BILINGUAL PUBLISHING GROUP

Journal of Computer Science Research

https://journals.bilpubgroup.com/index.php/jcsr

ARTICLE

Machine Learning Prediction of Fetal Health Status from Cardiotocography Examination in Developing Healthcare Contexts

***Olayemi, O. C.*[1], *Olasehinde, O. O.*[2*]**

[1] *Department of Computer Science, Teesside University, Middlesbrough, TS1 3BX, UK*

[2] *Department of Computer Science, University of Hundersfield, Hundersfield, HD1 3DH, UK*

ABSTRACT

Reducing neonatal mortality is a critical global health objective, especially in resource-constrained developing countries. This study employs machine learning (ML) techniques to predict fetal health status based on cardiotocography (CTG) examination findings, utilizing a dataset from the Kaggle repository due to the limited comprehensive healthcare data available in developing nations. Features such as baseline fetal heart rate, uterine contractions, and waveform characteristics were extracted using the RFE wrapper feature engineering technique and scaled with a standard scaler. Six ML models—Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), Categorical Boosting (CB), and Extended Gradient Boosting (XGB)—are trained via cross-validation and evaluated using performance metrics. The developed models were trained via cross-validation and evaluated using ML performance metrics. Eight out of the 21 features selected by GB returned their maximum Matthews Correlation Coefficient (MCC) score of 0.6255, while CB, with 20 of the 21 features, returned the maximum and highest MCC score of 0.6321. The study demonstrated the ability of ML models to predict fetal health conditions from CTG exam results, facilitating early identification of high-risk pregnancies and enabling prompt treatment to prevent severe neonatal outcomes.

Keywords: Neonatal; Mortality rate; Cardiotocography; Machine learning; Foetus health; Prediction; Features engineering

1. Introduction

Neonatal mortality, the death of a foetus between the 22nd week of gestation and the first week of birth is a major public health challenge in developing countries. Despite significant progress in reduc-

*CORRESPONDING AUTHOR:
Olasehinde, O. O., Department of Computer Science, University of Hundersfield, Hundersfield, HD1 3DH, UK; Email: o.olasehinde@hud.ac.uk

ARTICLE INFO
Received: 30 January 2024 | Revised: 3 March 2024 | Accepted: 4 March 2024 | Published Online: 23 March 2024
DOI: https://doi.org/10.30564/jcsr.v6i1.6242

CITATION
Olayemi, O.C., Olasehinde, O.O., 2024. Machine Learning Prediction of Fetal Health Status from Cardiotocography Examination in Developing Healthcare Contexts. Journal of Computer Science Research. 6(1): 43–53. DOI: https://doi.org/10.30564/jcsr.v6i1.6242

COPYRIGHT
Copyright © 2024 by the author(s). Published by Bilingual Publishing Group. This is an open access article under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License (https://creativecommons.org/licenses/by-nc/4.0/).

ing prenatal mortality rates in developed countries, the rates in developing countries remain high, with over 80% of all prenatal deaths occurring in these countries [1]. The high rate of maternal mortality remains a tenacious burden of developing countries. 213 million pregnancies were reported globally in 2012 [2], 89% of this statistic occurred in developing nations and 11% in developed nations. Lack of technology is a major obstacle to the provision of decent and adequate healthcare in developing countries. The Millennium Development Goals (MDGs) aim of reducing child mortality globally by 67% in 2015 was not achieved [3]. The World Health Organization (WHO) estimates that over two-thirds of prenatal deaths occur in low- and middle-income countries [4]. One factor that contributes to neonatal mortality is inadequate foetus monitoring during pregnancy [5]. Cardiotocography (CTG) is a non-invasive test used to assess foetus well-being during pregnancy. CTG exams are routinely performed during pregnancy, especially in high-risk cases, to monitor a foetus well-being. However, the interpretation of CTG exam results can be challenging, leading to subjective assessments and potential misdiagnosis. Therefore, developing a machine learning (ML) model to predict Foetus Health Status (FHS) from CTG exam results can help improve the accuracy and consistency of diagnoses and ultimately improve patient.

The CTG utilizes ultrasound technology to monitor the baby's heart rate. High-frequency sound waves, which are inaudible to the human ear, are emitted and detected by specialized machines [6].

Ultrasound travels freely through fluid and soft tissues, but when it encounters a solid surface, it bounces back as echoes. For instance, if it hits a solid valve or a gallstone, it will echo back strongly. The strength of the echoes varies depending on the density of the structure being hit. CTG monitoring employs a specific kind of ultrasound known as Doppler to measure moving structures, making it suitable for heart rate monitoring [7]. Meanwhile, the other plate on the CTG gauges the tension in the mother's abdomen to indicate when the uterus is contracting. The results from the CTG are either printed out or viewed electronically by the obstetrician who then decides the health status of the foetus and pregnancy in general. The data returned by the CTG exam have been standardized in their interpretation with numerous bodies adopting the same nomenclature for interpretation. Examples of fields returned by the CTG exam include defined risk, contractions (uterine activity), baseline foetus heart rate (FHR), baseline FHR variability, presence of accelerations, periodic or episodic decelerations, and changes or trends of FHR patterns over time.

