NANYANG TECHNOLOGICAL UNIVERSITY

AN6003 Project Report

Classifying Fetal Heart Health Condition from Cardiotocography Data Using Machine Learning

By Team 5 (Class B):

- Gan Jiujun (G2202718L)
- Huang Zhengyi (G2202329J)
- Lyu Jiaxin (G2202680H)
- Natashya Suryani Tiro (G2202005H)
- Wang Xinyu (G2202328A)
- Zhang Xinyu (G2201842L)

Nanyang Business School
2022

Table of Contents

List of Figures

List of Tables

Executive Summary

In a world puzzled by problems such as low fertility rates and declining population growth, it becomes more urgent to improve the survival rate of infants. Conventional diagnosis and examinations on fetal heart health conditions are unreliable, expensive, and time-consuming; therefore, our team comes up with a development and implementation plan of machine learning to resolve the problem. With the help of machine learning techniques, health professionals can make more accurate judgments based on cardiotocograph data and have an overall better grasp of fetal health status. Some previous researchers have attempted to tackle the issue using machine learning; however, their models perform poorly in detecting suspicious cases, and our team, through tireless efforts, successfully outperforms all of them after multiple trials to refine our models.

Our approach to the business question consists of obtaining the dataset, data pre-processing and modelling. The three machine learning models that we use are Logistic Regression for Multicategory Y, CART, and Random Forest. With Logistic Regression, we are proud to declare an overall accuracy of 98.79% and individual accuracy of 98.59% for the normal, 98.74% for the suspect, and 99.95% for the pathological. Pertaining to CART, the overall accuracy is 99.53%, and the individual accuracy for the normal, the suspect, and the pathological is 99.45%, 99.5%, and 100.0% respectively. Last but not least, our Random Forest model, compared with models by Alam and Mehbodniya, has an overall accuracy of 99.6% which is 2.09% and 5.1% higher than the model accuracies of preceding researchers correspondingly. Hence, our team has successfully constructed machine learning models that are even better than those built by research professionals across the world on the basis of the accuracy calculated from the same test set.

Out of the three models, Random Forest, with highest overall accuracy and impressively low false positive rate, is marginally better than the other models. After carrying out some research, we are well informed of the fact that there is a Women's Heart Clinic in NHCS, and many women visit it during pregnancy regularly for check-ups including measuring cardiotocography. Therefore, our machine learning models can play an important role to facilitate doctors at NHCS to make wiser decisions and reduce misdiagnosis.

To take advantage of this business opportunity, our team has also conceived an AI-enabled plan that is beyond a set of Machine Learning models and can be directly implemented here at NHCS. With the presence of our machine learning models, not only does a high accuracy builds up a reputation for being reliable, but it also attracts more patients from all over the world and even globally renowned doctors for a career opportunity. And it becomes a virtuous circle. Moreover, NHCS can save on costs that are incurred from misdiagnosis and potential legal fees; therefore, it is a good bargain for NHCS to adopt our plan from a financial standpoint. From the perspective of human beings as a whole, striving for a higher fertlity rate and improving the overall health of infants are truly aligned with NHCS's mission and core values.

1. Introduction

1.1 Business Problem

Currently, the business problem that hospitals around the world are facing, including the National Heart Centre Singapore (NHCS), is regarding traditional medical diagnosis and monitoring for fetal heart health that are sometimes still inaccurate, costly, and time-consuming. This research will focus on developing and recommend an implementation plan of Machine Learning (ML) in tackling this issue.

1.2 Background

Monitoring fetal health condition throughout pregnancy is one of the most difficult and complicated tasks in medical field. According to WHO, every day, there are about 810 women die as a result of pregnancy and delivery complications, with the majority of deaths occurring in low- and middle-income nations. One of the common complications behind high maternal mortality ratio (MMR) is improper monitoring of unborn baby condition and mother. In fact, congenital heart disease (CHD) caused 180.264 deaths among infants (Zimmerman *et al.*, 2020). In Singapore General Hospital (SGH), the incidence of CHD at birth was 9.7 per 1,000 live births, which is higher that reported in Australia, US, and Europe.

In order to assess fetal well-being and risk of pregnancy complications, cardiotocography (CTG) is a commonly used technical approach for continuously monitoring and recording the fetal heart rate (FHR) and uterine contractions. However, the interpretation of the information produced by CTG is currently not standardized. The incorrect interpretation of CTG can result in unnecessary surgical intervention, such as an increase in cesarean sections (Karabulut and Ibrikci, 2014). Other than that, CTG also allows early monitoring and intervention of embryonic hypoxia (low levels of oxygen) before death. However, more than 50% of death and long-term disablement due to hypoxia were caused by not recognizing the abnormal FHR pattern (Costa *et al.*, 2009).

Therefore, implementing timely and accurate machine learning in interpreting cardiotocography data can help health professionals making better medical diagnosis and decisions, and minimize human errors. Hence, it can reduce maternal and fetal mortality rates and complications during pregnancy and childbirth, and benefit populations in preserving national fertility rates in the current trend of declining fertility, especially in Singapore.

In financial aspect, misdiagnosis, and hospital cost in children with heart disease has huge contribution to the total hospital costs. 30% of annual healthcare spending in the US, which is around $750 billion is wasted on misdiagnosisClick or tap here to enter text.. In 2012, around 23% of total hospital cost (6,600 M$) is represented by the hospital cost of children with CHD (Faraoni, Nasr and DiNardo, 2016). Other than that, around 89.1% family with CHD child experienced at least one financial burden(McClung, Glidewell and Farr, 2018). Therefore, if the National Heart Centre Singapore (NHCS) implements machine learning as their diagnostic service for fetal heart health, it can assist in making accurate predictions at an early stage and can prevent the condition of mother and child from worsening. This can decrease the hospital cost tremendously, while also relieve families financial burden, which is aligned with the NHCS mission as a people-centered organization.

