scientific reports

Check for updates

OPEN

Ensemble machine learning framework for predicting maternal health risk during pregnancy

Alaa O. Khadidos[1,2], Farrukh Saleem[3], Shitharth Selvarajan[4,5]✉, Zahid Ullah[6] & Adil O. Khadidos[7]

Maternal health risks can cause a range of complications for women during pregnancy. High blood pressure, abnormal glucose levels, depression, anxiety, and other maternal health conditions can all lead to pregnancy complications. Proper identification and monitoring of risk factors can assist to reduce pregnancy complications. The primary goal of this research is to use real-world datasets to identify and predict Maternal Health Risk (MHR) factors. As a result, we developed and implemented the Quad-Ensemble Machine Learning framework to predict Maternal Health Risk Classification (QEML-MHRC). The methodology used a vacxsriety of Machine Learning (ML) models, which then integrated with four ensemble ML techniques to improve prediction. The dataset collected from various maternity hospitals and clinics subjected to nineteen training and testing tests. According to the exploratory data analysis, the most significant risk factors for pregnant women include high blood pressure, low blood pressure, and high blood sugar levels. The study proposed a novel approach to dealing with high-risk factors linked to maternal health. Dealing with class-specific performance elaborated further to properly understand the distinction between high, low, and medium risks. All tests yielded outstanding results when predicting the amount of risk during pregnancy. In terms of class performance, the dataset associated with the "HR" class outperformed the others, predicting 90% correctly. GBT with ensemble stacking outperformed and demonstrated remarkable performance for all evaluation measure (0.86) across all classes in the dataset. The key success of the models used in this work is the ability to measure model performance using a class-wise distribution. The proposed approach can help medical experts assess maternal health risks, saving lives and preventing complications throughout pregnancy. The prediction approach presented in this study can detect high-risk pregnancies early on, allowing for timely intervention and treatment. This study's development and findings have the potential to raise public awareness of maternal health issues.

Keywords Maternal health risk, Machine learning, Ensemble machine learning, Pregnancy complications

Common pregnancy-related problems include maternal depression, obesity, diabetes, high blood pressure, and anxiety. According to the World Health Organization, a woman dies every two minutes because of high blood pressure or any other pregnancy-related complications[1]. This fact emphasizes the risk that women face their pregnancy, which can lead to miscarriage or other complications during the postpartum period. It is critical to monitor the data generated in separate phases of pregnancy from various perspectives. Fortunately, there are computational tools that can help detect hidden patterns and predict the most common risk factors. For example, predictive modeling, natural language processing, pattern recognition, and image processing are major computing techniques that can be used to analyze the data collected during the pregnancy. Previous research has demonstrated that various maternal variables affect women's health and can lead to pregnancy instability[2].

[1]Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia. [2]Center of Research Excellence in Artificial Intelligence and Data Science, King Abdulaziz University, Jeddah, Saudi Arabia. [3]School of Built Environment, Engineering and Computing, Leeds Beckett University, Leeds LS1 3HE, UK. [4]Department of Computer Science, Kebri Dehar University, Kebri Dehar, Ethiopia. [5]School of Built Environment, Engineering and Computing, Leeds Beckett University, Leeds LS6 3QS, UK. [6]Information Systems Department, College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia. [7]Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia. ✉email: shitharths@kdu.edu.et

Obesity, for example, in women may increase the risk of gestational diabetes, which needs proper professional care and medication[3]. Furthermore, obesity can cause preeclampsia, which can lead to other complications such as high blood pressure[4]. Furthermore, other medical-related concerns such as age, heart rate, and body temperature can endanger the pregnancy and the lives of both the mother and the child. As a result, it is critical to check for symptoms of disease at every stage of pregnancy[5]. To ensure a healthy and safe birth, those risk factors must be addressed early on, appropriate medications administered, and all precautions taken in accordance with the expert guidance.

On the other hand, computer scientists are exploring several strategies to control this problem by utilizing data supplied by health agencies. In this regard, patient data, demographic data, medication lists, and patient behavior all play a significant role in the development of computing models. Machine Learning (ML) is a field of Artificial Intelligence (AI) that presented multiple solutions by finding significant relationships between patients' health records and pregnancy risk factors[6,7]. ML approaches can predict the best delivery mode[8], prediction of premature birth[9], and lower the maternal mortality rate[10]. The wide range of ML algorithms allows to use health-related datasets including health symptoms data[11,12], disease epidemiology[13], ultrasound reports[14], and prenatal medical imaging[15] for different purposes. The broad scope of ML algorithms, their capacity to uncover hidden patterns in datasets, and their ability to classify and forecast future transactions, make them an attractive tool for use with health-related datasets.

The research explored the use of predictive models for maternal health risk factors. The model implemented using two alternative approaches: standard and ensemble machine learning. The goal was to create a novel framework that could deal efficiently with accuracy, robustness, flexibility, and the bias-variance trade-off. Ensemble techniques, in particular, can perform better on complicated data with multi-class target variable and imbalanced classification of various categories. The proposed model can produce reliable results in terms of accuracy, stability, and avoiding overfitting problems, by combining multiple models' capabilities. Furthermore, while dealing with a multi-class classification problem, the model's performance must be appropriately evaluated. To evaluate the model's performance, two evaluation metrics used in this study: micro and macro weighted scores. This enables for the handling of class imbalances (to minimize bias towards larger classes), overall performance measures (to understand the model's general efficacy), and equal importance to all classes (ensures that the performance on smaller classes not overshadowed by performance on larger classes).

The main objective of this paper is to propose a fusion of state-of-the-art ML algorithms on the maternal risk factors dataset to predict the risk associated with pregnancy. The framework that can help the medical experts to understand the level of maternal risk associated with a specific case based on the patient's health history. The framework used combination of traditional and ensemble approaches to improve the performance, and to overcome issues like overfitting and class imbalances. Previous research employed ML algorithms to create a range of intelligent systems for discovering hidden patterns in medical images and other types of datasets[7]. Another research has found that heart rate, blood pressure, blood glucose level, and body temperature are common risk factors for maternal health[16]. As a result of the current development, appropriateness, and efficacy of the ML algorithm in predicting risk associated with pregnancy, this study developed an integrated framework QEML-MHRC that uses both traditional and ensemble ML techniques to predict risk during pregnancy. The research makes the following contributions to this field of study:

- Exploratory data analysis presented using a variety of ways to highlight the major qualities and correlations between the attributes.
- A novel QEML-MHRC framework for training machines utilizing classic ML algorithms like Decision Tree (DT), Random Forest (RF), Gradient Boosted Trees (GBT), and K-nearest Neighbor (KNN) by incorporating four ensemble techniques including Boosting, Bagging, Stacking, and Voting.
- As the problem addressed in this study is related to a multi-class attribute, the performance of the models evaluated using multi-class metrics. To do this, we first measured the evaluation criteria such as precision, recall, and F1 separately for each class, and then employed "Overall and Weighted" computations to understand overall performance.
- Analyzing performance using class breakdown status is a suitable measure, particularly for imbalanced classes. According to our understanding, this strategy was not used entirely in previous work on the same dataset[16]. The model performance for each class is significant when reviewing the overall performance of a multi-class dataset.
- A comparison of the machines trained in this study provided to demonstrate the effectiveness of the proposed work in this research.
- The results of this study clearly demonstrate the potential for dealing with complex dataset by reducing overfitting and improving accuracy.
- The number of strategies used in this study emphasizes the research's originality because they had not previously been used on a similar dataset.

In this study, a novel QEML-MHRC framework proposed for predicting maternal health risk during pregnancy. In comparison to conventional machine learning methods, it offered a number of novel features and developments. Firstly, the integration of multiple models provided a comprehensive analysis of the generated results. The number of techniques used for data analytics demonstrates a strong understanding and relationship between various aspects. In addition, micro and macro weighted scores used in this work to address the multi-classification problem. In future, the suggested system can give customized risk assessments based on individual patient data. Finally, cross validation and other specific factors applied with various techniques to increase model

performance. The findings of this study would be incorporated with decision-making system for healthcare practitioners.

Overall, this study focuses on one of the most important concerns affecting the lives of women and newborns, attempting to answer the question, "How can we predict the risk associated with pregnancy that can save women's lives, smooth childbirth, and reduce postpartum complications?" The remainder of the paper structured as follows: The next section discusses the most relevant work. "Materials and methods" section summarizes the materials and techniques. "Implementation of the Proposed Framework" and "Results, discussion, and comparison" sections address the suggested framework's implementation and results, respectively. Finally, the last section analyzes the study's findings and future directions.

Research background

The implementation of ML algorithms to medical data offers multiple answers to various health sectors[17–19]. The application of ML algorithms on healthcare industry offers a substantial amount of work in performing tasks such as diagnosis, treatment, patient care, and other operational efficiencies. Machine learning algorithms can give successful solutions in a variety of applications, including predictive analytics for various diseases, patient monitoring systems, disease identification automation, and the development of preventative and curative programs. Furthermore, the employment of machine learning techniques can aid in the discovery of a variety of solutions, such as risk assessments, treatment plans, drug discovery, and proper resource allocation. Table 1 depicts the application of ML algorithms in suggesting solutions for the healthcare industry.

This study addresses the concept of creating an optimal prediction model for maternal health risk. Several complications have been identified that can lead to major health concerns for the mother and child. For instance, gestational diabetes is a kind of diabetes that develops during pregnancy and results in an increase in blood glucose levels[30]. In addition, high blood sugar levels can lead to a number of complications, including the birth of overweight babies or premature birth[31]. Preeclampsia is another type of health condition that commonly develops in the middle of pregnancy and can harm the kidneys and blood sugar levels[32]. Other symptoms of preeclampsia include high protein levels in the urine and hypertension[33].

Several studies have been published in this field to highlight various prenatal concerns[2], including placental accrete, spontaneous abortion, preterm birth, and others. This impacts the number of complications and health issues that a woman may experience during her pregnancy. A tree-based optimization technique used with 95.2% accuracy to identify issues associated with placental invasion[34]. Another study suggested that blood pressure, blood sugar, and calcium levels might be used to predict the existence of preeclampsia[35]. Another study[36] employed machine learning techniques to present the issues associated with maternal health. The study emphasizes on the significance of pregnancy dangers and how to lower mortality rates in this condition.

The key aspect of this research is how to deal with maternal health risks. Another research highlighted the same issue by working with multiple datasets from the Bangladesh region[37]. The study focused on pregnancy-related difficulties for both the mother and the child. The linear regression model used for prediction, with several evaluation criteria, including root mean square error. When applied to given dataset, the model performed well, with an RMSE of 0.70. It also helped limit population growth and significant risks. Another study on the same topic presented the situation in the United States, revealing relatively high maternal death rates when compared to other developed countries[38]. The study discovered that diseases affecting the cardiovascular system had a significant impact on maternal fatalities.

