scientific reports

OPEN

Prediction of high-risk pregnancy based on machine learning algorithms

Xinyu Pi[1], Junzhi Wang[2✉], Liangliang Chu[2], Guochun Zhang[2] & Wenli Zhang[1]

This study explores the application of machine learning algorithms in predicting high-risk pregnancy among expectant mothers, aiming to construct an efficient predictive model to improve maternal health management. The study is based on the maternal health risk dataset (MHRD) from Bangladesh, covering multiple hospitals, community clinics, and maternal healthcare centers, and encompassing health data from 1014 pregnant women. Six machine learning algorithms—multilayer perceptron (MLP), logistic regression (LR), decision tree (DT), random forest (RF), eXtreme Gradient Boosting (XGBoost), and support vector machine (SVM)—are employed to construct predictive models. It is worth noting that MLP demonstrates superior performance compared with the other five algorithms. By applying the MLP method, the study successfully established an efficient pregnancy risk prediction model. The model evaluation results indicate that it has high accuracy in predicting pregnancy risks, with an overall accuracy rate of 82%, and particularly high accuracy in high-risk predictions, reaching 91%. With the computational support of an NVIDIA GPU RTX3050Ti, the model demonstrated excellent data processing capabilities, capable of predicting and processing 500 sets of data items per second. This study not only showcases the enormous potential of machine learning technology in the healthcare field, especially in the rapid and accurate identification of high-risk pregnancies, providing a powerful decision-support tool for medical professionals, but also offers significant reference value for future research in this area.

Keywords High-risk pregnancy, Machine learning, Multilayer perceptron, Nursing

Abbreviations
MHRD   Maternal health risk dataset
MLP   Multilayer perceptron
LR   Logistic regression
DT   Decision tree
RF   Random forest
XGBoost   EXtreme Gradient Boosting
SVM   Support vector machine
ROC   Receiver operating characteristic

During pregnancy, certain complications or pathogenic factors may threaten the life safety of the pregnant woman, fetus, and newborn, which are defined as high-risk pregnancies[1]. In recent years, researchers such as Mavreli et al.[2] have emphasized that early identification of high-risk pregnancies can effectively reduce the risk of perinatal mortality, pregnancy-related complications, and neonatal complications. However, due to a lack of professional knowledge and assessment skills, junior nursing staff may have deficiencies in clinical decision-making capabilities, which could affect the accuracy of pregnancy risk stratification[3]. It is noteworthy that significant progress has been made in the field of obstetrics and gynecology using machine learning technology. Researchers like Kovacheva[4] have predicted early and late pregnancy risks from clinical and genetic perspectives through machine learning and polygenic risk scores. Chu et al.[5] used machine learning methods to predict the risk of adverse events in pregnant women with congenital heart disease. These advances indicate that, with the development of smart medicine, machine learning technology has shown great advantages in the field of disease prediction and diagnosis[6,7].

[1]School of Nursing, Shandong First Medical University (Shandong Academy of Medical Sciences), Jinan 250117, Shandong, China. [2]Department of Obstetrics, The First Affiliated Hospital of Shandong First Medical University (Shandong Provincial Qianfoshan Hospital), Jinan 250014, Shandong, China. ✉email: 3307962366@qq.com

It is crucial to identify high-risk vulnerable populations as early as possible during pregnancy[8,9]. Developing innovative predictive methods is essential for identifying high-risk pregnant women[10,11]. This study introduces an innovative predictive model based on the MLP machine learning algorithm for identifying high-risk pregnancies. Compared with five other machine learning algorithms, it demonstrates superior performance, filling a gap in the existing research on MLP-based predictive models for high-risk pregnancies.

The model's performance is evaluated by assessing its accuracy, precision, recall, F1 score, and AUC in predicting high-, medium-, and low-risk pregnancies. This approach enables healthcare providers to detect high-risk pregnancies using basic clinical parameters, including age, systolic and diastolic blood pressure, blood glucose levels, body temperature, and heart rate. This facilitates the implementation of preventive measures. After prediction, pregnant women can receive targeted care based on their predicted risk levels, ensuring a smooth pregnancy and reducing the occurrence of complications[12]. Therefore, this study aims to explore the use of machine learning models for high-risk pregnancy risk prediction and evaluation, hoping to provide valuable reference for healthcare professionals in assessing and treating high-risk pregnant women.

