often result in abortion, is higher among young women, due in part to an unmet demand for family planning (17, 51). Limited access to contraception and family planning services was a major factor influencing the decision to terminate a pregnancy. Areas with high unmet need for family planning, especially in rural regions, saw higher rates of pregnancy loss, as women may not have had the opportunity to prevent unintended pregnancies in the first place (52).

Women who had media exposure were less likely to experience pregnancy loss compared to those without. This could be due to the fact that the media plays a vital role in broadcasting information about how and where to end a pregnancy. Additionally, women exposed to media may be better informed about abortion-related principles and are potentially less likely to face societal stigma (19, 53).

Women with lower incomes had higher odds of having an induced abortion compared to those with higher incomes (20, 54, 55), thus, the low contraceptive utilization among women with lower income may account for the higher odds of induced abortion in this study. In turn, studies have demonstrated that women who use contraceptives are less likely to have an induced abortion compared to those who do not (17, 20, 56).

This study showed that pregnant women who had secondary and above educational levels were more likely to undergo pregnancy loss as compared with those with no education. A similar study conducted in Ethiopia showed that the likelihood of pregnancy loss in uneducated women was 1.5 times lower than in women who attended elementary school, 1.5 times lower than in women who attended secondary school, and 1.8 times lower than in women who attended higher education (56). The reason for this is that women who have knowledge about the fertile period are 35% less likely to have an unsafe abortion compared to those with no knowledge of the fertile period. This is because educated women are more likely to use modern contraceptive methods to prevent unwanted pregnancies (57).

The goal of this study was to use supervised machine learning algorithms to predict pregnancy loss and identify the key determinants among reproductive-aged women in Sub-Saharan Africa. By leveraging the potential of machine learning models, we aimed to provide insights that can assist in improving maternal health policies, healthcare services, and intervention strategies in this SSA.

Strength and limitations

This study attempted to forecast pregnancy loss and more accurately evaluate the key predictors. Also, this study made use of the recent DHS data set of sub-Saharan Africa countries, which contains almost every demographic risk group that is vulnerable. However, this study has certain limitations because the DHS data collection is self-reported, which may have introduced some information biases.

Conclusion and recommendation

Machine learning can play a significant role in predicting pregnancy loss and understanding the underlying factors among reproductive-aged women in Sub-Saharan Africa. By leveraging data on demographics, health, and socio-economic status, predictive models can help policymakers and healthcare workers to better anticipate and address pregnancy-related complications, improve maternal health outcomes, and allocate resources more effectively. The finding is very important to identifying factors such as lack of prenatal care or socio-economic barriers that can help design more effective prevention programs. However, future research could explore the integration of more granular and real-time data sources, such as electronic health records and mobile health applications, to enhance prediction accuracy. Additionally, expanding the scope of machine learning models to account for cultural, legal, and regional variations in pregnancy-related outcomes could further improve the applicability and generalizability of these models across diverse populations in SSA.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: http://www.dhsprogram.com.

Ethical statement

The study was conducted after obtaining a permission letter from https://www.dhsprogram.com, following an online request to access the DHS data for Sub-Saharan countries. This request was approved after reviewing the submitted brief descriptions of the survey to the DHS program. The datasets were handled with the utmost confidentiality. This study utilized secondary data from the DHS of sub-Saharan countries. Ethics considerations regarding informed consent, confidentiality, anonymity, and privacy were addressed by the DHS office. We did not alter or manipulate the microdata except for the purposes of this study. There was no patient or public involvement in this research, and ethical clearance was not required.

Author contributions

TZY: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. SM: Data curation, Formal Analysis, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing. NB: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

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

Acknowledgments

We are extremely appreciative of the MEASURE DHS authority for providing us with all the information and materials we required to complete this task. We also want to express our sincere gratitude to our friends for their insightful critiques, supportive words, and recommendations, all of which have been extremely helpful in getting this work done.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Huss B. Well-being before and after pregnancy termination: the consequences of abortion and miscarriage on satisfaction with various domains of life. *J Happiness Stud.* (2021) 22(6):2803–28. doi: 10.1007/s10902-020-00350-5

2. Assefa N, Berhane Y, Worku A, Tsui A. The hazard of pregnancy loss and stillbirth among women in Kersa, east Ethiopia: a follow up study. *Sex Reprod Healthc.* (2012) 3(3):107–12. doi: 10.1016/j.srhc.2012.06.002

3. Organization WH. *Safe Abortion: Technical and Policy Guidance for Health Systems*. Geneva: World Health Organization (2003).