A study in rural Pakistan examined 7572 records of pregnancies and their outcomes. The study projected a fatality incidence of 238 per 100,000 pregnancies, with obstetric hemorrhage as the major cause. Furthermore, poverty, a lack of healthcare facilities, and a shortage of qualified birth attendants are major contributors to an increase in maternal mortality[39]. In the literature[40], machine learning models such as linear regression, random forest, and gradient boosting used to predict the MHR using public data collected from Kaggle. Blood pressure, blood glucose level, body temperature, and other variables are being investigated. The random forest model achieved 86% accuracy with a tenfold cross-validation strategy, while the LightGBM outperformed with 88%

References	Application	Dataset	ML Techniques	Performance
[20]	Sentiment Analysis of COVID-19 Tweets	Tweeter	Adaptive Neuro-Fuzzy Inference System	Accuracy: 0.916
[21]	COVID-19 Patient Health Prediction	Novel Corona Virus 2019 Dataset, Kaggle	Random Forest	Accuracy: 0.94
[22]	Chronic Diseases Detection Model	Kaggle	Decision Tree	Accuracy: 0.978
[23]	Medical Diagnosis	UCI	Multilayer Perceptron	Accuracy: 0.975
	Heart Disease Prediction	IEEE Data port	CART	Accuracy: 0.875
[24]	Sentiment Analysis of COVID-19 Tweets	Twitter	ABCML-SA	Accuracy: 0.983
[25]	Heart Disease Detection	Kaggle	Decision Tree	Accuracy: 0.90
[26]	Diabetes disease detection	Indian Demographic and Health Dataset	Random Forest	Accuracy: 0.99
[27]	Kidney Disease Prediction	Kaggle	LightGBM	Accuracy: 0.99
[28]	Cervical Cancer Disease Prediction	UCI	XG Boost	Accuracy: 0.94
[29]	Sentiment Classification for Healthcare Tweets	Tweeter	Bagging with KNN	Accuracy: 0.888

Table 1. Machine learning implementation on health datasets.

accuracy. The study underlines that using machine learning models to predict MHR can deliver better outcomes, thereby assisting health practitioners in lowering maternal mortality rates.

Previous work used variety of machine learning algorithms to categorize maternal health risk factors as low, medium, or high depending on certain characteristics[16]. This study conducts a comprehensive examination of MHR variables to assess the level of risk associated with pregnancy. The chosen dataset comprised variables such as blood pressure, heart rate, age, blood sugar level, and others. To forecast the risk factor and evaluate the accuracy, the authors used a Logistic Machine Tree, Naive Bayes, and other algorithms. To measure the prediction performance for multi-class variables, the study employed just accuracy as the assessment metric. If the class variable (risk level) is a multi-class attribute, it must be evaluated using multi-class problem-specific criteria. As described in the literature, in a multi-classification problem, weighted precision, recall, and F1-score are significant evaluation metrics to quantify the model's prediction accuracy in addition to class-wise precision, recall, and F1-score[41]. As a result, our study employed a similar dataset to predict MHR levels while addressing the issues raised in this section. Furthermore, multi-class evaluation criteria used to assess risk level classification and model performance. The following section describes a detailed description of the chosen dataset, attributes, machine learning methodologies, and proposed QEML-MHRC model.

Materials and methods

Overview of proposed framework

This section explains the overall methodological approaches employed in this study to predict MHR, as depicted in Fig. 1. We employed a variety of exploratory data analysis approaches to provide a thorough overview of the data, including the number of attributes, minimum and maximum values for each factor, description, correlation, and explanation using various visualization methods. In the second stage, a variety of preprocessing techniques applied to prepare the dataset for QEML-MHRC implementation. Finally, the proposed model implemented utilizing various ML and quad-ensemble techniques, as described in the following sections.

Practical and managerial implications of proposed work

Machine learning techniques used extensively in the healthcare industry for the analysis of vast amounts of data. It has various advantages when using machine learning algorithms to a maternal health risk dataset. Practically, the approach can assist medical practitioners in enhancing maternal and fetal health outcomes. Early and precise prediction can lead to prompt interventions, resulting in fewer difficulties for both mother and child. As discussed in this study, various other disorders, such as high blood sugar, obesity, and high blood pressure, may develop in the future if the situation is not professionally managed. Based on the findings of this study, physicians can improve monitoring and personalized care plans for patients who are at substantial risk during pregnancy. Furthermore, predictive modeling tools can assist healthcare practitioners in allocating resources more effectively

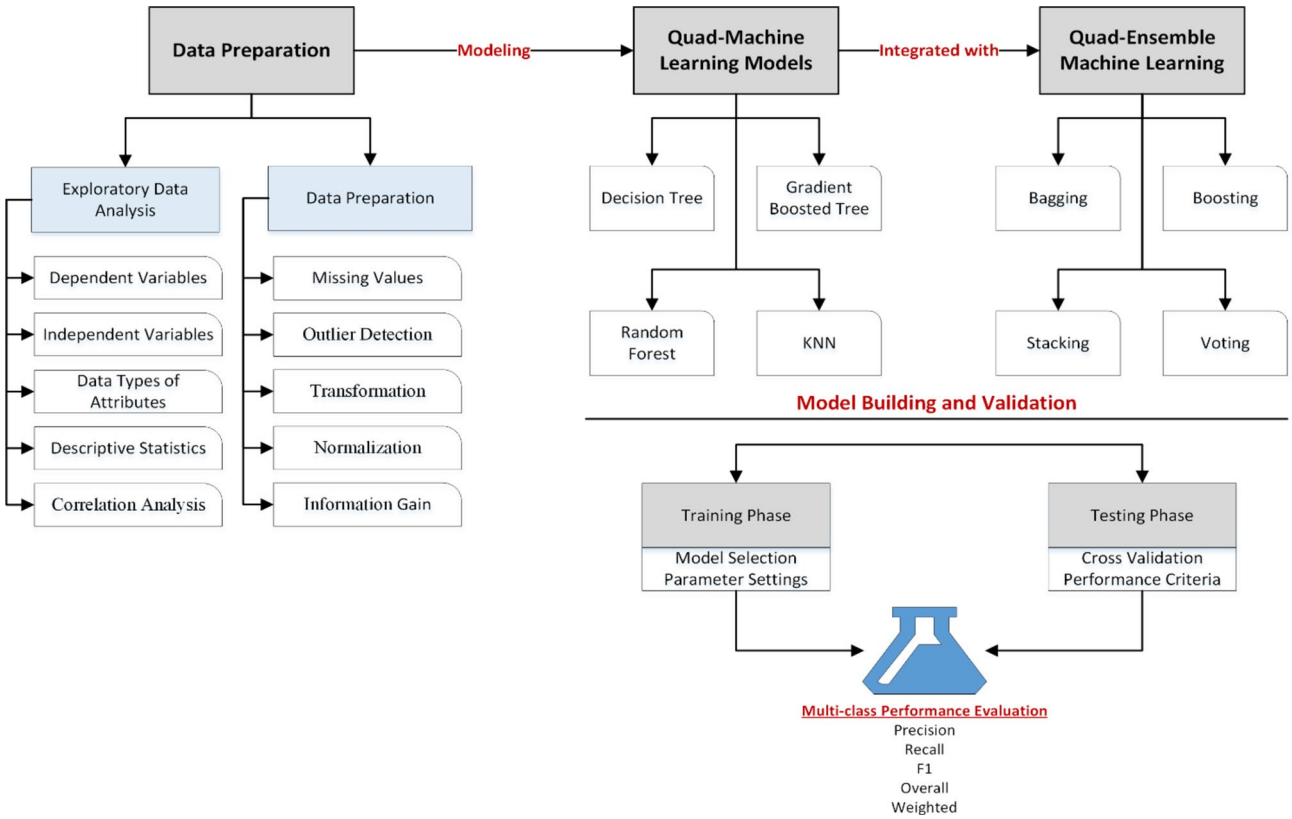

Fig. 1. Research methodology.

by identifying high-risk patients. Medicine, nursing staff, and other equipment might be assigned based on the probable patient's condition.

Enabling preventive measures based on prediction results may decrease the likelihood of serious health conditions, therefore reducing overall healthcare expenses. The proposed ML framework can be used for a variety of purposes. Predictive performance can improve diagnostic and treatment capabilities. The study presents the relationship between many features, which can establish a link between crucial blood pressure and sugar ranges and low or high risk. Different tests can help a health practitioner uncover pregnancy risks early on. Proper treatment and care can assist to lower the mortality rate and other complications.

In addition to the practical implications, the proposed work provides various managerial benefits. Predictive data can help hospital management plan strategies and manage resources and personnel more effectively. The identification of potential risk factors would result in the development of and updating of policies and guidelines for patients. Conducting informative seminars and delivering awareness campaign are some other benefits can be achieved through the outcomes of this study. In addition, the findings can encourage higher authorities to design training programs for medical personnel to keep them up to date on the most recent advances in the field of maternal healthcare. The framework can be integrated with an existing health information system to collect and analyze data in real-time using machine learning algorithms.

The study also underlines the importance of healthcare management ensuring that patients' data is managed ethically and confidentially. However, continuous monitoring and validation of predictive models can enhance overall accuracy and reliability over time. Moreover, a collaborative environment can be created in which multiple health organizations can share their findings for guidance and support. It may also help to refine the model with feedback from multiple organizations. Finally, investing in such a system can result in long-term benefits in terms of resource allocation, personnel development, improving preventive care guidelines, and lowering death rates. The appropriate balance between cost and technology management can eventually bring various health benefits to the patients as well as learning for the medical staff.

Dataset overview and exploratory data analysis

The research problem addressed in this study is related to women who encountered difficulties during their pregnancy. To address this issue, we proposed a model that can help doctors and medical practitioners to reduce the number of deaths and complications. The open dataset used in this work for model implementation collected from several hospitals in Bangladesh and is available online[42].

The dataset contains seven distinct features, including a class variable that indicates the level of risk associated with pregnancy. Table 2 discusses every attribute in detail. Six independent factors representing a variety of a patient's health issues considered to determine the level of risk (dependent variable), which further classified into three categories: Low Risk (LR), Medium Risk (MR), and High Risk (HR). The dataset contains a total of 1014 patients, with an average age of 30 years. Furthermore, Table 2 depicts the descriptive analysis of the variables using minimum, maximum, mean, and standard deviation (SD). Overall, blood sugar (BS) levels range from 6 to 19, with upper and lower blood pressure (BP) values ranging from 70 to 160 and 49 to 100, respectively.

In addition, Fig. 2 shows the number of patients in each class. According to the image, the dataset contains multi-class target attributes, which means that the model's performance should be evaluated accordingly. The sample size was determined to be sufficient for developing ML models capable of predicting maternal health risks for pregnant women and categorizing them based on various medical factors available in the dataset.

Figure 3 illustrates the relationship between independent variables and the target column. The analysis shows the total number of high-risk transactions for each attribute. For example, Fig. 3a reveals that patients aged 25 to 35 are at high risk, with almost 90 of the 1014 patients in the database falling into this age range. Furthermore, as shown in Fig. 3b, nearly 200 people are at high danger, with body temperatures ranging from 98 to 99. Figure 3c clearly shows that most pregnant women experienced difficulties with blood sugar level ranging from 7.5 to 12. Low and high blood pressures, on the other hand, are among the critical variables that may cause substantial problems during this period, as depicted in Fig. 3d and f.

This analysis also shows why MHR classified as low, medium, or high are not just based on a single feature. It influenced by age, blood pressure, body temperature, and blood sugar levels. However, all factors have been identified as significant and must be examined and monitored throughout the pregnancy. Based on the statistics, we can conclude that high blood pressure, low blood pressure, and high blood sugar levels are the most important

Attribute	Description	Min	Max	Mean	SD
Age	The age of the patient at the time of pregnancy	10	70	29.87	13.47
Systolic BP	The upper reading of blood pressure	70	160	113.19	18.40
Diastolic BP	The lower reading of blood pressure	49	100	76.46	13.86
BS	Blood sugar reading	6	19	8.72	3.29
Body temp	Body temperature of the patient	98	103	98.66	1.37
Heart rate	Hear beats per minute	60	90	74.30	8.08
Risk level	Target class: to identify the level of risks [LR: 406, MR: 336, HR: 272]				

Table 2. Descriptive statistics of dataset.

