Quality & Quantity
https://doi.org/10.1007/s11135-025-02210-x

# Predicting high risk pregnancies in Pakistan- a demographic assessment using predictive machine learning

**Sara Rizvi Jafree**[1] · **Mian Muhammad Mubasher**[2]

Accepted: 21 May 2025


**Abstract**
Pakistan is unable to meet its maternal and child health targets. Predictive machine learning has the potential to predict high risk pregnancies based on data from women who have had a miscarriage or stillbirth. This would help advise better healthcare plans at primary and tertiary level and help achieve Sustainable Development Goal targets in the country. The aim of this study was to evaluate several machine learning models to measure their ability to detect high risk pregnancies. The Pakistan Demographic Health Survey (2018) has been used which includes data from 15,068 women across Pakistan. Fourteen machine learning classifiers have been employed to predict high risk pregnancies, with the following evaluation metrics reported: precision, recall, false positive rate (FPR), accuracy, and F1-score. We find that five models have the highest overall performance: (i) Deep Neural Network, (ii) SELU Network, (iii) Multilayer Perceptron, (iv) Gradient Boosting, and (v) AdaBoost, exhibiting near good precision (73.0-76.0%), effective recall (83.0-86.0%), robust accuracy (89.0-90.0%), and decent F1-Scores (79.0-80.0%). This study recommends the integration of low-cost online models to predict high risk pregnancies as a critical tool to help achieve maternal health targets in the country.

**Keywords** Machine learning · High risk pregnancies · Pakistan · Artificial intelligence · Healthcare

✉ Sara Rizvi Jafree
sarajafree@fccollege.edu.pk

Mian Muhammad Mubasher
mubasher.pucit@pu.du.pk

[1] Department of Sociology, Forman Christian College University, Lahore, Pakistan

[2] Department of Information Technology, University of the Punjab, Lahore, Pakistan

Published online: 01 June 2025

## 1 Introduction

Over the years, Pakistan's health sector has made much progress in attempting to achieve maternal health targets, with some scholars agreeing that the Lady Health Worker program, which provides door-step services for maternal health, receives more attention and funding than other health issues in the country (Jafree and Barlow 2023). Another constructive effort to improve maternal and child health care has been the opening of Maternal and Child Care Health Units across the country, with patient databases transferred to the computerized Health Information System (Anwar et al., 2023). Moreover, technology has been used to introduce telemedicine and mobile health units in remote and rural areas of the country to support disadvantaged women for maternal health literacy and health access (Jafree et al. 2023). However, COVID-19 and regional instability have reversed much of the gains and maternal health outcomes in the country remain below the global targets (Ali et al. 2020). The maternal mortality and morbidity rates in Pakistan remain one of the highest in South Asia, with the maternal mortality ratio estimated at 186 deaths per 100,000 live births (Shaeen et al. 2022).

One integral factor contributing to maternal and child deaths is the significant prevalence of high-risk pregnancies, and the inability to detect such pregnancies on time (Habib et al. 2017). Studies in Pakistan have focused on interventions to improve pregnancy health outcomes, such as: medicine trials, training traditional birth attendants, and predicting psychosocial predictors for antenatal stress (Waqas et al. 2020; Mir et al. 2012; Jokhio et al. 2005). However, less attention has been given to making predictions for high-risk pregnancies (Nisar et al. 2016). Early detection of high-risk pregnancies is crucial for implementing preventive measures and providing timely medical and social interventions, which would significantly reduce maternal mortality and morbidity (Ramakrishnan et al. 2021). Recent scholarship suggests that majority of maternal and child deaths could be avoided with timely diagnosis of high-risk pregnancies using Artificial Intelligence and Machine Learning (Khan et al. 2022).

The global north is moving towards improving health services by using existing data to train algorithms using machine learning methods to predict high risk pregnancies and preterm birth, stillbirth, miscarriage, and fetal health (Ngiam and Khor 2019). Artificial Intelligence can be used to train a machine learning model on health and socio-demographic data, collected through health records, and to identify women who are at risk of pregnancy complications. These predictive models can help develop targeted interventions and prevent both maternal and child mortality. There are various basic models for classification and prediction purposes (Katarya and Srinivas 2020), and it is important first however to compare different machine learning models best suited for the available health dataset in Pakistan. Apart from the benefits of a low-cost solution, with predictions being generated rapidly on large data sets (Schadt et al. 2010), machine learning prediction provides solutions to detect high risk pregnancies without clinical interventions, which is an important concern in a country like Pakistan which is culturally conservative and frowns on too many mediations with women of reproductive years (Omer et al. 2021).

## 2 Literature review

Several scholars have recently leveraged machine learning to enhance maternal health outcomes (Islam et al. 2022). Select studies are discussed below in context to our research and to present evidence for the relevance of our study. In one study machine learning was used to predict early stillbirth, late stillbirth, and preterm birth pregnancies in the USA (Koivu and Sairanen 2020). Two data-sets were used, one for observation and the other for external validation, with algorithms such as logistic regression, artificial neural network, and gradient boosting decision tree used to construct individual classifiers. The results provided a solid foundation for risk prediction. In another recent study, from the USA, machine learning was used to detect patients at increased risk for hypertensive disorders during pregnancy (Shara et al. 2024). Electronic health records were used with the machine learning algorithm assessing risk factors selected by clinical experts in cardio-obstetrics. The algorithm was iteratively trained using relevant literature and current standards of risk identification. Use of predictive Artificial Intelligence showed stronger performance in early risk detection of myocardial infarction supporting its use for early detection of cardiovascular conditions during pregnancy.

At the same time developing countries have not been left behind and datasets have been used to show the potential of machine learning for improving maternal health. In a study in Zanzibar, researchers used program data from a community health worker program to predict if newly enrolled pregnant woman would deliver in a health facility (Fredriksson et al. 2022). Four machine learning methods- logistic regression, LASSO regularized logistic regression, random forest, and an artificial neural network, were used to correctly predict the delivery location for 68–77% of the women in the test set. The random forest model accurately identified 74.4% of women delivering at home. In yet another study two algorithms were combined to improve the accuracy and efficiency of risk classification in pregnant women (Ojo and Adedokun 2023). Artificial neural networks (ANN) and random forest (RF) algorithms were used. Data from Bangladesh was employed and was divided into training and testing sets, with 75% of the data used for training and 25% used for testing. Results showed that the proposed model achieved 95% accuracy, 97% precision, 97% recall, and an F1 score of 0.97 on the testing dataset; confirming that the deep hybrid model has the potential to improve the accuracy and efficiency of maternal health risk classification in pregnancy.

