A Comparison of Research for Machine Learning Approaches for Fetal Health Prediction

Shuqi Wu *

Biomedical Science, King's College London, London, United Kingdom

* Corresponding Author Email: k21218043@kcl.ac.uk

Abstract. Forecasting fetal health is essential for spotting issues throughout pregnancy. However, complicated data is often a challenge for conventional statistical techniques to cope with. This study used a publicly accessible fetal health dataset and optimized the dataset through preprocessing steps such as removing outliers, selecting relevant features, and normalizing the data. Subsequently, four models, including random forest (RF), support vector machine (SVM), lasso logistic regression (LLR), and neural network (NN), were selected to train the data. The results showed that the RF model performed best with an accuracy of 0.9494. The NN model performed the worst with an accuracy of 0.8752. At the same time, the results showed that RF can handle small datasets, provide feature importance, and resist overfitting. However, all models in this study have inherent limitations. Future research should focus on combining multiple models and using larger and more diverse datasets to further improve the prediction performance and effectiveness of clinical applications.

Keywords: Fetal Health Classification, Predictive Modelling, Machine Learning, Healthcare Analytics.

1. Introduction

Fetal health reflects the condition of the fetus and the mother during pregnancy. It will be monitored during the whole process of pregnancy and is the result of multi-factorial interaction. There are more than 33.3% of maternal deaths, 50% of stillbirths, and 25% of neonatal deaths during labor and delivery [1].

The necessity for more attention to recognizing and anticipating potential health concerns during pregnancy has become a hot topic in today's culture. This is crucial not just for improving neonatal survival, but also for lowering the long-term risk of chronic disease through early detection and treatment. Unfavorable conditions during fetal development might result in major health problems later in life, such as cardiovascular, renal, and metabolic illnesses [2, 3].

Many factors influence the health of the fetus, including physiological, genetic, and environmental factors. Diabetes, hypertension, and pre-eclampsia [4] are all maternal illnesses that can have an impact on fetal development. Genetic mutations can result in congenital cardiac problems or chromosomal abnormalities [5]. For example, cystic fibrosis, a genetic condition caused by mutations in the Cystic Fibrosis Transmembrane Conductance Regulator gene, causes newborns to be born with respiratory problems like coughing and obstructed airways [6]. Furthermore, environmental factors such as pollution, dietary inadequacies, and endocrine-disrupting chemicals may influence developmental processes throughout the prenatal period [7-10].

Despite the great advances in today's algorithms, some challenges remain. Existing models, including logistic regression, lead to inaccurate predictions due to the unbalanced datasets. Many models lack integration of different types of data, such as real-time monitoring inputs. In addition, while machine learning such as random forests (RF) is highly accurate, its computational demands make it difficult to apply to real-life applications [11].

The purpose of this paper is to evaluate and compare four models for predicting fetal health classification. This study aims to contribute to the improvement of fetal health prediction, which will ultimately improve neonatal birth rates and long-term health.

2. Method

2.1. Data Collection

This study mainly applied R for data processing and model development. Packages applied included caret, RF, etc.

The dataset used in this study is the cardiotocography (CTG) dataset, originally sourced from the UCI Machine Learning Repository [12]. The dataset has been validated in numerous studies, ensuring its reliability and relevance for medical data analysis.

The dataset contains 2126 records of features extracted from CTG exams and 21 features, aiming at predicting fetal health status using these features. The main variables are baseline fetal heart rate, fetal movements, accelerations and decelerations, histogram features of a period of fetal heart rate, etc. These features are commonly monitored in clinical observation of fetal health. The target variable is fetal health. It is a categorical variable, which stands for fetal health status, and is then classified by three expert obstetricians into 3 classes: 1 (Normal), 2 (Suspect), and 3 (Pathological).

2.2. Data Preprocessing

First off, the study examined the presence of missing values and outliers. Since the dataset had been preprocessed and there were no obvious missing values, but one outlier was found, the outlier was removed.

To streamline the analysis, a variable named "severe deceleration" with constant values was removed due to its lack of meaningful variation.