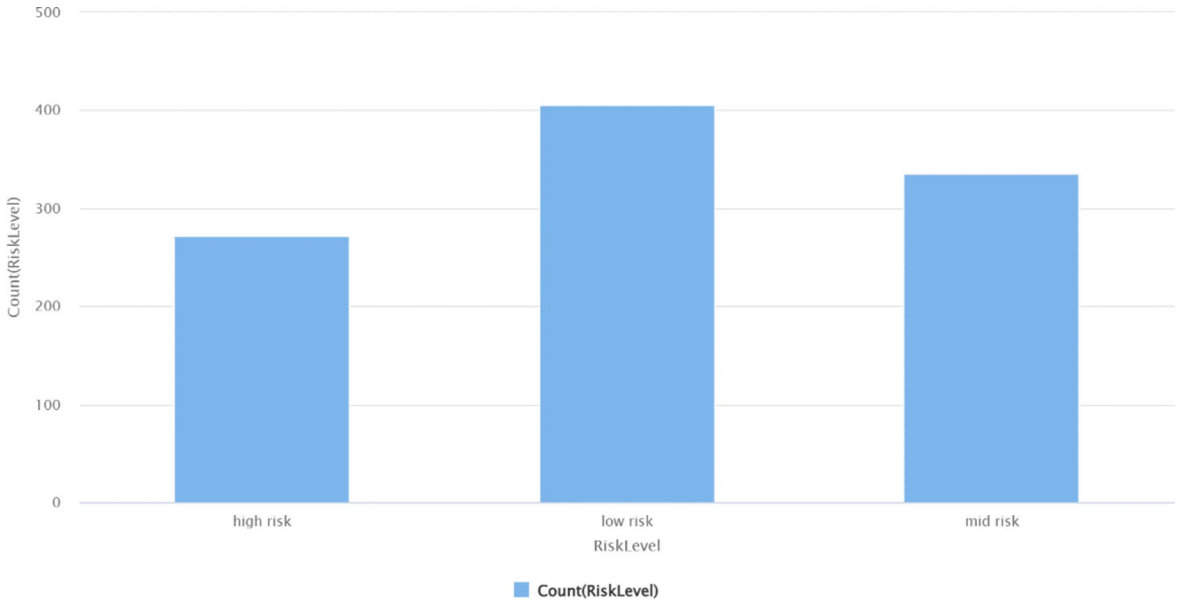

Fig. 2. Number of patients per class in the dataset.

risk factors for pregnant women. As a result, more than 260 of the 272 high-risk cases had difficulties associated with these characteristics. These statistics can assist doctors guide their patients correctly.

Furthermore, the correlation test performed to statistically validate the number of attributes in the selected dataset, with the results shown in Fig. 4. The correlation test, in particular, is important for identifying the relationship between variables and the impact of a single factor's change on other elements in the dataset. The results demonstrated a substantial link between various variables and the dataset associated with each variable. The correlation analysis reveals that the body temperature attribute is negatively correlated with almost all other features except heart rate. In addition, attributes such as Systolic BP and Diastolic BP is correlated negatively with body temperature and heart rate only. Apart from that, all features are associated with one another and can be used for classification problems as well as MHR prediction using ML.

Data preprocessing

Finally, prior to QEML-MHRC implementation, we examine the dataset's validity from several perspectives. As a result, various data preparation techniques used with the Rapid Miner (RM) tool. First, we double-checked the dataset to see if there were any missing values that might be modified. Second, an experiment conducted to identify outliers using the Euclidian distance function. Based on the distance computation using the k nearest neighbor approach, this distance function indicates the number of outliers in the provided dataset. The results of the outlier detection approach revealed that the dataset had no outliers, as shown in Fig. 5. Furthermore, different normalization techniques, such as z-transformation and range transformation, used to find the optimal QEML-MHRC framework implementation. Some attributes, in particular, include more than two decimal values, which have been eliminated for easier comprehension and reading of the dataset. The dataset already had specified classes for each transaction; therefore, no data labeling or further data transformation procedures were required, and it was ready to employ the proposed framework.

An overview of machine learning approaches

Decision tree (DT)

The decision tree (DT) is a prominent classification technique that is effective for analyzing data by segmenting it into tree-based structure. It is widely used, simple to apply, and particularly effective for classification and forecasting. Researchers and practitioners have recently employed this method for a range or purposes, including healthcare decision analytics[43], medical data[44], and predicting low-birth weight babies[45]. To investigate the possibility of improving the MHR prediction percentage, the DT algorithm integrated with ensemble approaches such as boosting, bagging, stacking, and voting.

Gradient boosted trees (GBT)

GBT is the next model employed in this study because of its wide implementation and applicability on medical datasets[46]. This is another example of a classification and regression decision tree model. This algorithm, which generates new predictions based on prior predictions, is also known as the forward learning ensemble approach. GBT is often used to predict class variables in medical datasets[47,48], demonstrating its effectiveness. As a result, this classifier used in the study to predict MHR values based on variety of factors.

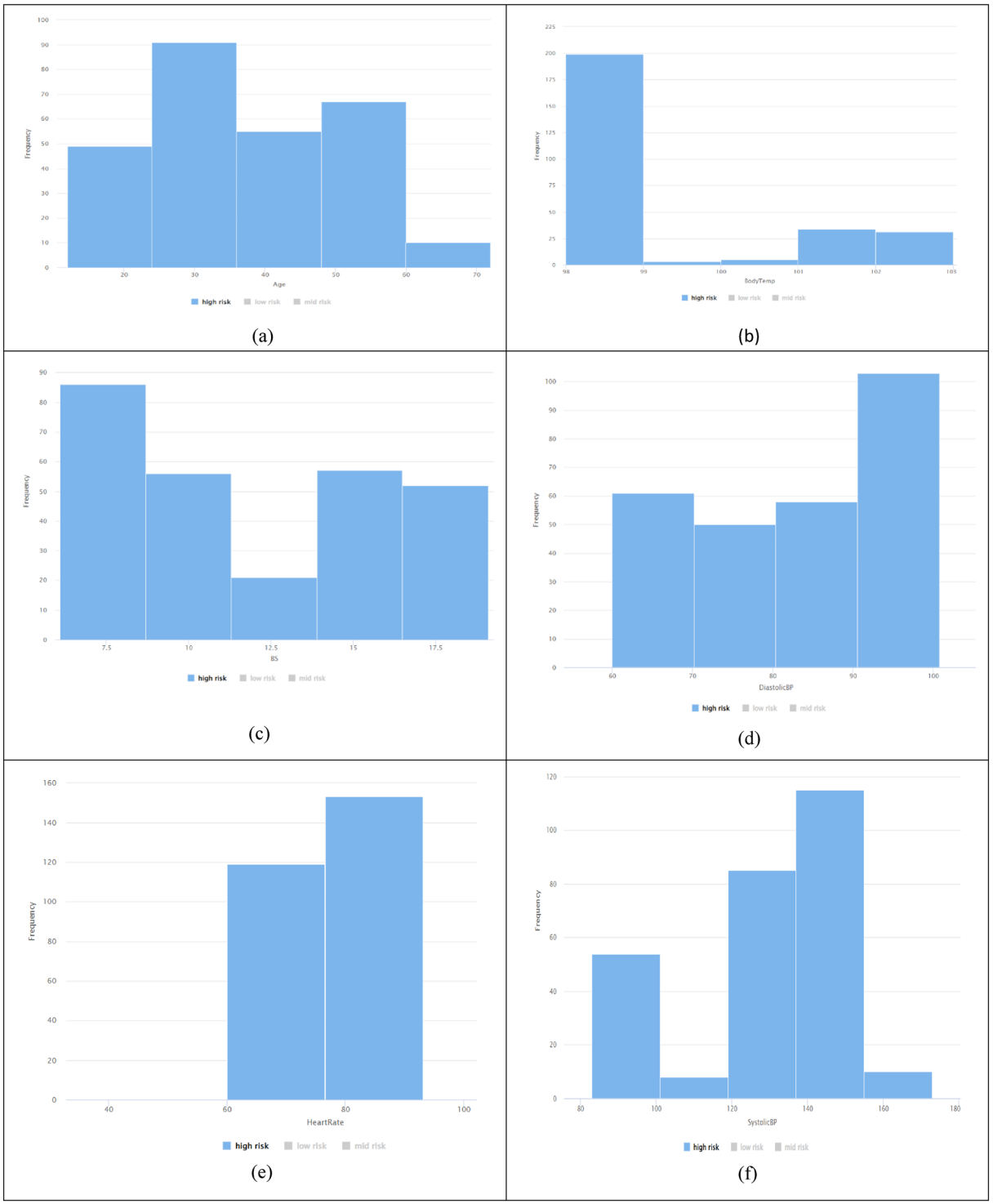

Fig. 3. Number of patients per class in the dataset. (**a**) Age with high risk, (**b**) body temp with high risk, (**c**) BS with high risk, (**d**) diastolic BP with high risk, (**e**) heart rate with high risk, (**f**) systolic BP with high risk.

Random forest (RF)

Random forest (RF) is another supervised learning technique capable of performing classification and regression tasks. This collective strategy, also known as the ensemble approach, has the advantage of simultaneously training and integrating multiple models into a single tree. The bagging or voting approach with random trees is frequently used in this algorithm[49]. This strategy combined with a number of approaches, including bagging, boosting, voting, and stacking[50]. This method used numerous times on the medical dataset dealing with diverse challenges[50,51]. The model initially used as an independent model in this investigation, utilizing the RM tool. Secondly, the experiment repeated using different ensemble methodologies to improve prediction performance.

Attributes	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate
Age	1	0.416	0.398	0.473	-0.255	0.080
SystolicBP	0.416	1	0.787	0.425	-0.286	-0.023
Diastolic...	0.398	0.787	1	0.424	-0.256	-0.046
BS	0.473	0.425	0.424	1	-0.103	0.143
BodyTemp	-0.255	-0.286	-0.256	-0.103	1	0.099
HeartRate	0.080	-0.023	-0.046	0.143	0.099	1

Fig. 4. Correlation analysis of features in the dataset.

Row No.	RiskLevel	outlier	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate
1	high risk	false	25	130	80	15	98	86
2	high risk	false	35	140	90	13	98	70
3	high risk	false	29	90	70	8	100	80
4	high risk	false	30	140	85	7	98	70
5	low risk	false	35	120	60	6.100	98	76
6	high risk	false	23	140	80	7.010	98	70
7	mid risk	false	23	130	70	7.010	98	78
8	high risk	false	35	85	60	11	102	86
9	mid risk	false	32	120	90	6.900	98	70
10	high risk	false	42	130	80	18	98	70
11	low risk	false	23	90	60	7.010	98	76
12	mid risk	false	19	120	80	7	98	70
13	low risk	false	25	110	89	7.010	98	77
14	mid risk	false	20	120	75	7.010	100	70

Fig. 5. Outlier detection analysis in the dataset.

k-nearest neighbor (KNN)

The KNN algorithm is a supervised classification algorithm that can be utilized as an ML approach. The *k* closest neighbor method involves comparing unknown data to *k* training examples. The measurement of distance used to match a specific example to the closest training example[52]. Because the dataset used in this study was of mixed type, the "Mixed Euclidean Distance" method was used to calculate the distance. The dataset predicted using KNN, both with and without ensemble methods. The classifier's performance explored further in the results section.

Bagging—first ensemble method

Bagging, an ensemble technique used in this study that can include multiple classification models. The working scenario of this technique is based on bootstrapping, which divides the initial data set into many training datasets known as bootstraps[53]. The primary reason for diving the datasets is to produce numerous models, which may subsequently be integrated to produce a strong learner. The experiment conducted with the MHR dataset using RM tool. Because the learner models in each sub-process will differ, this type of operator is known as an embedded operator.