Researchers from the Philippines used limited data from municipalities to compare multiple supervised machine learning algorithms to analyze and accurately predict high-risk pregnancies (Macrohon et al. 2022). Supervised learning algorithms such as Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbors, Naïve Bayes, and Multilayer Perceptron were evaluated by using 10-fold cross validation to obtain the best parameters with the best scores. The results were in favor of using machine learning for improved detection of high-risk pregnancies, and that Decision Tree showed better outcomes and attained a test score of 93.70%. Another study from Turkey used existing data of pregnant women to develop a diagnostic system with artificial intelligence for the early diagnosis of preeclampsia (Bulez et al. 2024). Artificial intelligence models were created using Python, scikit-learn, and TensorFlow, with results showing that the model achieved 73.7% sensitivity, 92.7% specificity, 90.6% accuracy, and an area under the curve value of 0.832. This

study also concluded that artificial intelligence is effective in the prediction and diagnosis of preeclampsia.

At the same time, some studies have highlighted concerns of predicting maternal health risks through predictive Artificial Intelligence. A study from India concluded that there are many dataset related challenges, such as missing data, incorrectly collected data, and improperly labeled variables, which can compromise predictive accuracy (Trivedi et al. 2019). In a similar vein, a systematic literature review confirmed that the reporting and methodological quality of machine learning-based prediction models for maternal health risks were poor, thus recommending that guidelines should be developed for the design, conduct, and reporting of such studies (Yang et al. 2023). Overall, review of these studies helped us recognize the growing role of machine learning in maternal health risk prediction, especially for a developing country like Pakistan plagued by low health budget and bleak maternal health outcomes.

### 2.1 Aim of study

In lieu of the above, this study aims to evaluate several machine learning models to measure their ability to detect high risk pregnancies based on prenatal stage variables. By analyzing large amounts of data, Artificial Intelligence algorithms can identify patterns that help predict and prevent potential complications during pregnancy. Identifying high risk pregnancies through machine learning can improve maternal and child healthcare targets in the primary and tertiary health sectors of Pakistan and assist in meeting SDGs. Such predictive models will further assist healthcare professionals in allocating resources effectively, providing targeted interventions, and developing personalizing care plans based on individual needs. Since it is very hard to predict beforehand the performance of an algorithm on a given dataset, it was important to compare several machine learning algorithms (Wolpert and Macready 1997). In this study, we included the following machine learning algorithms: Logistic Regression, Random Forest, Gradient Boosting, Ada Boost, Bayesian Network, Decision Tree, Multilayer Perceptron, Support Vector Machine, K Nearest Neighbor, Light GBM, Deep Neural Network, SELU Network, Average Ensemble, and Weighted Average Ensemble. All these models have been considered because no single model fits all scenarios. Some excel at handling class imbalances and missing values, while others are more computationally robust. Certain models offer greater interpretability, whereas others function more as black boxes. Ideally a model should be computationally robust, accurate, and noise tolerant (Liu et al. 2022), but this cannot be predetermined without actually assessing the models, which this study attempts to do.

## 3 Methodology

### 3.1 Data acquisition

We have used the approach called supervised learning, which involves training a machine learning model using historical data on maternal health, including medical records and demographic information. The model learns to recognize patterns in the data and make predictions based on those patterns. This allows the model to discover hidden patterns within

the data. By considering various factors like maternal age, body mass index (BMI), blood pressure, and maternal history such as antenatal practices, the model can identify clusters of women who may be at a higher risk for stillbirth or miscarriage.

The underlying raw data used to train the machine learning models is surveyed and acquired by the Demographic and Health Surveys (DHS) Program, 2018, (PDHS 2018). This nation-wide maternal and child health survey is collected by the National Institute of Population Studies (NIPS), with technical support from ICF and the Pakistan Bureau of Statistics, and financial support from the United States Agency for International Development. The women's survey is collected using a multistage, stratified sampling design and includes information data about health indicators of ever married women of reproductive years between 15 and 49 years, and collects a wide-range of data on (i) Socio-demographic characteristics (such as age, education), (ii) Pregnancy history and child mortality, (ii) Knowledge, use, and source of family planning methods, (iv) Antenatal, delivery, and postnatal care, (v) Vaccinations and childhood illnesses, (vi) Breastfeeding and infant feeding practices, (vii) Marriage and sexual activity, (viii) Fertility preferences, (ix) Women's work and husbands' background characteristics, (x) Knowledge, attitudes, and behaviour related to other health issues (e.g., smoking, tuberculosis, hepatitis), and (xi) Domestic violence.

In the 2018 PDHS survey, 15,068 women from all over Pakistan have been interviewed, including representation from urban and rural areas and all the provinces or semi-provinces, including: Punjab, Sindh, Khyber Pakhtunkhwa, and Balochistan; Gilgit Baltistan (GB); Azad Jammu and Kashmir (AJK), and the former Federally Administrated Tribal Areas (FATA). Women of reproductive years, who were either permanent residents of the selected households or visitors who stayed in the households the night before the survey were interviewed. The processing of the PDHS data being saved electronically begins simultaneously during the data collection process. All electronic data files are transferred to the NIPS central office in Islamabad and registered and checked for inconsistencies, incompleteness, and outliers. Secondary editing is carried out in the central office, which involves resolving inconsistencies and coding the open-ended questions. The PDHS core team members assist with the secondary editing, which secures the likelihood of the data being error-free and accurate. The final cleaning of the data set was carried out by the DHS Program data processing specialist. The final data files in SPSS files are available to researchers publicly and free of cost.

### 3.2 Data Preparation

In each interview of the PDHS data file at least 5,331 features have been logged. Supplementary File 1 summarizes the variables that were selected using the PDHS data, pertaining to high-risk pregnancies. The acquired data was decomposed in metadata (SPSS file) and raw data source (DAT files). The meta data (SPS file) can be used to load and investigate the raw data source (DAT file). For this study, PSPP, a free and open-source software (FOSS), has been used to prepare the data source. The metadata and raw data have been unified into a single SAV file using PSPP for further analysis and processing. Further data analytics is performed using, Google Colab, where a machine learning pipeline is developed in Python. The data SAV file was loaded into the Python environment using the pyreadstat library.