Second, in terms of feature selection, this study used correlation analysis to remove variables that had perfect multicollinearity. Features that had a significant correlation with the target variables such as fetal movement, deceleration, and baseline fetal heart rate values were retained.

To prepare for logistic regression analysis, a dichotomous variable representing the trend in the histogram of fetal heart rate, originally coded as 0 or 1, was transformed into two binary indicators to simplify the modeling process. Each new variable referred to the presence or absence of one of the original values, facilitating easier interpretation in the model.

To reduce complexity and increase interpretability, suspicious and pathological were combined into one category in the target variables, "potentially unhealthy fetus" (denoted by 1) and normal healthy fetus (denoted by 0).

2.3. Model Development and Training

In this study, to evaluate and compare the predictive performance of various models, four distinct models were used: RF, Lasso Logistic regression (LLR), Support Vector Machine (SVM), and Neural Network (NN). Each model was chosen for its unique strengths in different aspects.

RF builds many decision trees and aggregates their results for the final forecast. It excels at managing large amounts of complicated data while minimizing overfitting, resulting in high accuracy.

LLR improves logistic regression by introducing L1 regularization, which helps to pick key features and reduces overfitting. It simplifies the model by reducing some coefficients to zero, increasing its interpretability.

SVM can determine the optimal boundary for separating data into classes. It employs kernel functions to deal with high-dimensional spaces and non-linear interactions. It is renowned for its precision and clear judgment margins.

NN mimics the human brain by using layers of interconnected nodes to learn complicated patterns from data. They excel in handling vast and complicated datasets while maintaining high accuracy and flexibility in intricate relationships.

2.4. Model Evaluation

The model is evaluated using different metrics including accuracy, sensitivity, kappa, F1 score, Receiver Operating Characteristic (ROC) curve, and Area Under the Curve (AUC). These metrics allow for the assessment of classification capabilities and can be compared with others.

Accuracy is the ratio of the model's accurate predictions to all its forecasts. Sensitivity, also known as recall, is the measurement of the percentage of patients who have a positive result. This is important in medical applications because false negative results might have dangerous repercussions. Kappa assesses the degree of agreement between the actual and expected categories while considering random variation. By combining recall and precision into a single statistic, the F1 Score offers a balance between the two and provides information on how well the model performs on unbalanced datasets. ROC Curve provides a visual evaluation by showing the ratio of specificity to sensitivity across threshold settings. AUC measures the model's overall capacity to discriminate between positive and negative categories; numbers closer to 1 indicate a higher level of performance.

To address potential overfitting issues, 10-fold cross-validation was used in the training phase, which not only optimized the model parameters but also estimated the model's performance across different data subsets, ensuring the reliability and generalizability of the results.

3. Results

In this section, the performance of four predictive models will be shown, including LLR, RF, Stepwise Logistic Regression, Lasso, and Ridge Regression, on the fetal health dataset. The models are evaluated using Kappa, AUC sensitivity, and accuracy. The results are visualized through ROC curves, confusion matrices, and feature importance plots.