Boosting—second ensemble method

Boosting is a machine learning ensemble strategy that combines multiple models to get an effective model. AdaBoost (adaptive boosting) is a boosting technique that can be applied in conjunction with a variety of learning

algorithms. AdaBoost implementation in the RM tool is known as a meta-algorithm, and it can complete the process by including another algorithm as a sub-process. It runs and trains multiple models before combining weak learners to generate a single strong learner, which requires additional computation and execution time[54]. AdaBoost mostly used to examine the efficiency and precision of decision-making models with and without boosted approaches. The results and discussion section examines the overall analysis and effectiveness of the model.

Stacking—third ensemble method

Stacking is a technique for combining many models of several types to improve prediction performance. Stacking learning is based on multiple models rather than a single model. It is also known as a stacked generalization since it enables the combination of multiple classifiers in a single operation[55]. Stacking, as opposed to bagging and boosting, introduces a novel idea of ensemble learning by training the model with several classifiers and using a meta-learner for final output[56]. Because of its superior learning process and performance, the stacking technique applied in a variety of applications, including earthquake prediction[57], cancer images classification[55], and network intrusion detection[58].

The Rapid Miner tool employs a method that is divided into two parts: (i) the Base Learners and (ii) the Stacking Model Learner. In this work, the primary purpose of stacking is to conduct an assessment by integrating several models and to improve MHR predictions. We used a variety of base learners and meta-learners to evaluate, analyze, and compare the performance of various classifiers. We employed different scenarios to create the stacking model, picking four models (GTB, RF, DT, and KNN) as the base learners and one as the stacking learner model. The experiment repeated iteratively, with the stacking learner model replaced each time.

Ensemble method 4: voting

This ensemble method combines multiple machine-learning algorithms into a voting procedure. The voting method involves learning classifiers to vote by majority (for classification) and average (for regression). Finally, the class that received the most votes or average will be predicted[59]. The "Vote" function uses sample data from the input node to generates a classification model. The prediction approach employs a majority voting mechanism, with each classifier casting votes using the "Vote" operator. The unknown example will receive the most votes in each situation. The voting ensemble method combined with a variety of classifiers. To discover the most appropriate response, we conducted three voting trials, each with a different classifier. The following ML classifiers were employed in each experiment: Experiment 1 (GBT, DT, RF, and KNN), Experiment 2 (RF and GBT), and Experiment 3 (GBT, RF, and DT). The outcomes of each model are discussed in the results section.

Implementation of the proposed framework

This study conducted several experiments to predict maternal health risk utilizing several variables. The proposed work completed on a LENOVO Think Pad with an Intel Core i7 processor running at 2.80 GHz (8 CPUs) and 32 GB of RAM. In addition, the experiment conducted using the RM Studio tool, which is an open-source platform developed specifically for machine learning, deep learning, and data science activities[60]. Scholars from all over the world have utilized the tool extensively for ML model implementation and validation[61–64] particularly on healthcare industry datasets[65,66]. The dataset discussed in the prior section entirely loaded into RM tool. The dataset contained seven attributes, with one class variable and the remaining were independent variables.

MHR classification using individual ML model

To reduce delays in obtaining live data, the dataset imported into the RM repository. The RM tool provides the ability to directly load data and recover it later using the "Retrieve" operator, which has been renamed "MHR Dataset" as illustrated in Fig. 6. In the second stage, the "Multiply" operator used to create several copies of the dataset. The dataset then sent on to the "Cross Validation" process. We employed a tenfold cross-validation strategy, which is well-known for giving each transaction in the dataset an opportunity to be a part of the training, testing, and validation process. Furthermore, the k-fold validation strategy used several rounds by partitioning the dataset into k subsets and using one subset for testing and the remaining for training. As a result, k-fold cross-validation is a method for obtaining optimal results while reducing the likelihood of model overfitting[67]. For each ML model, we utilized four distinct cross-validation operators, as indicated in the image below. This operator is known as a nested operator, and it can train and test the machine as well as perform accuracy measurements.

Figure 7 depicts an inner view of each model implementation. The training and testing phases of the cross-validation operator further separated. The input training data linked to the RF model, and the "Apply Model" operator receives both the trained model and the testing dataset. The entire process repeated ten times to determine the ultimate accuracy of the model using the "Performance" operator. Each ML model executed using a similar approach. The outcome of this experiment explained further in the next section.

MHR classification using QEML model

Similarly, for MHR classification, we employed the QEML model. Four ensembles' approaches chosen to generate comparatively optimal classification results when implementing the model. Bagging, boosting, voting, and stacking are the four ensemble procedures used. "An overview of machine learning approaches" section discusses the description and significance of each ensemble strategy. RM provides several operators for employing the ensemble technique. Again, a cross-validation technique used for training and testing, with tenfold validation. Figure 8 displays four screenshots for each ensemble method's implementation. To begin, we employed bagging and boosting process with all ML models to compare the performance of all ML models. Besides that, stacking

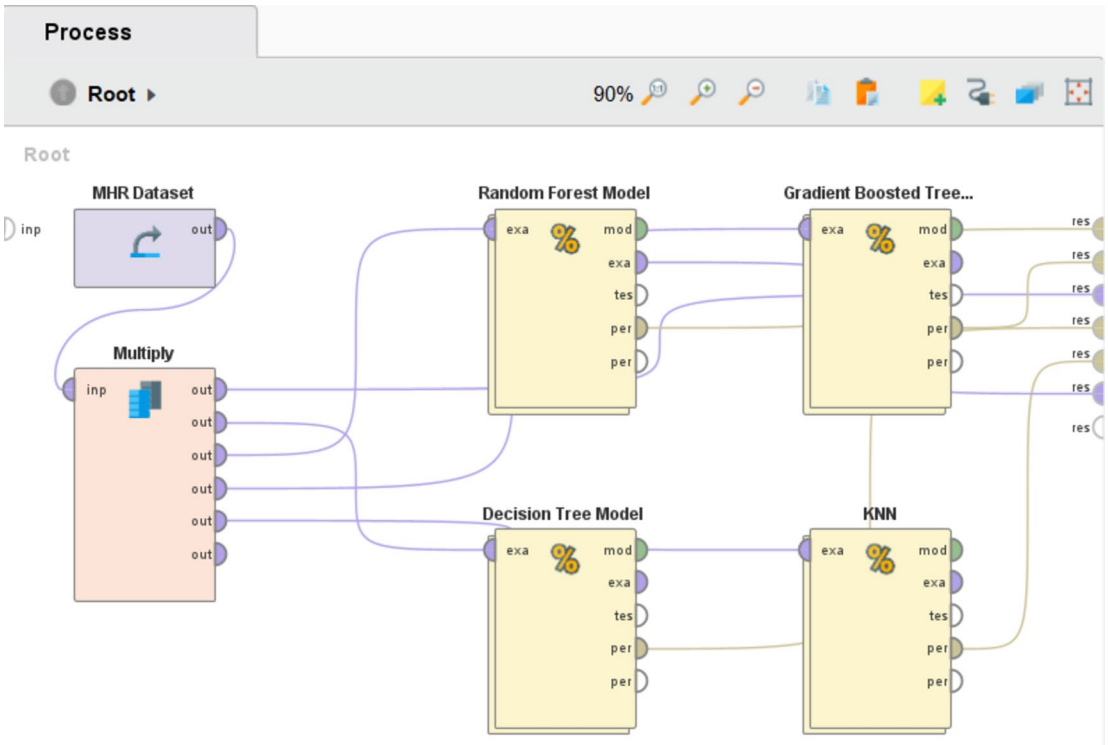

Fig. 6. Cross validation process for model implementation.

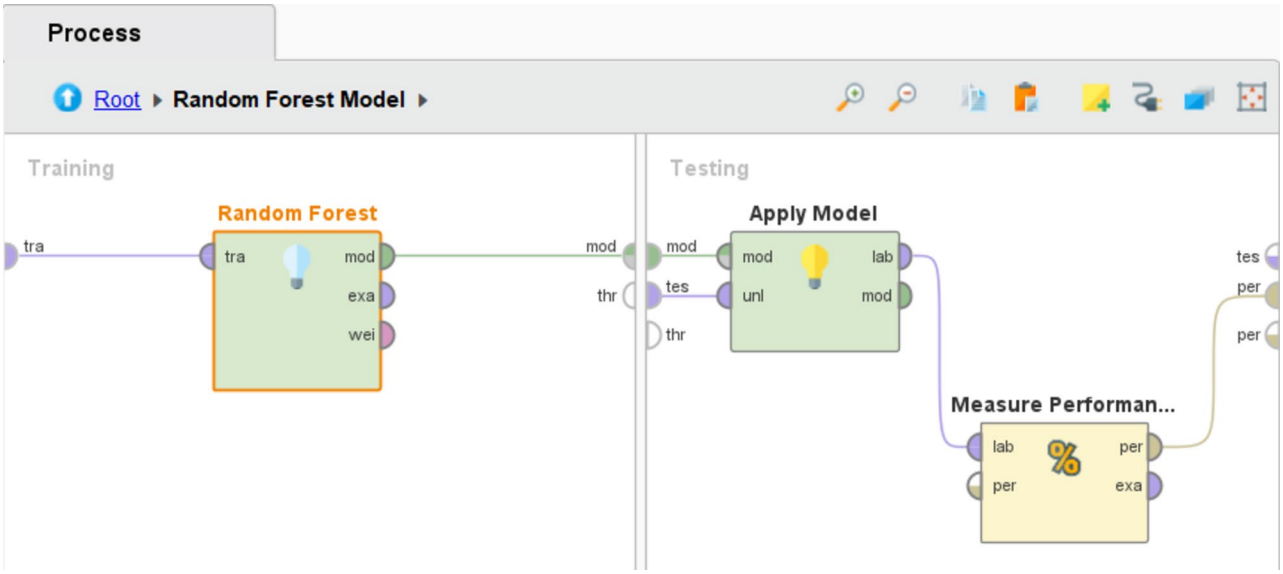

Fig. 7. Cross validation—training & testing phases.

used as a meta-learner model, and for this, we chose a different model combination as the base learner and employed a single model as the stacking model learner each time.

Figure 9 demonstrates the internal model implementation for each individual test. According to the diagram, each experiment divided into two phases: training and testing. In each execution, three primary operators used: (i) the ensemble technique, (ii) applying the model, and (iii) performance. Each ensemble operator placed in the training area and is also known as a nested operator since it contains another subprocess that uses the specific ML model for training. On the other hand, the "Apply Model" operator placed in the testing area and used to apply and evaluate the trained model to an unseen dataset. Finally, the "Apply Model" operator linked to the "Performance" operator to evaluates the model's performance based on a variety of criteria. Because the idea discussed in this study consists of a multi-class classification problem, the evaluation metric chosen accordingly. The outcomes of each experiment explained in more depth in the next section.

Fig. 8. QEML model implementation—outer view. (**a**) Bagging with all ML models, (**b**) boosting with all ML models, (**c**) stacking with different ML models combinations, (**d**) voting with different ML models combinations.

