Predicting Intrapartum Fetal Risk Status Using Cardiotocography Data Classification Models

Supervisor: Dr. Ashok Bhowmick

Presented By Fayruz Kibria

Ryerson University

CIND 820 W2021 April, 2022

Contents

Abstract

A non invasive technique developed in the 1960s [1], cardiotocography (CTG) is the process of continuously recording the fetal heart rate (FHR), fetal movements and uterine contractions, measured using doppler ultrasound transducers and pressure sensors [2]. Intrapartum monitoring of the FHR alongside the aforementioned parameters can be used to determine whether the fetus is at risk of hypoxemia during labour. CTG pattern classification is a wide spread screening test carried out specially during labour to detect and prevent possible hypoxic-ischemic encephalopathy, cerebral palsy, and fetal death [3]. The International Federation of Gynaecology and Obstetrics (FICO) guidelines classify CTG findings into 3 categories: normal, suspect and pathological. The classifications are guided by observing FHR variability, accelerations and decelerations in relation to uterine contractions and fetal movements [4]. The interpretation of data is done by a obstetrician or skilled personnel [5]. However this surveillance method presents inter and intra observer variability [1]. While FHR classification is considered a reliable indicator of fetal well being, it is reported that up to 50% of FHR patterns classified as pathological can be explained by other physiological changes such as fetal behavioural stages and associated movements [2]. The high false positive rate for CTG monitoring intrapartum may lead to increased rates of avoidable operative deliveries. This poses increased risk of post operative maternal health complications, and amounts to increased cost imparted to the healthcare system [6]. Hence the goal of this project is to improve specificity without sacrificing sensitivity of detecting pathological state. This information can also assist health care workers in low- and middle-income countries to better triage pregnant women [5]. For this project the multivariate CTG dataset of 2126 pregnant women obtained from University of California Irvine Machine Learning Repository with 23+ features were used [7]. The dataset was classified by three expert obstetricians to define the fetal

risk. Then a final consensus label was assigned to each instance. Decision tree classifier (DT) [8], Random forest classifier (RF) [9] and Support vector machine classifier (SVM / SVC) [10] models were created and hyperparameters were tuned and model validated using a 10 fold cross validation technique, on a train set that was created using a 70/30 test train data split. Upon comparing the outputs of all the models, the best test scores obtained were, test accuracy = 0.934, (validation accuracy = 0.933), target class test f1 score = 0.907, target class test precision = 0.891, and target class recall = 0.925, using a DT classifier with data imbalance of 0.890, which was sampled using the SMOTENC technique on data with no min max normalization. When compared to a DT with no up sampling and hyperparameter tuning the improvement in scoring was as follows, test accuracy +0.039, test target class recall +0.113 and test target class f1 score +0.080. The model was trained in 2 minutes. However, there was no statistically significant difference found in the maximum test accuracies of the DT, RF, SVM model when compared pairwise using McNamar's test (P=0.978). The highest accuracy reported, using this dataset in the literature was found to be 0.990 using bagging and RF model [11] [12] in WEKA [13] which is much higher than the reported values in this study.

Introduction

For this literature review the following key words: cardiotocography, electronic fetal monitoring, intrapartum monitoring, intrapartum fetal heart rate, CTG guidelines, and machine learning intrapartum, artificial intelligence, labour monitoring, doppler fetal monitoring, were used in google scholar, PUBMED and, IEEE Xplore to review relevant publications. While most attention was paid to the more recent articles (year 2014+) the development of the field was also researched by reviewing older publications. Initial research was concentrated on understanding the signal acquisition methods and their relation to the clinical and physiological processes, and then reviewing clinical guidelines. Then later the focus was shifted to understanding recent developments in the field of machine learning (ML) and artificial intelligence (AI) involving neonatal risk analysis using CTG data and whether any clinical trials have used any such techniques.

Electronic fetal monitoring (EFM) is the name coined in 1960s that represented the continuous monitoring of fetal heart rate (FHR) and uterine contraction (UC) signals during labour using electronic systems [14]. The main reason to monitor is to predict / pre-empt fetal hypoxemia and acidemia, that may lead to brain or other organ injury. In other parts of the world this EFM became known as cardiotocography (CTG), the Greek word kardia means heart and tokos means labour and childbirth. CTG is a more accurate term since it includes the signals that are acquired from maternal contraction. Figure 1 shows the first commercially available fetal monitor released in 1968 [14]. The signal is recorded onto graph papers moving at a certain predetermined velocity, which can then be used to transform x axis to time units.

Figure 1: The Hewlett-Packard 8020A, the first commercially available fetal monitor, released in 1968 [14]. The signal is recorded onto graph papers moving at a certain predetermined velocity, which can then be used to transform x axis to time units.

The CTG data is made up of two parts where doppler ultrasound is used to monitor and record the fetal heart rate, and pressure sensors are used to record uterine contractions.

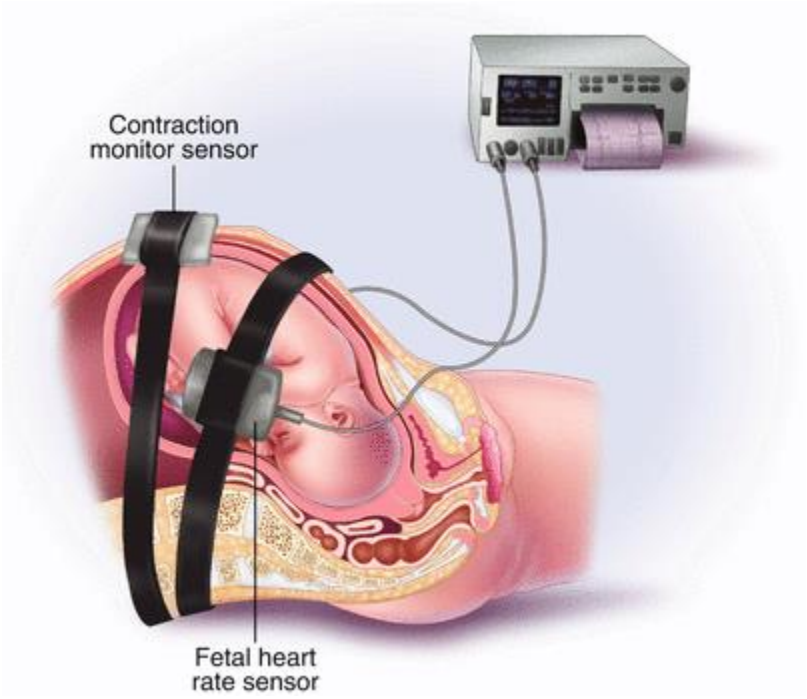


Figure 2: Diagram showing positioning of the doppler ultrasound transducer (fetal heart rate sensor) and pressure sensor (contraction monitor) on the mother [15].

Figure 2 shows how the sensors are placed on the mother during monitoring [15].

A typical display of the reading is shown in Figure 3. The red signals at the top display fetal heart rate and the blue signal at the bottom are UCs. Solid blocks of various colors above and below the signal tracings are indicative of certain calculated features detected at that point of the tracing [16]. A succinct description of how doppler ultrasound is converted into FHR signal can be found elsewhere [17].

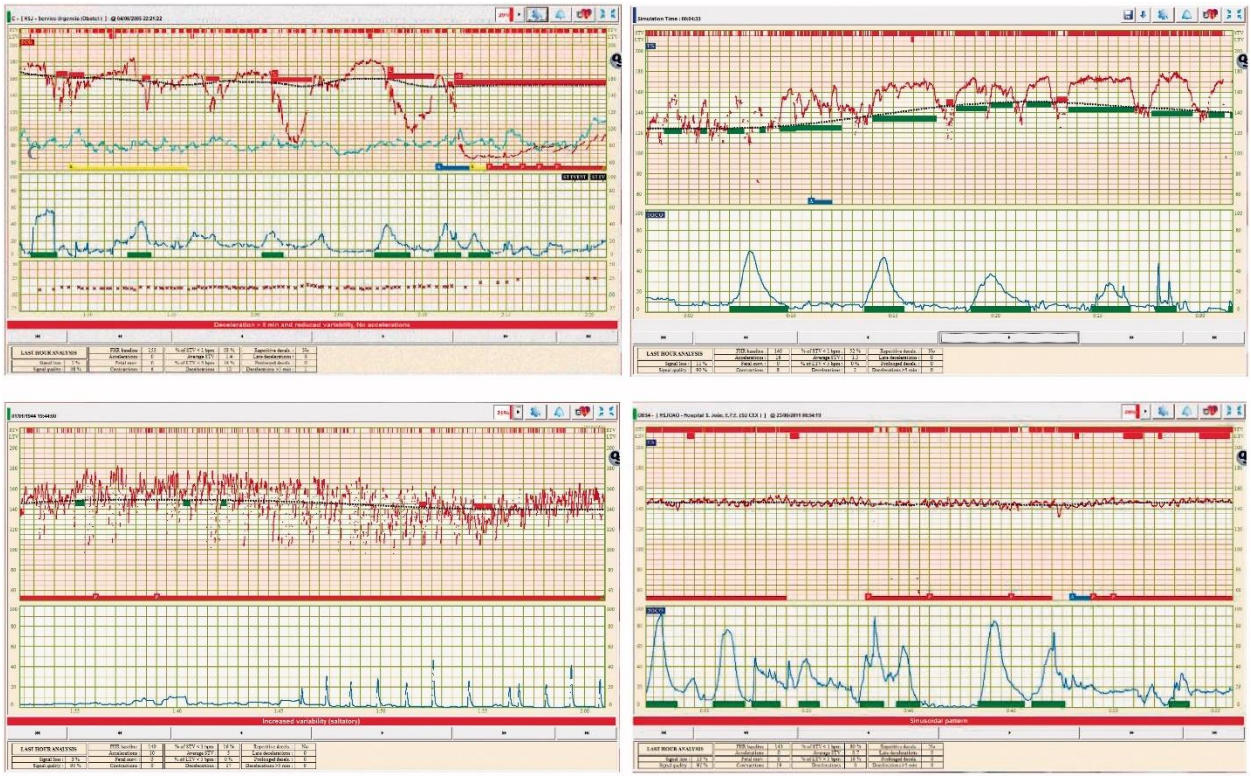


Figure 3: Tracings of CTG data from monitors. Red signals at the top are the FHR and blue at the bottom are UC. Solid dots above and below the signals show the detection of certain parameters as calculated by the monitor [16].