4. Fathalla MF. Safe abortion: the public health rationale. *Best Pract Res Clin Obstet Gynaecol.* (2020) 63:2–12. doi: 10.1016/j.bpobgyn.2019.03.010

5. Organization WH. *WHO Recommendations on Self-Care Interventions: Self-Management of Medical Abortionm*. Geneva: World Health Organization (2022).

6. Riley T. ADDING IT UP-Investing in Sexual and Reproductive Health 2019. (2020).

7. Ayalew HG, Liyew AM, Tessema ZT, Worku MG, Tesema GA, Alamneh TS, et al. Prevalence and factors associated with unintended pregnancy among adolescent girls and young women in Sub-Saharan Africa, a multilevel analysis. *BMC Women's Health.* (2022) 22(1):1–9. doi: 10.1186/s12905-022-02048-7

8. Bearak J, Popinchalk A, Ganatra B, Moller A-B, Tunçalp Ö, Beavin C, et al. Unintended pregnancy and abortion by income, region, and the legal status of abortion: estimates from a comprehensive model for 1990–2019. *Lancet Glob Health.* (2020) 8(9):e1152–e61. doi: 10.1016/S2214-109X(20)30315-6

9. Tsui AO, Brown W, Li Q. Contraceptive practice in Sub-Saharan Africa. *Popul Dev Rev.* (2017) 43(Suppl Suppl 1):166. doi: 10.1111/padr.12051

10. Fawcus SR. Maternal mortality and unsafe abortion. *Best Pract Res Clin Obstet Gynaecol.* (2008) 22(3):533–48. doi: 10.1016/j.bpobgyn.2007.10.006

11. Singh S. Global consequences of unsafe abortion. *Women's Health.* (2010) 6(6):849–60. doi: 10.2217/WHE.10.70

12. Kibira SP, Karp C, Wood SN, Desta S, Galadanci H, Makumbi FE, et al. Covert use of contraception in three Sub-Saharan African countries: a qualitative exploration of motivations and challenges. *BMC Public Health.* (2020) 20:1–10. doi: 10.1186/s12889-020-08977-y

13. Schivone GB, Blumenthal PD. *Contraception in the Developing World: Special Considerations. Seminars in Reproductive Medicine.* New York: Thieme Medical Publishers (2016).

14. Tyrer LB. Obstacles to use of hormonal contraception. *Am J Obstet Gynecol.* (1994) 170(5):1495–8. doi: 10.1016/S0002-9378(94)05010-6

15. Aliyu AA, Dahiru T. Factors associated with safe disposal practices of child's faeces in Nigeria: evidence from 2013 Nigeria demographic and health survey. *Niger Med J.* (2019) 60(4):198–204. doi: 10.4103/nmj.NMJ_3_19

16. Cramer DW, Wise LA. The epidemiology of recurrent pregnancy loss. *Semin Reprod Med.* (2000) 18(4):331–40. doi: 10.1055/s-2000-13722

17. Ahinkorah BO. Individual and contextual factors associated with mistimed and unwanted pregnancies among adolescent girls and young women in selected high fertility countries in Sub-Saharan Africa: a multilevel mixed effects analysis. *PLoS One.* (2020) 15(10):e0241050. doi: 10.1371/journal.pone.0241050

18. Xue T, Guan T, Geng G, Zhang Q, Zhao Y, Zhu T. Estimation of pregnancy losses attributable to exposure to ambient fine particles in south Asia: an epidemiological case-control study. *Lancet Planet Health.* (2021) 5(1):e15–24. doi: 10.1016/S2542-5196(20)30268-0

19. Yogi A, K C P, Neupane S. Prevalence and factors associated with abortion and unsafe abortion in Nepal: a nationwide cross-sectional study. *BMC Pregnancy Childbirth.* (2018) 18:376. doi: 10.1186/s12884-018-2011-y