Methods
2.1 Research process
This study harnessed the open-source MHRD from Bangladeshi medical institutions, covering multiple hospitals, community clinics, and maternal healthcare centers, and encompassing data on 1014 pregnant women. Records of pregnant women aged 10–18 years were excluded due to ethical concerns and data sparsity, ensuring the dataset's clinical validity. Risk levels are stratified into low, medium, and high categories. Post-data preprocessing, the dataset is randomly divided into a training set of 362 and a test set of 90, adhering to an 8:2 ratio. To curb overfitting in the MLP model, early stopping is employed[13]. The model's predictive accuracy is gauged through a confusion matrix and Receiver Operating Characteristic (ROC) curve, with the research flowchart delineated in Fig. 1.
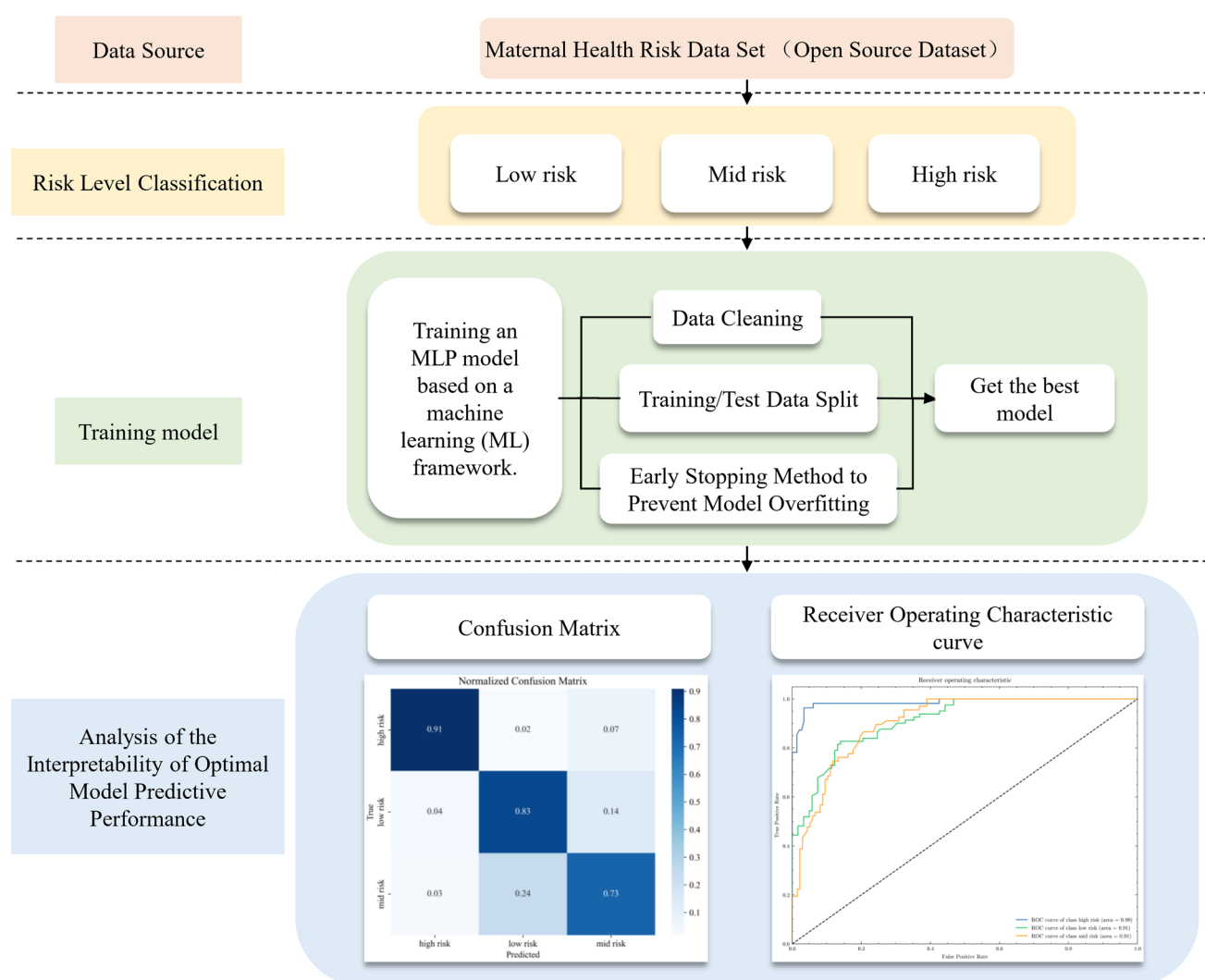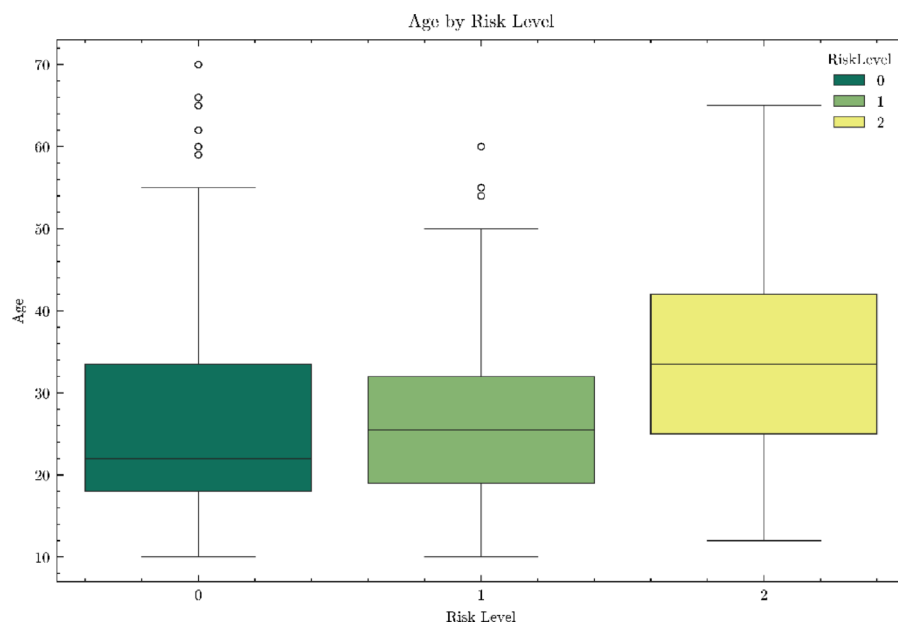