Some of the main organizations involved in overseeing the CTG guidelines include Society of Obstetricians and Gynaecologists of Canada (SOGC), American College of Obstetricians and Gynecologists (ACOG), National Institute for Health and Care Excellence (NICE), and International Federation of Obstetrics and Gynecology (FIGO).

Current Challenges and Opportunities in CTG / EFM

Guidelines around CTG usage and interpretation are established and most recent and notable ones are [4], [18]. While wide scale adoption of standardized communication terminologies has progressed in the past decade, the adoption of new technologies in clinical practices comes with its own challenges [19]. The high liability of exposure in obstetrics can limit the upgrade from well established practices to newer improved patient care options [19], [20]. Another special case in obstetrics is the fact that majority of the patients are healthy. So, the use of screening techniques have unique applications in this field. Obstetric data is highly imbalanced because problematic cases are rarer compared to unproblematic ones. A high overall accuracy does not mean much for assessing EFM techniques, this will mostly quantify the healthy patients that are predicted as healthy. A lack of specificity in EFM assessments can result in disproportionate intrapartum interventions, which can lead to unnecessary complications, and expenditures. Hence any ML algorithm that is used for automatic risk assessment screening, needs to have a high positive predictive value (PPV), sensitivity and specificity. Active ML and AI research in this field is ongoing and is focused on increasing sensitivity, specificity and PPV [20]. Clinical trials and recent articles have not yet reported improvement of neonatal outcomes over conventional methods using computerized programs[21], [20]. Most research in this realm has been in calculating the same features that were for long being interpreted visually, like baseline FHR, decelerations, accelerations. However, ML/AI has the capability to assess many more parameters from the signals that would not be available to a visual interpreter. While a lot of research is currently focused on this subject matter [22]–[24] and many innovative approaches are emerging, for now it is the prevailing consensus that ML/AI and computer analysis is not still at a point where they can effect better neonatal outcomes [25].

More prevalent public EFM data is needed [26] and increased cross discipline collaboration between clinical practitioners and biomedical researchers, and scientific reforming of CTG interpretation standards should go hand in hand, to create competitive, predictive monitors with high PPV and efficacy.

This work is the author's introductory undertaking to analyze, create and compare basic ML models using a medium sized public dataset from 2126 pregnant women obtained from University of California Irvine Machine Learning Repository [27] (based on FHR and UC signal features) to predict intrapartum fetal risk. While other work has been done using this dataset [5], [28]–[30] for analysis of CTG. Most of them report that more data and additional parameters are necessary to create a model that can eventually lead clinical to practice.

Pathophysiology of CTG signals in the Dataset

CTG interpretation is heavily dependent on recognizing FHR patterns in relation to UC and fetal movement (FM). The signals themselves do not tell the whole story though, they should always be interpreted in conjunction with all other available external and clinical information [2]. CTG signal characteristics as collected form the data source publications [7], [16], [31] and their physiological basis [32] are explained below:

Baseline FHR is the first step to determining all other features relative to it during FHR morphological classifications. It is expressed as a function of FHR histogram distribution during fetal rest and the prevalence of abnormal short term variability (STV) [7]. A condition where the heart rate is high >170 bpm is called tachycardia and a where it is low <100 bpm is called bradycardia, both of which may signal fetal distress.

FHR variability has two components, short and long term. STV is said the have a physiological basis in the interplay of the sympathetic and parasympathetic nervous system of the fetus. It is the variability in beat-to-beat measurements of the heart rate and near term is predominantly connected to the parasympathetic tone. A drop in STV below 1 beat per minute (bpm) is considered an abnormal STV (ASTV). Long term variability (LTV) of FHR has its physiological origin in the sympathetic nervous system and varies over a range of a minutes. A drop in healthy variation below 5 bpm over a 1-minute duration is considered an abnormal LTV (ALTV). High percentage of ASTV and ALTV can be an indicator fetal distress. A flat sinusoidal pattern where there is no STV is considered very ominous and imminent fetal risk can be inferred.

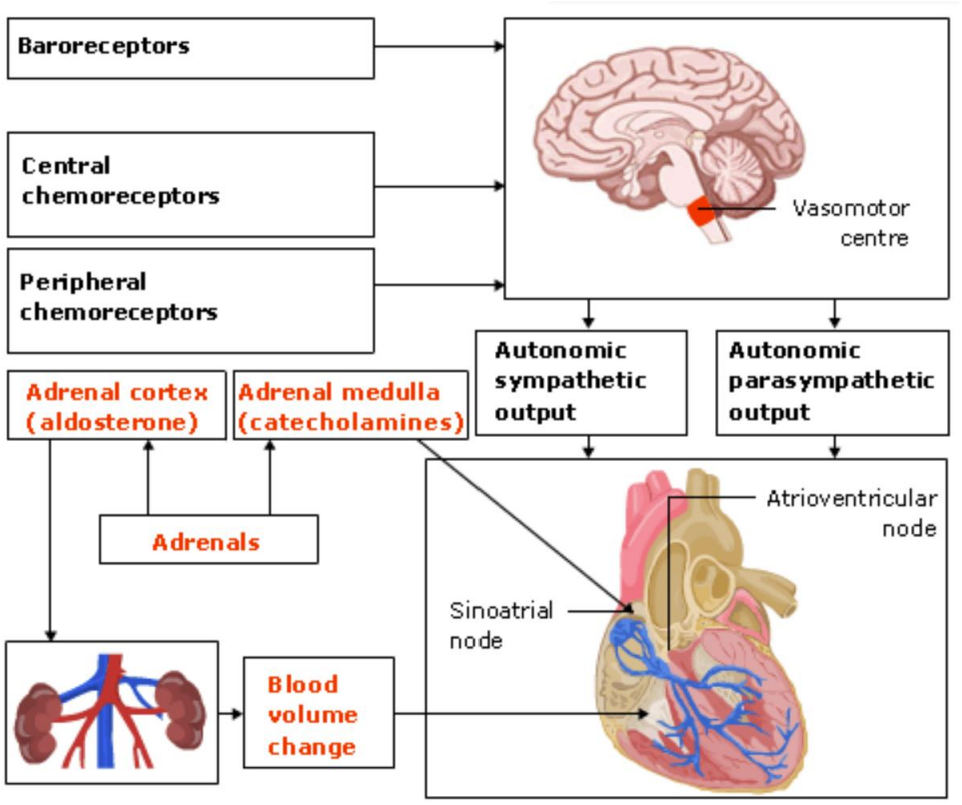


Figure 4: A complicated interplay of intrinsic and extrinsic factors that control the heartrate[32].

Figure 4 shows the complex process of several intrinsic and extrinsic factors that control FHR baseline value and the variability. In addition, Figure 5 shows how a normal level of variability is present in the fetus and varies with periods of movement and quiescence.

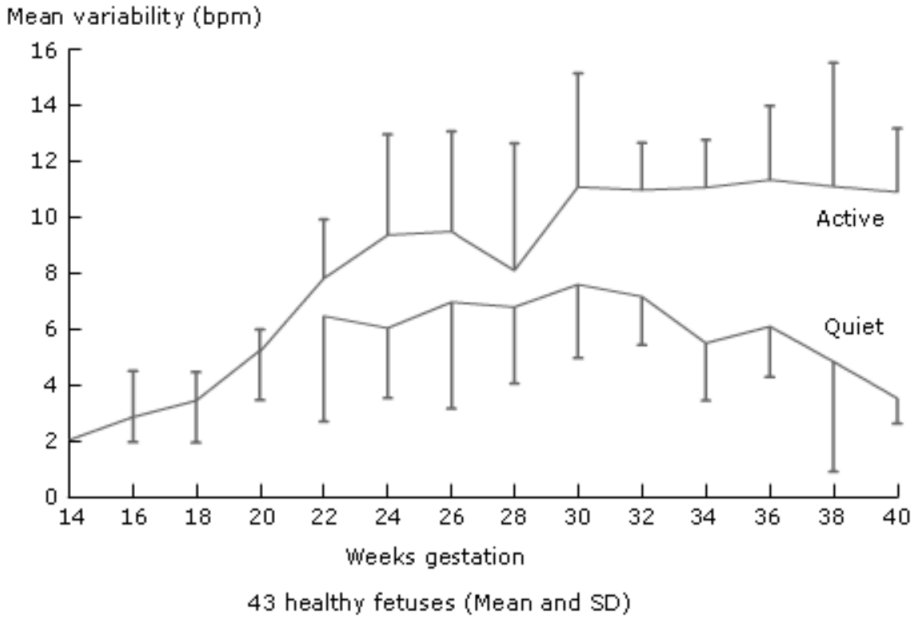


Figure 5: Variation in FHR with gestational stage. By week 38 there is marked distinction in baseline variability. It is very low when the fetus is quiet, and high when the fetus is active [32].

Accelerations are increases in FHR above the baseline by at least 15 bpm and lasting for 15-120 seconds. Accelerations are linked to fetal movement, and their presence is considered assuring of a healthy fetus. However, an absence of acceleration (when no other compounding suspect signs are present) is not necessarily problematic since the fetus has many cycles of sleep and vigilance.

Decelerations are defined as decreases in FHR below baseline lasting at least for 15 seconds and with an amplitude exceeding at least 15 bpm. Decelerations are considered mild/light if not lasting over 120 seconds, prolonged if they last between 120 to 300 seconds, and severe if they exceed 300 seconds. A repetitive deceleration is when they occur more than three times in 10 minutes, are

accompanying more than 80% of contractions or exceeding 50% of the tracing recording. They are also signs of fetal distress. Decelerations are considered normal when they occur and recover concurrently with UCs (this kind of heart rate drop is also called early decelerations). This is related to the fetal head compression during contractions. Late decelerations manifesting after UCs are considered pathological and is an indicator of placental insufficiency. Variable decelerations with no relation to UC are considered to originate form cord compression and is pathologic.

Uterine contractions are measured as a relative change in pressure during contraction by the sensor placed on the uterine fundus. Contractions when present occur normally 3-5 times over a 10-minute period.

Fetal movement can be detected by monitors or be reported using maternal perception. Fetal movement is an important parameter because it may lead to FHR accelerations which may be interpreted as tachycardia if the signal is considered in isolation or bradycardia if loss of signal quality due to motion has resulted. In general, a healthy pattern of fetal movement is assuring. As part of normal development, the fetus commonly manifests eye movement, body/limb movements which increases FHR.