According to the American College of Obstetricians and Gynaecologists (2020), the benefits of CTG include but are not limited to [5]:

1) It allows Early Detection of Foetus Distress: CTG is an effective tool for detecting Foetus distress, which can occur due to a variety of reasons, including problems with the placenta, lack of oxygen to the baby, or infections. Early detection of Foetus distress can help healthcare providers take appropriate measures to prevent complications.

2) It helps in monitoring high-risk pregnancies, such as those involving preterm labour, multiple pregnancies, or women with pre-existing medical conditions.

3) It helps healthcare providers keep a close eye on the health of the foetus and make timely decisions to prevent complications.

4) It helps in assessing the well-being of the foetus.

5) It is a non-invasive and safe test that poses no threat to the mother and the foetus.

6) It ensures peace of mind for the mother, knowing that their baby's health is monitored closely.

ML is defined as a field of artificial intelligence that focuses on the implementation of algorithms that learn with the aid of data [8]. Supervised and unsupervised learning are the two major types of ML: Supervised MLis trained on labelled data, where the desired output is known, and the algorithm learns to predict the output based on the input data, classification and regression are examples of supervised learning. Unsupervised ML, on the other hand, is trained on unlabelled data, where the desired output is not

known, and the algorithm must find patterns and relationships within the data on its own, clustering and association are examples of unsupervised learning [9]. The importance of data in the learning process of ML algorithms cannot be overstated. The ability of ML algorithm to recognize patterns and make accurate predictions is directly related to the amount and quality of data that is available for training [10]. Data are the foundation on which ML models are built, and the more, and diverse the available data to train the ML algorithms, the more accurate and robust the model becomes.

ML algorithms are computer programs that learn from data to make predictions or decisions based on that knowledge inferred during the learning process, the performance of ML models depends on several factors, including the quality and quantity of data used to train the ML algorithm, the choice of the ML algorithms implemented, the selection of features of the dataset used in training the, and the turning of the hyper-parameter of the algorithms [9,10].

Data pre-processing and modelling are two crucial steps in the creation of an ML model. These steps improve the quality of the data and prepare it for the ML process. The resulting model performs better and is more accurate, while the cost, time, and complexity of the model-building process are decreased. It involves cleaning, removing duplicate records, and filling in missing values, correcting invalid values, transforming, and organizing the data for model training. Olayemi et al. (2022) employed data pre-processing and ensemble method to improve prediction accuracy of diagnoses of knee osteoarthritis risk in adults. Data modelling entails choosing and transforming a pre-processed dataset's features into a format appropriate for ML in order to increase the effectiveness and efficiency of ML models, it aims to increase the dataset's quality used for ML [11]. Data transformation involves converting the data into a suitable format for ML and analysis. This includes techniques such as scaling, normalization, and feature engineering. The sensitivity of ML algorithm to the scale and distribution of the input data make data transformation necessary for ML and data analysis [12].

Feature selection identifies the most relevant features of a dataset that are most useful for predicting the target variable [13], it helps to reduce the complexity of the model and improves its performance. Feature selection was applied to improve ML model performance improvement for the diagnoses of breast cancer [14]. In data modelling, the reduced features of the pre-processed and transformed dataset are used to train MLalgorithms to build an MLmodel, this process involves the validation and the turning of the dataset to obtain an optimal result, and the built model is then used to evaluate a new and unknown test dataset. ML models have been widely and increasingly applied to analyse datasets to identify patterns that may be difficult for humans and build ML models for the diagnoses and prediction of diseases and infections, such as cancer, heart diseases, Parkinson's disease and the prediction of FHS.

Most of the previous research carried out on the FHS either focuses on systematics review, or the prediction of the FHS using a clinical dataset or a small sample of a dataset, furthermost of the research on the prediction of FHS from the CTG examination result does not carry out the feature selection of the relevant risk factors (features) correlated to the determination of the FHS before it was used to build the prediction models. As a result of its bias towards the majority class in the imbalance dataset, the accuracy metric, which is frequently used to assess the performance of FHS predictive models, was also found to produce misleading results. In order to reduce the neonatal mortality rate in developing countries, this study investigated and selected importantt features extracted from the CTG examination result to train and build models for six ML algorithms: Logistic Regression (LR) and Decision Tree (DT), and four ensemble learning models; Random Forest (RF), Gradient Boosting (GB), Categorical Boosting (CB), and Extended Gradient Boosting (XGB). Each of the models was assessed using the accuracy; precision, Matthew's correlation coefficient (MCC) and the F1 score metrics. This study has the following contribution to knowledge:

1) Class distribution of the dataset can affect

the performances of an ML model's training and evaluation.