Fig. 1. Research flowchart.
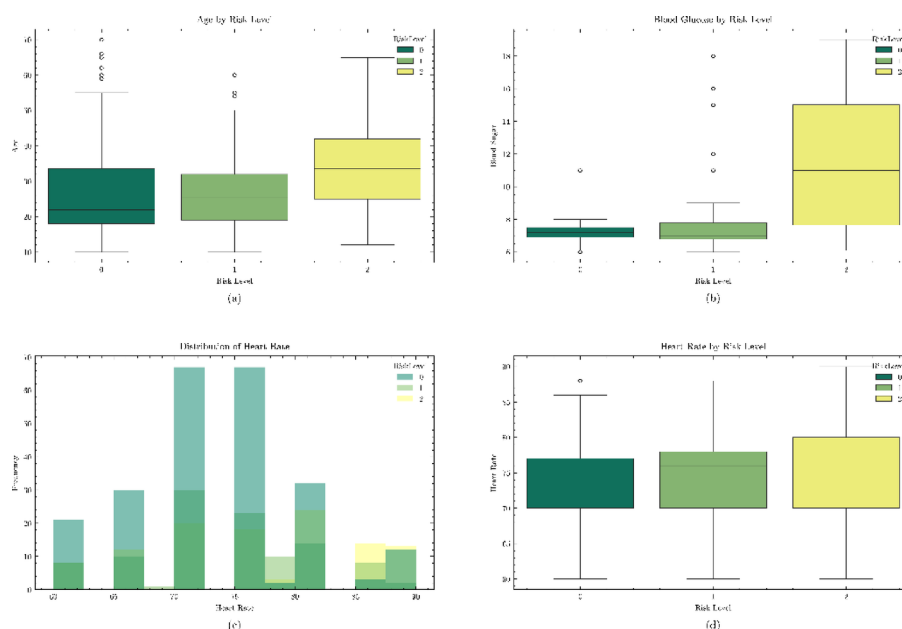
Fig. 2. Risk level distribution.