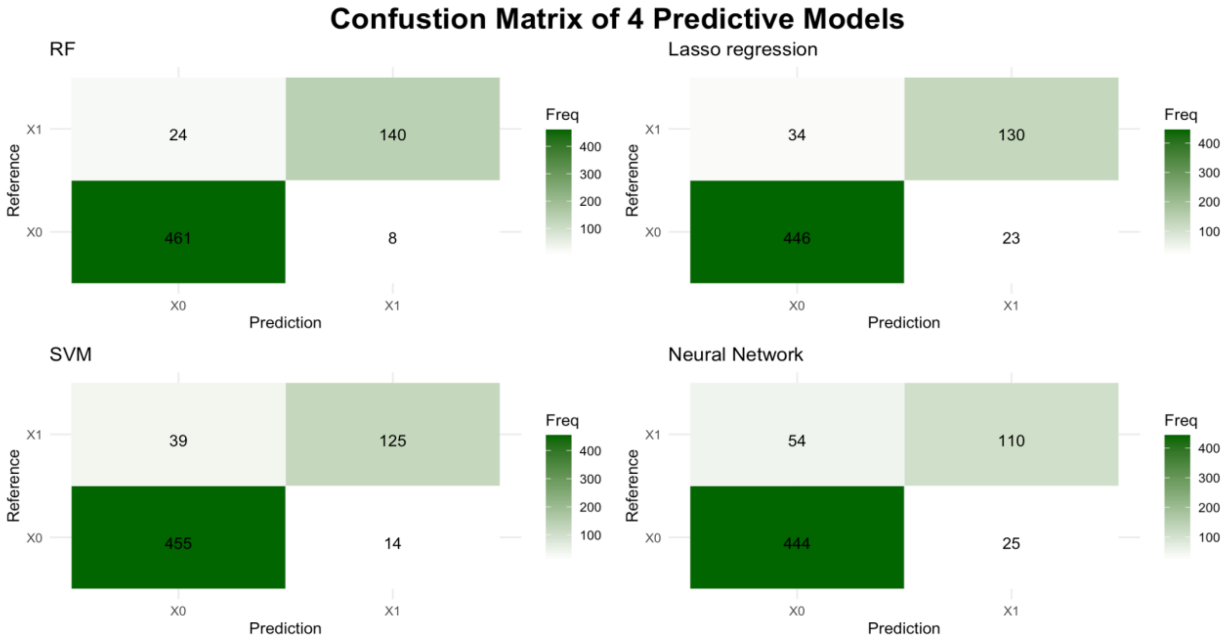

Figure 1. A heat map of the confusion matrix of four different models (Photo/Picture credit: Original).

As illustrated in Figure 1, the confusion matrix for the RF shows fewer classifications in the suspect class, compared to the LLR model which also tends to misclassify some suspect cases as normal.

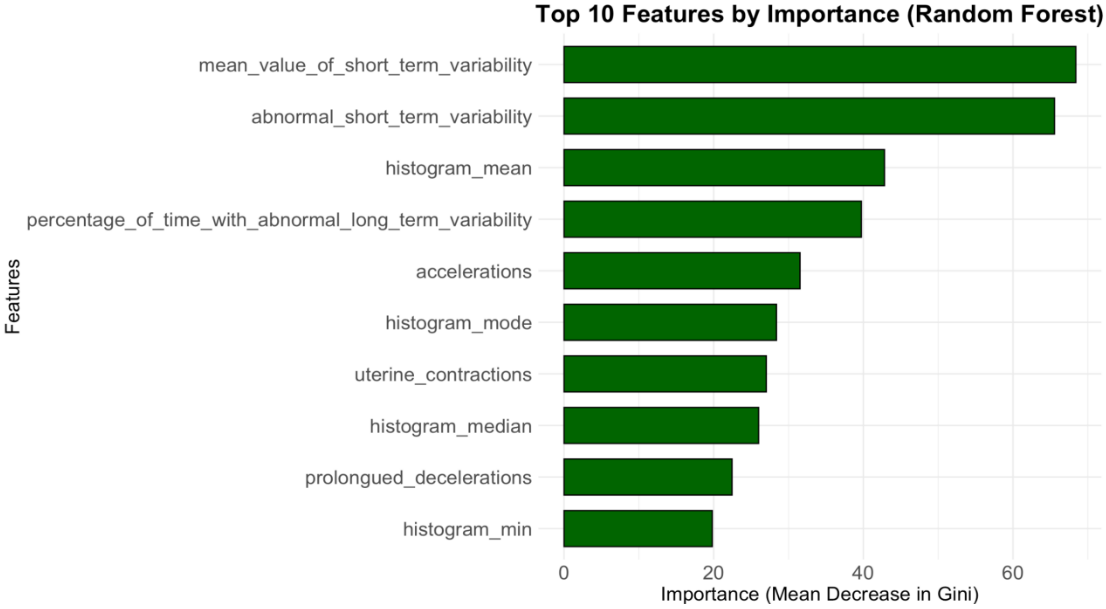

Figure 2. Feature importance in RF (Photo/Picture credit: Original).