When considering morphology of the FHR signal all complicated physiological pathways should be considered. Based on the possible onset and progression of hypoxia the FHR signal changes are manifested in different ways, as Figure 6 illustrates.

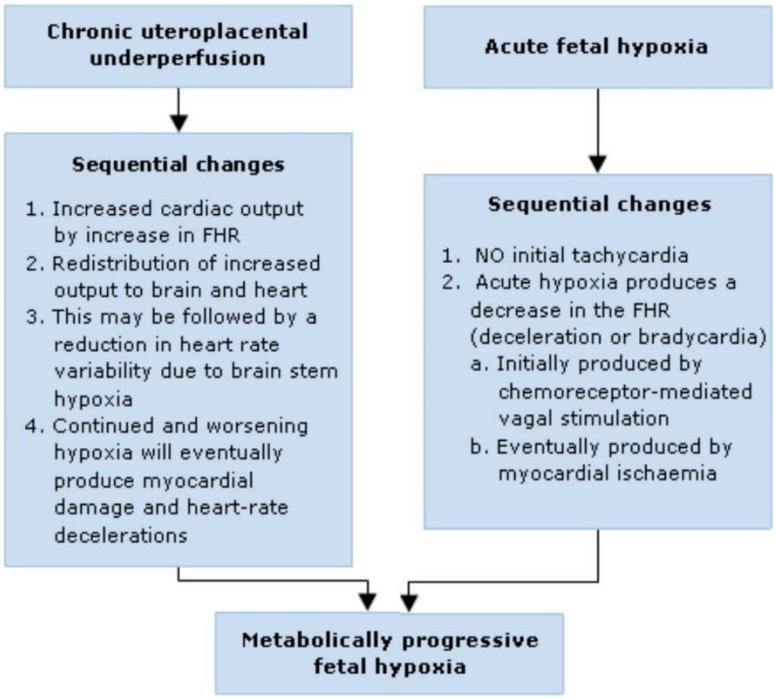


Figure 6: Sequential changes in FHR with chronic and acute hypoxia [32].

After considering all the physiology, and morphology of fetal risk hypoxia, a classification can be assigned as **normal**, **suspect** (where close monitoring is needed for early intervention should any pathological development occur) or **pathologic** (where immediate recourse is warranted) to the FHR signal. In short, the classification of fetal risk status is named 'NSP'.

Even though this dataset has a lot of information the unavailability of any raw signal data is limiting in terms of application of new parameter extraction. It is also does not any time serios information so no comparative analysis between UC and the FHR features in temporal distance can be done. Lastly there were no information on repetitive deceleration in this dataset which limits understanding of this parameter. This dataset also does not have any other clinical parameters and neonatal outcomes are unavailable which limits analysis into neonatal clinical outcomes.

Data Preparation Methodology and Statistics of Class Imbalance

Figure 7 shows the overall methodology [33] that was followed to prepare a clean and consistent version of the raw dataset [7], [27] for analysis. There were several iterations and adjustments as more understanding of the data was gained during the process.

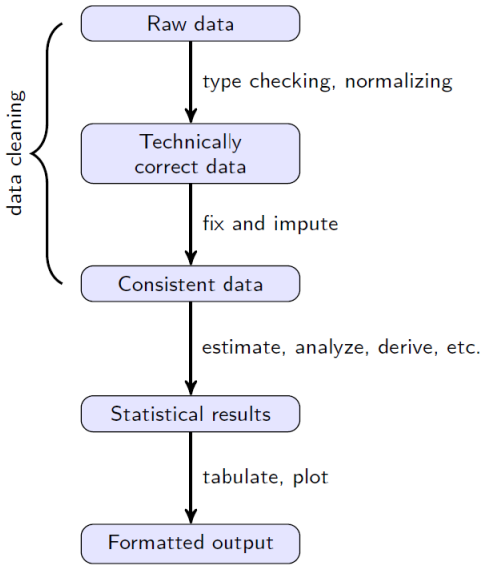


Figure 7: Data preparation methodology [33].

In order to quantify class imbalance in a dataset, the formula [34] below was used:

$$imbalance = \sum_{n=1}^{C} \left| \frac{1}{C} - \frac{n_c}{T_{data}} \right| \quad (1)$$

where C is the total number of classes, n_c is the count of instances in each class, T_{data} is the total count of instances in the dataset. When *imbalance* is 0 it implies equal number of instances in all classes and when the value is 1 it means only one class has all the instances. This statistic will be used in this report to balance data.

Data Descriptive Statistics

The raw dataset imported from the repository has the following attribute names. A list of names and their description is shown in Table 1. The data was imported into R Studio [35] for analysis. All code related to this report can be found on GitHub repository (Appendix A).

Exam Data File	**FileName**	of CTG examination
	Date	of the examination
EFM Measurements	**b**	start instant
	e	end instant
	LBE	baseline value (medical expert)
	LB	baseline value (SisPorto)
	AC	accelerations (SisPorto)
	FM	foetal movement (SisPorto)
	UC	uterine contractions (SisPorto)
	ASTV	percentage of time with abnormal short-term variability (SisPorto)
	mSTV	mean value of short-term variability (SisPorto)
	ALTV	percentage of time with abnormal long-term variability (SisPorto)
	mLTV	mean value of long-term variability (SisPorto)
	DL	light decelerations
	DS	severe decelerations
	DP	prolonged decelerations
	DR	repetitive decelerations
FHR Histogram	**Width**	histogram width
	Min	low freq. of the histogram
	Max	high freq. of the histogram
	Nmax	number of histogram peaks
	Nzeros	number of histogram zeros
	Mode	histogram mode
	Mean	histogram mean
	Median	histogram median
	Variance	histogram variance
	Tendency	histogram tendency: -1=left asymmetric; 0=symmetric; 1=right asymmetric
Classification	**A**	calm sleep
	B	REM sleep
	C	calm vigilance
	D	active vigilance
	E / (SH)	shift pattern (A or Susp with shifts)
	AD	accelerative/decelerative pattern (stress situation)
	DE	decelerative pattern (vagal stimulation)
	LD	largely decelerative pattern
	FS	flat-sinusoidal pattern (pathological state)
	SUSP	suspect pattern
	CLASS	Class code (1 to 10) for classes A to SUSP
	NSP	Normal=1; Suspect=2; Pathologic=3

Table 1: List of attributes in the raw data. Attributes belong to 4 broad categories. Exam file information, EFM measurements, histogram of the measured FHRs, morphology classification of the FHR, and classification of neonatal risk category.

Upon importing the raw data any incomplete instances with null values were deleted. These rows were not part of the original data, rather just some post processing information and a blank row between data and header. The columns FileName and Date were dropped. After which the dataset had 2126 rows and 37 columns. And there were no missing values. Columns Tendency, A, B, C, D, E, AD, DE, LD, FS, SUSP, CLASS, NSP were changed to factor. Following Table 2 reports the descriptive statistics for the remaining attributes

	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
b	2126	878.44	894.08	538.0	763.43	794.67	0.0	3296.0	3296.0	0.83	-0.54	19.39
e	2126	1702.88	930.92	1241.0	1622.78	782.81	287.0	3599.0	3312.0	0.66	-0.82	20.19
LBE	2126	133.30	9.84	133.0	133.26	10.38	106.0	160.0	54.0	0.02	-0.30	0.21
LB	2126	133.30	9.84	133.0	133.26	10.38	106.0	160.0	54.0	0.02	-0.30	0.21
AC	2126	2.72	3.56	1.0	2.05	1.48	0.0	26.0	26.0	1.66	3.11	0.08
FM	2126	7.24	37.13	0.0	1.06	0.00	0.0	564.0	564.0	9.41	104.28	0.81
UC	2126	3.66	2.85	3.0	3.41	2.97	0.0	23.0	23.0	0.83	1.28	0.06
ASTV	2126	46.99	17.19	49.0	46.88	20.76	12.0	87.0	75.0	-0.01	-1.05	0.37
MSTV	2126	1.33	0.88	1.2	1.22	0.74	0.2	7.0	6.8	1.66	4.68	0.02
ALTV	2126	9.85	18.40	0.0	5.17	0.00	0.0	91.0	91.0	2.19	4.23	0.40
MLTV	2126	8.19	5.63	7.4	7.71	4.60	0.0	50.7	50.7	1.33	4.11	0.12
DL	2126	1.57	2.50	0.0	1.03	0.00	0.0	16.0	16.0	1.82	3.13	0.05
DS	2126	0.00	0.06	0.0	0.00	0.00	0.0	1.0	1.0	17.33	298.43	0.00
DP	2126	0.13	0.46	0.0	0.00	0.00	0.0	4.0	4.0	4.23	19.15	0.01
DR	2126	0.00	0.00	0.0	0.00	0.00	0.0	0.0	0.0	NaN	NaN	0.00
Width	2126	70.45	38.96	67.5	68.63	46.70	3.0	180.0	177.0	0.31	-0.90	0.84
Min	2126	93.58	29.56	93.0	92.97	40.03	50.0	159.0	109.0	0.12	-1.29	0.64
Max	2126	164.03	17.94	162.0	163.11	16.31	122.0	238.0	116.0	0.58	0.63	0.39
Nmax	2126	4.07	2.95	3.0	3.76	2.97	0.0	18.0	18.0	0.89	0.50	0.06
Nzeros	2126	0.32	0.71	0.0	0.17	0.00	0.0	10.0	10.0	3.91	30.26	0.02
Mode	2126	137.45	16.38	139.0	138.49	14.83	60.0	187.0	127.0	-0.99	2.99	0.36
Mean	2126	134.61	15.59	136.0	135.51	14.83	73.0	182.0	109.0	-0.65	0.92	0.34
Median	2126	138.09	14.47	139.0	138.72	14.83	77.0	186.0	109.0	-0.48	0.66	0.31
Variance	2126	18.81	28.98	7.0	12.38	8.90	0.0	269.0	269.0	3.22	15.08	0.63

	Level	Count	Prop
Tendency	1	846	39.79
Tendency	0	1115	52.45
Tendency	-1	165	7.76
A	0	1742	81.94
A	1	384	18.06
B	0	1547	72.77
B	1	579	27.23
C	0	2073	97.51
C	1	53	2.49
D	0	2045	96.19
D	1	81	3.81
E	0	2054	96.61
E	1	72	3.39
AD	0	1794	84.38
AD	1	332	15.62
DE	0	1874	88.15
DE	1	252	11.85
LD	0	2019	94.97
LD	1	107	5.03
FS	1	69	3.25
FS	0	2057	96.75
SUSP	0	1929	90.73
SUSP	1	197	9.27

	Level	Count	Prop
CLASS	9	69	3.25
CLASS	6	332	15.62
CLASS	2	579	27.23
CLASS	8	107	5.03
CLASS	10	197	9.27
CLASS	7	252	11.85
CLASS	1	384	18.06
CLASS	3	53	2.49
CLASS	5	72	3.39
CLASS	4	81	3.81
NSP	2	295	13.88
NSP	1	1655	77.85
NSP	3	176	8.28

Table 2: Data descriptive statistics reported in R.