The subsequent section discusses the findings of each experiment. We presented the results using confusion matrix to understand positive and negative values, as well as actual and predicted values. The label column divided into three categories: "HR-High Risk", "LR-Low Risk", and "MR-Medium Risk". As a result, the results for each class discussed using precision, recall, and F1 values. The precision value is an evaluation metric that can be used to examine the results and determine the correctness of a model by counting true positive values divided by total positive values[68] and can be measured using following formula:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

The recall factor is the second evaluation criteria employed in this study. The recall value derived by dividing all true positive values by the total number of true positive and false negative values[69] and can be calculated using the formula below:

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

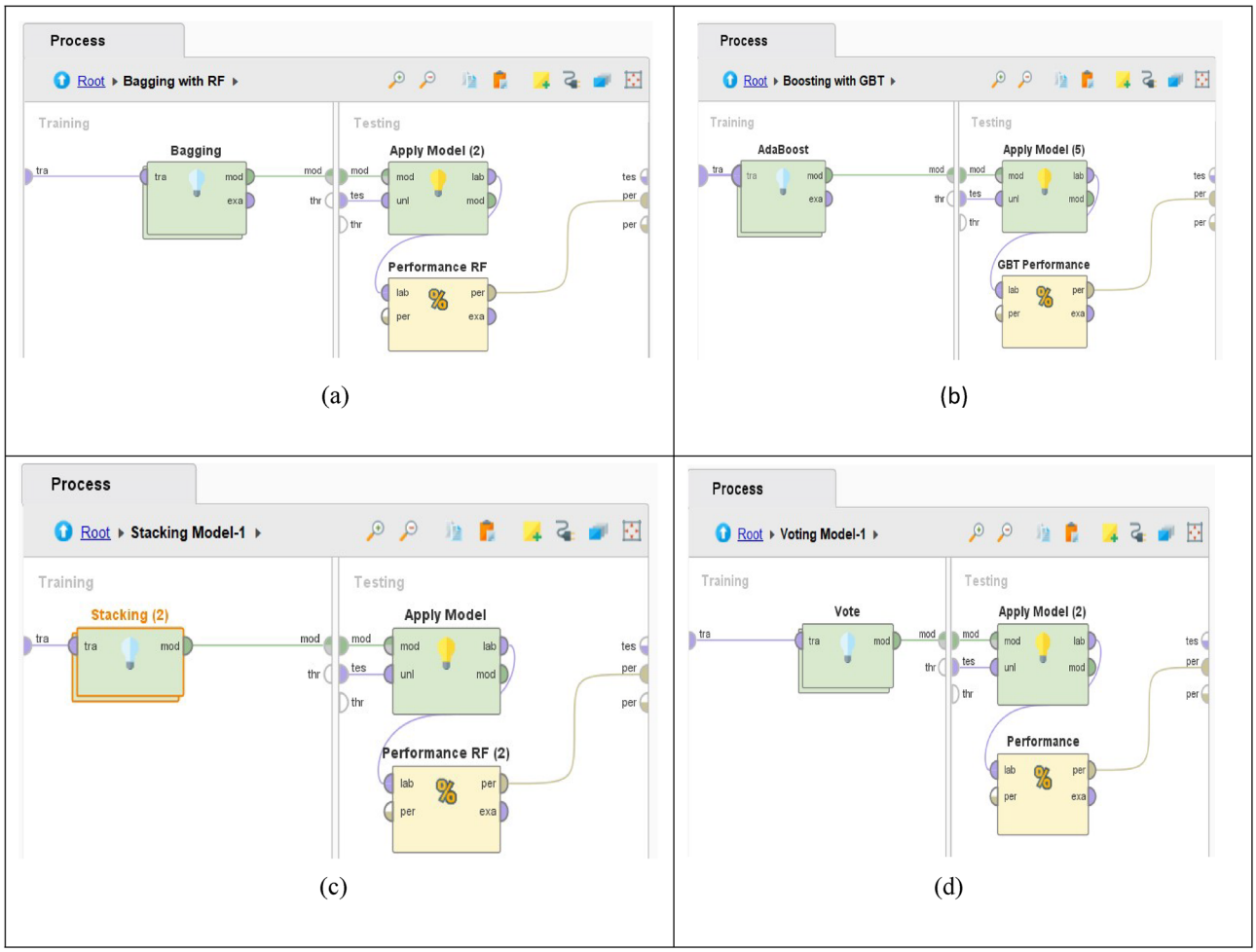

Fig. 9. QEML model implementation—inner view. (**a**) Bagging ensemble approach—inner view, (**b**) boosting ensemble approach—inner view, (**c**) stacking ensemble approach—inner view, (**d**) voting ensemble approach—inner view.

The results were subsequently evaluated using a third common assessment tool known as the F1 score. It is another valuable metric for assessing the performance of ML models. This measure combines the output of recall and precision values and can be measured using the formula:

$$F1Score = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{3}$$

Overall, the three classes represented the patients' level of risk (High, Low, and Medium) in relation to the other independent variables included in this investigation. Due to the multi-class classification challenge, we determined the recall, accuracy, and F1 values for each class. The outcomes of each experiment discussed in the five subsections that follow.

Results, discussion, and comparison

This study aims to forecast maternal health risk utilizing several factors to assist healthcare providers counsel pregnant women and reduce complexity throughout the pregnancy. The dataset used in this study includes information from several test reports as well as demographic factors. The findings of this study are crucial to understanding the usage of real-world datasets obtained from various health organizations. We utilized various machine learning algorithms for prediction, and the integration of ensemble approaches on the dataset yielded the best results. We predict a patient's level of risk during pregnancy using a set of data values connected with different parameters. Four machine learning models, DT, RF, GBT, and KNN, used to predict the number of patients who fell into specific risk categories, such as High, Low, and Medium. The risk level, which calculated based on the values recorded under each independent variable, describes the concerns that may arise in a patient. In addition, quad-ensemble models utilized to improve prediction performance, including bagging, boosting, stacking, and voting. Because risk level (class variable) defined as a multi-class feature, the performance evaluation presented using class-level precision, recall, F1, and weighted scores to help readers understand the results[69,70].

MHR classification without ensemble

During this phase, four independent experiments conducted using DT, RF, GBT, and KNN machine learning models. The primary purpose is to determine how effectively each algorithm predicts the risk level for each patient. Table 3 displays the results of all models assessed using various assessment metrics. We displayed the findings for each class and presented the values within each category. Displaying results for each class is a typical approach to understand the class wise performance instead of presenting overall accuracy[70]. Previous research reported overall accuracy, which can lead to incorrect interpretation, while micro averages might provide greater understanding for each class. Because it is possible for one class to have 100% accuracy while another has less, this can have an impact on overall accuracy. As a result, in a multi-class classification task, overall accuracy cannot offer forecast performance for each class separately[71].

Table 3 demonstrates that each model achieves precision values greater than 0.84 for the "HR" class. On the other hand, the "MR" class achieved the lowest precision (0.6772). Furthermore, GBT had the highest recall value in the "HR" class (0.919). Overall, the results show that the model utilized in this study can be used to develop a system that predicts the risk associated with pregnant women. The table shows that all models performed well, with scores greater than 0.75 in any class/metric. The precision value, which is better than 0.75 for each model, is very notable because it indicates the outcome of correct prediction. Overall, we can conclude that the DT (0.75) model performs the lowest for this classification task, whereas the GBT (0.85) model has the highest weighted precision, recall, and F1 values compared with any model.

MHR classification with ensemble bagging

The ensemble bagging approach implemented in the second phase of this investigation. Bagging is a type of ensemble method that can be integrated with other ML models to improve prediction performance. According to Table 4, DT has the lowest F1 score of any classifier for class "MR" (0.64), whereas KNN has the lowest weighted F1 score (0.71). Conversely, GBT calculates the maximum F1, recall, and precision values for class "HR" and reports them as 0.90, 0.91, and 0.89, respectively. The GBT model achieved the highest weighted values across all classes.

On the contrary, KNN (0.72) had the lowest prediction performance. In comparison, the best precision performance for classes "HR", "LR", and "MR" attained by RF (0.90), GBT (0.88), and GBT (0.77), respectively. It allows us to use several methods for MHR classification, however GBT with bagging is the most efficient because it computes the highest values for all classes. As GBT used an ensemble strategy, combining it with a bagging approach improved its performance significantly.

MHR classification with ensemble boosting

As shown in Table5, this section illustrates the performance of the models when paired with the ensemble boosting approach. This approach produced comparable results as bagging. However, we used this procedure to evaluate the level of performance and determine the feasibility of both methods in a single study. However, some performance measurements are lower than in the bagging approach. GBT with boosting, for example, returns a lower weighted precision value (0.849) than GBT with bagging (0.853). Similarly, using the boosting approach,

Classifier	Class	Count	TP	TN	FN	FP		Precision	Recall	F1
DT	HR	272	232	704	38	40		0.859259259	0.852941176	0.856088561
	LR	406	319	499	109	87		0.745327	0.785714286	0.76498801
	MR	336	214	576	102	122		0.677215	0.636904762	0.656441718
	Total	1014	765	1779	249	249	Overall →	0.754438	0.754438	0.754438
							Weighted →	0.753319158	0.75443787	0.753457236
RF	HR	272	243	714	28	29		0.896678967	0.893382353	0.895027624
	LR	406	342	519	89	64		0.79350348	0.842364532	0.817204301
	MR	336	236	602	76	100		0.756410256	0.702380952	0.728395062
	Total	1014	821	1835	193	193	Overall →	0.809664694	0.809664694	0.809664694
							Weighted →	0.808888499	0.80966469	0.808652072
GBT	HR	272	250	713	29	22		0.896057348	0.919117647	0.907441016
	LR	406	337	561	47	69		0.877604	0.830049261	0.853164557
	MR	336	278	605	73	58		0.792023	0.827380952	0.809315866
	Total	1014	865	1879	149	149	Overall →	0.853057	0.853057	0.853057
							Weighted →	0.854195807	0.85305720	0.853194179
KNN	HR	272	243	695	47	29		0.837931034	0.893382353	0.864768683
	LR	406	309	525	83	97		0.788265306	0.761083744	0.77443609
	MR	336	241	587	91	95		0.725903614	0.717261905	0.721557
	Total	1014	793	1807	221	221	Overall →	0.782051282	0.782051282	0.782051282
							Weighted →	0.780923639	0.78205128	0.781145215

Table 3. Performance of the models—without ensemble.

Bagging	Class	Count	TP	TN	FN	FP		Precision	Recall	F1
With DT	HR	272	237	700	42	35		0.849462366	0.871323529	0.860254083
	LR	406	323	499	109	83		0.747685185	0.795566502	0.770883055
	MR	336	205	580	98	131		0.676567657	0.610119048	0.641627543
	Total	1014	765	1779	249	249	Overall →	0.75443787	0.75443787	0.75443787
							Weighted →	0.75142079	0.75443787	0.75202612
With RF	HR	272	242	717	25	30		0.906367041	0.889705882	0.897959184
	LR	406	335	515	93	71		0.782710	0.825123153	0.803357314
	MR	336	235	594	84	101		0.736677	0.699404762	0.717557252
	Total	1014	812	1826	202	202	Overall →	0.800789	0.800789	0.800789
							Weighted →	0.800626943	0.800788955	0.800302963
With GBT	HR	272	248	712	30	24		0.892086331	0.911764706	0.901818182
	LR	406	331	567	41	75		0.889784946	0.815270936	0.850899743
	MR	336	283	597	81	53		0.777472527	0.842261905	0.808571429
	Total	1014	862	1876	152	152	Overall →	0.850098619	0.850098619	0.850098619
							Weighted →	0.853186331	0.850098619	0.850532388
With KNN	HR	272	201	185	35	71		0.851694915	0.738970588	0.791338583
	LR	406	296	316	121	110		0.709832134	0.729064039	0.719319563
	MR	336	226	294	135	110		0.626038781	0.672619048	0.648494
	Total	1014	723	795	291	291	Overall →	0.713017751	0.713017751	0.713017751
							Weighted →	0.720120211	0.713017751	0.715169297

Table 4. Performance of the models—With Bagging.