1.3 Key Business Questions

Key business questions that will be answered through this research are:

- Can machine learning improve medical diagnosis accuracy and serve as decision support system for health professionals?
- What are the best practices to implement machine learning in healthcare?
- What insights and competitive advantages can be obtained by adding machine learning into NHCS workflow?
- Will it be profitable for NHCS to implement machine learning in interpreting cardiotocography data?

1.4 Objective

The main objective of this research is to develop machine learning models to help health professionals in interpreting cardiotocography data and implement it in NHCS. Through this implementation, health professionals can make better medical diagnosis and decisions, and minimize human errors which can decrease the maternal and fetal mortality rates and positively impact profitability of NHCS.

2. Literature Review

Throughout the years, there are several previous research that studied about machine learning techniques to interpret CTG data. Table 1 shows the research that used the identical dataset that will also be used in this research. Artificial neural network-based classifier shown good overall performance in classifying normal and pathologic cases. However, it has poor performance in detecting suspicious cases Click or tap here to enter text.. The highest accuracy from previous study is 97.51% by using random forest Click or tap here to enter text.. Based on previous studies results, there is still room for improvement. Hence, this research aims at developing a model with higher performance while also providing an implementation plan.

Table 1 Literature Review

References	Problem	Methods	Findings
(Alam *et al.*, 2022)	Traditionally, obstetricians have evaluated CTG data artificially, which takes time and is inaccurate. As a result, developing a fetal heart health categorization model is critical, as it has the potential to save not only time but also medical resources in the diagnostic process.	Random Forest, Logistic Regression, Decision Tree, Support Vector Classifier, Voting Classifier, K-Nearest Neighbours	Random forest model produces the best result with 97.51% accuracy
(Mehbodniya *et al.*, 2022)	One of the most common complications behind high maternal mortality ratio is improper monitoring of unborn baby condition and mother.	Support Vector Classifier, Random Forest, Multi-layer Perceptron, K-Nearest Neighbours	Random forest model produces the best result with 94.5% accuracy and support of 635. The next best performer is SVM, with 93% accuracy and same F1-score. All the algorithms show same support value,

References	Problem	Methods	Findings
	This research deploys various ML algorithms to predict fetal heart health from the cardiotocographic (CTG) to improve process of monitoring baby condition.		which indicates the purity of classes.
(Karabulut and Ibrikci, 2014)	Many typical findings are included in CTG, and obstetricians make clinical decisions about the state of the fetus considering these findings. However, the interpretation of the information provided by CTG is not standardized. The deficient interpretation of CTG leaded to unnecessary surgical intervention.	Decision Tree Based Adaptive Boosting (AdaBoost) Method	The most prominent result belongs to decision tree based AdaBoost algorithm by 0.034 MAE, 0.861 kappa statistics and 95.01% accuracy, meaning that 2020 of 2126 samples are perfectly predicted.
(C, M.Chitradevi and Geetharamani, 2012)	The predictive capacity of the methods for interpreting a typical cardiotocography data remains controversial and still inaccurate.	Artificial Neural Network (ANN)	The performance of the algorithms in terms of Rand Index was good and always greater than 0.9. And it gives good precision, recall and f-score for normal (0.9663, 0.991, 0.9784) as well as pathological (0.9706, 0.9745, 0.9724) records, but giving poor performance in the case of suspicious (0.5897, 0.3688, 0.4514) records. Arrived results obviously show that supervised machine learning based methods can be used for the classification of CTG data.

3. Approach

There are three primary steps of approach used in developing the ML model as shown in Figure 1.

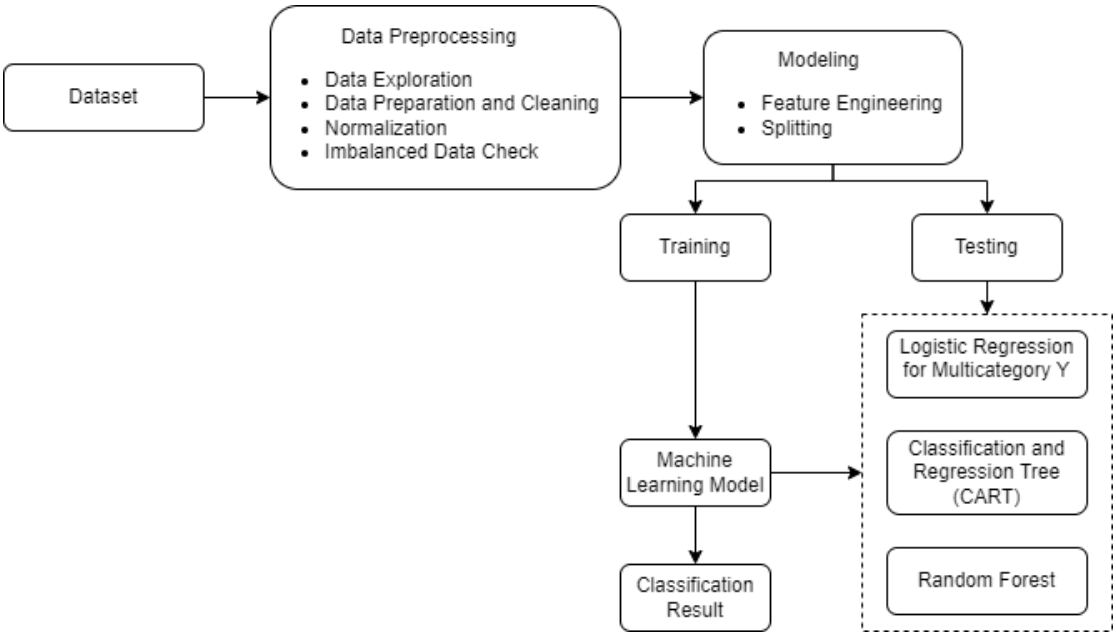

Figure 1 Research Approach Flowchart

3.1 Data Preprocessing

Data preprocessing is important to be done before modelling so issues in the dataset can be detected. Example of issues are as follows:

- Missing or incomplete records
- Outliers or Anomalies
- Improperly formatted / structured data
- Inconsistent values and non-standardized categorical variables
- Duplicates

If the dataset contains issues, data cleaning will be needed to improve data quality issues. Some data cleaning alternatives that can be used are estimating the values by using the mode of the column or mode in the relevant subgroup, using model such as logistic regression, linear regression, CART, or others depend on the data type, or ignore cases with missing values. After that, normalization of data will be used to scale the features. Other than that, if the data is unbalanced, it can be balanced by using over sampling technique.

3.2 Modeling

In the modeling step, it is important to do feature engineering where the features utilized in the analysis will be selected, aggregated, or transformed from the raw data. After that, the input data will be split randomly into a training dataset and a test dataset. The machine learning model than will be trained by using the train set. To gain the best result, this project will test several models in R programming language and compare its performance that is based on the same test set. The models used are as follows.

3.2.1 Logistic Regression for Multicategory Y

When the dependent variable is categorical, Logistic Regression is the proper regression analysis to conduct. Logistic Regression is used to describe the relationship between multiple

independent variables and the dependent variable, and to assign new data to different categories based on all the x variables. We measure the relationship by evaluating P values and coefficients in regression analysis. The mathematical relationship between individual independent variable and the dependent variable is given by the Logistic Regression coefficients, and the P value for every independent variable shows whether the relationship is statistically significant. Moreover, Logistic Regression is a type of predictive modeling. When it comes to predicting, a cutoff value is set. With the help of sigmoid function, the outcome of a logistic function is mapped to a probability between one and zero. When it is greater than a threshold, the data is assigned to a certain category, and when it is smaller than the threshold, it is assigned to a different category.

3.2.2 Classification and Regression Tree (CART)

CART is a technique that can process continuous and nominal attributes as targets and predictors by using a binary recursive partitioning procedure. The data are divided into two children starting at root node, and then each of the children is further divided into grandchildren. This can be seen in Figure 2. The trees are grown to maximal size and stops when no more splits are feasible. The cost-complexity pruning method is then used to prune the maximal-sized tree back to the root (split by split). The split that is contributing the least to the overall effectiveness of the tree on training data is the next split to be pruned. The goal of the CART mechanism is to build a series of nested trimmed trees, each of which is a potential candidate of the optimal tree. By comparing the predicted accuracy of each tree in the pruning sequence to independent test data, the "right sized" or "honest" tree is found (Steinberg, 2009). Advantages of CART are it is simple to understand, interpret and visualize, it implicitly performs variable screening and feature selection, it can handle both numerical and categorical data or multi-output problems, the performance is not affected by non-linear relationships between parameters do. The disadvantages are it is prone to overfitting, can be unstable, can create biased trees if some classes dominate, and cannot guarantee to return globally optimal decision tree.

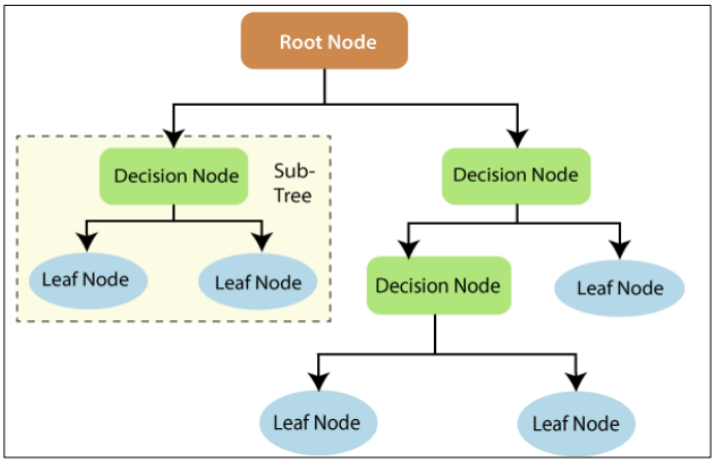

Figure 2 Flowchart of Decision Tree Classifier

3.2.3 Random Forest

Random Forest is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned.

In a Random Forest, every decision tree consists of decision nodes, leaf nodes, and a root node. The leaf node of each tree is the final output produced by that specific decision tree. The selection of the final output follows the majority-voting system. In this case, the output chosen by the majority of the decision trees becomes the final output of the forest system. Figure 3 shows a simple Random Forest classifier.