2) Precision and Accuracy metrics only can give a wrong evaluation of ML Models, especially for imbalanced datasets.

3) Feature selection can improve the performance of an ML model.

4) Matthew's correlation coefficient (MCC) is a better metric than the F1-score to evaluate ML models, while the F1-score does not consider TN in its computation, MCC considered all four confusion matrix indices.

2. Literature review

According to Fawole et al., there is a significant correlation associated withthe level of education, lack of antenatal, high-risk, parity, mode of delivery and maternal mortality [15]. Akbulut et al. compared the performance of nine ML models (Averaged Perceptron, Boosted Decision Tree, Bayes Point Machine, Decision Forest, Decision Jungle, Locally-Deep Support Vector Machine, Logistic Regression, Neural Network, Support Vector Machine) for the prediction of FHS based on accuracy, F1-score, AUC measures metrics from a clinical dataset compiled from 96 pregnant women's responses to a maternal questionnaire. The decision tree recorded the highest validation accuracy of 87.5% and prediction accuracy of 87.5% from the testing of a new dataset of 16 pregnant women records. The small dataset employed in the study limits the generalizability of its results [16]. Sundar et al. applied a supervised artificial neural network to classify foetus heart rate and uterine contractions into one of three categories: normal, suspicious and pathological [17].

The results of the work by Alfirevic et al. [7] showed that the proposed method achieved a classification accuracy of 92.2%, F1-score metrics better especially when the class is balanced or in balance and the cost of misclassification is very high. The work of Batra et al. applied five algorithms, including decision trees (DT), support vector machines (SVM), random forests, neural networks, and gradient boosting, to evaluate foetus distress. The study claimed to have recorded an accuracy of 99.25%, which is higher than what was obtained in previous research. However, the source and size of the dataset were not mentioned, and the F1-score would have been a better metric to measure the performance of its model [18]. The study of Agrawal and Mohan also reported the ability of ML to predict the foetus health rate and do so accurately [19]. The systematic review of 16 selected studies by the American College of Obstetricians and Gynaecologists in 2020 revealed the potential of ML models to predict prenatal mortality with receiver operating curve (AUC) of over 0.80 and limitation of small sample sizes and lack of external validation [5].

A machine learning (ML) model was developed in a study by Park et al. to distinguish between normal and abnormal fetal cardiotocography (CTG) data. 17,592 fetal CTG records were obtained from three teaching hospitals, which were divided into training and validation sets. The model achieved an average area under the receiver operating characteristic curve (AUROC) of 0.73 and an area under the precision-recall curve (AUPRC) of 0.40 in the external validation dataset [20]. It's noteworthy that the study did not carry out a feature selection process to identify relevant features for fetal health status model building and prediction. However, the findings underscored the efficacy of ML techniques in predicting health conditions, as supported by existing literature.

3. Methodology

The dataset utilized for this project was obtained from the Kaggle dataset itcontained two files: the training dataset with 1488 instances (records) and 22 attributes (fields), and one of the attributes is the target class, it was used for the training and validation of our models. We observed a target class bias in the training set as follows: normal value (1158), suspect value (202), and pathological value (123), we minimise this level of bias by combining both suspects and pathological values together as abnormal. The second file is the testing dataset, it was used to test our validated models, it consists of 21 attributes and 638 instances, the target class has a normal value

(317), suspect value (199), and pathological value (121), and for the purpose of evaluation of our models, we combined the suspects and pathological values together as abnormal (320). Both the training and testing datasets do not have any missing values. The name of the features of both datasets and their descriptions is depicted in **Table 1**.

Table 1. Attributes and description dataset utilized.

Index	Feature
1	Baseline value
2	Accelerations
3	Foetus movement
4	Uterine contractions
5	Light decelerations
6	Severe decelerations
7	Prolongued decelerations
8	Abnormal short-term variability
9	Mean value of short-term variability
10	Percentage of time with abnormal long-term variability
11	Mean value of long-term variability
12	Histogram width
13	Histogram min
14	Histogram max
15	Histogram number of peaks
16	Histogram number of zeros
17	Histogram mode
18	Histogram mean
19	Histogram median
20	Histogram variance
21	Histogram tendency
22	Foetus health Status

We first applied scaling to the dataset to reduce the values of each column in the dataset to reduce the computation power required; a standard scaler was used in scaling the dataset. Then we plotted the heat map of the correlation between the columns of the dataset in **Figure 1**, to get an insight into the important columns that are important in the dataset target class.

We then defined a Python reusable function for the training and validation of different ML algorithms. The ML models used in our experiment are Logistic Regression (LR) and Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), Categorical Boosting (CB), and Extended Gradient Boosting (XGB). The function takes 6 arguments, two of which were the recursive feature elimination (REF) and n (number of features from 2 to 21). The function uses a FOR loop to iterate over different models, fit the model and validate it. The validation for each model is stored in a dictionary and returned as a data frame (tabular data).