Figure 2 indicates the top 10 features contributing to the RF model. The most significant features include the mean value of short-term variability, abnormal short-term variability, mean of histogram, and percentage of time with abnormal long-term variability.

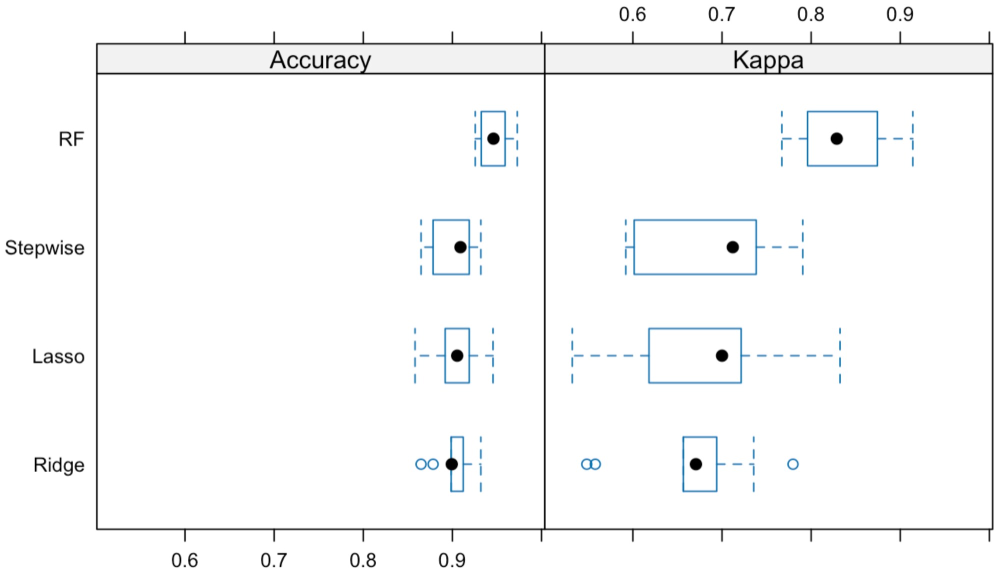

Figure 3. Box plot of RF, LLR, SVM, NN (Photo/Picture credit: Original).

Figure 3 presents the cross-validation box plots for four models, demonstrating that RF not only achieves the highest average accuracy but also has a lower variance. This result indicates better stability across different data splits.

Table 1. Comparison of five different models

Model	Sensitivity	Kappa	AUC	Accuracy	F1 score
RF	0.9829	0.8640	0.9183	0.9494	0.9665
Lasso Logistic	0.9510	0.7602	0.8718	0.9100	0.9399
SVM	0.9701	0.7705	0.8662	0.9163	0.9450
NN	0.9467	0.6551	0.8087	0.8752	0.9183

Performance Metrics Table (Table 1) summarizes key evaluation metrics for the models. These metrics collectively provide a robust evaluation of model performance, with RF demonstrating the best results across sensitivity, Kappa, AUC, accuracy, and F1 Score. In contrast, NN showed the lowest performance.

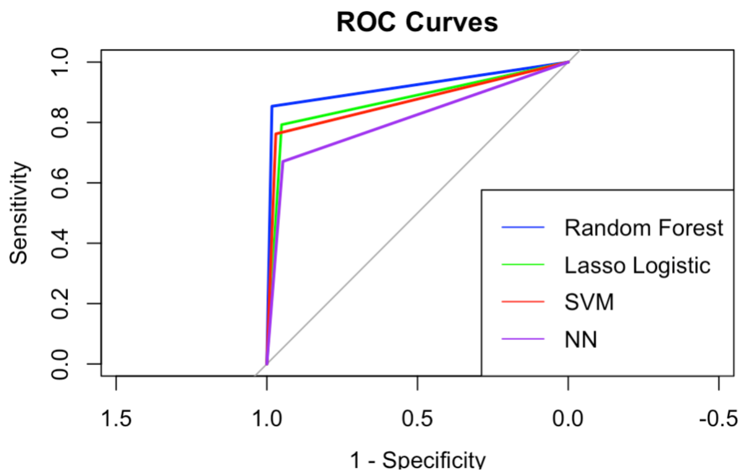

Figure 4. ROC curves for RF, LLR, SVM, NN (Photo/Picture credit: Original).