Fig. 3. Distribution chart of risk levels for different ages, blood sugar, and heart rates.

Data distribution

The input features include maternal age, systolic blood pressure, diastolic blood pressure, blood glucose levels, body temperature, heart rate, and risk level. These features were selected based on their established medical relevance to pregnancy risk assessment. After data cleaning and deduplication, a total of 452 data entries were obtained in this study, including 234 low-risk entries, 106 medium-risk entries, and 112 high-risk entries. The data distribution is shown in Fig. 2.

In Fig. 3, there are three box plots and one histogram. Figure 3a, b, d are box plots representing the distribution of age, blood sugar, and heart rate across different risk levels in the dataset, respectively. Figure 3c is a histogram showing the distribution of risk levels corresponding to different heart rates. From Fig. 3a–d, it is evident that individuals with a high pregnancy risk level have higher age, blood sugar, and heart rate compared to those with medium and low risk levels, which aligns with medical objective principles.

Machine learning methods

Six machine learning algorithms—MLP, LR, DT, RF, XGBoost, and SVM—are employed to construct predictive models. It is worth noting that MLP demonstrates superior performance compared with the other five algorithms. This choice is based on the comprehensive evaluation of each algorithm's performance in terms of accuracy, precision, and other relevant metrics. In this study, we've adopted a multifaceted strategy to refine the accuracy and generalization of our MLP model, which includes data preprocessing, the application of the SMOTE algorithm to address class imbalance, and the implementation of early stopping to prevent overfitting. The MLP model consists of three hidden layers, with 256, 128, and 64 neurons respectively. All hidden layers utilize the ReLU activation function and are followed by a SoftMax output layer. To prevent overfitting, Dropout layers with a dropout rate of 0.5 are inserted after the first two hidden layers. The model was trained using the Adam optimizer with a learning rate of 0.001 and a cross-entropy loss function. The batch size was set to 32, and the maximum number of training epochs was 10,000. Training would be halted if the validation loss did not improve for 300 epochs.

Our data cleaning process was essential for ensuring data quality, where we removed erroneous, duplicate, or irrelevant information from the open-source dataset from Bangladesh, including outliers such as records of pregnant women aged between 10 and 18 years. We quantified risk levels numerically, assigning values of 2, 1, and 0 to high, medium, and low risks, respectively, to facilitate subsequent analysis. To accurately assess model performance, we divided the dataset into a training set, comprising 80% of the data, and a test set, comprising 20%, using stratified random sampling to maintain the distribution of each category across both sets. To combat overfitting and enhance the model's ability to learn from the minority class, we applied the SMOTE algorithm and introduced Dropout layers within the MLP model. We applied SMOTE exclusively to the training set post-split to avoid data leakage The test set remained unmodified to ensure unbiased evaluation. Additionally, we employed early stopping with a patience parameter P, which monitors performance on the test set and halts training if there's no improvement for P consecutive epochs, reverting the model to its best-performing state. This approach not only prevents overfitting but also ensures that the model retains its peak performance.

For model interpretability analysis, we utilized a confusion matrix and ROC curve to assess the performance of the optimal model. The confusion matrix provides a clear view of the model's predictive accuracy, while the ROC curve offers an intuitive representation of the model's overall performance by comparing true positive and false positive rates across various thresholds. These methods validate the model's risk level predictions and enhance the interpretability of the results, thereby ensuring the reliability of our predictions.

Experimental platform

This study was conducted on a computer equipped with an NVIDIA GPU RTX3050Ti, a CPU model AMD Ryzen 7 5800H, and 1.5TB of disk space. All experiments were based on the Python programming language. To build and train the MLP model, TensorFlow (2.18.0) and Keras (3.6.0) libraries were used. Data processing and analysis benefited from Pandas (2.2.3) and NumPy (1.24.4), while data visualization was performed using Matplotlib (3.5.1). Model optimization was achieved using Sklearn (1.5.2).

Results

Table 1 presents a comparison of the performance parameters of the MLP model with five other algorithms. Among these five algorithms, the one using LR for high-risk pregnancy prediction achieved the lowest performance, with an accuracy of 0.61 and precision of 0.59. In contrast, the RF algorithm demonstrated the highest performance for high-risk pregnancy prediction, with an accuracy of 0.78 and precision of 0.79. Compared with these five algorithms, the MLP model constructed in this study exhibited superior performance, achieving an accuracy of 0.81 and precision of 0.82.