### 3.3 Data preprocessing

Once the complete dataset was loaded in machine learning integrated development environment, preprocessing was performed which involved data cleaning and conversion. Initially, the dataset contained several variables with missing values in the target variable- 'last pregnancy outcome' (total sample size = 13,451); which had five response categories: (1) Live birth ($n$ = 11,644); (2) Stillbirth ($n$ = 237); (3) Miscarriage ($n$ = 1,366); (4) Opted to abort baby ($n$ = 203). Therefore, the total sample size for the pooled dataset included only women who had listed a stillbirth or miscarriage in their last pregnancy at the time of the interview, which was 1,603 women. Salient socio-demographics showing representation of sample across Pakistan is summarized in Supplementary File 2. The target variable was converted to a Boolean format, where miscarriage and stillbirth were mapped to 'True' and all other categories to 'False'. This transformation is performed to facilitate the attribute ranking and binary classification task. To ensure the completeness of the dataset, all missing values in all the features and selected independent variables were replaced with minus one (-1). This imputation method is chosen to simplify the analysis and to maintain consistency across the dataset.

### 3.4 Attribute ranking

For attribute evaluation purposes the Fisher score algorithm has been used. The Fisher Score, also called the Fisher Discriminant Ratio, is a feature selection method. By calculating the ratio of variation between classes to variance within classes, it assesses the relative relevance of each feature. Features that optimize this ratio are chosen to develop machine learning models because they are thought to be more discriminative (Yan et al. 2022). Once the score for each attribute is computed these scores are normalised in zero-one range.

### 3.5 Attribute selection

The Fisher's score algorithm was executed on the dataset, with the procedure returning the high-ranked features listed in Table 1. It is observed that, with most other variables pertaining to prenatal assistance from different types of healthcare providers showing high scores. In addition, the total number of children ever born, delivery by caesarean section, births in last five years, weight of mother, place of delivery, also show significant scores. Taking of blood pressure, iron tablets, blood and urine samples also show significant scores. It must be noted that in a culturally conservative climate with majority disadvantaged women, unless there are pregnancy complications, women do not visit doctors or opt for tests during pregnancy (Ahmed et al. 2020). Having a health card, ever receiving vaccination, pregnancy order number, pregnancy losses, mothers age, and timing on decision for caesarean section have lower scores comparatively.

### 3.6 Experimental design and model Building

In this study, performance of fourteen (14) machine learning models / classifier have been evaluated (Fig. 1). These classification algorithms include: (1) Logistic Regression; (2) Random Forest; (3) Gradient Boosting; (4) AdaBoost; (5) Bayes Net; (6) J48; (7) Multilayer

**Table 1** Top predictors of high-risk pregnancies with their relative fisher score

| Predictors | Fisher's Score Value |
|---|---|
| Seeking prenatal care from Lady Health Worker | 0.97 |
| Seeking assistance from Community Midwife | 0.96 |
| Seeking prenatal care from Traditional Birth Attendant | 0.96 |
| Size of child at birth | 0.93 |
| Seeking assistance from Lady Health Visitor | 0.92 |
| Seeking prenatal care from Nurse | 0.85 |
| Seeking prenatal care from Homeopath | 0.83 |
| Total children ever born | 0.79 |
| Last birth a caesarean section | 0.79 |
| Seeking prenatal care from Hakim | 0.79 |
| Births in last five years | 0.77 |
| Birth weight in kilograms (of mother) | 0.76 |
| Seeking assistance from Dai | 0.72 |
| Place of delivery | 0.72 |
| During pregnancy: blood pressure taken | 0.66 |
| During pregnancy: given or bought iron tablets | 0.64 |
| During pregnancy: blood sample taken | 0.61 |
| During pregnancy: urine sample taken | 0.60 |
| Number of tetanus injections before birth | 0.55 |
| Pregnancy order number | 0.50 |
| Has health card | 0.50 |
| Number of pregnancy losses | 0.32 |
| Ever had vaccination | 0.18 |
| Respondent's current age | 0.13 |
| Timing on decision for caesarean section | 0.09 |

Perceptron; (8) Support vector machine (SVM); (9) K-nearest neighbors (KNN); (10) Light gradient boosting machine (LGBM); 11. Deep Neural Network; 12. Scaled Exponential Linear Unit (SELU) Network; 13. Averaged Ensemble; and 14. Weighted Average Ensemble. Figure 1, presents step by step methodology to conduct the study. In the first phase the dataset is preprocessed and necessary cleaning and imputation steps are performed. Later the data has been examined with the help of visualization to better understand the input and output features. Later to systematically identify good feature which can help in predicting the high-risk pregnancy Pearson coloration has been deployed and thirty (30) top performing features are selected for model building. Once input feature is identified the models have been trained and evaluated. A more detailed expression on the methodology is presented at the end of this section starting from data preprocessing section.

All the machine learning models are a piece of computation which takes input vector and maps it to the target variable by incorporating model parameters / weights which are learned from training dataset. In our case the input vector carries variables related to individual pregnancy and output is a Boolean variable indicating whether the pregnancy would be high-risk or not. It is worth mentioning that a machine learning model can be represented by the mathematical abstraction: $f : X \rightarrow Y$