Please note the combined morphological classification variable CLASS should not be confused with the feature variable of this investigation which is NSP.

Figure 8 Below shows the histogram of the baseline FHR calculated by expert opinion LBE and the one obtained from the EFM, LE. The colour fill is based on the target variable NSP. The two distributions present very similar to the eye.

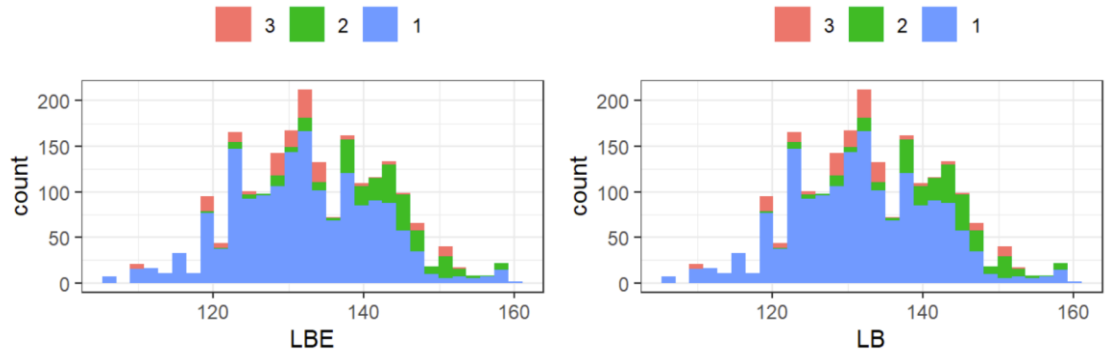


Figure 8: Baseline FHR histogram, LBE is based on expert opinion and LB is taken from the EFM.

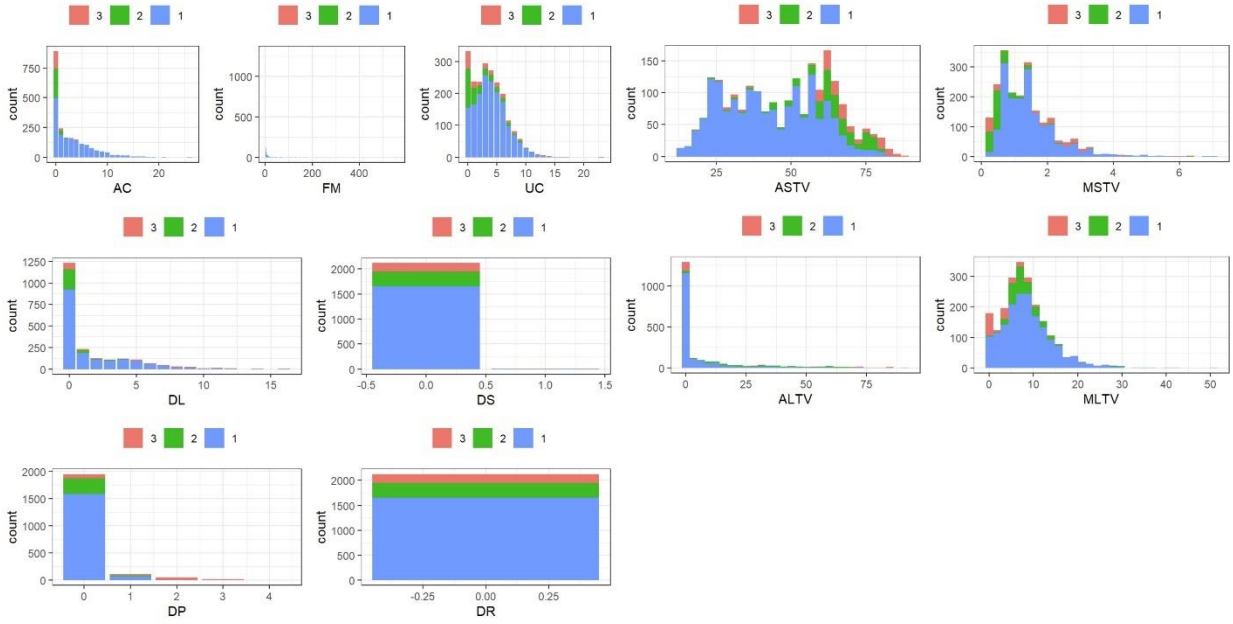


Figure 9: Histogram of the numeric variables, where fill is the NSP classifications.

Figure 9 shows the histogram of the numerical values reported from the EFM. The data does not have normal distribution. And for DS, DP there are very few counts reported. Upon close inspection of Table 2 it is noted that there are no events reported for DR.

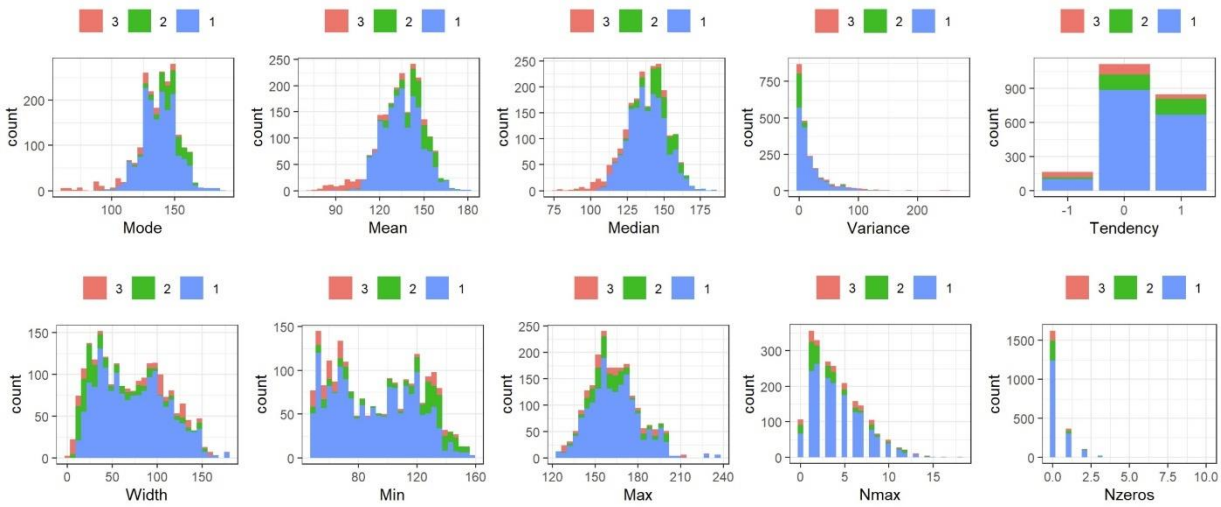


Figure 10: Distribution of the variables related to the FHR histogram.

In Figure 10 the distribution of the variables related to the histogram of the FHR is shown. While mean, mode and median of the histograms show a trend towards normal distribution, the other parameters are not normally distributed. There were more counts of the right asymmetric description over the left asymmetric description in the tendency variable. A visual inspection of the mode, mean and median distributions show that NSP class 3 is more prevalent in the values that are more peripheral. Hence no outliers were removed form this dataset as plausibly more extreme values may be more indicative of a riskier fetal state.

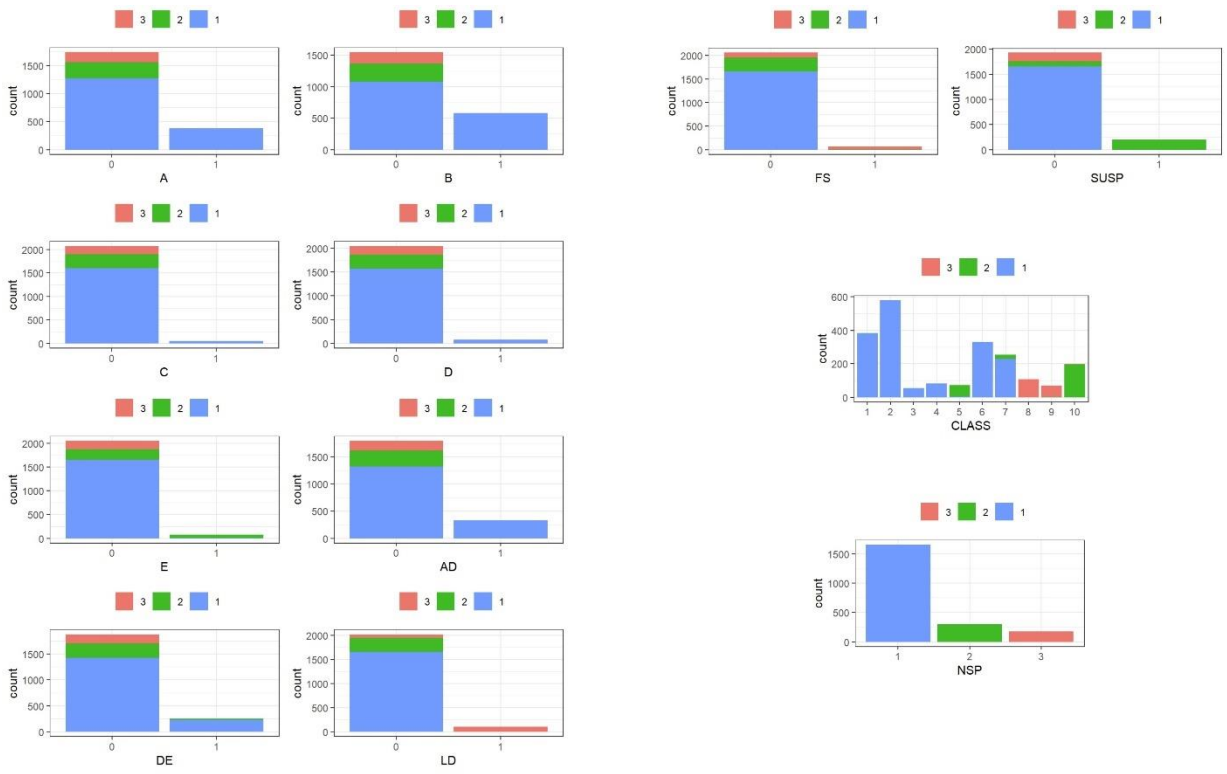


Figure 11: Distribution off categorical variables for FHR signal morphology and neonatal risk.