Table 2 illustrates that the MLP model performs well across different risk categories, particularly in the high-risk category, achieving an accuracy, recall, and F1 score of 0.91. The performance in the medium-risk category is slightly lower, with an accuracy and recall of 0.80 and 0.83, respectively, and an F1 score of 0.81. The low-risk category has slightly lower accuracy, recall, and F1 scores of 0.77, 0.73, and 0.75, respectively. The overall accuracy is 0.82, indicating excellent overall performance of the model.

The normalized confusion matrix shows the performance of the MLP model in predicting risk levels in high-risk pregnancies, covering high, medium, and low risk levels. The model excels in high-risk prediction with an accuracy of 91%, with misclassification rates of 2% for low risk and 7% for medium risk. The accuracy for low-risk prediction is 83%, with a 14% misclassification rate for medium risk. Medium-risk prediction accuracy stands at 73%, with misclassification rates of 24% for low risk and 3% for high risk. Overall, the model is reliable in predicting pregnancy risk for pregnant women (Fig. 4).

Algorithm	Accuracy	Precision	Recall	F1-score
MLP	0.81	0.82	0.82	0.82
LR	0.61	0.59	0.61	0.60
DT	0.73	0.75	0.73	0.74
RF	0.78	0.79	0.78	0.78
XGBoost	0.76	0.77	0.76	0.76
SVM	0.63	0.61	0.63	0.62

Table 1. Comparison of the performance of MLP with five other machine learning algorithms.

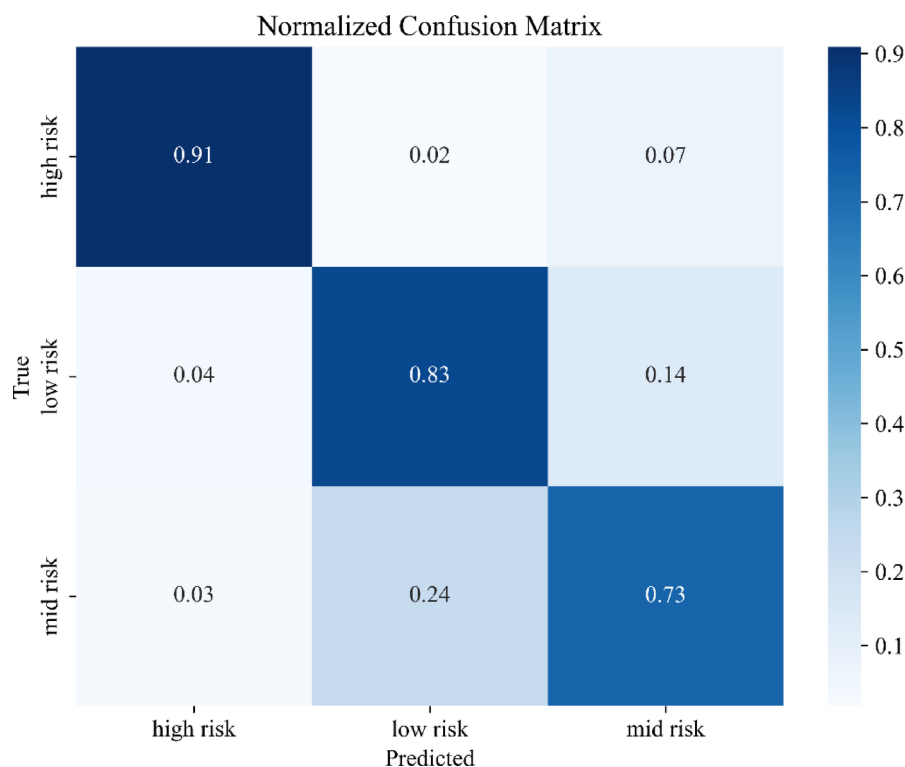|	Precision	Recall	F1-score	AUC	Sample size
High risk	0.91	0.91	0.91	0.99	22
Mid risk	0.80	0.83	0.81	0.91	21
Low risk	0.77	0.73	0.75	0.91	47
Total	0.82	–	–	–	90

Table 2. This is prediction performance of MLP model on pregnancy risk levels.