Boosting	Class	Count	TP	TN	FN	FP		Precision	Recall	F1
With DT	HR	272	235	699	43	37		0.845323741	0.863970588	0.854545455
	LR	406	322	497	111	84		0.743649	0.793103448	0.767580453
	MR	336	207	582	96	129		0.683168	0.616071429	0.647887324
	Total	1014	764	1778	250	250	Overall →	0.753452	0.753452	0.753452
							Weighted →	0.750881746	0.75345168	0.751246714
With RF	HR	272	244	715	27	28		0.900369004	0.897058824	0.898710866
	LR	406	344	514	94	62		0.785388128	0.84729064	0.815165877
	MR	336	233	606	72	103		0.763934426	0.693452381	0.72698908
	Total	1014	821	1835	193	193	Overall →	0.809664694	0.809664694	0.809664694
							Weighted →	0.809122205	0.80966469	0.80835802
With GBT	HR	272	252	714	28	20		0.9	0.926470588	0.913043478
	LR	406	329	564	44	77		0.882037534	0.810344828	0.844672657
	MR	336	278	595	83	58		0.770083102	0.827380952	0.797704448
	Total	1014	859	1873	155	155	Overall →	0.847140039	0.847140039	0.847140039
							Weighted →	0.849758541	0.84714004	0.847449329
With KNN	HR	272	203	190	40	69		0.835390947	0.746323529	0.788349515
	LR	406	296	316	126	110		0.701421801	0.729064039	0.714975845
	MR	336	220	293	129	116		0.630372493	0.654761905	0.642336
	Total	1014	719	799	295	295	Overall →	0.709072978	0.709072978	0.709072978
							Weighted →	0.713815332	0.70907298	0.710587849

Table 5. Performance of the models—With Boosting.

KNN's precision value reduced to 0.71. Aside from that, the performance of DT and RF with boosting estimated using a method that is almost equivalent to the bagging approach.

MHR classification with ensemble stacking

The proposed QEML-MHRC framework considers stacking as the third ensemble model. This method is important since it allows you to integrate numerous ML models instead of just one, which can lead to improve performance. To find the best set of models, we ran several scenarios and built a stacking approach with a range of ML models. Table 6 displays the outcomes of all the experiments conducted during this phase. Performance

Base learners	Stacking model learner	Class	Count	TP	TN	FN	FP		Precision	Recall	F1
GBT RF DT KNN	GBT	HR	272	246	714	28	26		0.897810219	0.904411765	0.901098901
		LR	406	340	565	43	66		0.88772846	0.837438424	0.861850444
		MR	336	282	603	75	54		0.789915966	0.839285714	0.813852814
		Total	1014	868	1882	146	146	Overall →	0.856015779	0.856015779	0.856015779
								Weighted →	0.858021596	0.85601578	0.856474089
GBT RF DT KNN	RF	HR	272	245	714	28	27		0.897435897	0.900735294	0.899082569
		LR	406	334	522	86	72		0.795238095	0.822660099	0.808716707
		MR	336	244	601	77	92		0.760124611	0.726190476	0.742770167
		Total	1014	823	1837	191	191	Overall →	0.811637081	0.811637081	0.811637081
								Weighted →	0.811016864	0.81163708	0.811104752
GBT RF DT KNN	DT	HR	272	242	701	41	30		0.855123675	0.889705882	0.872072072
		LR	406	323	511	97	83		0.769047619	0.795566502	0.782082324
		MR	336	228	595	83	108		0.733118971	0.678571429	0.704791345
		Total	1014	793	1807	221	221	Overall →	0.782051282	0.782051282	0.782051282
								Weighted →	0.780231703	0.78205128	0.780610374
GBT RF DT KNN	KNN	HR	272	203	703	39	69		0.838842975	0.746323529	0.789883268
		LR	406	295	479	129	111		0.695754717	0.726600985	0.710843373
		MR	336	216	546	132	120		0.620689655	0.642857143	0.631578947
		Total	1014	714	1728	300	300	Overall →	0.704142012	0.704142012	0.704142012
								Weighted →	0.709263736	0.70414201	0.705780261

Table 6. Performance of the models—with stacking.

analysis performed by combining all ML models as base learners and selecting each model as a stacking model learner during the training phase. Except for KNN (0.70), all stacking model learners, GBT (0.85), RF (0.81), and DT (0.78), outperformed bagging and boosting. Stacking's overall improvement emphasizes the significance of a better MHR prediction approach.

Four ensemble models developed utilizing GBT, RF, DT, and KNN as stacking model learners, as shown in the table below. Precision for class "MR" was lower for all models, including GBT (0.78), RF (0.76), DT (0.73), and KNN (0.62), affecting weighted scores significantly. This could be due to similar values or a lack of variation in data values between the "LR" and "MR" classes. As a result, the findings revealed the importance of class-wise performance analysis in multi-class classification problems, which were not well addressed in prior work[70]. As shown in Table 6, the overall accuracy cannot provide a comprehensive analysis if the number of records for each class varies. GBT outperformed all other models as a stacking meta learner, with the highest weighted scores for precision (0.8580), recall (0.8560), and F1 (0.8564). It also improved the results achieved from bagging and boosting. Finally, with the stacking method, GBT outperformed KNN (0.70) by more than 16%.

MHR classification with ensemble voting

Table 7 presents the performance analysis of the ensemble voting approach. Voting is another strategy for combining multiple models in a single experiment. This is the final approach used for the proposed QEML-MHRC framework. The number of experiments conducted to enhance prediction for the MHR classification problem. Three scenarios were developed for this purpose, and all of them enhanced performance as compared to single model's performances. For example, in previous results tables (from 3 to 6), the KNN was the worst-performing model as an individual, but after integrating it with other models using the voting approach, it improved the performance by 11%. It supports the idea of utilizing a voting technique here, where we may combine multiple models to benefit from each one and create a meta-learner process. Second, GBT and RF outperformed in terms of precision (0.83), while the other two models had 0.81 and 0.82, respectively. The class-wise performance also shows that class "MR" has improved significantly. As previously discussed, combining GBT with RF increases the correct prediction of class "MR" and achieves the highest precision value (0.84). It implies that the voting approach can improve the correct risk classification of pregnant women using different attributes.

Final discussion

This research provides a thorough ML architecture to address a multi-class classification task involving maternal health risk. The obtained data demonstrate that varied levels of risk can be observed in women during pregnancy. The dataset divided into three risk categories: low, medium, and high. The performance analysis performed using multi-class evaluation metrics to improve the work conducted in the previous study[70]. The study proposed a methodology that provides several advantages over previous studies. A thorough analysis performed on the dataset to determine the relationship between each attribute. The concept of using a unique set of machine learning algorithms to improve prediction accuracy. Previously, the model's performance provided based on model accuracy, which does not apply to this classification problem when the target variable includes more than two mutually exclusive classes. When dealing with a multi-class classification challenge, assessment metrics

Voting algorithms	Class	Count	TP	TN	FN	FP		Precision	Recall	F1
GBT DT RF KNN	HR	272	245	713	29	27		0.894160584	0.900735294	0.897435897
	LR	406	353	507	101	53		0.777533	0.869458128	0.820930233
	MR	336	225	617	61	111		0.786713	0.669642857	0.723472669
	Total	1014	823	1837	191	191	Overall →	0.811637	0.811637	0.811637
							Weighted →	0.811859721	0.81163708	0.809158832
GBT RF	HR	272	251	711	31	21		0.890070922	0.922794118	0.906137184
	LR	406	364	514	94	42		0.794759825	0.896551724	0.842592593
	MR	336	231	646	43	105		0.843065693	0.6875	0.757377049
	Total	1014	846	1871	168	168	Overall →	0.834319527	0.834319527	0.834319527
							Weighted →	0.836333188	0.83431953	0.831400981
GBT RF DT	HR	272	246	714	28	26		0.897810219	0.904411765	0.901098901
	LR	406	356	516	92	50		0.794642857	0.876847291	0.833723653
	MR	336	235	621	57	101		0.804794521	0.699404762	0.748407643
	Total	1014	837	1851	177	177	Overall →	0.825443787	0.825443787	0.825443787
							Weighted →	0.825680807	0.82544379	0.823526304

Table 7. Performance of the models – With Voting.

should be calculated both averaged and per class. The study incorporates an idea for working with individual ML algorithm as well as four different types of ensemble approaches that were not covered in earlier research. A wide range of ensemble techniques used to address a variety of issues, including bias and variance reduction. It also efficiently solves overfitting concerns[72]. As a result, this study offered a state-of-the-art by training multiple ML models as base learners and to improve the prediction performance utilizing meta-learner.

The QEML-MHRC framework applied for processing the data using four different ensemble methodologies, in addition to implementing the ML models individually (without ensemble). The implementation was wide, seeking to identify potential improvements by employing ensemble methods. The work incorporates numerous ML models, including RF, DT, GBT, and KNN, which are then used multiple times via ensemble approaches. The ML models chosen based on their performance when applied to medical datasets[16,70,73]. As a result, we used appropriate evaluation criteria to assess class achievement. Precision, recall, and F1 values calculated using class-wise, overall, and weighted equations.

Figure 10 compares the weighted precision values obtained by all investigations using a three-dimensional line graph. The diagram depicts a summary of all ML models in each category. This comparison provides a summary of the model's performance and capacity to predict maternal health risks in pregnant women. Among all experiments, the stacking method clearly delivers the best performance. We also utilized different meta-learners for stacking, with GBT (0.85) outperformed others. Even after combining it with multiple models, we can infer that KNN is the worst performer, implying that it is inadequate for the used MHR dataset.

The weighted recall is the next evaluation metric used to assess the model's performance. Figure 11 displays the weighted recall comparison of all models within each category. It demonstrates that the recall values for all ML models are identical, except for a little rising curve achieved by GBT. Again, ensemble stacking outperformed all others, as multiple models combined for creating each stacking model. Techniques such as Decision Tree, and Gradient Boosting are highly effective in dealing with high-dimensional data, specially when applied using stacking approach. On the other side, KNN scored the lowest in every category. The KNN algorithm's

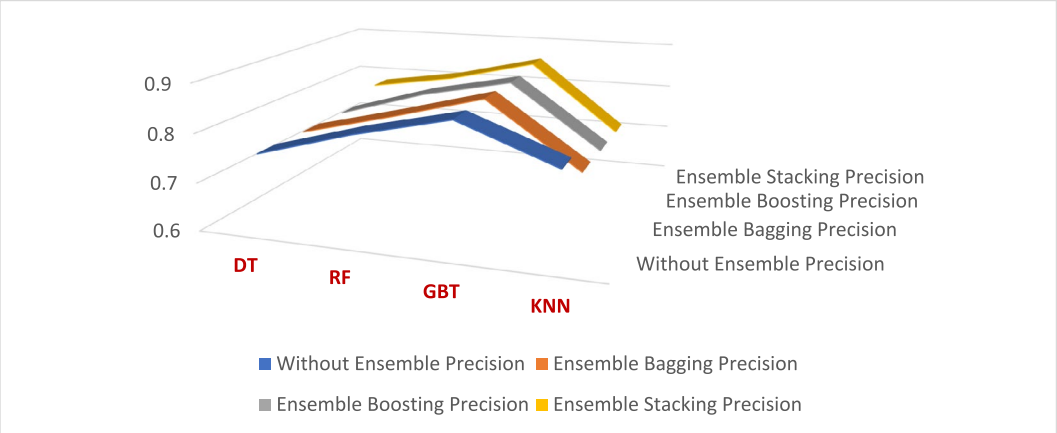

Fig. 10. Weighted precision comparisons for all experiments.