The FHR signal morphological classification variables and the neonatal risk classification variables are presented in Figure 11. There is a very large class imbalance in this dataset for the target variable NSP. There are many more NSP class 1 (blue in Figure 9, Figure 10, and Figure 11) instances than class 2 (green) and 3 (red) combined.

The feature DR was investigated further, and it had zero variance and hence the column was removed from any further analysis. And the feature LBE and LB were investigated and found to be exact duplicates and hence LBE was removed from any further analysis.

Exploratory Data Analysis

The has two parts, the machine reported data on the various parameters of the CTG and the (human) expert classified morphological data. We focus on the machine reported data for classification with some preliminary investigations involving the morphological categorical data.

CTG Recording Length Normalization

Highlighted in blue in Table 1 are attributes b and e. They represent the start and stop instance numbers of the recording. Based on which the number of data points ‘nPoints’ (proportional to the length of recording time) collected for each case can be calculated. This task was completed in R and the new column nPoints was added to the dataframe. Figure 12 shows the distribution of recording length (nPoints distribution). The signal acquisition frequency is not known hence the number of recordings points can not be directly converted to length of recording time.

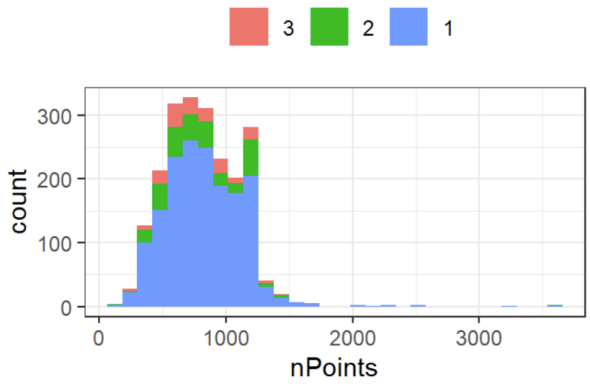


Figure 12: Histogram of the length of recording of FHR signal.

Over these varying periods of recording lengths, the number of data points where, fetal movement, uterine contractions, FHR accelerations, light decelerations, severe decelerations, and prolonged decelerations, were detected are reported in columns FM, UC, AC, DL, DS, and DP respectively.

Since the recorded length can influence the number of calculated feature points that are reported, the feature point counts were normalized by the signal record length count to observe their effect. The calculated new columns were added to the dataframe as nFM, nUC, nAC, nDL, nDS, nDP and Table 3 shows the descriptive statistics. The original columns. Original columns FM, UC, AC, DL, DS, and DP were kept in a separate dataframe that did not have any normalization applied to it. The normalized and non normalised data would be investigated separately to see how the predictive performance of models vary.

	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
nPoints	2126	824.44	305.05	799.0	817.31	306.16	117.0	3599.00	3482.00	1.35	8.72	6.62
nAC	2126	0.00	0.00	0.0	0.00	0.00	0.0	0.02	0.02	1.21	0.78	0.00
nFM	2126	0.01	0.05	0.0	0.00	0.00	0.0	0.48	0.48	7.80	64.05	0.00
nUC	2126	0.00	0.00	0.0	0.00	0.00	0.0	0.01	0.01	0.16	-0.65	0.00
nDL	2126	0.00	0.00	0.0	0.00	0.00	0.0	0.02	0.02	1.72	2.49	0.00
nDS	2126	0.00	0.00	0.0	0.00	0.00	0.0	0.00	0.00	17.79	320.02	0.00
nDP	2126	0.00	0.00	0.0	0.00	0.00	0.0	0.01	0.01	4.27	20.01	0.00

Table 3: Descriptive statistics of the new calculated time normalized attributes.

Min-Max Normalization

The numerical variables in the time normalized dataframe were then scaled using the min-max scaling method in R. Table 4shows the descriptive statistics of the min-max normalized variables stratified by NSP group.

	group: 1											group: 2											group: 3										
	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
LB	0.48	0.18	0.48	0.48	0.19	0.00	1.00	1.00	0.07	-0.12	0.00	0.66	0.15	0.69	0.67	0.14	0.26	0.98	0.72	-0.45	0.15	0.01	0.48	0.17	0.48	0.47	0.11	0.07	0.85	0.78	0.24	0.03	0.01
ASTV	0.41	0.21	0.39	0.40	0.26	0.00	0.95	0.95	0.14	-1.00	0.01	0.67	0.16	0.68	0.68	0.12	0.08	0.89	0.81	-1.12	1.58	0.01	0.70	0.19	0.71	0.73	0.10	0.08	1.00	0.92	-1.14	1.32	0.01
MSTV	0.18	0.12	0.16	0.16	0.11	0.00	1.00	1.00	1.85	5.70	0.00	0.06	0.10	0.03	0.05	0.02	0.00	0.90	0.90	4.79	30.30	0.01	0.20	0.18	0.22	0.19	0.26	0.00	0.90	0.90	0.63	0.51	0.01
ALTV	0.06	0.13	0.00	0.02	0.00	0.00	0.80	0.80	3.03	9.64	0.00	0.32	0.22	0.30	0.31	0.28	0.00	0.75	0.75	0.27	-1.13	0.01	0.25	0.37	0.00	0.20	0.00	0.00	1.00	1.00	0.97	-0.88	0.03
MLTV	0.17	0.11	0.16	0.16	0.10	0.00	1.00	1.00	1.31	3.91	0.00	0.16	0.07	0.14	0.15	0.05	0.00	0.58	0.58	2.13	8.33	0.00	0.07	0.08	0.06	0.06	0.10	0.00	0.42	0.42	1.35	2.29	0.01
Width	0.40	0.21	0.38	0.39	0.24	0.04	0.98	0.94	0.33	-0.82	0.01	0.26	0.22	0.16	0.23	0.13	0.03	0.83	0.80	1.11	0.06	0.01	0.43	0.28	0.51	0.43	0.31	0.00	1.00	1.00	-0.20	-1.32	0.02
Min	0.38	0.25	0.37	0.37	0.33	0.00	1.00	1.00	0.15	-1.15	0.01	0.58	0.28	0.71	0.61	0.18	0.00	0.96	0.96	-0.81	-0.80	0.02	0.31	0.31	0.15	0.29	0.18	0.00	0.95	0.95	0.65	-1.32	0.02
Max	0.37	0.15	0.35	0.36	0.14	0.00	1.00	1.00	0.57	0.84	0.00	0.35	0.14	0.32	0.34	0.10	0.03	0.67	0.64	0.72	0.05	0.01	0.35	0.19	0.31	0.34	0.19	0.05	0.93	0.88	0.50	-0.38	0.01
Nmax	0.23	0.16	0.22	0.22	0.16	0.00	1.00	1.00	0.84	0.45	0.00	0.18	0.17	0.11	0.16	0.08	0.00	0.89	0.89	1.44	1.70	0.01	0.25	0.18	0.22	0.23	0.25	0.00	0.78	0.78	0.62	-0.19	0.01
Nzeros	0.03	0.07	0.00	0.02	0.00	0.00	1.00	1.00	3.52	27.36	0.00	0.02	0.08	0.00	0.01	0.00	0.00	0.80	0.80	6.04	47.22	0.00	0.03	0.07	0.00	0.02	0.00	0.00	0.30	0.30	2.00	3.50	0.01
Mode	0.62	0.11	0.61	0.62	0.11	0.20	1.00	0.80	0.16	0.47	0.00	0.68	0.09	0.69	0.69	0.07	0.24	0.84	0.61	-1.09	2.90	0.01	0.43	0.21	0.49	0.44	0.22	0.00	0.79	0.79	-0.36	-0.88	0.02
Mean	0.57	0.12	0.57	0.57	0.12	0.14	1.00	0.86	0.10	-0.12	0.00	0.66	0.10	0.67	0.67	0.08	0.23	0.91	0.68	-0.86	1.68	0.01	0.37	0.21	0.31	0.36	0.24	0.00	0.78	0.78	0.31	-1.13	0.02
Median	0.56	0.12	0.56	0.56	0.12	0.26	1.00	0.74	0.12	-0.12	0.00	0.64	0.10	0.65	0.65	0.08	0.13	0.89	0.76	-0.97	2.77	0.01	0.39	0.18	0.36	0.39	0.18	0.00	0.76	0.76	0.20	-0.71	0.01
Variance	0.06	0.08	0.03	0.05	0.04	0.00	0.66	0.66	2.36	7.04	0.00	0.03	0.07	0.00	0.01	0.01	0.00	0.43	0.43	3.69	13.94	0.00	0.19	0.23	0.14	0.15	0.20	0.00	1.00	1.00	1.44	1.88	0.02
nPoints	0.20	0.09	0.20	0.20	0.09	0.01	1.00	0.99	1.26	7.68	0.00	0.20	0.09	0.19	0.20	0.10	0.00	1.00	1.00	2.10	15.26	0.01	0.19	0.08	0.18	0.19	0.08	0.01	0.46	0.45	0.36	-0.17	0.01
nAC	0.21	0.21	0.15	0.18	0.23	0.00	1.00	1.00	0.90	0.15	0.01	0.01	0.04	0.00	0.00	0.00	0.00	0.26	0.26	3.19	12.24	0.00	0.02	0.05	0.00	0.01	0.00	0.00	0.25	0.25	2.75	6.71	0.00
nFM	0.02	0.09	0.00	0.00	0.00	0.00	1.00	1.00	9.35	95.13	0.00	0.02	0.09	0.00	0.01	0.00	0.00	0.89	0.89	9.04	83.44	0.01	0.05	0.18	0.00	0.00	0.00	0.00	0.78	0.78	3.40	9.74	0.01
nUC	0.32	0.18	0.33	0.32	0.20	0.00	1.00	1.00	0.04	-0.49	0.00	0.16	0.18	0.10	0.13	0.15	0.00	0.76	0.76	0.84	-0.41	0.01	0.25	0.24	0.22	0.23	0.33	0.00	0.96	0.96	0.63	-0.62	0.02
nDL	0.13	0.19	0.00	0.09	0.00	0.00	1.00	1.00	1.59	2.07	0.00	0.03	0.10	0.00	0.01	0.00	0.00	0.90	0.90	4.38	24.71	0.01	0.24	0.27	0.15	0.20	0.22	0.00	0.95	0.95	0.86	-0.47	0.02
nDS	0.00	0.02	0.00	0.00	0.00	0.00	0.66	0.66	40.61	1648.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	NaN	NaN	0.00	0.03	0.15	0.00	0.00	0.00	0.00	1.00	1.00	5.20	25.64	0.01
nDP	0.01	0.05	0.00	0.00	0.00	0.00	0.52	0.52	5.54	32.86	0.00	0.02	0.07	0.00	0.00	0.00	0.00	0.52	0.52	4.82	23.24	0.00	0.24	0.25	0.21	0.21	0.32	0.00	1.00	1.00	0.61	-0.70	0.02

Table 4: Descriptive statistics of the min-max normalized numerical variables, stratified by NSP group.