REF repeatedly trains the model on thenumber of selected subsets of the features, and assesses their performance on a validation set, until the optimal subset of features with the best performance is identified. We applied ten-fold cross validation using the whole features of the training dataset for the training and validation of our models, and, then passed as arguments to the function aforementioned, to obtain the baseline metric table. We then performed RFE, varying the features selected between 2 to 21 featuresto see the performances of the selected features, and compare with the baseline models using the function aforementioned. We obtained the optimal number of features that returned optimal validation scores for each metric that were average, and plotted out the bar chart that denotes the optimal number of features with optimal scores for each model. The models with the optimal validation scores were then used to evaluate the testing dataset.

4. Evaluation metrics

Evaluation metrics are measurable metrics used to assess the efficacy and performance of ML models in terms of their predictions or classifications. The confusion matrix provides the foundation for the computation of these evaluation metrics, and it presents the counts of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) to evaluation model performance. This work considered four evaluation metrics of accuracy, precision, F1-score and the Matthews Correlation Coefficient (MCC) for the evaluation of the foetus health status prediction model.

Accuracy metrics is a straightforward, easy-to-understand, and widely used measure; it calculates the

proportion of correctly predicted instances over the total number of instances. It is often used as a baseline metric for initial model evaluation in addition to other metrics. It provides a general sense of the model's predictive capability and is useful when the target classes are well-balanced. It is not suitable when dealing with imbalanced or biased models. Precision metrics measure the accuracy of positive predictions made by a model. Precision is importantwhen the quality of positive predictions is emphasized, such as when false positives are not desirable. It contributes to determining the model's capacity to prevent false positives and create accurate positive predictions. A high precision score suggests a low rate of false positives, indicating that the model is accurate for predicting favourable situations.

The F1 scores and Matthews Correlation Coefficient (MCC) are binary classification evaluation metrics. The F1-score combines accuracy and recall into a single value, providing a balanced measure of the model's performance by considering both the ability to make accurate positive predictions (precision) and the ability to capture all positive instances (recall), when both precision and recall are important and no extreme class imbalance exists. When dealing with imbalanced datasets, other metrics such as the Matthews Correlation Coefficient (MCC) will provide more suitable insights into the model's performance. The F1 value ranges between 0 and 1, with 1 indicating perfect precision and recall, and 0 indicating the worst performance. The MCC, on the other hand, is a balanced measure that takes into consideration all four values in the confusion matrix, making it appropriate for imbalanced datasets, or when the cost of false positives and false negatives differs significantly. It provides an in-depth evaluation of a model's ability to accurately predict both positive and negative instances in the presence of class imbalance. MCC evaluation values range between –1 and +1, where +1 indicates a perfect prediction, 0 denotes a random prediction, and –1 represents a total disagreement between predictions and actual labelled values.

5. Results and discussion

From the heat map plot in **Figure 1**, we see that accelerations, uterine concentration, the mean value of short-time variability, the mean value of long-time variability, histogram mode and histogram mean had a negative correlation to the foetus health rate, while prolonged accelerations, abnormal short-term variability, and percentage of time with abnormal long-time variability had a positive correlation with the foetus health rate status.

The baseline model's validation result is shown in **Table 2**, the model's validation performs appreciably well on the dataset via ten-fold cross validation. From **Table 2**, the XGB model recorded the highest F1 (0.8683) and MCC scores (0.5178), though it recorded the same precision score as GB Model (0.9406), it performs better than GB based on the F1 score (0.8683) and MCC score (0.5178). Precision score alone cannot give a good measurement of ML models performance; it does not take into account false negative predictions, in the case of our research, foetus instances which are normal but being predicted as abnormal.

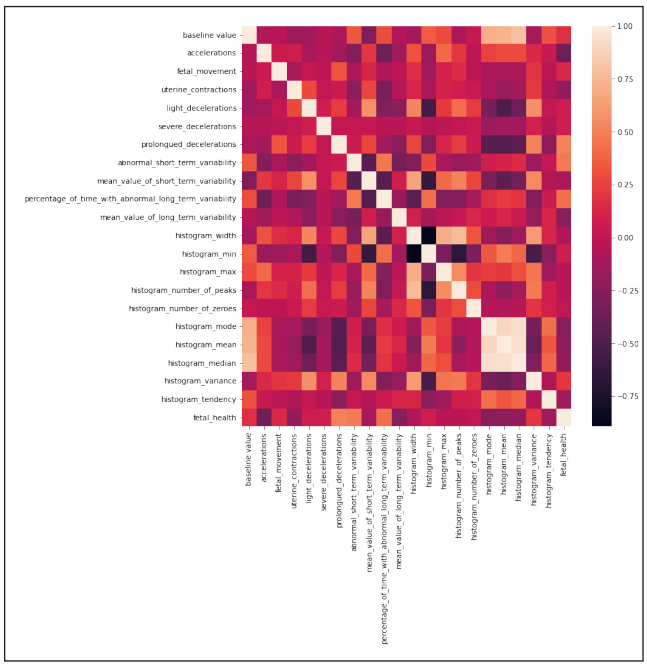

Figure 1. Heat map of the feature correlation.