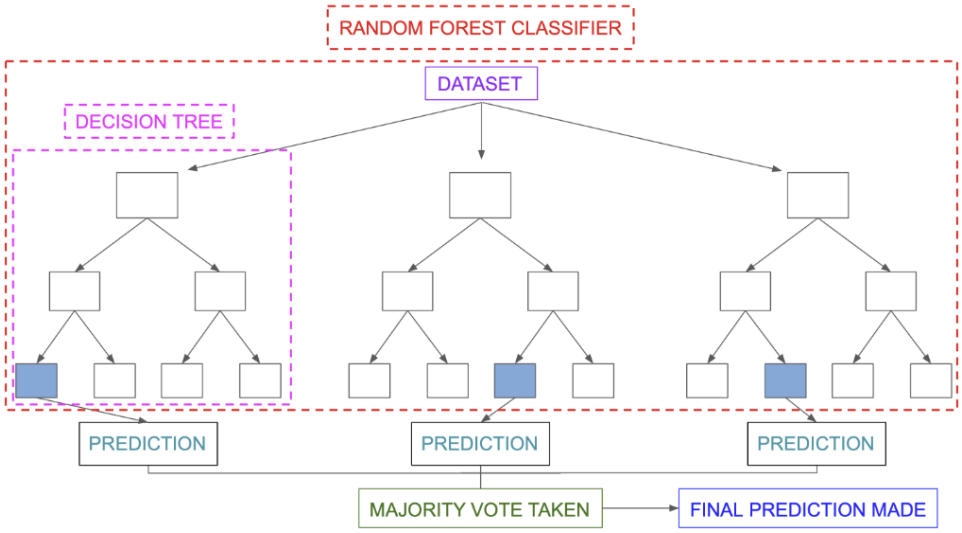

Figure 3 Random Forest Classification

Random Forest corrects for decision trees' habit of overfitting to their training set. Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance. This comes at the expense of a small increase in the bias and some loss of interpretability, but generally greatly boosts the performance in the final model.

3.3 Model Evaluation

The model evaluation includes the following steps:

a. Determine the “best” model

To evaluate each model, confusion matrix will be used to show the true negative rate, true positive rate, false negative rate, and false positive rate. To decide the best model for this problem, evaluation can be done by comparing the success metrics (accuracy, sensitivity, specificity, and others) between alternative models.

b. Evaluate whether model is suitable for production

The key questions in this step are:

- Given the test data, does the model answer the business questions with adequate confidence?
- Does trying other alternative approaches is needed?
- Does collecting more data, undertaking more feature engineering, or experimenting with other algorithms are needed?

c. Interpret the Model

To explain the model interpretation clearly, visualization dashboard can be used to interact with model and result explanations.

4. Results and Data Analysis

4.1 Dataset Information

The data that is used for this project is fetal cardiotocography data. The data was provided in September 2010 by the Biomedical Engineering Institute and the Faculty of Medicine at the University of Porto, Portugal. These datasets were obtained on a regular basis in 1980 and again between 1995 and 1998, resulting in an ever-growing collection. The dataset contains 2126 fetal cardiotocography (CTG) specimens, which include fetal heartbeat and uterine contractions during pregnancy and labor. Three professional obstetricians classified the CTGs, and a consensus risk of heart disease classification label was issued to each of them. There are 2 types of classification done in the dataset which are the morphologic pattern and the fetal state (Normal, Suspect, Pathologic). The fetal state would be the primary focus of classification in this project.

Figure 4 shows that the dataset is imbalanced, and is dominated by normal cases. Therefore, the SMOTE was employed to balance the data. Figure 5 shows the total number of normal, suspect and pathological data after balancing.

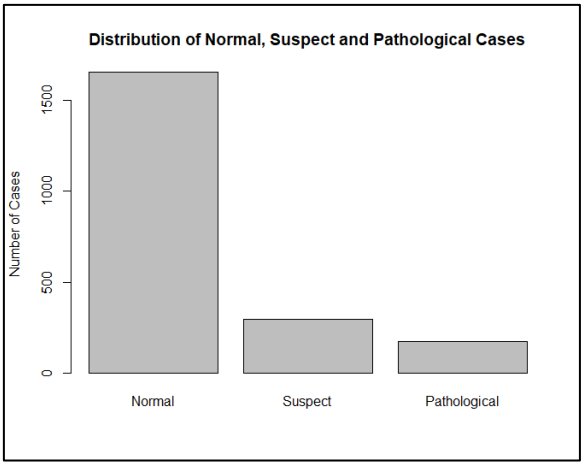

Figure 4 Distribution of Normal, Suspect and Pathological Cases

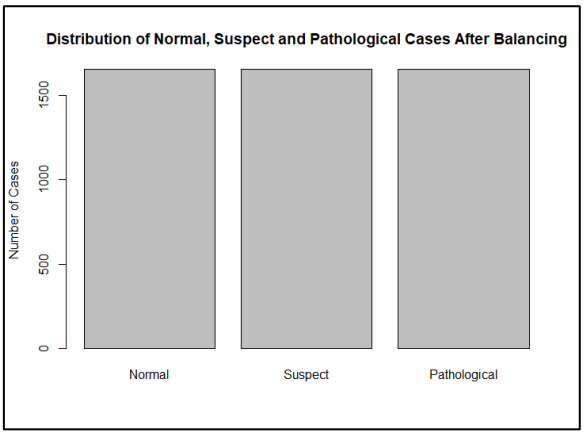

Figure 5 Distribution of Normal, Suspect and Pathological Cases After Balancing

4.2 Analysis of the Model

4.2.1 Logistic Regression for Multicategory Y

Figure 6 shows the confusion matrix and success metrics statistics of the Logistic Regression model that was run on test set. The overall accuracy is 98.79%. The individual accuracy of each class is 98.59% for normal, 98.74% for suspect, and 99.95% for pathological. There are 9 normal cases that is predicted as suspect, 1 normal case that is predicted as pathological and 8 suspect cases that is predicted as normal. However, this model predicts perfectly in pathological cases which is the most critical case since it needs to be quickly identified and treated. The total number of correct predictions is 1473, with 18 incorrect forecasts. Hence, the Logistic Regression model gives a really good performance with high accuracy in interpreting CTG data and classifying fetal heart health.

```
Confusion Matrix and Statistics

              Reference
Prediction     Normal Suspect Pathological
  Normal          487       8            0
  Suspect           9     489            0
  Pathological      1       0          497

Overall Statistics

               Accuracy : 0.9879
                 95% CI : (0.981, 0.9928)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9819

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Normal Class: Suspect Class: Pathological
Sensitivity                 0.9799         0.9839              1.0000
Specificity                 0.9920         0.9909              0.9990
Pos Pred Value              0.9838         0.9819              0.9980
Neg Pred Value              0.9900         0.9919              1.0000
Prevalence                  0.3333         0.3333              0.3333
Detection Rate              0.3266         0.3280              0.3333
Detection Prevalence        0.3320         0.3340              0.3340
Balanced Accuracy           0.9859         0.9874              0.9995
```