20. Ahinkorah BO. Socio-demographic determinants of pregnancy termination among adolescent girls and young women in selected high fertility countries in Sub-Saharan Africa. *BMC Pregnancy Childbirth.* (2021) 21(1):1–8. doi: 10.1186/s12884-020-03485-8

21. Ahinkorah BO. Intimate partner violence against adolescent girls and young women and its association with miscarriages, stillbirths and induced abortions in Sub-Saharan Africa: evidence from demographic and health surveys. *SSM Popul Health.* (2021) 13:100730. doi: 10.1016/j.ssmph.2021.100730

22. Venkataramanan S, Sadhu AKR, Gudala L, Reddy AK. Leveraging artificial intelligence for enhanced sales forecasting accuracy: a review of AI-driven techniques and practical applications in customer relationship management systems. *Aust J Mach Learn Res App.* (2024) 4(1):267–87.

23. Sokolov AN, Pyatnitsky IA, Alabugin SK. Research of classical machine learning methods and deep learning models effectiveness in detecting anomalies of industrial control system. *Global Smart Industry Conference (GloSIC); 2018: IEEE* (2018).

24. Islam MN, Mustafina SN, Mahmud T, Khan NI. Machine learning to predict pregnancy outcomes: a systematic review, synthesizing framework and future research agenda. *BMC Pregnancy Childbirth.* (2022) 22(1):1–19. doi: 10.1186/s12884-022-04594-2

25. Hevner AR, March ST, Park J, Ram S. Design science in information systems research. *MIS Q.* (2004) 28:75–105. doi: 10.2307/25148625

26. Rutstein SO, Rojas G. Guide to DHS statistics. *Calverton, MD: ORC Macro.* (2006) 38:78.

27. Bhatnagar R. Machine learning and big data processing: a technological perspective and review. *The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018)* (2018).

28. Kang M, Tian J. *Machine Learning: Data Pre-processing. Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things*. Hoboken, NJ: Wiley (2018). p. 111–30.

29. Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B, Tabona O. A survey on missing data in machine learning. *J Big Data.* (2021) 8(1):1–37. doi: 10.1186/s40537-021-00516-9

30. Venkatesh B, Anuradha J. A review of feature selection and its methods. *Cybernetics and Information Technologies.* (2019) 19(1):3–26. doi: 10.2478/cait-2019-0001

31. Kira K, Rendell LA. *A Practical Approach to Feature Selection. Machine Learning Proceedings 1992.* Elsevier (1992). p. 249–56.

32. Engels R, Theusinger C. Support for data transformation in machine learning applications. *Proceedings of the 10th ECML Workshop on Upgrading Learning to the Meta-level: Model Selection and Data Transformation* (1998).

33. Liu L, Wong AK, Wang Y. A global optimal algorithm for class-dependent discretization of continuous data. *Intelligent Data Analysis.* (2004) 8(2):151–70. doi: 10.3233/IDA-2004-8204

34. Croft T, Marshall AM, Allen CK, Arnold F, Assaf S, Balian S, et al. *Guide to DHS Statistics: DHS-7 (version 2).* Rockville, MD: ICF (2020).

35. Croft TN, Marshall AM, Allen CK, Arnold F, Assaf S, Balian S. *Guide to DHS Statistics.* Rockville: ICF (2018). p. 645.

36. Ramyachitra D, Manikandan P. Imbalanced dataset classification and solutions: a review. *Int J Comp Bus Res.* (2014) 5(4):1–29.

37. Elreedy D, Atiya AF. A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. *Inf Sci (NY).* (2019) 505:32–64. doi: 10.1016/j.ins.2019.07.070

38. Fernández A, Garcia S, Herrera F, Chawla NV. SMOTE For learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *J Artif Intell Res.* (2018) 61:863–905. doi: 10.1613/jair.1.11192

39. Yehuala TZ, Derseh NM, Tewelgne MF, Wubante SM. Exploring machine learning algorithms to predict diarrhea disease and identify its determinants among under-five years children in east Africa. *J Epidemiol Glob Health.* (2024) 14:1–11. doi: 10.1007/s44197-024-00259-9

40. Deeba F, Patil SR. Utilization of machine learning algorithms for prediction of diseases. *2021 Innovations in Power and Advanced Computing Technologies (I-PACT)* (2021).