To investigate the effect of min max normalization, both the normalized and non normalized dataframes were used for modeling. Unless otherwise mentioned the dataset referred to is the non normalized one.

Kruskal-Wallis Test

To get more insight on the variation of median of all the variables when grouped by the NSP class, a Kruskal-Wallis test was performed in R.

key	statistic	p.value	parameter	method
DP	538.	1.17e-117	2	Kruskal-Wallis rank sum test
ASTV	533.	1.46e-116	2	Kruskal-Wallis rank sum test
AC	441.	1.37e- 96	2	Kruskal-Wallis rank sum test
MSTV	383.	5.83e- 84	2	Kruskal-Wallis rank sum test
ALTV	367.	1.78e- 80	2	Kruskal-Wallis rank sum test
LB	254.	5.92e- 56	2	Kruskal-Wallis rank sum test
Median	189.	9.89e- 42	2	Kruskal-Wallis rank sum test
UC	160.	2.13e- 35	2	Kruskal-Wallis rank sum test
Mean	153.	5.61e- 34	2	Kruskal-Wallis rank sum test
Variance	140.	5.00e- 31	2	Kruskal-Wallis rank sum test

key	statistic	p.value	parameter	method
nPoints	5.75	5.63e- 2	2	Kruskal-Wallis rank sum test
Max	8.82	1.22e- 2	2	Kruskal-Wallis rank sum test
Nzeros	12.2	2.26e- 3	2	Kruskal-Wallis rank sum test
FM	16.3	2.83e- 4	2	Kruskal-Wallis rank sum test
Min	22.0	1.70e- 5	2	Kruskal-Wallis rank sum test
Nmax	49.4	1.87e-11	2	Kruskal-Wallis rank sum test
Width	52.1	4.93e-12	2	Kruskal-Wallis rank sum test
DS	55.5	9.04e-13	2	Kruskal-Wallis rank sum test
Tendency	71.3	3.22e-16	2	Kruskal-Wallis rank sum test
MLTV	86.6	1.58e-19	2	Kruskal-Wallis rank sum test

Table 5: KW test data to understand groupwise variation of the machine reported data. (Left) 10 variables who have the most significant intergroup difference, and (right) 10 variables with the least intergroup variability, when grouped by NSP.

In Table 5 (left) the variables DP (prolonged deceleration), ASTV (abnormal short term variability), AC (acceleration), MSTV (mean value of short term variability), ALTV (abnormal long term variability), LB (baseline heart beat rate) are all very important parameters in determining fetal behaviour. So, the effect of fetal state is most impacted their distribution and vice versa. On the right side of Table 5 we see that the variables nPoints (recording length), Max (max heart rate), Nzeros (number of heart rate histogram zeros), FM (fetal motion) are least affected by NSP state. While they do seem least important it is to be noticed that their values may be scarce and hence not best represented in this kind of analysis. All variables except for nPoints were statistically significant.

Correlation

A Pearson's correlation plot (Figure 13) in R for all the numerical variables shows that the variable nPoints is not strongly correlated to any of the other numerical variables. Strong correlation is found between Mean, Max and Mode and LB. Variance and MSTV were strongly correlated.

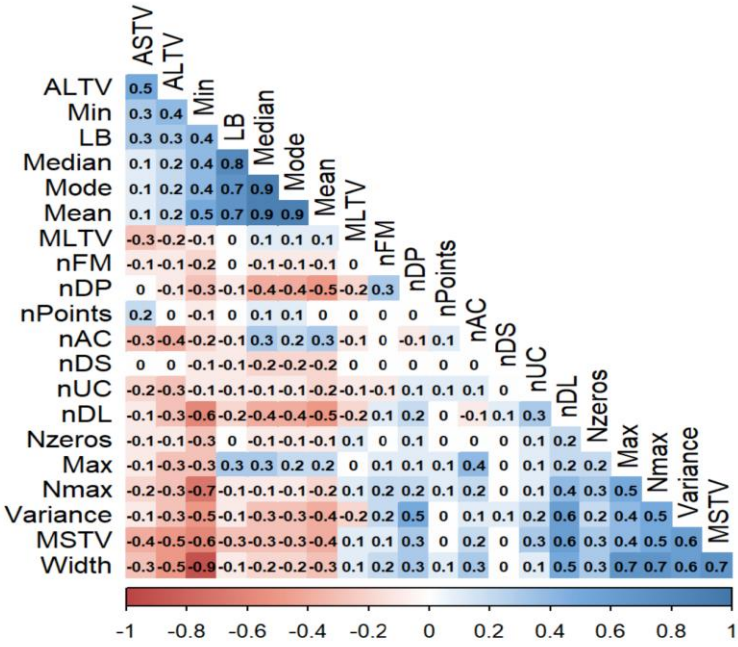


Figure 13: Pearson's correlation matrix between all the numerical variables in the dataset.

A Spearman's rank correlation plot showing the strength of correlation between the machine reported variables and the feature variable were plotted in R. Figure 14 shows the strength of the correlation of ranks for the variables. To highlight a few the Variable NSP has negative relation with AC, UC, MSTV, Variance (these variables indicate normal state within an accepted range). And NSP state a positive relation with ASTV, ALTV, DP which are indicators of distress.

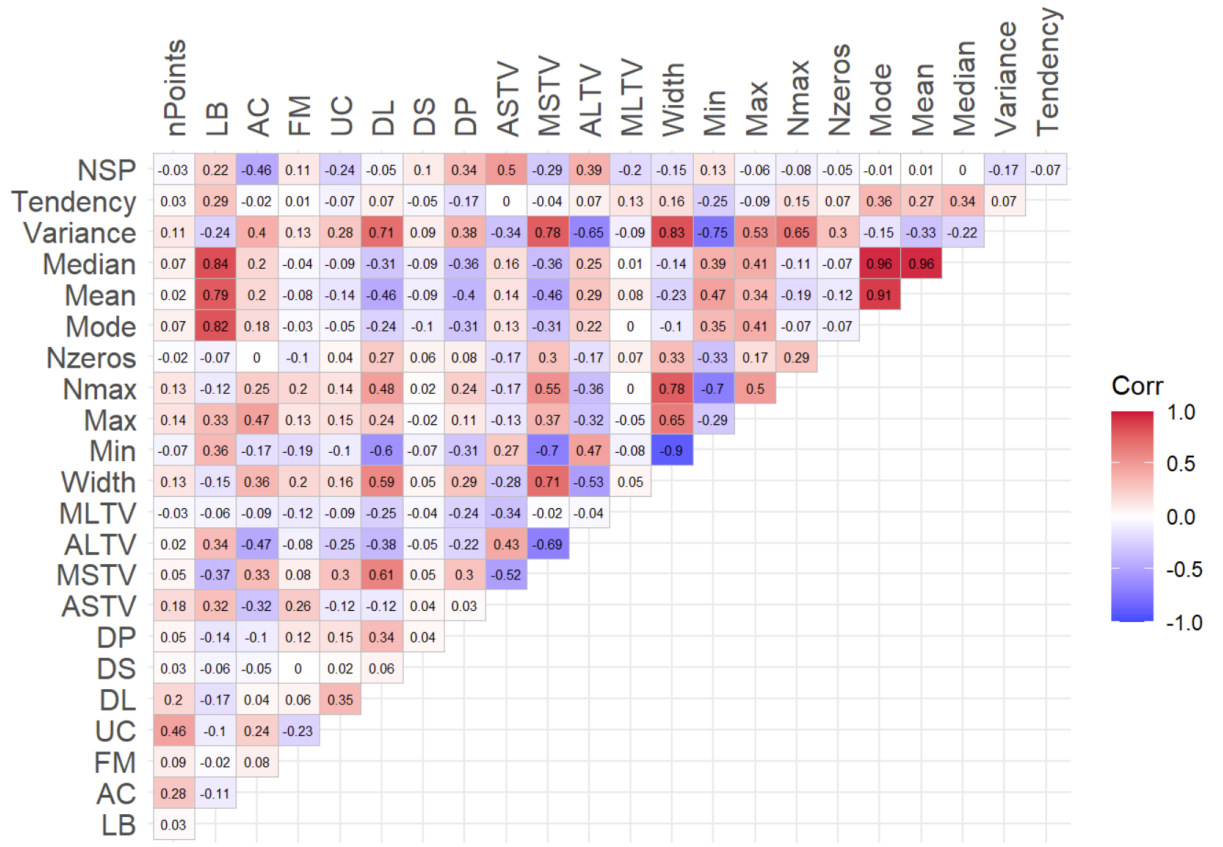


Figure 14: Spearman's correlation for machine reported variables, and their relation to the class NSP.

Also, for complete understanding the morphological variables were also compared using Spearman's correlation, as shown in Figure 15. The variables LD (largely decelerative pattern), FS (flat sinusoidal pattern), SUSP (a suspicious state but not necessarily pathological), CLASS are all highly correlated with NSP.