Figure 4 shows ROC curves with sensitivity (y-axis) vs. 1 - specificity (x-axis). The RF model (red line) demonstrates the best performance, with its curve closest to the top left corner and an AUC approaching 0.9665. The SVM model (blue line) also performs well, with an AUC of approximately 0.9450. Lasso Regression (red) and NN (purple) show comparatively lower performance, with their curves further from the top left.

The performance of four models—RF, SVM, LLR, and NN—was evaluated using various metrics. RF consistently outperformed the other models, obtaining the highest AUC, accuracy, sensitivity, and F1 scores, which demonstrates its excellent category differentiation and overall robustness. In comparison, SVM showed competitive performance but did not surpass RF in all metrics. LLR and NN had lower accuracy, Kappa, and F1 scores, and were less effective in classification. Overall, RF was found to be the most effective fetal health classification model, showing an optimal balance of sensitivity and specificity across all assessment metrics.

4. Discussion

This study evaluated several models for fetal health classification, including RF, SVM, LLR, and NN. The results provide insights into the strengths and limitations of each model, in line with findings from existing literature.

RF demonstrated the highest performance, achieving superior accuracy and AUC, which supports the paper highlighting RF's robustness in handling smaller datasets [13]. Subasi et al. conducted a study predicting chronic kidney disease (CKD) and found that RF outperformed other algorithms, emphasizing its robustness, particularly with limited datasets [14]. Besides, RF can manage overfitting and provide feature importance without being overly sensitive to noisy data, reinforcing its strong results in the study. In addition, its built-in feature importance estimate proved to be very useful in identifying key factors related to fetal health.

SVM also performed well, achieving competitive accuracy, though slightly lower than RF. SVM strengthens in handling high-dimensional spaces effectively. However, it often requires large datasets to reach its full potential, as noted in prior research [15]. In the study, with a relatively smaller dataset, SVM's performance was strong but did not surpass RF, supporting the view that SVM may need a more extensive dataset to maximize its accuracy.

LLR, while useful for feature selection and model simplification, showed lower performance compared to RF and SVM [16]. In a study of constructing miRNA-target regulatory networks, Stanhope et al. have highlighted its strengths in specificity and multiple target evaluation [17]. However, as research suggests, Lasso prioritizes overall model prediction at the expense of accurately interpreting the contribution of individual features, which may be a drawback in clinical applications where feature interpretability is critical [18]. In this study, while Lasso successfully simplified the model by selecting key features, its accuracy was limited, and its coefficients were less interpretable.

NN demonstrated variable performance, with lower accuracy and AUC. This reflects the need for larger datasets and careful tuning of parameters to prevent overfitting [19]. While NN has the theoretical advantage of capturing complex non-linear relationships, it did not outperform the simpler models in this study, likely due to the limited size of the dataset.

Overall, RF proved to be the most robust model for predicting fetal health, reflecting its favorable evaluation in previous research. SVM also performed well but required large datasets. Lasso Regression and NN showed limitations in this context, consistent with documented challenges. Future research might explore advanced techniques such as XGBoost, which have shown promise in complex non-linear relationships and could potentially address some of the limitations observed with the current models [20].

5. Conclusion

This study compared four machine learning models—RF, SVM, LLR, and NN—for fetal health prediction. Findings show that RF delivered the best performance, achieving the highest accuracy, AUC, and sensitivity. This confirms the robustness of RF in handling median-sized datasets and complex feature interactions without overfitting. SVM also demonstrated strong performance, though it is likely to perform better with larger datasets. Lasso Regression, while valuable for model simplification and feature selection, exhibited lower predictive accuracy, highlighting its limitation in estimating individual feature contributions reliably. NN, despite its theoretical ability to capture non-linear patterns, performed sub-optimally due to the dataset size.