Varying the number of features to be selected for each of the six models and validating the perfor-

mance, we saw that for a lesser number of selected features, the models (though performing well) were not performing as high as they were when the baseline model was validated. On getting to 15 features, we begin to see a similar or slightly better performance compared to the baseline models. **Tables 3–5** show the performance of the best two, seven and fifteen features selected by RFE respectively. From **Table 4**, it could be observed that with the seven optimal features, the Categorical Boosting (CB) Model's performance is higher than the rest of the models, though it recorded a lower precision than GB and XGB models; it has the highest F1-score and accuracy values. CB model maintainsits superiority with the fifteen features models in **Table 5**, it recorded the highest value across all the evaluation metrics, and RF came second in performance ahead of XGB across the four metrics of evaluation.

In another experimental analysis performed, we wrapped each of the ML algorithms with the RFE to determine the features that returned the optimal testing dataset validation MCC's score. **Table 6** reveals the number of features, and the maximum MCC's score recorded by it, only 8 out of 21 features selected by GB returned its maximum MCC score of 0.6255, while CB with 20 of 21 features returned an MCC score of 0.6321, **Table 6** represents in the bar chart in **Figure 2**. One advantage of wrapper features selection techniques is its ability to select the subset of features that gives the optimal validation score, unlike filtered methods that only rank features based on their importance to the target class.

Table 2. Baseline model metrics.

Models	MCC	F1-score	Accuracy	Precision
LR	0.1956	0.8030	0.7116	0.8562
DT	0.4030	0.8348	0.7586	0.8881
GB	0.5092	0.8665	0.8009	0.9406
RF	0.4714	0.8565	0.7868	0.9269
XGB	0.5178	0.8683	0.8041	0.9406
CATB	0.4762	0.8571	0.7884	0.9247

Table 3. Models optimal performance of two selected features by RFE.

Model	Features selected	MCC	F1-score	Accuracy	Precision
LR	2,18	0.1387	0.6847	0.5972	0.6370
DT	9,18	0.3035	0.7414	0.6708	0.6872
GB	8,10	0.0592	0.6320	0.5455	0.5685
RF	8,18	0.1866	0.7087	0.6238	0.6667
XGB	7,9	0.1774	0.7047	0.6191	0.6621
CB	8,18	0.1751	0.7055	0.6191	0.6644

Table 4. Models optimal performance ofthe seven selected features by RFE.

Model	Features selected	MCC	F1-score	Accuracy	Precision
LR	1,2,7,8,17,18,20	0.1425	0.7698	0.6600	0.8208
DT	1,2,8,9,10,11,18	0.2989	0.8039	0.7163	0.8493
GB	1,2,7,8,9,10,18	0.3430	0.8185	0.7351	0.8699
RF	8,9,10,11,13,17,18	0.3578	0.8196	0.7457	0.8703
XGB	2,4,8,9,10,17,18	0.3490	0.8190	0.7367	0.8676
CB	1,2,7,8,9,10,18	0.4486	0.8345	0.7680	0.8516

Table 5. Models optimal performance ofthe fifteen selected features by RFE.

Model	Features selected	MCC	F1-score	Accuracy	Precision
LR	1,2,3,4,7,8,9,10,13,14,15,17,18,19,20	0.2098	0.8182	0.7304	0.8836
DT	1,2,4,7,8,9,10,11,12,13,14,15,17,18,19	0.3474	0.8265	0.7414	0.8973
GB	1,2,4,7,8,9,10,11,13,14,15,16,17,18,19	0.4001	0.8384	0.7602	0.9064
RF	1,2,4,7,8,9,10,11,12,13, 14,17,18,19,20	0.5945	0.8496	0.7774	0.9155
XGB	1,2,3,4,7,8,9,10, 12,13,14,17,18,19,21	0.5415	0.8414	0.7649	0.9087
CB	1,2,4,7,8,9,10,11,12,13,14,17,18,19,20	0.6120	0.8611	0.7931	0.9338

Table 6. Testing dataset validation of models with the optimal number of selected feature subsets by RFE.