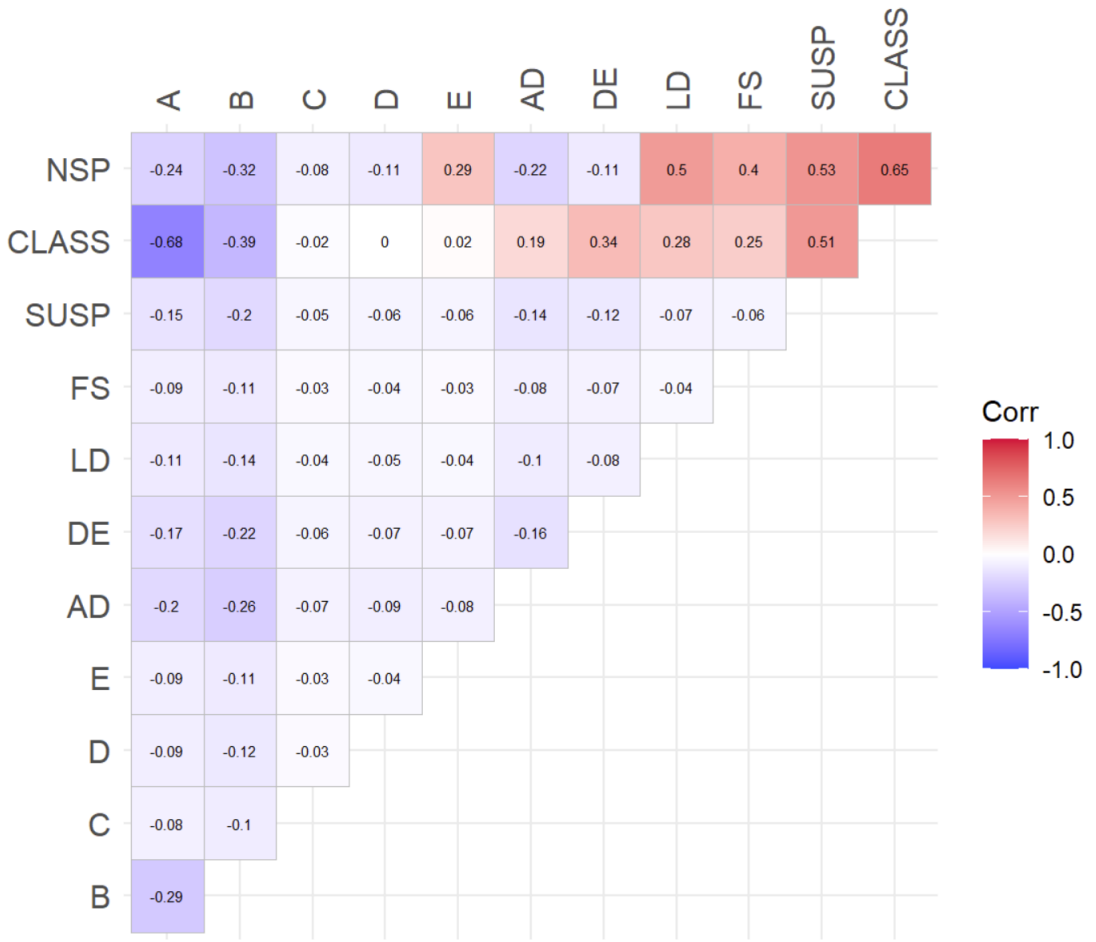


Figure 15: Spearman's correlation between NSP class and the morphological variables.

Multicollinearity and Variable Grouping

Since the correlation matrix showed high correlation amongst some of the variables. A further visual inspection was done to visualize multicollinearity in the data using 'pairs.panel' method in R. The grouping for comparison was based on 5 different types. Baseline characteristics, acceleration related characteristics, decelerative characteristics, variability characteristics and uterine contraction characteristics.

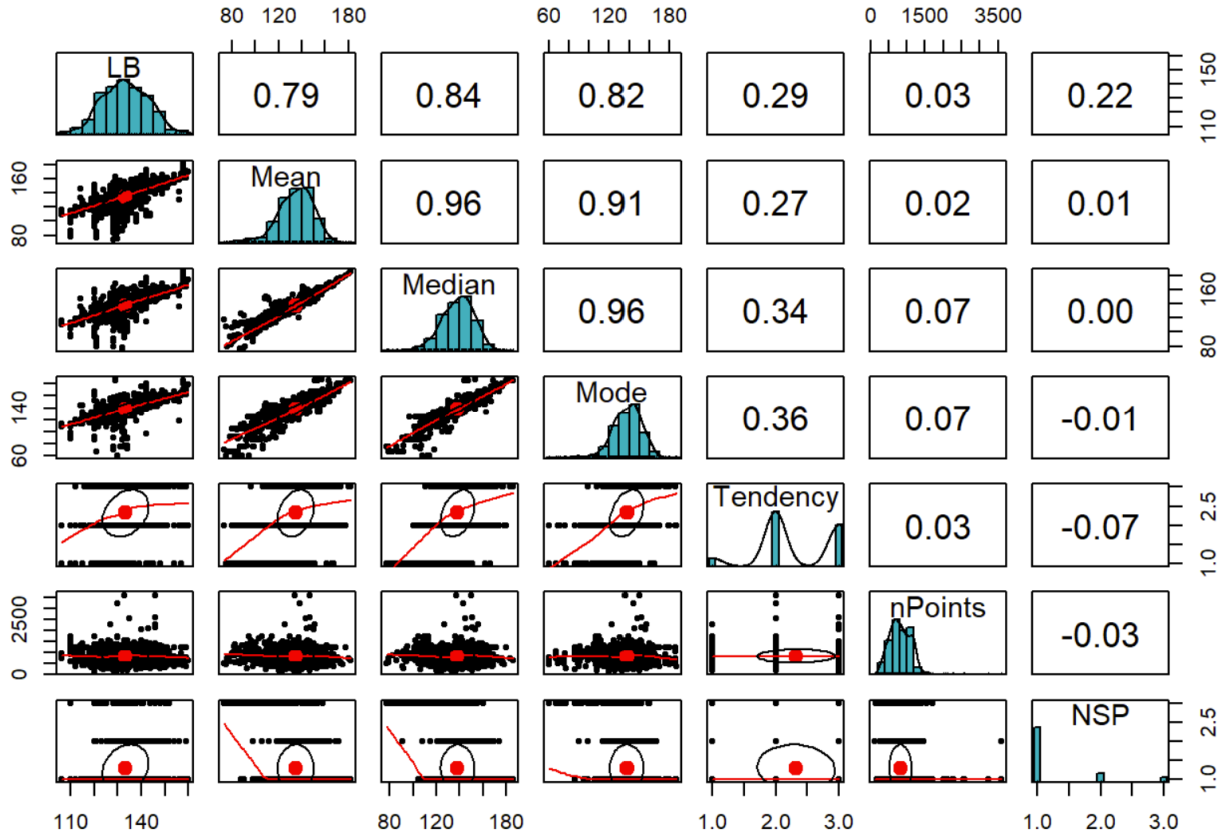


Figure 16: Baseline characteristics under pairs panel.

Figure 16 shows pairwise comparison of all the baseline characteristics. We can see multicollinearity between Mean, Median and Mode, LB is also strongly positively correlated to these three features. NSP correlated negatively with Mean, Median and mode, at low cut off values. The categorical variable Tendency (kind of a surrogate for skew) positively correlated with the three central value indicators. When the recording length was above a certain value there were no NSP class 3 categories.

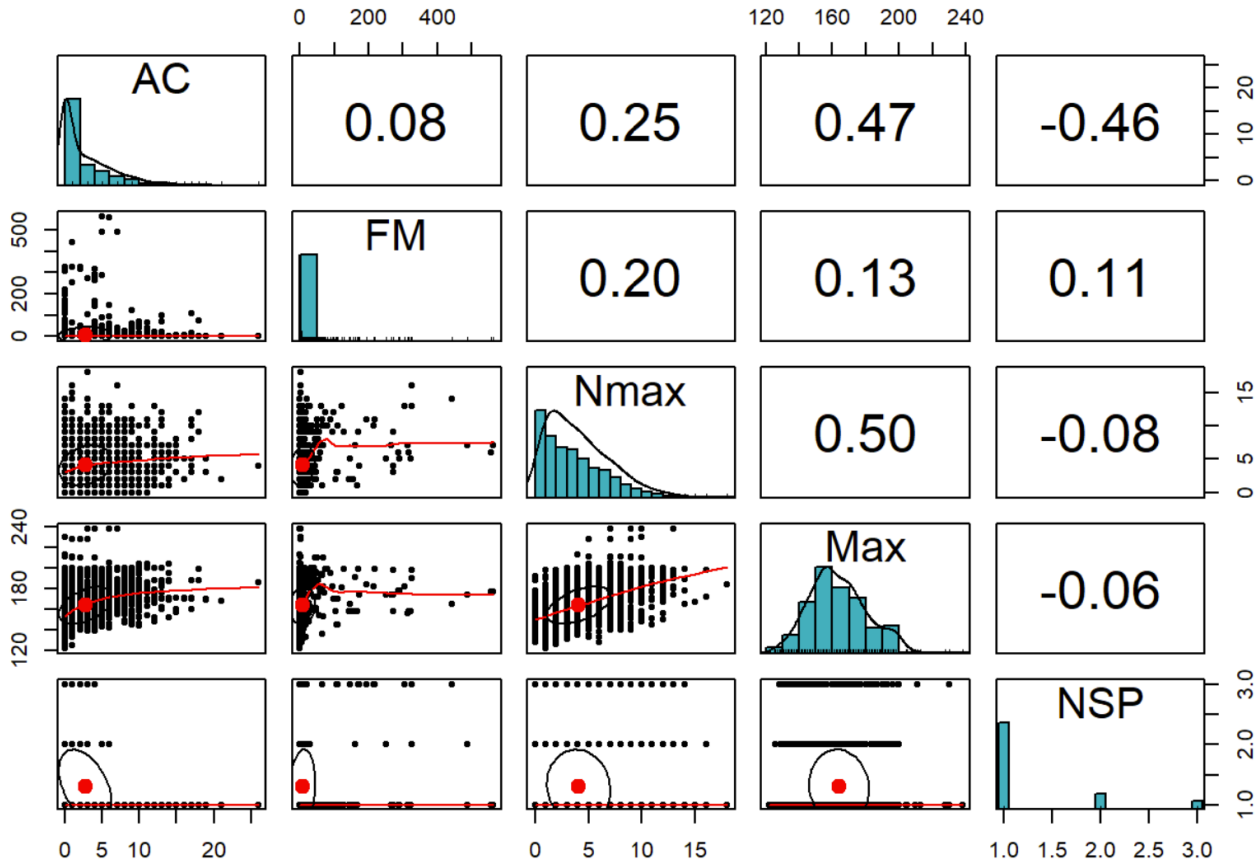


Figure 17: Accelerative characteristics grouping pairs panel.