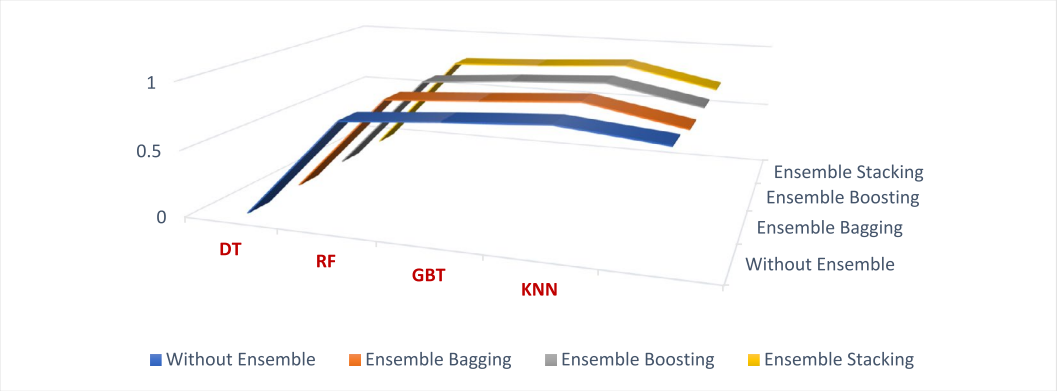

Fig. 11. Weighted recall comparisons for all experiments.

underperformance could be attributed to the fact that it focuses mostly on measuring distances between data points. Sometimes the number of data values in each dimension makes accurate classification challenging. It can also be improved by adjusting the value of "K" to better suit the specific dataset.

In this research, the final evaluation criteria used is known as weighted F1. Figure 12 illustrates the comparing scores. The study emphasized the importance of numerous techniques for predicting maternal health risk in women based on several factors and findings of the experiments. Overall, the ensemble stacking with GBT (stacking model learner) outperformed the model, scoring 0.86 for all classes (low, medium, and high) associated with maternal health risk factors.

The F1 score is a valuable metric in machine learning specially when dealing with multi-classification problem. It integrates precision and recall values into a single metric, which provides more comprehensive view of model performance. Precision and recall calculated using the ratio of true positives, false positives, and false negatives predictions, whereas the calculation of F1-score is based on harmonic mean of precision and recall values. It considers both false positives and false negatives prediction in a single metric. The use of F1-score is particularly important when you need a balance between precision and recall, specially when an uneven class distribution may bias simpler metrics such as accuracy. Using a combination of metrics provides a better understanding of a model's performance. In a multi-classification problem, the accuracy alone is insufficient to assess the model's performance. As a result, this study employed weighted scores per class to better comprehend each category's performance. F1-score is often considered more balanced and unbiased metric than other single metrics like as precision and recall. High precision does not always imply that the model is good; similarly, high recall can indicate that the model performance is good, but it does not account for false positives predictions. On the other hand, the F1-score provides a balance between precision and recall, ensuring that neither false positives nor false negatives prioritized. Therefore, this study focused and presented all relevant evaluation metrics such as true positive rate, true negative rate, false positive rate, false negative rate, precision, recall, and F1-score to understand the model's performance comprehensively.

The findings clearly indicate that the proposed QEML-MHRC framework employs ensemble ML approaches, which have numerous advantages over individual ML models. Firstly, it reduces forecast variance by averaging the findings of multiple models. It further mitigates the impact of abnormalities discovered in the single training

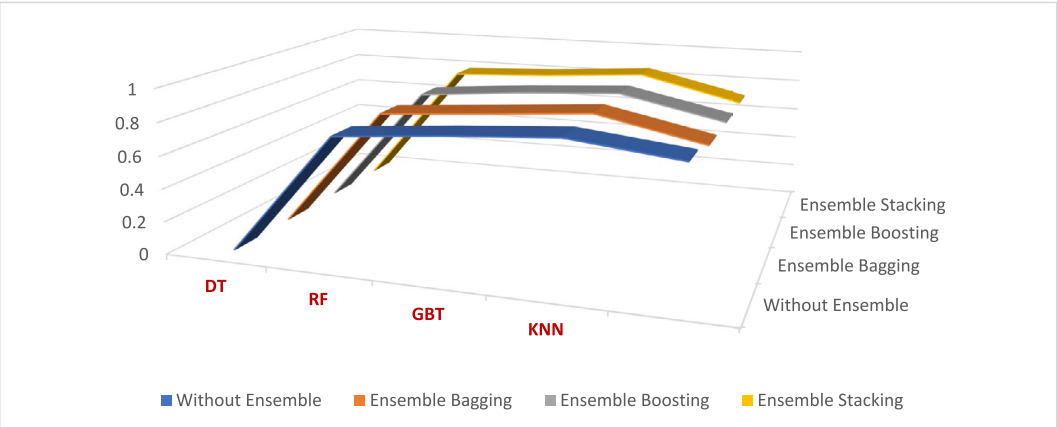

Fig. 12. Weighted F1 comparisons for all experiments.

dataset. The concept is further enhanced by using tenfold cross-validation procedures, automatically eliminating concerns such as overfitting and dataset bias. Secondly, boosting was another ensemble strategy utilized in this study to reduce the number of errors caused by other models. Boosting is a method that works across multiple iterations to reduce bias and variation, resulting in effective and accurate predictions. Moreover, ensembles incorporate models with multiple structures and learning algorithms, allowing the model to be trained and learn all the patterns in the data. For example, stacking is the third techniques applied in the research, which utilizes many models as a base learner. It further connects the output to a meta learner that integrates their predictions to improve overall performance. The use of ensembles also provides an additional advantage by demonstrating the ability to generalize to previously unseen data, which is useful in this situation where the data is complex in nature and the target variable has multiple classes. Model training is strengthened by employing different parameters, reducing the risk of depending on a single, potentially overfitted model.

Conclusion

Maternal health risk identification is critical, particularly in reducing the number of maternal deaths. This study investigated the issue using real-world data acquired from various hospitals with patients during their pregnancy. The dataset includes multi-class attributes for categorizing the level of risk associated with each patient. According to the maternal health exploratory data analysis, the most important variables driving high risk for pregnant women are high blood pressure, low blood pressure, and high blood sugar levels. Furthermore, all variables in the dataset are strongly correlated and have been shown to help predict maternal health risks. To address the challenge of dealing with multi-class attributes, we proposed the QEML-MHRC framework, which made up of various ML models and implemented using four different ensemble techniques. To provide an effective learning environment, we trained the model using ensemble techniques. In terms of class performance, the dataset associated with the "HR" class had the highest accuracy and other metrics, as well as a correct prediction performance of 0.90. GBT with the ensemble stacking approach outperformed and demonstrated outstanding performance for all evaluation measures (0.86) for all classes available in the dataset.

The study's findings can help doctors and consultants predict maternal health concerns and reduce maternal death rates. The study provided an innovative approach for dealing with the patients experiencing difficulties throughout pregnancy. The suggested approach has demonstrated exceptional accuracy in predicting the extent of risk engagement utilizing several criteria. The application of advanced predictive modeling approaches assures that the findings are applicable across groups and can address gaps in maternal health outcomes. Authors identifies that the dataset has a limited number of features; however, using a large, diverse dataset that includes additional factors such as demographic and socioeconomic factors can improve the idea presented in this study. Furthermore, collaborating with specialist in other domains (e.g., obstetrics, and public health) can help to improve data dimensions. The authors can enhance the datasets in the future by collaborating with additional medical organizations. Modifications to the dataset could help to improve the performance of the suggested system.

Data availability

"The datasets analysed during the current study are available in the "UCI Machine Learning Repository". The dataset can be accessed through the web link https://archive.ics.uci.edu/dataset/863/maternal+health+risk".

Received: 23 November 2023; Accepted: 2 September 2024
Published online: 14 September 2024

References

1. World Health Organization. A woman dies every two minutes due to pregnancy or childbirth: UN agencies.
2. Bertini, A., Salas, R., Chabert, S., Sobrevia, L. & Pardo, F. Using machine learning to predict complications in pregnancy: A systematic review. *Front. Bioeng. Biotechnol.* **9**, 1385 (2022).
3. Giouleka, S. *et al.* Obesity in pregnancy: A comprehensive review of influential guidelines. *Obstet. Gynecol. Surv.* **78**(1), 50–68 (2023).
4. Poniedziałek-Czajkowska, E., Mierzyński, R. & Leszczyńska-Gorzelak, B. Preeclampsia and obesity—The preventive role of exercise. *Int. J. Environ. Res. Public Health* **20**(2), 1267 (2023).
5. Bogren, M., Denovan, A., Kent, F., Berg, M. & Linden, K. Impact of the helping mothers survive bleeding after birth learning programme on care provider skills and maternal health outcomes in low-income countries—An integrative review. *Women Birth* **34**(5), 425–434 (2021).
6. Varghese, B. *et al.* Integrated metabolomics and machine learning approach to predict hypertensive disorders of pregnancy. *Am. J. Obstet. Gynecol. MFM* **5**(2), 100829 (2023).
7. Aljameel, S. S. *et al.* Prediction of preeclampsia using machine learning and deep learning models: A review. *Big Data Cogn. Comput.* **7**(1), 32 (2023).
8. Ullah, Z., Saleem, F., Jamjoom, M. & Fakieh, B. Reliable prediction models based on enriched data for identifying the mode of childbirth by using machine learning methods: Development study. *J. Med. Internet Res.* **23**(6), e28856 (2021).
9. Rawashdeh, H. *et al.* Intelligent system based on data mining techniques for prediction of preterm birth for women with cervical cerclage. *Comput. Biol. Chem.* **85**, 107233 (2020).
10. Patel, S. S. Explainable machine learning models to analyse maternal health. *Data Knowl. Eng.* **146**, 102198 (2023).
11. Ullah, Z. *et al.* Detecting high-risk factors and early diagnosis of diabetes using machine learning methods. *Comput. Intell. Neurosci.* **2022**, 1–10 (2022).
12. Alsolami, F. *et al.* A unified decision-making technique for analysing treatments in pandemic context. *Comput. Mater. Contin.* **73**, 2591–2618 (2022).
13. Saleem, F., Al-Ghamdi, A.S.A.-M., Alassafi, M. O. & AlGhamdi, S. A. Machine learning, deep learning, and mathematical models to analyze forecasting and epidemiology of COVID-19: A systematic literature review. *Int. J. Environ. Res. Public Health* **19**(9), 5099 (2022).