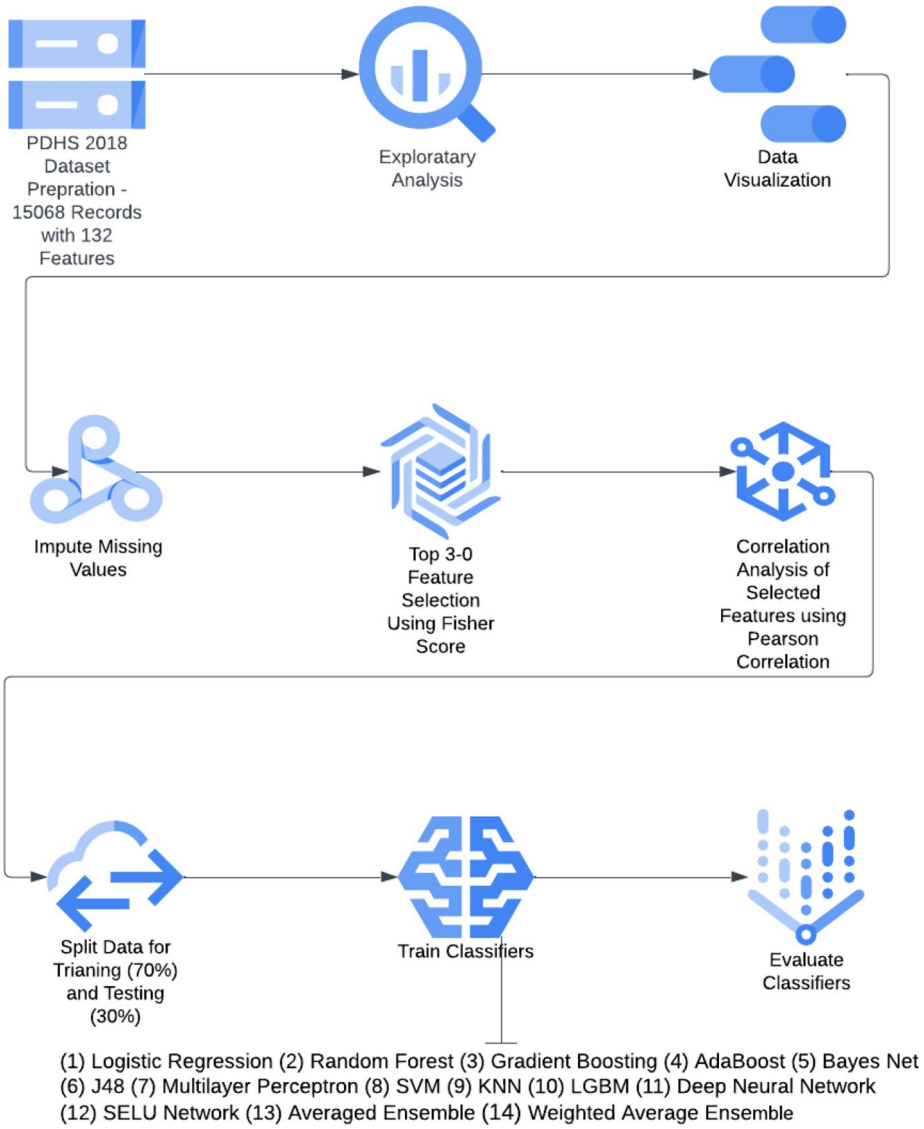


**Fig. 1** High risk pregnancy prediction model

Each model performs computation on the input vector and maps it onto an appropriate output. In the following sections, computation of each model is presented and briefly explained. In all the following equations $X$ denotes the input feature vector and $\widehat{y}$ denotes the predicted output of the model. These equations represent a mathematical abstraction of the computations performed by the respective classifiers, formulated based on their standard principles as described in the literature. While these exact equations may not appear in prior work, they reflect the fundamental computations underlying each classifier. Relevant references are provided for each model.

Equation 1 - Logistic Regression (Cox 1958)

$$\widehat{y} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)}}$$

In Eq. 1, $\beta$ represents parameters of the model.

Equation 2 – Random Forest (Breiman et al. 2001)

$$\widehat{y} = majority\_vote\{h_1(X), h_2(X), \ldots, h_n(X)\}$$

In Eq. 2, $h_i(X)$ represents an individual decision tree, with the mathematical formulation of individual decision tree is presented in Eq. 2.

Equation 3 – Gradient Boosting (Freidman, 2001)

$$\widehat{y} = \sum \widehat{y}_i$$

$$\widehat{y}_i = \widehat{y}_{i-1} + \eta . h_i(X)$$

In Eq. 3, the sum of prediction of all the decision tree at step $i$ for all the decision trees $h_i(X)$ is predicted as final output, in this model parameter $\eta$ refers to the learning rate of the model.

Equation 4 – AdaBoost (Freund and Schapire 1997)

$$\widehat{y} = sign\left(\sum_{m=1}^{M} \alpha_m h_m(X)\right)$$

In Eq. 4, $\alpha_m$ is the weight assigned to the prediction of an individual decision tree $h_m(X)$.

Equation 5 – Bayesian Network (Peal 1985)

$$\widehat{y} = argmax_y P(y|X)$$

In Eq. 5, the model maximizes the conditional probability $P$ to predict the outcome.

Equation 6 – J48 (C4.5 Decision Tree) (Quinlan 1996)

$$\widehat{y} = argmax_{i \in C} N_i$$

$$N_i = \sum_{j=1} \delta(y_j, i)$$

In Eq. 6, $N_i$ represents the count of instances of class $i$ in the lead node. Additionally, $\delta(y_{j,i})$ is an indicator function which outputs 1 if the class label of instance $j$ is equal to $i$, otherwise it outputs 0.

Equation 7 – Multilayer Perceptron (MLP) (Rumelhart et al. 1986)

$$\widehat{y} = \sigma(W^{(L)} a^{(L-1)} + b^{(L)})$$

In Eq. 7, $L$ represents the last layer of the multilayer perceptron architecture where as $\sigma$ is the activation function. Additionally, the vector $W$ and $b$ represents weights and biases of the architecture.

Equation 8 – Support Vector Machine (Cortes and Vapnik 1995)

$$\widehat{y} = sign\left(\sum {}_{i=1}^{N} \alpha_i y_i K\left(X_i, X\right) + b\right)$$

In Eq. 8, $K$ is kernel function and $\alpha_i$ and $b$ are parameters of the model.

Equation 9 – K-Nearest Neighbour (Cover and Hart 1967)

$$\widehat{y} = argmax_{y \in C}\left(\sum {}_{j=1}^{k} \delta\left(y_j, y\right)\right)$$

In Eq. 9, $\delta$ is an indicator function to determine $y_j$ is equal to label $y$ or not and $C$ is the set of all the possible label C.

Equation 10 – LightGBM (Ke et al. 2017)

$$\widehat{y} = \sum \widehat{y}_m$$

$$\widehat{y}_m = \widehat{y}_{m-1} + \eta\, .h_m\left(X\right)$$

In Eq. 10, $\widehat{y}_m$is prediction of model at the mth iteration and $\eta$ is the weight of the model.

Equation 11 – Deep Neural Network (LeCun et al. 2015)

$$\widehat{y} = h^{(L)}\left(X\right)$$

$$h^{(l)}\left(X\right) = f^{(l)}(W^{(l)}.h^{(l-1)}\left(X\right) + b^{(l)})$$

In Eq. 11, $f$ is an activation function, $W$and $b$ are weights and biases respectively, $L$ represents the output layer and $l$ represents the intermediate layer(s).

Equation 12 – SELU Network (Klambauer et al. 2017)

$$f^{(l)}\left(z\right) = \lambda \begin{cases} z & if\ z > 0 \\ \alpha\ \left(e^z - 1\right) & if\ z \leq\ 0 \end{cases}$$

The SELU has the same computation as of Deep Neural Network presented in Eq. 11 except it uses SELU as an activation function. Definition of SELU activation function is presented in Eq. 12. In this computation $\alpha$ is a scaling constant.