Figure 6 Confusion Matrix and Statistics of Logistic Regression Model

Additional insight that is gained from the Logistic Regression model is that there are 23 variables that is statistically significant for patients to be diagnosed as suspect and there are 11 variables that is statistically significant for patients to be diagnosed as pathological. These variables give different weights in diagnosing patients.

Table 2 Variables coefficient Based on the Logistic Regression Model

	Dependent variable:	
	Suspect	**Pathological**
LBE	0.959*	3.859
LB	0.959*	3.859
AC	-2.024	2.317
FM	1.502***	1.590

UC	0.831***	-0.271
ASTV	1.969***	-15.420
MSTV	-0.088	-0.150
ALTV	0.211	7.092
MLTV	-0.430	-8.979
DL1	0.111	-14.303
DL2	-1.491	22.130
DL3	-1.771	23.497
DL4	-4.167	2.870***
DL5	-64.054***	-38.545
DL6	-16.788	-33.221
DL7	-6.022*	-17.223***
DL8	-26.929***	-22.210***
DL9	-25.574***	-17.150
DL10	-5.353*	-0.857***
DL11	-25.612***	-14.374***
DL12	-27.838***	-34.050***
DL14	2.906***	-1.113
DL16	11.162	-21.598
DS1	32.670***	-13.757***
DP1	2.996**	2.809
DP2	31.945	11.461
DP3	88.399	71.594
DP4	13.937***	35.040
Width	0.604**	10.125
Min	-0.173	-15.905
Max	1.105	-4.778
Nmax	-0.056	-10.479
Nzeros	0.123	2.851
Mode	-5.591***	-33.366
Mean	20.772***	-33.780
Median	-16.315***	54.074
Variance	4.665***	-16.413
Tendencysymmetric	2.935	-0.449
Tendencyright assymetric	4.078	-4.851
A1	9.643	-32.675***
B1	-45.051***	-26.593
C1	-36.998	-10.368***
D1	-28.652***	-54.967***
E1	22.088	1.622
AD1	-1.220	-52.180
DE1	14.952	-61.150***
LD1	-5.157	87.056
FS1	19.110	90.992
SUSP1	34.677	27.101
Constant	-16.608	-31.163
Akaike Inf. Crit.	306.020	306.020
Note:		*p<0.1; **p<0.05; ***p<0.01

4.2.2 Classification and Regression Tree (CART)

Figure 7 shows the confusion matrix and success metrics statistics of the CART model that was run on test set. Here, the overall achieved accuracy is 99.53%. The individual accuracy of each class is 99.45% for normal, 99.5% for suspect, and 100% for pathological. There are 4 normal cases that is predicted as suspect, and 3 suspect cases that is predicted as normal. However, this model predicts perfectly in pathological cases which is the most critical case since it needs to be quickly identified and treated. The total number of correct predictions is 1484, with 7 incorrect forecasts. Hence, the CART model gives a really good performance with high accuracy in interpreting CTG data and classifying fetal heart health.

Additional insight that is gained from the CART model is that there are 20 out of 33 CTG variables that contribute to the classification of fetal heart health. The percentage of importance can be seen on Table 3 Variables Importance Based on the CART Model. The average importance of the 20 variables is 5% and there is no variable with extreme importance or overly dominated other variables. This proves that it is very difficult to interpret CTG data manually with 20 combinations of variables and this shows that machine learning is extremely helpful in diagnosing fetal heart health.

```
Confusion Matrix and Statistics

              Reference
Prediction     Normal Suspect Pathological
  Normal          493       3            0
  Suspect           4     494            0
  Pathological      0       0          497

Overall Statistics

               Accuracy : 0.9953
                 95% CI : (0.9904, 0.9981)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.993

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Normal Class: Suspect Class: Pathological
Sensitivity                 0.9920         0.9940              1.0000
Specificity                 0.9970         0.9960              1.0000
Pos Pred Value              0.9940         0.9920              1.0000
Neg Pred Value              0.9960         0.9970              1.0000
Prevalence                  0.3333         0.3333              0.3333
Detection Rate              0.3307         0.3313              0.3333
Detection Prevalence        0.3327         0.3340              0.3333
Balanced Accuracy           0.9945         0.9950              1.0000
```

Figure 7 Confusion Matrix and Statistics of CART Model

Table 3 Variables Importance Based on the CART Model

Variables	Percentage of Importance	Accumulative Importance
Suspect pattern	10%	10%
Flat-sinusoidal pattern (pathological state)	9%	19%
Largely decelerative pattern	8%	27%
Mean value of short-term variability (Sisporto)	8%	35%
Histogram mean	7%	42%
Percentage of time with abnormal short-term variability (Sisporto)	7%	49%