14. Diniz, P. H. B., Yin, Y. & Collins, S. Deep learning strategies for ultrasound in pregnancy. *Eur. Med. J. Reprod. Health* **6**(1), 73 (2020).
15. Yousefpour Shahrivar, R., Karami, F. & Karami, E. Enhancing fetal anomaly detection in ultrasonography images: A review of machine learning-based approaches. *Biomimetics* **8**(7), 519 (2023).
16. Ahmed, M., Kashem, M. A., Rahman, M. & Khatun, S. Review and analysis of risk factor of maternal health in remote area using the internet of things (IoT). In *ECCE2019: Proceedings of the 5th International Conference on Electrical, Control & Computer Engineering, Kuantan, Pahang, Malaysia, 29th July 2019* 357–365 (Springer, 2020)
17. Alshammari, W. & Saleem, F. A ML framework for early detecting the likelihood of cardiovascular disease in a patient using multi-attributes. *Int. J. Adv. Res. Comput. Commun. Eng.* **11**(9), 73–80 (2022).
18. Alsolami, F. J. *et al.* Impact assessment of COVID-19 pandemic through machine learning models. *Comput. Mater. Contin.* **68**(3), 2895. https://doi.org/10.32604/cmc.2021.017469 (2021).
19. Oh, W. & Nadkarni, G. N. Federated learning in health care using structured medical data. *Adv. Kidney Dis. Health* **30**(1), 4–16 (2023).
20. Mohammed, S. S., Menaouer, B., Zohra, A. F. F. & Nada, M. Sentiment analysis of COVID-19 tweets using adaptive neuro-fuzzy inference system models. *Int. J. Softw. Sci. Comput. Intell. (IJSSCI)* **14**(1), 1–20 (2022).
21. Iwendi, C. *et al.* COVID-19 patient health prediction using boosted random forest algorithm. *Front. Public Health* **8**, 357 (2020).
22. Srivastava, A., Samanta, S., Mishra, S., Alkhayyat, A., Gupta, D. & Sharma, V. Medi-Assist: A decision tree based chronic diseases detection model. In *2023 4th International Conference on Intelligent Engineering and Management (ICIEM)* 1–7 (IEEE, 2023)
23. Mahoto, N. A. *et al.* A machine learning based data modeling for medical diagnosis. *Biomed. Signal Process. Control* **81**, 104481 (2023).
24. Fakieh, B., AL-Ghamdi, A. A., Saleem, F. & Ragab, M. Optimal machine learning driven sentiment analysis on COVID-19 twitter data. *Comput. Mater. Contin.* **75**(1), 81–97 (2023).
25. Hartono, A. *et al.* Machine learning classification for detecting heart disease with K-NN algorithm, decision tree and random forest. *Eksakta Berk. Ilm. Bid. MIPA* **24**(4), 513–522 (2023).
26. Thotad, P. N., Bharamagoudar, G. R. & Anami, B. S. Diabetes disease detection and classification on Indian demographic and health survey data using machine learning methods. *Diabetes Metab. Syndr. Clin. Res. Rev.* **17**(1), 102690 (2023).
27. Farjana, A. *et al.* Predicting chronic kidney disease using machine learning algorithms. In *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)* 1267–1271 (IEEE, 2023).
28. Kumawat, G. *et al.* Prognosis of cervical cancer disease by applying machine learning techniques. *J. Circuits Syst. Comput.* **32**(01), 2350019 (2023).
29. Menaouer, B., Zahra, A. F. & Mohammed, S. Multi-class sentiment classification for healthcare tweets using supervised learning techniques. *Int. J. Serv. Sci. Manag. Eng. Technol. (IJSSMET)* **13**(1), 1–23 (2022).
30. Song, X., Wang, C., Wang, T., Zhang, S. & Qin, J. Obesity and risk of gestational diabetes mellitus: A two-sample Mendelian randomization study. *Diabetes Res. Clin. Pract.* **197**, 110561 (2023).
31. Johns Hopkins Med. Gestational diabetes mellitus (GDM) (accessed 10 Feb 2023); [Online]. Available https://www.hopkinsmedicine.org/health/conditions-and-diseases/diabetes/gestational-diabetes
32. Chang, K.-J., Seow, K.-M. & Chen, K.-H. Preeclampsia: Recent advances in predicting, preventing, and managing the maternal and fetal life-threatening condition. *Int. J. Environ. Res. Public Health* **20**(4), 2994 (2023).
33. Malm, G. *et al.* Maternal serum vitamin D level in early pregnancy and risk for preeclampsia: A case-control study in Southern Sweden. *PLoS One* **18**(2), e0281234 (2023).
34. Sun, H. *et al.* Identification of suspicious invasive placentation based on clinical MRI data using textural features and automated machine learning. *Eur. Radiol.* **29**, 6152–6162 (2019).
35. Jhee, J. H. *et al.* Prediction model development of late-onset preeclampsia using machine learning-based methods. *PLoS One* **14**(8), e0221202 (2019).
36. Lakshmi, B. N., Indumathi, T. S. & Ravi, N. A comparative study of classification algorithms for risk prediction in pregnancy. In *TENCON 2015–2015 IEEE Region 10 Conference* 1–6 (IEEE, 2015).
37. Sultana, M. I., Lovely, M. L. S., & Hasan, M. M. Building prediction models for maternal mortality rate in Bangladesh. In *2019 5th International Conference on Advances in Electrical Engineering (ICAEE)* 375–380 (IEEE, 2019).
38. Wang, S., Rexrode, K. M., Florio, A. A., Rich-Edwards, J. W. & Chavarro, J. E. Maternal mortality in the United States: Trends and opportunities for prevention. *Annu. Rev. Med.* **74**, 199–216 (2023).
39. Anwar, J., Torvaldsen, S., Morrell, S. & Taylor, R. Maternal mortality in a rural district of Pakistan and Contributing Factors. *Matern. Child Health J.* **27**, 1–14 (2023).
40. Özsezer, G & Mermer, G. Prevention of maternal mortality: Prediction of health risks of pregnancy with machine learning models. Available at *SSRN 4355295* (2023).
41. Baig, A. R. *et al.* Light-Dermo: A lightweight pretrained convolution neural network for the diagnosis of multiclass skin lesions. *Diagnostics* **13**(3), 385 (2023).
42. Ahmed, M. *Maternal health risk data set data set*, UCI Machine Learning Repository; (accessed 05 February 2023) [Online]. Available https://archive.ics.uci.edu/ml/datasets/Maternal+Health+Risk+Data+Set
43. Leemans, S. J. J., Partington, A., Karnon, J. & Wynn, M. T. Process mining for healthcare decision analytics with micro-costing estimations. *Artif. Intell. Med.* **135**, 102473 (2023).
44. Silva, M. D. B., de Oliveira, R. D. V. C., da Alves, S. B. D. & Melo, E. C. P. Predicting risk of early discontinuation of exclusive breastfeeding at a Brazilian referral hospital for high-risk neonates and infants: A decision-tree analysis. *Int. Breastfeed. J.* **16**(1), 1–13 (2021).
45. Arayeshgari, M., Najafi-Ghobadi, S., Tarhsaz, H., Parami, S. & Tapak, L. Machine learning-based classifiers for the prediction of low birth weight. *Healthc Inform. Res.* **29**(1), 54–63 (2023).
46. Priscila, S. S. & Kumar, C. S. Classification of medical datasets using optimal feature selection method with multi-support vector machine. In *Advancements in Smart Computing and Information Security: First International Conference, ASCIS 2022, Rajkot, India, November 24–26, 2022, Revised Selected Papers, Part I* 220–232 (Springer, 2023).
47. Zou, S. & Wu, Z. A narrative review of the application of machine learning in venous thromboembolism. *Vascular* **32**, 698. https://doi.org/10.1177/17085381231153216 (2023).
48. Kazijevs, M. & Samad, M. D. Deep imputation of missing values in time series health data: A review with benchmarking. Preprint at arXiv:2302.10902 (2023).
49. Kiangala, S. K. & Wang, Z. An effective adaptive customization framework for small manufacturing plants using extreme gradient boosting-XGBoost and random forest ensemble learning algorithms in an Industry 4.0 environment. *Mach. Learn. Appl.* **4**, 100024 (2021).
50. Deshpande, H. S. & Ragha, L. A hybrid random forest-based feature selection model using mutual information and F-score for preterm birth classification. *Int. J. Med. Eng. Inform.* **15**(1), 84–96 (2023).
51. Chaula, R. B. & Justo, G. N. A robust random forest prediction model for mother-to-child hiv transmission based on individual medical history. *Tanzania Journal of Engineering and Technology*, vol. 41, no. 3, (2023)@@
52. Soleymani, F., Masnavi, H. & Shateyi, S. Classifying a lending portfolio of loans with dynamic updates via a machine learning Technique. *Mathematics* **9**(1), 17 (2021).

53. Zhao, C., Peng, R. & Wu, D. Bagging and boosting fine-tuning for ensemble learning. *IEEE Trans. Artif. Intell.* **5**, 1728 (2023).
54. RM Documentation. AdaBoost; (accessed 20 May 2021) [Online]. Available https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/ensembles/adaboost.html#:~:text=AdaBoost%2C short for Adaptive Boosting,instances misclassified by previous classifiers.
55. Xiong, Y., Ye, M. & Wu, C. Cancer classification with a cost-sensitive Naive Bayes stacking ensemble. *Comput. Math. Methods Med.* **2021**, 5556992 (2021).
56. Chao, L. I., Wen-Hui, Z., Ran, L. I., Jun-Yi, W. & Ji-Ming, L. Research on star/galaxy classification based on stacking ensemble learning. *Chin. Astron. Astrophys.* **44**(3), 345–355 (2020).
57. Cui, S., Yin, Y., Wang, D., Li, Z. & Wang, Y. A stacking-based ensemble learning method for earthquake casualty prediction. *Appl. Soft. Comput.* **101**, 107038 (2021).
58. Zhang, H., Li, J. L., Liu, X. M. & Dong, C. Multi-dimensional feature fusion and stacking ensemble mechanism for network intrusion detection. *Futur. Gener. Comput. Syst.* **122**, 130–143 (2021).
59. Dogan, A. & Birant, D. A weighted majority voting ensemble approach for classification. In *2019 4th International Conference on Computer Science and Engineering (UBMK)* 1–6 (IEEE, 2019, September)
60. Rapid Miner Team. Rapid Miner; (accessed 01 March 2023) [Online]. Available https://rapidminer.com/
61. Saleem, F., Ullah, Z., Fakieh, B. & Kateb, F. Intelligent decision support system for predicting student's E-learning performance using ensemble machine learning. *Mathematics* **9**(17), 2078 (2021).
62. Burlaka, R. *Testing the fraud detection algorithms of online chess platform and exploring ways to improve them using data mining techniques*. (2023).
63. Mirbod, M. & Dehghani, H. Smart trip prediction model for metro traffic control using data mining techniques. *Procedia Comput. Sci.* **217**, 72–81 (2023).
64. Alsolami, F. J., Saleem, F. & Abdullah, A. L. Predicting the accuracy for telemarketing process in banks using data mining. *Comp. It. Sci* **9**, 69–83 (2020).
65. Khounraz, F. *et al.* Prognosis of COVID-19 patients using lab tests: A data mining approach. *Health Sci. Rep.* **6**(1), e1049 (2023).
66. Kumbhar, C. & Hussain, A. Prediction of Diabetics in the Early Stages Using Machine-Learning Tools and Microsoft Azure AI Services. In *Machine Learning, Blockchain, and Cyber Security in Smart Environments* 59–80 (Chapman and Hall/CRC, 2023).
67. Winoto, A. A. & Roy, A. F. V. Model of predicting the rating of bridge conditions in Indonesia with regression and K-fold cross validation. *Int. J. Sustain. Constr. Eng. Technol.* **14**(1), 249–259 (2023).
68. Eltrass, A. S., Tayel, M. B. & Ammar, A. I. Automated ECG multi-class classification system based on combining deep learning features with HRV and ECG measures. *Neural Comput. Appl.* **34**(11), 8755–8775 (2022).
69. Dritsas, E. & Trigka, M. Supervised machine learning models for liver disease risk prediction. *Computers* **12**(1), 19 (2023).
70. Ahmed, M. & Kashem, M. A. IoT based risk level prediction model for maternal health care in the context of Bangladesh. In *2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI)* 1–6 (IEEE, 2020).
71. Lango, M. & Stefanowski, J. What makes multi-class imbalanced problems difficult? An experimental study. *Expert Syst. Appl.* **199**, 116962 (2022).
72. Ganaie, M. A., Hu, M., Malik, A. K., Tanveer, M. & Suganthan, P. N. Ensemble deep learning: A review. *Eng. Appl. Artif. Intell.* **115**, 105151 (2022).
73. 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), 1–19 (2022).

Author contributions

Data curation: Farrukh Saleem and Zahid Ullah; Writing original draft: Shitharth S; Supervision: Adil O. Khadidos; Alaa O. Khadidos; Project administration: Adil O. Khadidos; Alaa O. Khadidos; Conceptualization: Shitharth S; Methodology: Shitharth S; Validation: Adil O. Khadidos Visualization: Adil O. Khadidos; Resources: Farrukh Saleem and Zahid Ullah; Overall Review & Editing: Shitharth S All authors reviewed the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.