Equation 13 – Averaged Ensemble (Opitz and Maclin 1999)

$$\widehat{y} = \frac{1}{M} \sum {}_{m=1}^{M} \widehat{y}_m\left(X\right)$$

In this technique (Averaged Ensemble) an average result of all the participating classifiers is predicted.

Equation 14 – Weighted Average Ensemble (Dietterich 2000)

$$\widehat{y} = \sum_{m=1}^{M} w_m \widehat{y}_m (X)$$

In Eq. 14, the $w_m$ is weight of the m[th] model.

In this study, fourteen (14) different models are trained, out of which three (3) are neural network-based models, these models differ on the base of different activation functions being used by different neural network layers and also differ in complexity sue to the number of layers being employed. The Multilayer Perceptron (MLP) (Eq. 7) and the Deep Neural Network (Eq. 11) uses Rectified Linear Unit (ReLU) in hidden layers and Sigmoid on the output layer as an activation function. They key difference is the ability of MLP to automatically decide the architecture of the neural network. The SELU Network (Eq. 12) uses Scaled Exponential Linear Unit on the hidden layers. The main advantage of SELU is its ability to self-normalization which prevents vanishing and exploding gradients but generally SELU networks are computationally complex in comparison to the ReLU networks.

All these fourteen (14) models are trained using 70% of the data and later prediction of these models is compared with the remaining 30% of the data in the testing phase. These models are evaluated using five (5) performance metrices, namely- Train Error, Test Error, Generalization Gap, Precision, Recall, Accuracy, False Positive (FP) Rate and F1-Score (Table 2).

A simple explanation of these evaluation metrics is as follows: Train Error is the proportion of incorrect predictions made by the model on the training dataset, while Test Error is the proportion of incorrect predictions on the test dataset. The Generalization Gap is the difference between Test Error and Train Error, indicating how well the model generalizes to unseen data. Precision is the proportion of true positive predictions among all positive predictions. Recall is the proportion of true positive predictions among all actual positive instances. FP Rate (False Positive Rate) is the proportion of false positive predictions among all actual negative instances. Accuracy is the proportion of true positive and true negative predictions among all instances. F1-Score is the harmonic mean of Precision and Recall. In the context of this study the True instances are referring to the high-risk pregnancies and False instances are referring to the normal pregnancies.

**Table 2** Metrices for model evaluation

| Metric | Definition |
|---|---|
| Train Error | $\frac{TP_{train}+TN_{train}}{TP_{train}+\mathrm{TN}_{train}+FP_{train}+FN_{train}}$ |
| Test Error | $1 - \mathrm{Accuracy}$ |
| Generalization Gap | *Train Error– Test Error* |
| Precision | $\frac{TP}{TP+FP}$ |
| Recall | $\frac{TP}{TP+FN}$ |
| Accuracy | $\frac{TP+TN}{TP+\mathrm{TN}+FP+FN}$ |
| FP Rate | $\frac{FP}{TN+FP}$ |
| F1-Score | $\frac{2\times\ Precision\times\ Recall}{Precision+Recall}$ |

Notes- TP refers to the True Positives; FP refers to the False Positives; FN refers to the False Negative; T refers to the True instances; and F refers to the False instances

### 3.7 Train-test split

The dataset was initially divided into training and testing sets using a 70:30 split ratio. The training set comprised 70% of the data for model learning, while the remaining 30% was reserved for performance evaluation. The original dataset exhibited significant class imbalance, with 11,644 instances of normal pregnancies (77.3%) and only 3,423 instances of high-risk pregnancies (22.7%). To mitigate this imbalance and improve classifiers' performance, the SMOTEENN technique was applied to the training data. SMOTEENN is a hybrid method that combines the Synthetic Minority Over-sampling Technique (SMOTE) with Edited Nearest Neighbors (ENN). SMOTE generates synthetic examples of the minority class to balance the dataset, while ENN removes noisy or ambiguous samples near class boundaries, resulting in a cleaner and more balanced training set (Batista et al. 2004).

## 4 Model training and evaluation

Each of the fourteen (14) classifiers included in this study have been trained on the training dataset and evaluated on the testing dataset. The threshold for classifying a positive outcome is set at 0.5. The python code for data preprocessing, attribute selection, train-test split, and model training and evaluation and evaluation is provided in the Appendix A.

## 5 Results and discussion

The evaluation metrics have been able to highlight the strengths and weaknesses of each method in predicting high risk pregnancy outcomes. The distribution of outcomes is also analyzed to ensure a balanced representation in both training and testing datasets. The detailed results for each classifier are summarized in Table 3; Fig. 2.

**Table 3** Classification and algorithm analysis with fourteen predictors using the PDHS dataset

| Classifier | Train Error | Test Error | Gen Gap | Precision | Recall | Accuracy | FP Rate | F1-Score |
|---|---|---|---|---|---|---|---|---|
| Logistic Regression | 0.1 | 0.16 | 0.06 | 0.6 | 0.85 | 0.84 | 0.17 | 0.71 |
| Random Forest | 0 | 0.1 | 0.1 | 0.77 | 0.83 | 0.90 | 0.07 | 0.80 |
| Gradient Boosting | 0.02 | 0.1 | 0.08 | 0.76 | 0.86 | 0.90 | 0.08 | 0.80 |
| AdaBoost | 0.04 | 0.11 | 0.07 | 0.73 | 0.86 | 0.89 | 0.09 | 0.79 |
| Bayes Net | 0.25 | 0.29 | 0.05 | 0.43 | 0.84 | 0.71 | 0.33 | 0.57 |
| J48 | 0 | 0.11 | 0.11 | 0.73 | 0.84 | 0.89 | 0.09 | 0.78 |
| Multilayer Perceptron | 0.07 | 0.1 | 0.03 | 0.75 | 0.84 | 0.90 | 0.08 | 0.79 |
| SVM | 0.25 | 0.29 | 0.05 | 0.43 | 0.84 | 0.71 | 0.33 | 0.57 |
| KNN | 0 | 0.13 | 0.13 | 0.67 | 0.83 | 0.87 | 0.12 | 0.74 |
| LGBM | 0 | 0.1 | 0.09 | 0.77 | 0.84 | 0.90 | 0.08 | 0.8 |
| Deep Neural Network | 0.07 | 0.1 | 0.03 | 0.76 | 0.83 | 0.90 | 0.08 | 0.79 |
| SELU Network | 0.09 | 0.1 | 0.02 | 0.75 | 0.83 | 0.90 | 0.08 | 0.79 |
| Averaged Ensemble | 0.02 | 0.09 | 0.08 | 0.77 | 0.84 | 0.91 | 0.08 | 0.8 |
| Weighted Average Ensemble | 0.01 | 0.09 | 0.09 | 0.78 | 0.83 | 0.91 | 0.07 | 0.8 |