Variables	Percentage of Importance	Accumulative Importance
Histogram median	6%	55%
Percentage of time with abnormal long-term variability (Sisporto)	6%	61%
Prolonged decelerations	6%	67%
Percentage of time with abnormal long-term variability (Sisporto)	6%	73%
Histogram mode	5%	78%
Histogram width	5%	83%
Histogram variance	5%	88%
Mean value of long-term variability (Sisporto)	5%	93%
Low freq. Of the histogram	2%	95%
High freq. Of the histogram	1%	96%
Baseline value (medical expert)	1%	97%
Decelerative pattern (vagal stimulation)	1%	98%
Number of histogram peaks	1%	99%
Baseline value (Sisporto)	1%	100%

4.2.3 Random Forest

Figure 8 shows the confusion matrix and success metrics statistics of the Random Forest model that was run on the test set. Here, the overall achieved accuracy is 99.6%. The individual accuracy of each class is 99.55% for both normal and suspect, and 100% for pathological. There are 3 normal cases that are predicted as suspect, and 3 suspect cases that are predicted as normal. However, this model predicts perfectly in pathological cases which is the most critical case since it needs to be quickly identified and treated. The total number of correct predictions is 1485, with 6 incorrect forecasts. Hence, the Random Forest model gives a really good performance with high accuracy in interpreting CTG data and classifying fetal heart health.

Additional insight that is gained from the Random Forest model is that 26 out of 33 variables contribute to the classification of fetal heart health. The importance measured in two different ways can be seen on Figure 9. *Suspect pattern* shows significantly higher importance than other variables. *Flat-sinusoidal pattern*, *largely decelerative pattern* and *percentage of time with abnormal short-term* also have great contribution to the classification. Figure 8 indicates the importance of different variables respectively and the sequence most probably followed when a certain tree is generated.

```
Confusion Matrix and Statistics

              Reference
Prediction     Normal Suspect Pathological
  Normal          494       3            0
  Suspect           3     494            0
  Pathological      0       0          497

Overall Statistics

               Accuracy : 0.996
                 95% CI : (0.9913, 0.9985)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.994

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Normal Class: Suspect Class: Pathological
Sensitivity                 0.9940         0.9940              1.0000
Specificity                 0.9970         0.9970              1.0000
Pos Pred Value              0.9940         0.9940              1.0000
Neg Pred Value              0.9970         0.9970              1.0000
Prevalence                  0.3333         0.3333              0.3333
Detection Rate              0.3313         0.3313              0.3333
Detection Prevalence        0.3333         0.3333              0.3333
Balanced Accuracy           0.9955         0.9955              1.0000
```

Figure 8 Confusion Matrix and Statistics of Random Forest Model

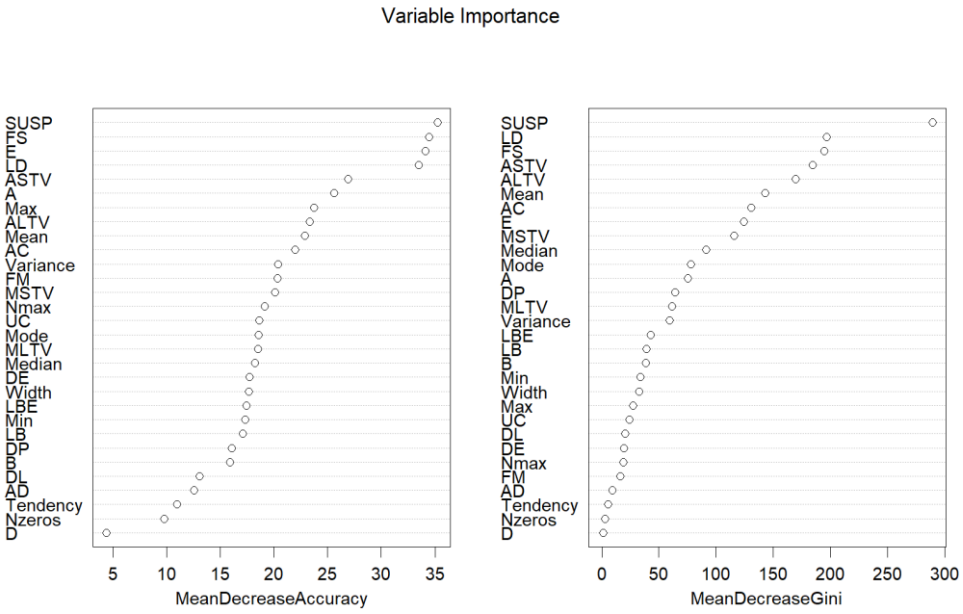

Figure 9 Variable Importance

4.2.4 Model Comparison

Through the training and testing of Logistic Regression, CART and Random Forest, we can see that all three models have extremely great performance in classification accuracy. Random Forest shows a slightly better accuracy performance, which is 99.6%, and has both a low

specificity (true negative rate) and a low fall-out (false positive rate). This can answer our first key business question: machine learning can improve medical diagnosis accuracy and serve as decision support system for health professionals?