This research highlights the strengths of RF for clinical applications in fetal health monitoring, where accurate predictions are critical for early intervention. Future research could explore advanced models to further enhance prediction performance. Integrating machine learning in healthcare could improve diagnostic accuracy and patient care, leading to improved maternal and fetal health.

References

[1] World Health Organization. (2020, Dec 15). Monitoring childbirth in a new era for maternal health. World Health Organization. https://www.who.int/news/item/15-12-2020-monitoring-childbirth-in-a-new-era-for-maternal-health.

[2] Calkins, K., Devaskar, S. U. (2011, June 17). Fetal origins of adult disease. Current Problems in Pediatric and Adolescent Health Care. https://www.sciencedirect.com/science/article/pii/S1538544211000265.

[3] Tarantal, A. F., Berglund, L. Obesity and lifespan health-importance of the fetal environment. MDPI. 2014.

[4] Tenenbaum-Gavish, K., Hod, M. (2013, June 11). Impact of maternal obesity on Fetal Health. Karger Publishers. https://karger.com/fdt/article/34/1/1/136587/Impact-of-Maternal-Obesity-on-Fetal-Health.

[5] Gobbo, G. F. D., Konwar, C., Robinson, W. P. (2019, September 25). The significance of the placental genome and methylome in fetal and maternal health - human genetics. SpringerLink. https://link.springer.com/article/10.1007/s00439 - 019 - 02058 - w.

[6] Bell, S. C., Mall, M. A., Gutierrez, H., Macek, M., Madge, S., Davies, J. C., ... & Ratjen, F. The future of cystic fibrosis care: a global perspective. The Lancet Respiratory Medicine, 2020, 8 (1), 65 - 124.

[7] Duttaroy, A. K. Influence of maternal diet and environmental factors on fetal development. MDPI. 2023.

[8] Hamelmann P. Doppler Ultrasound Technology for Fetal Heart Rate Monitoring: A Review in IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, 2020, 67 (2), 226 - 238.

[9] Abouelyazid, M., Xiang, C. Machine learning-assisted approach for fetal health status prediction using Cardiotocogram Data. International Journal of Applied Health Care Analytics. 2021.

[10] Sufriyana, H., Husnayain, A., Chen, Y., Kuo, C., Singh, O., Yeh, T., Wu, Y., & Su, E. C. Comparison of multivariable logistic regression and other machine learning algorithms for prognostic prediction studies in Pregnancy Care: Systematic Review and Meta-Analysis. JMIR Medical Informatics, 2020, 8(11), e16503.

[11] Dixit, R. R. (2022, January 22). Predicting Fetal Health using Cardiotocograms: A Machine Learning Approach. https://research.tensorgate.org/index.php/JAAHM/article/view/38.

[12] Campos, D., Bernardes, J. Cardiotocography [Dataset]. UCI Machine Learning Repository. 2020.

[13] Rigatti, Steven J. RF. Journal of Insurance Medicine, 2017, 47 (1), 31 – 39.

[14] Subasi, A. Diagnosis of chronic kidney disease by using RF. IFMBE Proceedings, 2017, 589 – 594.

[15] Belgiu, M., Lucian, D. RF in remote sensing: a review of applications and future directions.” ISPRS Journal of Photogrammetry and Remote Sensing, 2016, 114 (114), 24 – 31.

[16] Akinsola, J. E. T. Supervised Machine Learning Algorithms: Classification and Comparison. ResearchGate, 8 June 2017, www.researchgate.net/publication/318338750_Supervised_Machine_Learning_Algorithms_Classification_and_Comparison.

[17] Yiming, L. A lasso regression model for the construction of microrna-target regulatory networks. Bioinformatics, 2011, 27 (17), 2406 – 2413.

[18] Ranstam, J., Cook J. A. LASSO regression. British Journal of Surgery, 2018, 105 (10), 1348 – 1348.

[19] Jin, W. The improvements of BP neural network learning algorithm. IEEE Xplore, 2000.

[20] Tianqi, C., Carlos, G. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16, 2016, 785 – 794.