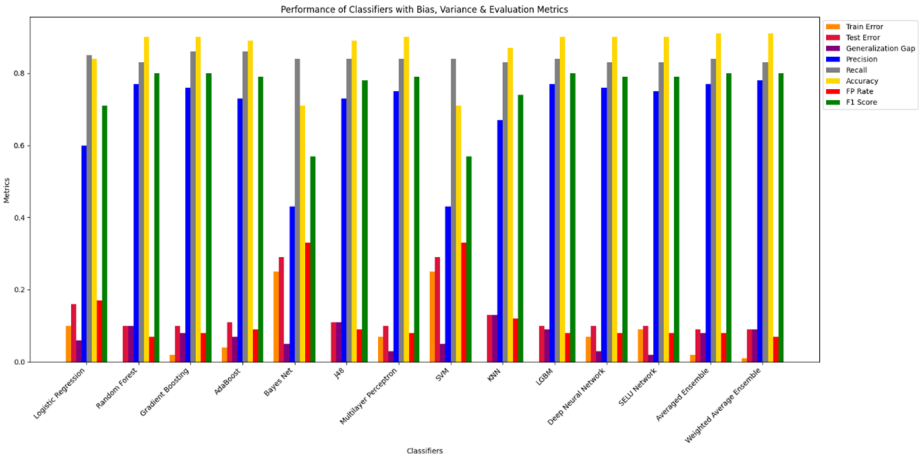


**Fig. 2** Classification and algorithm analysis with fourteen predictors using the PDHS dataset

The evaluation metrics have been able to highlight the strengths and weaknesses of each method in predicting high-risk pregnancy outcomes. The distribution of outcomes is also analyzed to ensure a balanced representation in both training and testing datasets. We find that the following three models have the highest overall performance: (i) Deep Neural Network, (ii) SELU Network, and (iii) Multilayer Perceptron. These models exhibit low train and test errors (around 9–10%), and very small generalization gaps (~0.005–0.007), indicating strong stability and minimal overfitting. They also show high precision (around 78–81%) and balanced recall (~78–80%), resulting in robust accuracy (~91%) and an F1-Score of about 80%. They maintain a very low false positive rate, making them the most suitable choices if the goal is to achieve a balance of high precision and recall along with excellent overall accuracy.

Although Gradient Boosting has slightly higher train error (~2%) compared to the top three models, it maintains a low-test error (~9.6%) and a moderate generalization gap (~7.6%). It demonstrates relatively high recall (around 85%) and high accuracy (~91%), with a decent F1-Score of 80%. This model is a strong alternative if you can accept a minor trade-off in precision (around 76%) for a higher recall. This model also offers a balance between precision and recall, making it a viable option if you're looking for a model that performs well in both aspects.

Closely following Gradient Boosting and AdaBoost is KNN, which shows good precision (~67%), good recall (~83%), and high accuracy (~90%), with a decent F1-Score (~74%). However, KNN exhibits a larger generalization gap (~13%), indicating some risk of overfitting. Similarly, Logistic Regression, Random Forest, and LGBM show adequate to good precision (~67–77%), high accuracy (~90–91%), and decent F1-Scores ranging from 78% (Logistic Regression and Random Forest) to 79% (LGBM), with moderate generalization gaps (4.6-9.6%).

The Weighted Average Ensemble model also shows adequate to good precision (~77%), and with relatively high recall (~83%), high accuracy (~91%), and a decent F1-Score of 80%. The Averaged Ensemble, J48, and SVM models exhibit above average precision (ranging from 75 to 79%), however, SVM has lower recall (58%), affecting its F1-Score

(65%). Comparatively, the Averaged Ensemble model out of the three has better recall (80%), higher accuracy (91%), and a decent F1-Score of 80%. These ensemble models have low train errors but moderate generalization gaps (~8–9%), suggesting reasonable but not exceptional generalization. Finally, we find that Bayes Net demonstrates poor precision (43%) and fair accuracy (71%), despite having decent recall (84%). Its F1-Score of 57% and relatively high false positive rate make it a less favorable choice due to its significant trade-offs in precision and overall accuracy.

While deep learning models (DNN, SELU, MLP) achieve high precision and accuracy with very low generalization gaps, their lower recall suggests they might miss some high-risk cases, which is problematic in medical applications. They are also computationally expensive, making deployment challenging in resource-limited settings. Tree-based models like Gradient Boosting and AdaBoost provide a better balance, offering higher recall at a slight cost to precision, but with slightly higher generalization gaps. However, their interpretability is limited compared to simpler models like Logistic Regression, which surprisingly performs competitively despite its simplicity and moderate generalization gap. Ensemble models do not significantly outperform individual classifiers, suggesting that merely combining similar models is not always beneficial. SVM struggles with recall, likely due to class imbalance, while Bayes Net has a high false positive rate, making it unreliable for precise predictions. Overall, the best model depends on the specific goal: tree-based models are preferable for minimizing false negatives, deep learning models for maximizing precision (if computational resources allow), and simpler models like Logistic Regression when interpretability and moderate generalization are crucial.

As a developing region, which is not competitive in the technology and artificial intelligence industry (Khan et al. 2023), Pakistan faces the grave danger of falling further behind on its targets to achieve health equity and satisfactory maternal and child health ratios. There is need to invest in the implementation of predictive machine learning models to secure health targets, including financial injections for research, technology advancement, skill development, and pilot testing of projects before upscale (Khalid et al. 2022). There is also need for investment in governance and ethics of predictive artificial intelligence in the country, and assurance for data security and health privacy, especially for data related to women and household information (Gilani et al. 2023). Investment, technological growth and innovation, and regulation will require multiple stakeholders and sectors to come together including the state, researchers, academic institutions, and regulatory authorities.

## 6 Conclusion and future direction

Pakistan is in dire need to develop low-cost and rapid solutions to solve multiple health challenges facing the country, inclusive of targeting to reduce maternal and child health mortality ratios. This study concludes that the integration of low-cost online models to predict high risk pregnancies is a critical and effective tool to help achieve health targets in the country. Several machine learning models were used in this study and results advise that for optimal performance in predicting high risk pregnancies, the following five (5) models ae recommended: (i) SELU Network, (ii) Multilayer Perceptron, (iii) Deep Neural Network, (iv) Gradient Boosting, and (v) Weighted Average Ensemble. These five models offer the best combination of precision, recall, accuracy, and F1-Score. The following four models

may also be considered as strong alternatives if slight variations in precision or recall are acceptable- KNN, LGBM, Random Forest, and AdaBoost. The most important implication of this study is that further evaluation of machine learning models must be conducted for other salient health areas in the country such as predicting risks to child mortality, multimorbidity, infectious diseases (specifically tuberculosis and malaria), diabetes, hypertension, coronary artery disease, malignancies, and mental health (Qidwai 2017).