Fig. 4. Confusion matrix of MLP model predictions for different risk levels on the test set.

The ROC curve analysis shows that the MLP model achieves an AUC value of 0.99 for the high-risk category when predicting pregnancy risk levels, indicating near-perfect identification. The ROC curve closely aligns with the top left corner, demonstrating the ability to achieve a high true positive rate at a very low false positive rate (Fig. 5).

Discussion

Accurate prediction and treatment options for pregnancy risk are crucial

This study aims to develop a machine learning model for predicting high-risk pregnancy in expectant mothers. By analyzing the health data of pregnant women, the model can accurately assess pregnancy risk levels, demonstrating excellent accuracy and efficiency[14]. This tool can support clinical nurses in making more precise risk assessments during evaluations, thereby enabling appropriate interventions to reduce pregnancy risks.

High-risk pregnancy refers to a pregnancy state that faces higher risks due to various high-risk factors (such as personal health issues, pregnancy complications, adverse environmental impacts, etc.)[1,2] during the pregnancy period. Such pregnancy conditions increase maternal and neonatal morbidity and mortality rates. Taking China as an example, with the development of society, the implementation of the universal two-child policy, and the popularization of assisted reproductive technology, the incidence of high-risk pregnancies is increasing year by year. Compared to normal pregnancies, high-risk pregnancies significantly increase the risk of adverse pregnancy outcomes, including preterm birth, low birth weight, neonatal complications, and death. Moreover, the diagnosis of high-risk pregnancy may lead to reduced coping ability, decreased well-being, and increased psychiatric symptoms in pregnant women, including stress, depression, and anxiety[15].

Therefore, accurately and early identifying pregnancy risks and implementing effective management measures are crucial for improving maternal and infant health. Nurses, as the first medical team members to contact pregnant women and collect their basic information, play a critical role in the early identification and management of high-risk pregnancies[16]. However, since the health status of pregnant women may change over time, continuous monitoring by nurses is required, which is time-consuming and labor-intensive. Therefore,
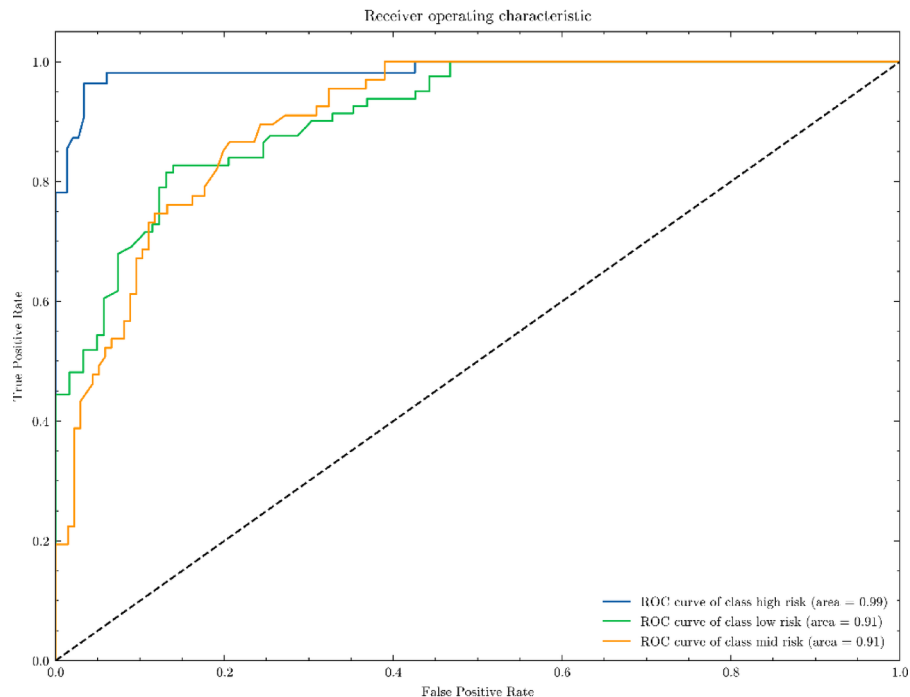
Fig. 5. ROC Curves for predictions of different risk levels on the test set by the MLP model.

developing a simple yet accurate tool for identifying high-risk pregnancies is particularly important for ensuring the health and safety of mothers and newborns.