Model	Features selected	No of features	Maximum MCC score recorded
LR	1,2,7,8,10,17,18,19,20	9	0.4914
DT	1,2,4,8,9,10,11,12,13,17,18	11	0.5069
GB	1,2,7,8,9,10,14,18	8	0.6255
RF	1,2,4,7,8,9,10,11,12,13, 14,17,18,19,20	15	0.5945
XGB	1,2,3,4,7,8,9,10, 12,13,14,17,18,19,21	15	0.5615
CB	1,2,4,5,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21	20	0.6321

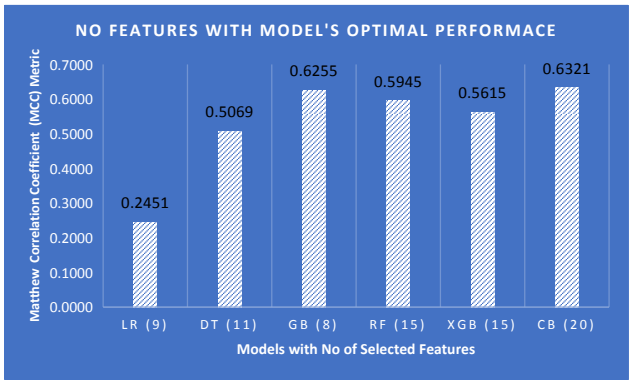

Figure 2. Best features elected by each model with the MCC score.

Figures 3–8 show the optimal features selected for each model by RFE with their ranking.

From the result of the work, we can see that the features selected vary greatly for different models, and with the RFE feature selection the model's performance was increased compared to when the models were trained using the whole data. It is important to note that the models used are not deterministic in nature, therefore a rerun of the exact experiments can lead to a slightly different result based on some stochastic algorithm but our observations were made based on numerous reruns and an average of the results on each rerun. From the above plots, we can see that the features selected vary greatly for different models, and with feature selection the model's performance was increased compared to when the models were trained using the whole data. We can also see that balancing of target class distribution can improve the performances of ML models on validation and testing datasets.

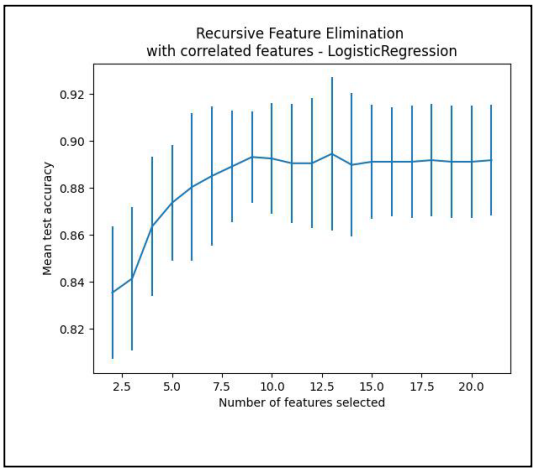

Figure 3. Mean of F1-score of the number of features subset selected on logistic regression.

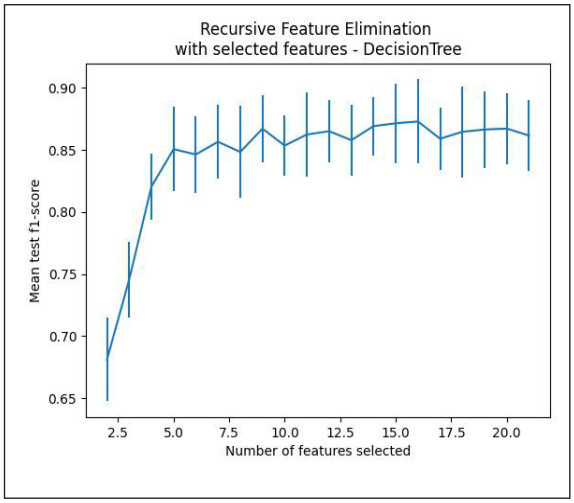

Figure 4. Mean of F1-score of the number of features subset selected on decision tree.

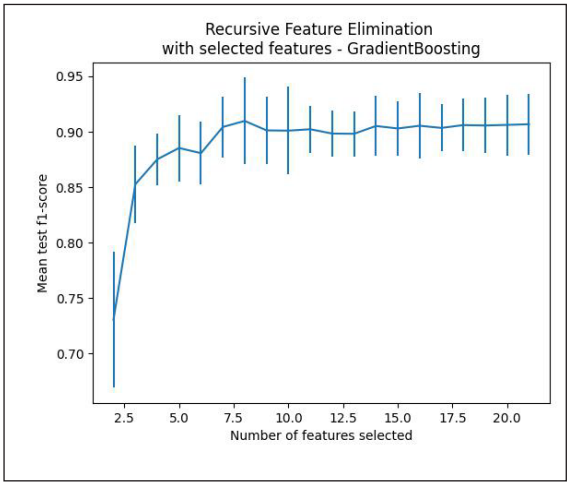

Figure 5. Mean of F1-score of the number of features subset selected on gradient boosting.