This study has its limitations. This is a proof-of-concept cross-sectional study and thus can predict associations but not causality and cannot measure changes over time. Future researchers may want to use temporal data. We also did not aim to investigate differences across socio-demographic groups, such as provinces, ethnicities, and region (urban versus rural) and there may be differences in high-risk pregnancies based on demographics. Furthermore, predictive machine learning interventions for maternal and child healthcare rely on the data collected by healthcare providers or trained collectors, who are recording data based on the availability of computer equipment and devices for recording and monitoring (Kiragu 2014). There may be potential problems with limited time of respondent, missing data, and inaccurate information collected.

There is also concern about the ever-evolving nature of technology and machine learning, which creates uncertainty in professionals who are not eager to depend on it entirely (Deka and Kim 2024), and also the matter of input-output nature of machine learning prediction which does not explain causation or consider explanatory variables (Yarkoni and Westfall 2017). We must also consider that apart from the PDHS data, the Pakistan health sector collects computerized data at different levels- primary to tertiary, through the National Health Data Center (NHDC 2024). However, researchers' face obstacles in accessing this data (Saxena and Muhammad 2018), and we recommend that this data should be made available to independent researchers and research centers to support the government for predictive machine learning results to secure SDG goals.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s11135-025-02210-x.

**Acknowledgements** We acknowledge all the officers involved in challenging nation-wide data collection, in this case the Pakistan Demographic Health Survey. We are grateful to senior peers Dr Rubeena Zakar and Dr Ali Raza for providing us feedback for this project; and to our respective institutes (Forman Christian College University and Punjab University) for providing us with access to technology to complete this project. We are also thankful to the reviewers for their valuable revision recommendations.

**Author contributions** SRJ conceptualized the project, oversaw the completion of the entire study, analyzed the data, and prepared the manuscript. MMM analyzed the data and generated the results. Both authors agreed on the final version of the manuscript submitted.

**Funding** None to declare.

**Data availability** The data is available with the principal investigator (SRJ) based on request.

## Declarations

**Ethics approval** This study involves secondary data analysis. Data was retrieved from a publically accessible dataset (Demographic Health Survey) so no permissions and ethics approval was required.

**Informed consent** Not applicable.

**Conflict of interest** None to declare.

## References


Ahmed, J., Alam, A., Khokhar, S., Khowaja, S., Kumar, R., Greenow, C.R.: Exploring women's experience of healthcare use during pregnancy and childbirth to understand factors contributing to perinatal deaths in Pakistan: A qualitative study. PloS One, **15**(5), e0232823. (2020)

Ali, I., Sadique, S., Ali, S.: COVID-19 significantly affects maternal health: A rapid-response investigation from Pakistan. Front. Global Women's Health. **1**, 591809 (2020)

Anwar, J., Torvaldsen, S., Morrell, S., Taylor, R.: Maternal mortality in a rural district of Pakistan and Contributing Factors. Matern. Child Health J. **27**(5), 902–915 (2023)

Batista, G.E., Prati, R.C., Monard, M.C.: A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsl. **6**(1), 20–29 (2004)

Breiman, L., Random, Forests: Mach. Learn. **45**, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

Bulez, A., Hansu, K., Cagan, E.S., Sahin, A.R., Dokumaci, H.O.: Artificial intelligence in early diagnosis of preeclampsia. Niger. J. Clin. Pract. **27**(3), 383–388 (2024)

Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**, 273–297 (1995). https://doi.org/10.1007/BF00994018

Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Trans. Inf. Theory. **13**(1), 21–27 (1967)

Cox, D.R.: The regression analysis of binary sequences. J. Royal Stat. Soc. Ser. B: Stat. Methodol. **20**(2), 215–232 (1958)

Deka, G.C., Kim, S.: Artificial Intelligence and Machine Learning for Open-world Novelty. Elsevier, United Kingdom (2024)

Dietterich, T.G.: *Ensemble methods in machine learning.* In International workshop on multiple classifier systems (pp. 1–15). Springer, Berlin Heidelberg. (2000)

Fredriksson, M., Fulcher, I.R., Russell, A., Li, X., Tsai, D., Seif, S., Mpembeni, R., Hedt-Gauthier, B.: Machine learning for maternal health: Predicting delivery location in a community health worker program in Zanzibar. Front. Digit. Health. **4**, Article 855236. (2022). https://doi.org/10.3389/fdgth.2022.855236

Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232 (2001)

Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**(1), 119–139 (1997)

Gilani, S.H., Rauf, N., Zahoor, S.: Artificial intelligence and the rule of law: A critical appraisal of a developing sector. Pakistan J. Social Res. **5**(02), 743–750 (2023)

Habib, M.A., Raynes-Greenow, C., Nausheen, S., Soofi, S.B., Sajid, M., Bhutta, Z.A., Black, K.I.: Prevalence and determinants of unintended pregnancies amongst women attending antenatal clinics in Pakistan. BMC Pregnancy Childbirth. **17**, 1–10 (2017)

Islam, M.N., Mustafina, S.N., Mahmud, T., Khan, N.I.: Machine learning to predict pregnancy outcomes: A systematic review, synthesizing framework and future research agenda. BMC Pregnancy Childbirth. **22**(1), 348 (2022)

Jafree, S.R., Barlow, J.: Systematic review and narrative synthesis of the key barriers and facilitators to the delivery and uptake of primary healthcare services to women in Pakistan. BMJ Open., **13**(10), e076883. (2023)

Jafree, S.R., Muzammil, A., Burhan, S.K., Bukhari, N., Fischer, F.: Impact of a digital health literacy intervention and risk predictors for Multimorbidity among poor women of reproductive years: Results of a randomized-controlled trial. Digit. Health. **9**, 20552076221144506 (2023)

Jokhio, A., Winter, H., Cheng, K.: An intervention involving traditional birth attendants and perinatal and maternal mortality in Pakistan. N. Engl. J. Med. **352**(20), 2091–2099 (2005)

Katarya, R., Srinivas, P.: Predicting heart disease at early stages using machine learning: a survey. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)* (pp. 302–305). IEEE (2020), July