Artificial intelligence-assisted diagnosis and treatment of pregnancy risk becomes possible

High-risk pregnancies pose significant health risks to mothers and newborns, involving various complex factors such as advanced maternal age, multiple pregnancies, and pregnancy complications, leading to increased rates of difficult labor and cesarean sections, as well as an increase in neonatal health issues[17]. These situations place high professional demands on medical and nursing staff, emphasizing the need for rigorous and professional measures. The World Health Organization points out that most deaths related to pregnancy and childbirth can be prevented through timely identification and response to pregnancy risks[18]. Timely intervention in pregnancy risks and quality healthcare services are key.

In 2017, China implemented a pregnancy risk assessment and management system, adopting a five-color grading system to assess maternal risk levels and adjust medical resource allocation accordingly, effectively reducing adverse outcomes of high-risk pregnancies[19]. However, traditional risk assessment methods have issues with large workloads and low efficiency.

The machine learning method based on the MLP proposed in this study demonstrates remarkable performance in predicting high-risk pregnancies. Compared with other methods, it achieves an accuracy rate of up to 91% in predicting high-risk cases. Moreover, the method is highly efficient, with the capability to process up to 500 sets of data per second. This method not only improves the efficiency and accuracy of risk assessment but also reduces the workload on medical staff and the demand for medical resources, showing great potential in optimizing pregnancy risk management using advanced technology.

Additionally, by utilizing the SMOTE algorithm[20] and early stopping mechanisms[21], our model has demonstrated remarkable generalization capabilities. It effectively addresses the issue of data imbalance and prevents overfitting. This ensures that the model performs well not only on the training data but also maintains high accuracy and reliability when applied to unseen data.

This study is innovative not only in its technical approach but also in its practical application. By accurately predicting the risks associated with high-risk pregnancies, our model serves as an important decision-support tool for clinical practice. It enables healthcare professionals to take preemptive interventions, thereby reducing the risk of complications for both mothers and infants[12].

Conclusions

This study successfully developed a machine learning based high-risk pregnancy prediction model, which estimates pregnancy risk by analyzing the health data of pregnant women, with an accuracy rate of up to 91%, and can accurately assess the level of pregnancy risk. At the same time, it also has a fast prediction speed, which not only improves the efficiency of pregnancy risk assessment, but also reduces the workload of medical personnel and the demand for medical resources, demonstrating the huge potential of using advanced technology to optimize pregnancy risk management.

Data availability
The data is provided in the supplementary information files. The dataset link is as follows, and you can download it by clicking the link:https://www.kaggle.com/datasets/csafrit2/maternal-health-risk-data.