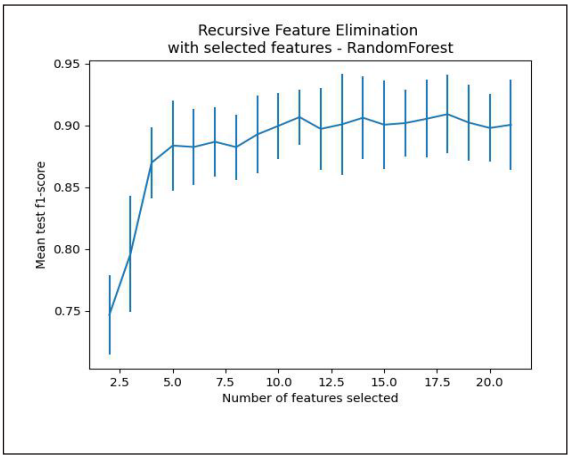

Figure 6. Mean of F1-score of the number of features subset selected on random forest.

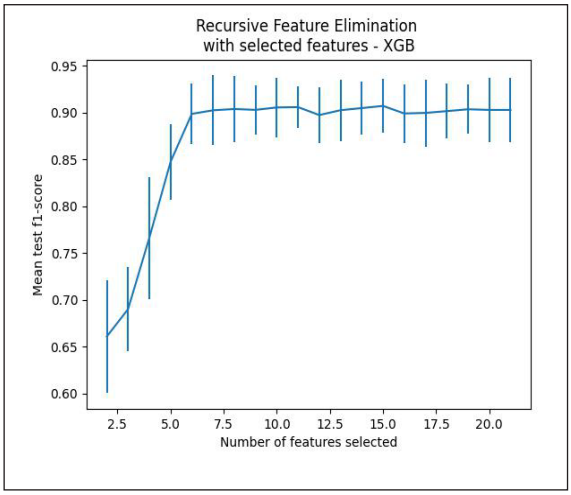

Figure 7. Mean of F1-score of the number of features subset selected on extended gradient boosting.

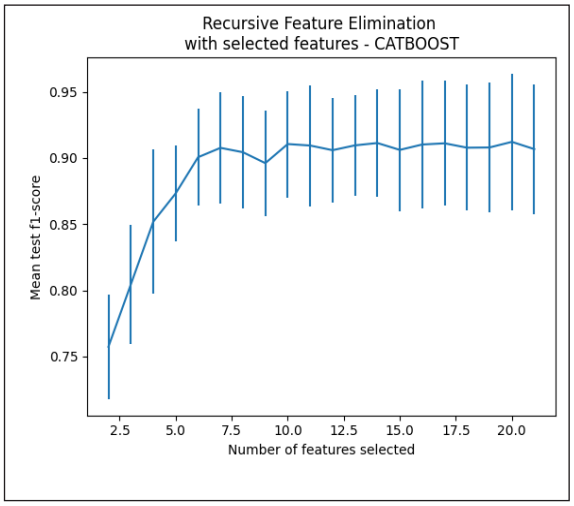

Figure 8. Mean of F1-score of the number of features subset selected on extended categorical boosting.

6. Conclusions and future works

This research underscores the potential of machine learning methodologies for addressing neonatal mortality rates by accurately predicting fetal health status from CTG examination results. By leveraging advanced algorithms and feature engineering techniques, we have shown that ML models can enhance the accuracy and consistency of fetal health assessments, particularly in settings with limited healthcare resources.

It is important to note that the models used are not deterministic in nature, therefore a rerun of the exact

experiments can lead to a slightly different result based on some stochastic algorithm but our observations were made based on numerous reruns and an average of the results on each rerun. This paper has shown that the prediction of foetus health rate from CTG is all important and contributes to the success of ML models depending on the methodology of the model. ML models can predict the FHR accurately from the features returned by readings from a CTG exam, and in training such ML models; feature selection is a useful action to perform in the process of increasing the performance of the models. We have also seen that some of the persistently important features across different models are Histogram mean, mean value of short-term variability, accelerations and baseline value.

Moving forward, efforts should focus on enhancing model interpretability, expanding datasets to include diverse demographics and medical features, and exploring stacked ensemble learners for further performance improvements. Ultimately, the application of ML in fetal health prediction holds promise for improving maternal and neonatal healthcare outcomes, thereby contributing to the global effort to reduce neonatal mortality rates in developing countries.

Author Contributions

The collaborative efforts of the author made significant contributions across various aspects of predicting fetal health status (FHS) from cardiotocography (CTG) examination results in developing healthcare contexts. The first author focused on framing the problem, conducting literature reviews, designing the methodology, analyzing data, interpreting results, and discussing implications. The second author contributed to explaining CTG examination principles, reviewing relevant literature, implementing the methodology, analyzing data, interpreting results, and discussing conclusions. Together, the authors synthesized research findings, proposed future directions, and aimed to enhance maternal and neonatal healthcare outcomes in developing countries.

Conflict of Interest

The authors affirm that they have no financial or personal associations with individuals or entities that might unduly influence their work or affect their objectivity. This study utilized a publicly accessible, anonymized dataset, ensuring the confidentiality of sensitive information. All aspects of the research process, including decisions and interpretations, were conducted impartially and autonomously, free from external biases.