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Liu, T.Y.: LightGBM: A highly efficient gradient boosting decision tree. Adv. Neural. Inf. Process. Syst., 30. (2017)

Khalid, S., Khan, M.A., Mazliham, M.S.U., Alam, M.M., Aman, N., Taj, M.T., Jehangir, M.: Predicting risk through artificial intelligence based on machine learning algorithms: A case of Pakistani nonfinancial firms. Complexity. **20221**, 6858916 (2022)

Khan, M., Khurshid, M., Vatsa, M., Singh, R., Mona Duggal, and, Singh, K.: On AI approaches for promoting maternal and neonatal health in low resource settings: A review. Front. Public. Health. **10**, 1864 (2022)

Khan, A.N., Jabeen, F., Mehmood, K., Soomro, M.A., Bresciani, S.: Paving the way for technological innovation through adoption of artificial intelligence in Conservative industries. J. Bus. Res. **165**, 114019 (2023)
Kiragu, A.: Maternal morbidity in Kenya: Measurement, contributions and limitations of DHS data. Quetelet J. **2**(2), 121–145 (2014)
Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S.: Self-normalizing neural networks. Adv. Neural. Inf. Process. Syst., 30. (2017)
Koivu, A., Sairanen, M.: Predicting risk of stillbirth and preterm pregnancies with machine learning. Health Inform. Sci. Syst. **8**(1), 14 (2020)
LeCun, Y., Bengio, Y., Hinton, G.: Deep Learn. Nat. **521**(7553), 436–444 (2015)
Liu, H., Wang, Y., Fan, W., Liu, X., Li, Y., Jain, S., Tang, J.: Trustworthy Ai: A computational perspective. ACM Trans. Intell. Syst. Technol. **14**(1), 1–59 (2022)
Macrohon, J.J.E., Villavicencio, C.N., Inbaraj, X.A., Jeng, J.H.: A semi-supervised machine learning approach in predicting high-risk pregnancies in the Philippines. Diagnostics. **12**(11), 2782 (2022)
Mir, A., Wajid, A., Gull, S.: Helping rural women in Pakistan to prevent postpartum hemorrhage: A quasi experimental study. BMC Pregnancy Childbirth, **12**(120). (2012)
Ngiam, K.Y., Khor, W.: Big data and machine learning algorithms for health-care delivery. Lancet Oncol. **20**(5), e262–e273 (2019)
NHDC: National Health Data Center. Retrieved: (2024). https://www.nih.org.pk/health-data-center
Nisar, Y.B., Aurangzeb, B., Dibley, M.J., Alam, A.: Qualitative exploration of facilitating factors and barriers to use of antenatal care services by pregnant women in urban and rural settings in Pakistan. BMC Pregnancy Childbirth. **16**, 1–9 (2016)
Ojo, A.I., Adedokun, A.O.: Deep hybrid model for maternal health risk classification in pregnancy: Synergy of ANN and random forest. Front. Artif. Intell. **6**, Article 1213436. (2023). https://doi.org/10.3389/frai.2023.1213436
Omer, S., Zakar, R., Zakar, M.Z., Fischer, F.: The influence of social and cultural practices on maternal mortality: A qualitative study from South Punjab, Pakistan. Reproductive Health. **18**(1), 97 (2021)
Opitz, D., Maclin, R.: Popular ensemble methods: An empirical study. J. Artif. Intell. Res. **11**, 169–198 (1999)
PDHS: Pakistan Demographic and Health Survey, The National Institute of Population Studies, (2018). https://dhsprogram.com/pubs/pdf/FR354/FR354.pdf
Peal, J.: *Bayesian networks: A model of self-activated memory for evidential reasoning*. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 7). (1985)
Qidwai, W.: Growing disease burden in Pakistan: Status, challenges, and opportunities. J. Coll. Physicians Surg. Pakistan. **27**(11), 671–673 (2017)
Quinlan, J.R.: Improved use of continuous attributes in C4. 5. J. Artif. Intell. Res. **4**, 77–90 (1996)
Ramakrishnan, R., Rao, S., Jian-Rong He: Perinatal health predictors using artificial intelligence: A review. Women's Health. **17**, 17455065211046132 (2021)
Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature. **323**(6088), 533–536 (1986)
Saxena, S., Muhammad, I.: Barriers to use open government data in private sector and ngos in Pakistan. Inform. Discovery Delivery. **46**(1), 67–75 (2018)
Schadt, E.E., Linderman, M.D., Sorenson, J., Lee, L., Nolan, G.P.: Computational solutions to large-scale data management and analysis. Nat. Rev. Genet. **11**(9), 647–657 (2010)
Shaeen, S.K., Tharwani, Z.H., Bilal, W., Islam, Z., Essar, M.Y.: Maternal mortality in Pakistan: Challenges, efforts, and recommendations. Annals Med. Surg., **81**. (2022)
Shara, N., Mirabal-Beltran, R., Talmadge, B., Falah, N., Ahmad, M., Dempers, R., Crovatt, S., Eisenberg, S., Anderson, K.: Use of machine learning for early detection of maternal cardiovascular conditions: Retrospective study using electronic health record data. JMIR Cardio. **8**, e53091 (2024). https://doi.org/10.2196/53091
Trivedi, A., Mukherjee, S., Tse, E., Ewing, A., Lavista Ferres, J.: Risks of using Non-verified open data: A case study on using machine learning techniques for predicting pregnancy outcomes in India. ArXiv Preprint (2019). *arXiv*:1910.02136.
Waqas, A., Zubair, M., Zia, S.: Psychosocial predictors of antenatal stress in Pakistan: perspectives from a developing country. *BMC Res Notes, 13*(160). (2020)
Wolpert, D.H., Macready, W.G.: No free lunch theorems for optimization. IEEE Trans. Evol. Comput. **1**(1), 67–82 (1997)
Yan, T., Wang, D., Zheng, M., Xia, T., Pan, E., Xi, L.: Fisher's discriminant ratio based health indicator for locating informative frequency bands for machine performance degradation assessment. Mech. Syst. Signal Process. **162**, 108053 (2022)

Yang, Q., Fan, X., Cao, X., Hao, W., Lu, J., Wei, J., Ge, L.: Reporting and risk of bias of prediction models based on machine learning methods in preterm birth: A systematic review. Acta Obstet. Gynecol. Scand. **102**(1), 7–14 (2023)
Yarkoni, T., Westfall, J.: Choosing prediction over explanation in psychology: Lessons from machine learning. Perspect. Psychol. Sci. **12**(6), 1100–1122 (2017)