References
1. Phillips, S. E. et al. Improving care beyond birth: A qualitative study of postpartum care after high-risk pregnancy. *J. Womens Health* **33**, 1720–1729. https://doi.org/10.1089/jwh.2024.0108 (2024).
2. Mavreli, D., Theodora, M. & Kolialexi, A. Known biomarkers for monitoring pregnancy complications. *Expert Rev. Mol. Diagn.* **21**, 1115–1117. https://doi.org/10.1080/14737159.2021.1971078 (2021).
3. Dewi, N. A., Yetti, K. & Nuraini, T. Nurses' critical thinking and clinical decision-making abilities are correlated with the quality of nursing handover. *Enferm. Clin.* **31**, S271–S275. https://doi.org/10.1016/j.enfcli.2020.09.014 (2021).
4. Kovacheva, V. P. et al. Preeclampsia prediction using machine learning and polygenic risk scores from clinical and genetic risk factors in early and late pregnancies. *Hypertension* **81**, 264–272. https://doi.org/10.1161/HYPERTENSIONAHA.123.21053 (2024).
5. Chu, R. et al. Predicting the risk of adverse events in pregnant women with congenital heart disease. *J. Am. Heart Assoc.* **9**, e016371. https://doi.org/10.1161/JAHA.120.016371 (2020).
6. Montgomery-Csobán, T. et al. Machine learning-enabled maternal risk assessment for women with pre-eclampsia (the PIERS-ML model): A modelling study. *Lancet Digit. Health* **6**, e238–e250. https://doi.org/10.1016/S2589-7500(23)00267-4 (2024).
7. Bowe, A. K., Lightbody, G., Staines, A., Murray, D. M. & Norman, M. Prediction of 2-year cognitive outcomes in very preterm infants using machine learning methods. *JAMA Netw. Open* **6**, e2349111–e2349111. https://doi.org/10.1001/jamanetworkopen.2023.49111 (2023).
8. Deng, Z., Duan, L. & Wang, K. Advancing comprehensive care for high-risk pregna-ncies: Integrating social and clinical perspectives. *Am. J. Obstet. Gynecol.* https://doi.org/10.1016/j.ajog.2025.02.002 (2025).
9. Correa-de-Araujo, R. & Yoon, S. S. Clinical outcomes in high-risk pregnancies due to advanced maternal age. *J. Womens Health* **30**, 160–167. https://doi.org/10.1089/jwh.2020.8860 (2021).
10. Cooray, S. D. et al. Development, validation and clinical utility of a risk prediction model for adverse pregnancy outcomes in women with gestational diabetes: The PeRSonal GDM model. *EClinicalMedicine* https://doi.org/10.1016/j.eclinm.2022.101637 (2022).
11. Li, S. et al. Improving preeclampsia risk prediction by modeling pregnancy trajectories from routinely collected electronic medical record data. *NPJ Digit. Med.* **5**, 68. https://doi.org/10.1038/s41746-022-00612-x (2022).
12. Li, M. et al. Healthy dietary patterns and common pregnancy complications: A prospective and longitudinal study. *Am. J. Clin. Nutr.* **114**, 1229–1237. https://doi.org/10.1093/ajcn/nqab145 (2021).
13. An, L., Wang, L. & Li, Y. HEA-Net: Attention and MLP hybrid encoder architecture for medical image segmentation. *Sensors* **22**, 7024. https://doi.org/10.3390/s22187024 (2022).
14. AlMashrafi, S. S., Tafakori, L. & Abdollahian, M. Predicting maternal risk level using machine learning models. *BMC Preg. Childb.* **24**, 820. https://doi.org/10.1186/s12884-024-07030-9 (2024).
15. Williamson, S. P., Moffitt, R. L., Broadbent, J., Neumann, D. L. & Hamblin, P. S. Coping, wellbeing, and psychopathology during high-risk pregnancy: A systematic review. *Midwifery* **116**, 103556. https://doi.org/10.1016/j.midw.2022.103556 (2023).
16. Tahmasebi, T. High-risk pregnancy: Managing complications for a safe and healthy outcome. *J. Preg. Neonatal Med.* **8**, 226. https://doi.org/10.35841/aapnm-8.5.226 (2024).
17. Nazari, M., Moayed Rezaie, S., Yaseri, F., Sadr, H. & Nazari, E. Design and analysis of a telemonitoring system for high-risk pregnant women in need of special care or attention. *BMC Preg. Childb.* **24**, 817. https://doi.org/10.1186/s12884-024-07019-4 (2024).
18. Lattof, S. R. et al. Developing measures for WHO recommendations on antenatal care for a positive pregnancy experience: A conceptual framework and scoping review. *BMJ Open* **9**, e024130. https://doi.org/10.1136/bmjopen-2018-024130 (2020).
19. Liu, Y., Luo, R. & Huang, A. The distribution of pregnant women with different pregnancy risks—4 cities, China, 2019. *China CDC Week.* **3**, 50. https://doi.org/10.46234/ccdcw2021.016 (2021).
20. Kosolwattana, T. et al. A self-inspected adaptive SMOTE algorithm (SASMOTE) for highly imbalanced data classification in healthcare. *Bio Data Min.* **16**, 15. https://doi.org/10.1186/s13040-023-00330-4 (2023).
21. Nguyen, M. H., Abbass, H. A. & McKay, R. I. Stopping criteria for ensemble of evolutionary artificial neural networks. *Appl. Soft Comput.* **6**, 100–107. https://doi.org/10.1016/j.asoc.2004.12.005 (2005).

Author contributions
P.X.Y: Prepared the first draft. W.J.Z and Z.G.C: Revised the manuscript. C.L.L: Provide assistance in data analysis, Z.W.L: Edited and finalized the manuscript. All authors contributed to editorial changes in the manuscript. All authors read and approved the final manuscript.

Declarations

Competing interests
The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-00450-3.

Correspondence and requests for materials should be addressed to J.W.

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.