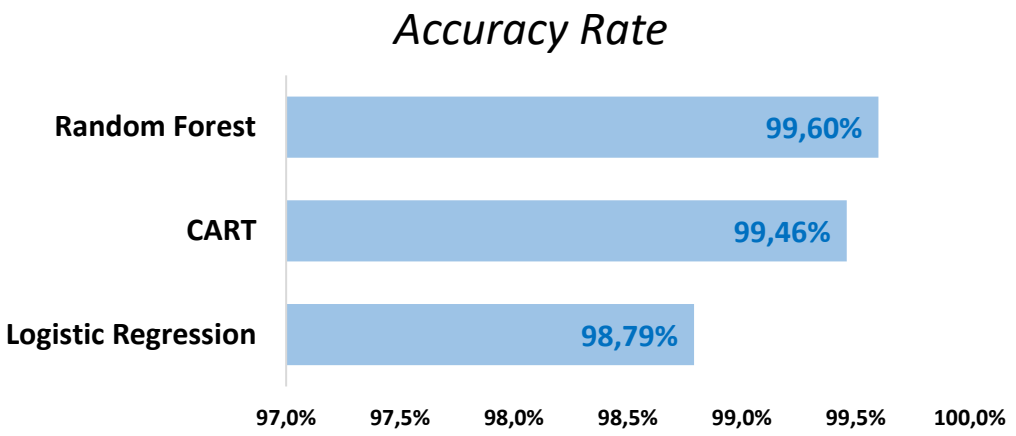

Figure 10 Variable Importance

Moreover, we can also find that Random Forest model is suitable for production. Given the test data, the model shows its competence in predicting fetal heart health conditions with adequate confidence. The accuracy of the model is so high that we do not have to collect more data, undertake more feature engineering, or experiment with other algorithms. Given the CTG indexes needed, the Random Forest model can provide us with a fast and accurate prediction of potential fetal heart disease.

4.3 Implementation Plan and Implications at NHCS

Currently in NHCS, there is a designated clinic for women which is called Women's Heart Clinic. The clinic provides comprehensive treatments that are tailored to the unique requirements of women. This includes end-to-end services for pregnant women covering from prevention to diagnosis, treatments, and post-surgery care. In this clinic, many women during pregnancy will visit obstetrician and / or cardiologist for regular check-ups which cover various test including CTG. It is in this process where machine learning will play an important role in interpreting CTG data. Figure 11 illustrates the implementation of AI assessment in hospital process level and data flow level.

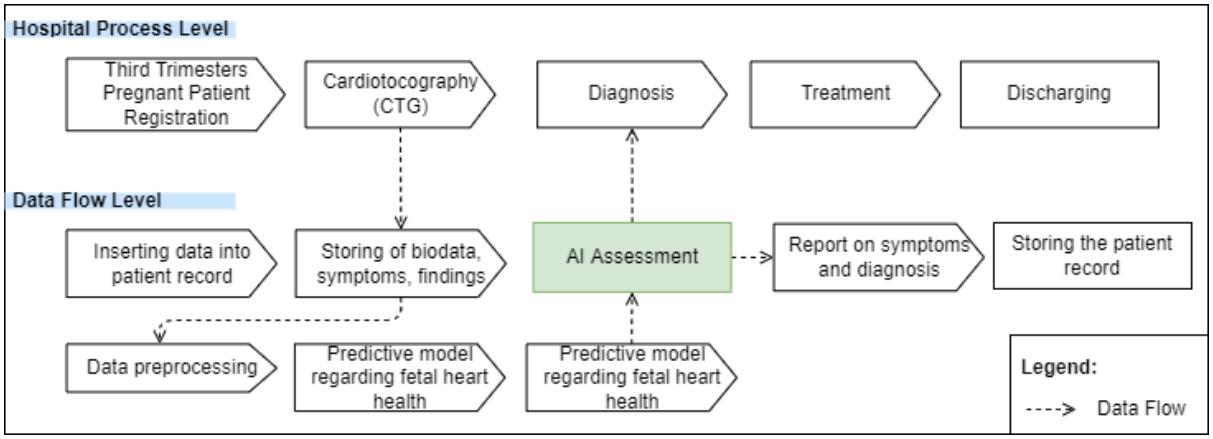

Figure 11 Implementation of AI Assessment Flowchart on Hospital Process and Data Flow Level

Before deploying the ML model directly, it is crucial to comprehend the difficulties in providing care that are related to the clinical use case; simply concentrating on the ability to effectively perform a prediction task is insufficient for improving care. It is important to take a more holistic view of what AI enabled solutions would look like beyond just a set of ML models. Therefore, this study also recommends process to build AI enabled system for interpreting CTG data in NHCS, which is as follows.

1. Understand the Users and Potential Issues
    - Map processes and key drivers of antenatal care provided by NHCS
    - Map thoughts, actions, and feelings of users (obstetricians, nurses, cardiologists)
    - Interview and observe users (obstetricians, nurses, cardiologists)
2. Design Intervention
    - Brainstorm with stakeholders
    - Design processes, workflows, and teams involved in antenatal care provided by NHCS
    - Build ML enabled information systems and digital tools
    - Establish data requirements and pipelines (including initial set of historical data and continuous data to train and improve)
    - Establish clinical utility and protocols of using ML
3. Implementation
    - Change management process
    - Educate and train users
    - Real time user feedback and observation
4. Evaluation
    - Assess implementation, process, and clinical outcomes
    - Monitor ML model performance
    - Decide to iterate, maintain, scale, or retire intervention

As shown in Figure 12, the implications of using ML can positively affect all aspects of NHCS value chain. In acquire / enroll aspect, using ML with high accuracy is a competitive advantage which increases credibility and reputation of NHCS, hence attract new patients and repeating patients. In prevention and diagnosis phase, ML can reduce misdiagnosis or underdiagnosis. For example, ML can accurately detect abnormal FHR pattern which allows early monitoring and intervention of fetal diseases such as embryonic hypoxia. This can reduce maternal and fetal mortality rates and complications during pregnancy and childbirth. Other than that, ML can deliver accurate prediction result quickly which saves a lot of time in diagnosing and can help health professionals decide on the right treatments and take quick actions if emergency cases appear. Regarding claims / payment, both NHCS and patients will be positively benefitted. Using ML can reduce NHCS financial cost that is caused by misdiagnosis and medical legal costs. On top of that, this also can relieve family's financial burden through early and accurate detection and intervention. All in all, considering investment to implement ML in interpreting CTG data and its benefits, it is clearly profitable for NHCS to adopt machine learning in its business.