References

[1] Lawn, J.E., Cousens, S., Zupan, J., 2005. 4 million neonatal deaths: When? Where? Why? Lancet. 365(9462), 891–900.
DOI: https://doi.org/10.1016/S0140-6736(05)71048-5

[2] Sedgh, G., Singh, S., Hussain, R., 2014. Intended and unintended pregnancies worldwide in 2012 and recent trends. Studies in Family Planning. 45(3), 301–314.
DOI: https://doi.org/10.1111/j.1728-4465.2014.00393.x

[3] The Millennium Development Goals Report [Internet]. United Nations; 2015. Available from: https://www.un.org/millenniumgoals/2015_MDG_Report/pdf/MDG%202015%20rev%20(July%201).pdf

[4] World Health Organization [Internet]. Perinatal Mortality [Accessed 2023 Mar 31]. Available from: https://www.who.int/publications-detail-redirect/9789240068759

[5] Boraas, C.M., Sanders, J.N., Schwarz, E.B., et al., 2021. Risk of pregnancy with levonorgestrel-releasing intrauterine system placement 6–14 days after unprotected sexual intercourse. Obstetrics and Gynecology. 137(4), 623–625.
DOI: https://doi.org/10.1097/AOG.0000000000004118

[6] Corso, J.F., 2005. Bone-conduction thresholds for sonic and ultrasonic frequencies. The Journal of the Acoustical Society of America. 35(11), 1738–1743.
DOI: https://doi.org/10.1121/1.1918804

[7] Alfirevic, Z., Gyte, G., Cuthbert, A., et al., 2017. Continuous CTGas a form of electronic foetus monitoring (EFM) for foetus assessment during labour. Cochrane Database of Systematic Reviews.
DOI: https://doi.org/10.1002/14651858.CD006066.pub3

[8] Géron, A., 2022. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media, Inc.: Sebastopol.

[9] Jordan, M.I., Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. Science. 349(6245), 255–260.
DOI: https://doi.org/10.1126/science.aaa8415

[10] Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep learning. MIT Press: Cambridge.

[11] Olayemi, O.C., Olasehinde, O.O., Alowolodu, O.D., et al. 2022. Ensemble learning improvement of clinical diagnoses of knee osteoarthritis risk in adults. Int. J. Intell. Inf. Syst. 11(4), 51-64. Available from: https://www.researchgate.net/profile/Olayemi-Olasehinde/publication/362324563_Ensemble_Learning_Improvement_of_Clinical_Diagnoses_of_Knee_Osteoarthritis_Risk_in_Adults/links/62e358c64246456b55f11d2f/Ensemble-Learning-Improvement-of-Clinical-Diagnoses-of-Knee-Osteoarthritis-Risk-in-Adults.pdf

[12] Raschka, S., Mirjalili, V., 2017. Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2. Packt Publishing Ltd.: Birmingham.

[13] Han, J., Kamber, M., 2011. Data mining: Concepts and techniques (3rd ed.) Morgan Kaumann Publisher: Burlington.

[14] Jia, N., Zhao, W., Liu, X., 2019. Dempster-Shafer theory-based hierarchical saliency detection. International Journal of Imaging Systems and Technology. 29(3), 329–338.
DOI: https://doi.org/10.1002/ima.22322

[15] Fawole, A.O., Shah, A., Fabanwo, A.O., et al., 2012. Predictors of maternal mortality in institutional deliveries in Nigeria. African Health Sciences. 12(1).

[16] Akbulut, A., Ertugrul, E., Topcu, V., 2018. Foetus health status prediction based on maternal clinical history using ML techniques. Computer Methods and Programs in Biomedicine. 163, 87–100.
DOI: https://doi.org/10.1016/j.cmpb.2018.06.010

[17] Sundar, C., Chitradevi, M., Geetharamani, G., 2012. Classification of cardiotocogram data using neural network based machine learning technique. International Journal of Computer Applications. 47(14), 19–25.
DOI: https://doi.org/10.5120/7256-0279

[18] Batra, A., Chandra, A., Matoria, V. (editors), 2017. Cardiotocography analysis using conjunction of machine learning algorithms. 2017 International Conference on Machine Vision and Information Technology (CMVIT); 2017 Feb 17–19; Singapore.New York: IEEE.p. 1–6.
DOI: https://doi.org/10.1109/CMVIT.2017.27

[19] Agrawal, K., Mohan, H. (editors), 2019. Cardiotocography analysis for fetal state classification using machine learning algorithms. 2019 International Conference on Computer Communication and Informatics (ICCCI); 2019 Jan 23–25; Coimbatore, India.New York: IEEE.p. 1–6.
DOI: https://doi.org/10.1109/ICCCI.2019.8822218

[20] Park, T.J., Chang, H.J., Choi, B.J., et al.,2022. Machine learning model for classifying the results of fetal cardiotocography conducted in high-risk pregnancies. Yonsei Medical Journal. 63(7), 692–700.
DOI: https://doi.org/10.3349/ymj.2022.63.7.692