41. Deo RC. Machine learning in medicine. *Circulation.* (2015) 132(20):1920–30. doi: 10.1161/CIRCULATIONAHA.115.001593

42. Shehab M, Abualigah L, Shambour Q, Abu-Hashem MA, Shambour MKY, Alsalibi AI, et al. Machine learning in medical applications: a review of state-of-the-art methods. *Comput Biol Med.* (2022) 145:105458. doi: 10.1016/j.compbiomed.2022.105458

43. Lee SM, Nam Y, Choi ES, Jung YM, Sriram V, Leiby JS, et al. Development of early prediction model for pregnancy-associated hypertension with graph-based semi-supervised learning. *Sci Rep.* (2022) 12(1):15793. doi: 10.1038/s41598-022-15391-4

44. Shah D, Patel S, Bharti SK. Heart disease prediction using machine learning techniques. *SN Comput Sci.* (2020) 1:1–6. doi: 10.1007/s42979-020-00365-y

45. Song Y-Y, Ying L. Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry.* (2015) 27(2):130. doi: 10.11919/j.issn.1002-0829.215044

46. Ali J, Khan R, Ahmad N, Maqsood I. Random forests and decision trees. *Int J Comput Sci Res.* (2012) 9(5):272.

47. Ali ZA, Abduljabbar ZH, Taher HA, Sallow AB, Almufti SM. Exploring the power of eXtreme gradient boosting algorithm in machine learning: a review. *Acad J Nawroz Univ.* (2023) 12(2):320–34. doi: 10.25007/ajnu.v12n2a1612

48. Poola I. The best of the machine learning algorithms used in artificial intelligence. *International Journal of Advanced Research in Computer and Communication Engineering ISO 3297: 2007 Certified.* (2017) 6(10). doi: 10.17148/IJARCCE.2017.61032

49. Lubo-Robles D, Devegowda D, Jayaram V, Bedle H, Marfurt KJ, Pranter MJ. Machine learning model interpretability using SHAP values: application to a seismic facies classification task. *SEG International Exposition and Annual Meeting* (2020).

50. Henshaw SK. Unintended pregnancy in the United States. *Fam Plann Perspect.* (1998):24–46. doi: 10.2307/2991522

51. Gunawardena N, Fantaye AW, Yaya S. Predictors of pregnancy among young people in Sub-Saharan Africa: a systematic review and narrative synthesis. *BMJ Glob Health*. (2019) 4(3):e001499. doi: 10.1136/bmjgh-2019-001499

52. Sensoy N, Korkut Y, Akturan S, Yilmaz M, Tuz C, Tuncel B. Factors affecting the attitudes of women toward family planning. *Fam Plann*. (2018) 13(33):2. doi: 10.5772/intechopen.73255

53. Onukwugha FI, Magadi MA, Sarki AM, Smith L. Trends in and predictors of pregnancy termination among 15–24 year-old women in Nigeria: a multi-level analysis of demographic and health surveys 2003–2018. *BMC Pregnancy Childbirth*. (2020) 20(1):550. doi: 10.1186/s12884-020-03164-8

54. Hailegebreal S, Enyew EB, Simegn AE, Seboka BT, Gilano G, Kassa R, et al. Pooled prevalence and associated factors of pregnancy termination among youth aged 15–24 year women in east Africa: multilevel level analysis. *PLoS One*. (2022) 17(12):e0275349. doi: 10.1371/journal.pone.0275349

55. Bennett SM, Litz BT, Lee BS, Maguen S. The scope and impact of perinatal loss: current status and future directions. *Prof Psychol*. (2005) 36(2):180. doi: 10.1037/0735-7028.36.2.180

56. Yemane GD, Korsa BB, Jemal SS. Multilevel analysis of factors associated with pregnancy termination in Ethiopia. *Ann Med Surg*. (2022) 80:104120. doi: 10.1016/j.amsu.2022.104120

57. Woldeamanuel BT. Assessment of determinant factors of pregnancy termination among women of reproductive age group in Ethiopia: evidence from 2016 Ethiopian demographic and health survey. *Int J Sexual Reprod Health Care*. (2019) 2(1):010–5. doi: 10.17352/ijsrhc.000005