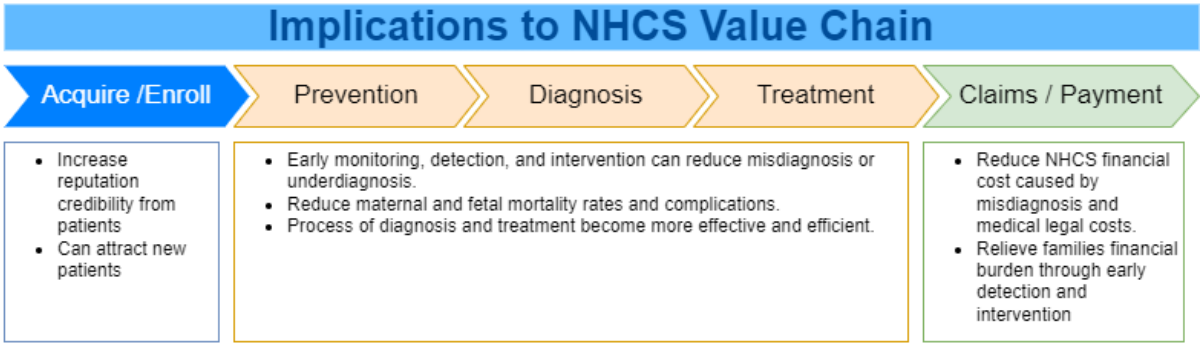

Figure 12 Implications of ML to NHCS Value Chain

5. Conclusion and Recommendation

For fetal heart health, CTG is useful for obstetricians to detect and monitor fetal heart condition. However, interpretation of the CTG data merely through visual analysis could result in misdiagnosis and financial burden. So, utilizing machine learning in interpreting CTG will have tremendous benefit in solving healthcare and business issues.

In this study, three machine learning models were developed and tested, which are Logistic Regression, CART and Random Forest. According to the results, Logistic Regression, CART and Random Forest based classifier could identify Normal, Suspicious and Pathologic condition, from the nature of CTG data with very high accuracy. Random Forest performed best out of three with highest accuracy (99.6%) with both a low specificity and a low fall-out value. Not to mention, Random Forest in this study also outperformed models from previous research.

To answer the key business questions, we have also developed AI-enable system implementation plan in the Women's Heart Clinic at NHCS. Through this implementation, NHCS will be benefited in all aspects of its value chain as machine learning will attract new customers, reduce misdiagnosis and mortality rates, improve efficiency in diagnosis and treatment. This can be directly impacted NHCS financially as it will reduce misdiagnosis and medical legal cost, while also relieve family's financial burden.

Even though Random Forest provided excellent performance, however in the real medical situation, hybrid models using statistical and other machine learning techniques, and combining CTG data with other data record of patients could potentially improve prediction or even diversify the target prediction. This would be worth to explore in future study. Furthermore, outlier value could possibly exist in future dataset. Therefore, developing machine learning model that is less sensitive to outliers would also be worth to explore for future research.

Bibliography

Alam, M.T. *et al.* (2022) 'Comparative Analysis of Different Efficient Machine Learning Methods for Fetal Health Classification', *Applied Bionics and Biomechanics*, 2022. Available at: https://doi.org/10.1155/2022/6321884.

C, Sundar., M.Chitradevi, M.C. and Geetharamani, G. (2012) 'Classification of Cardiotocogram Data using Neural Network based Machine Learning Technique', *International Journal of Computer Applications*, 47(14), pp. 19–25. Available at: https://doi.org/10.5120/7256-0279.

Costa, A. *et al.* (2009) 'Prediction of neonatal acidemia by computer analysis of fetal heart rate and ST event signals', *American Journal of Obstetrics and Gynecology*, 201(5), pp. 464.e1-464.e6. Available at: https://doi.org/10.1016/j.ajog.2009.04.033.

Faraoni, D., Nasr, V.G. and DiNardo, J.A. (2016) 'Overall Hospital Cost Estimates in Children with Congenital Heart Disease: Analysis of the 2012 Kid's Inpatient Database', *Pediatric Cardiology*, 37(1), pp. 37–43. Available at: https://doi.org/10.1007/s00246-015-1235-0.

Karabulut, E.M. and Ibrikci, T. (2014) 'Analysis of Cardiotocogram Data for Fetal Distress Determination by Decision Tree Based Adaptive Boosting Approach', *Journal of Computer and Communications*, 02(09), pp. 32–37. Available at: https://doi.org/10.4236/jcc.2014.29005.

McClung, N., Glidewell, J. and Farr, S.L. (2018) 'Financial burdens and mental health needs in families of children with congenital heart disease', *Congenital Heart Disease*, 13(4), pp. 554–562. Available at: https://doi.org/10.1111/chd.12605.

Mehbodniya, A. *et al.* (2022) 'Fetal health classification from cardiotocographic data using machine learning', *Expert Systems*, 39(6). Available at: https://doi.org/10.1111/exsy.12899.

Steinberg, D. (2009) *Chapter 10 CART: Classification and Regression Trees*. Available at: https://www.researchgate.net/publication/265031802.

The Human Cost and Financial Impact of Misdiagnosis How Significant a Problem Is Misdiagnosis? (2016). Available at: http://goo.

Zimmerman, M.S. *et al.* (2020) 'Global, regional, and national burden of congenital heart disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017', *The Lancet Child and Adolescent Health*, 4(3), pp. 185–200. Available at: https://doi.org/10.1016/S2352-4642(19)30402-X.