In Figure 17 we see that the accelerative (healthy) indicator variables mostly correlated negatively with NSP. Except for fetal motion. Since the distribution of FM is very skewed it is hard to assess this variable.

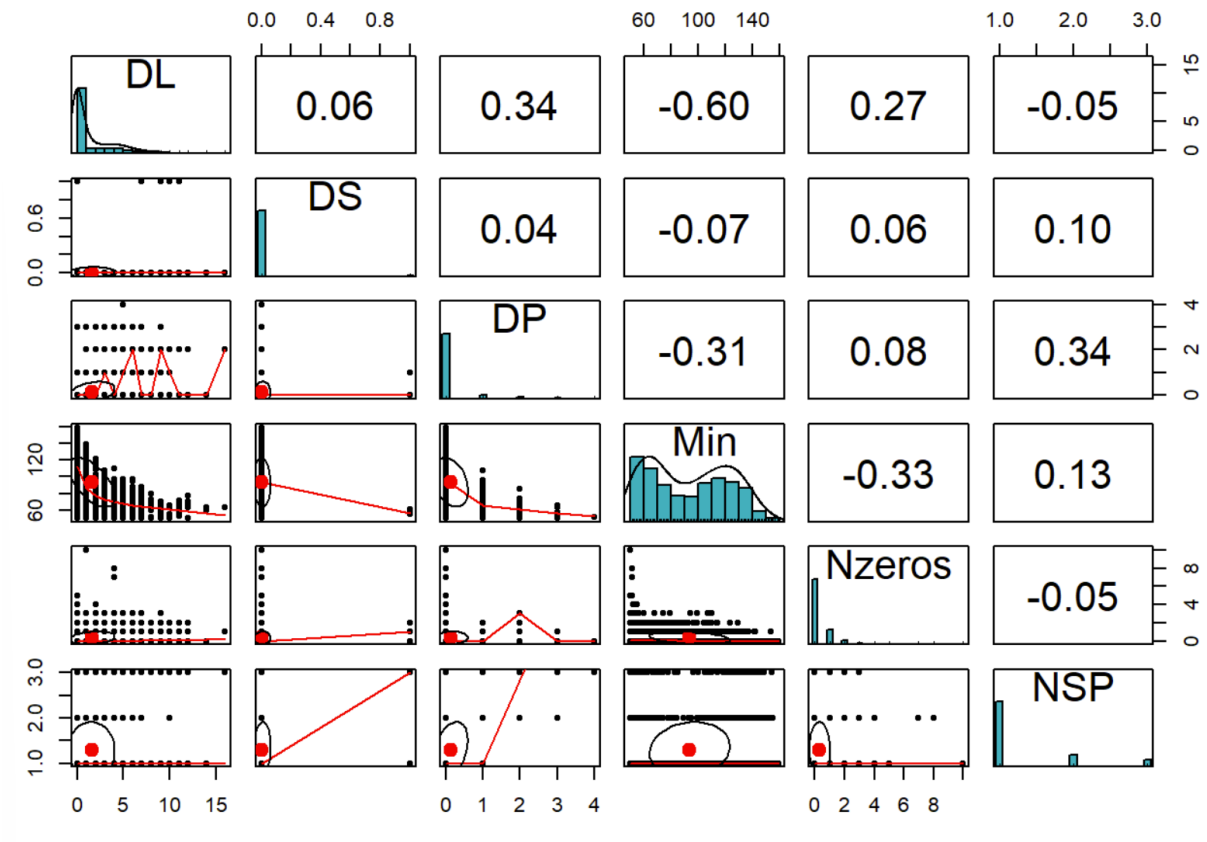


Figure 18: Decelerative grouping pairs panel.

The decelerative grouping pairs show that NSP increases as DS and DP increases. There was slightly negative correlation between DL and NSP. Light deceleration LD caused by vagal stimulation can be a normal response to uterine contraction.

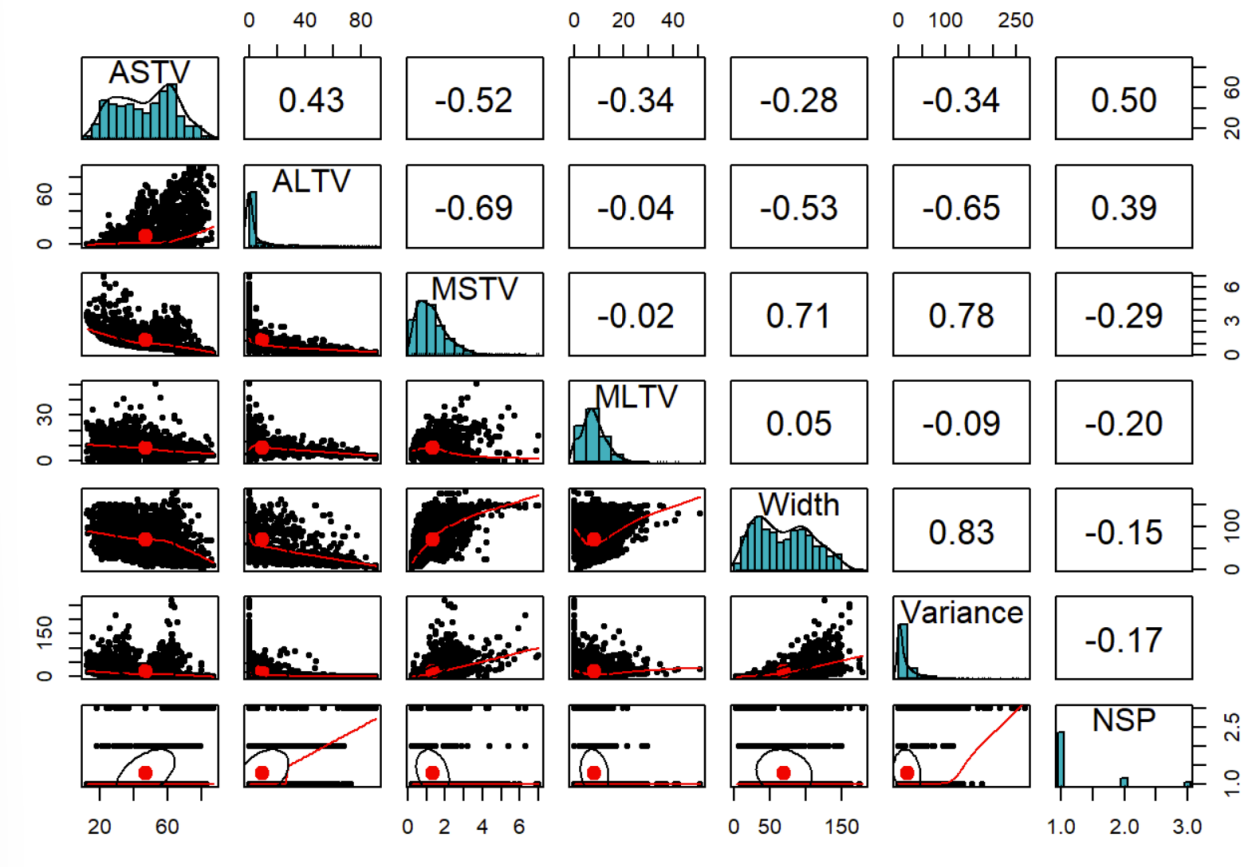


Figure 19: Variability group pairs panel.

Most of the variability features have some trends related to each other. NSP class 3 chances increase markedly with increase in Variance and ALTV.

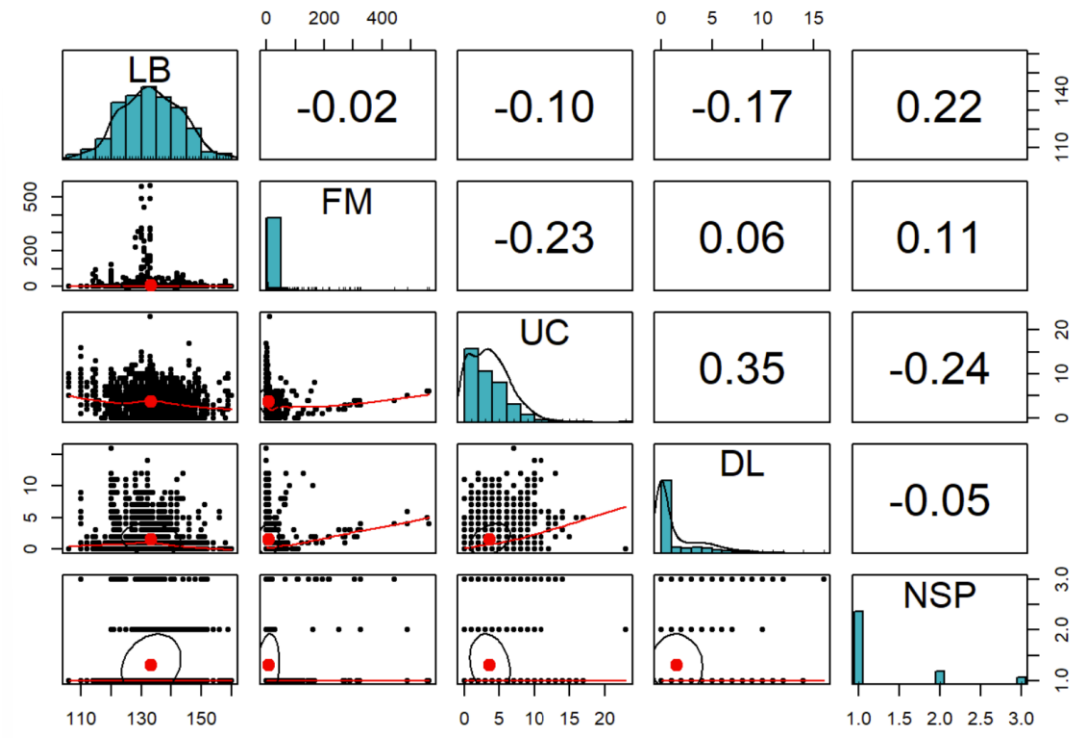


Figure 20: Uterine contraction variable with some related variables for comparison investigation.

Uterine contractions UC positively correlate with light decelerations LD and heart rate LB and fetal motion FM have a maxima at the normal physiological level of contraction as shown in Figure 20.

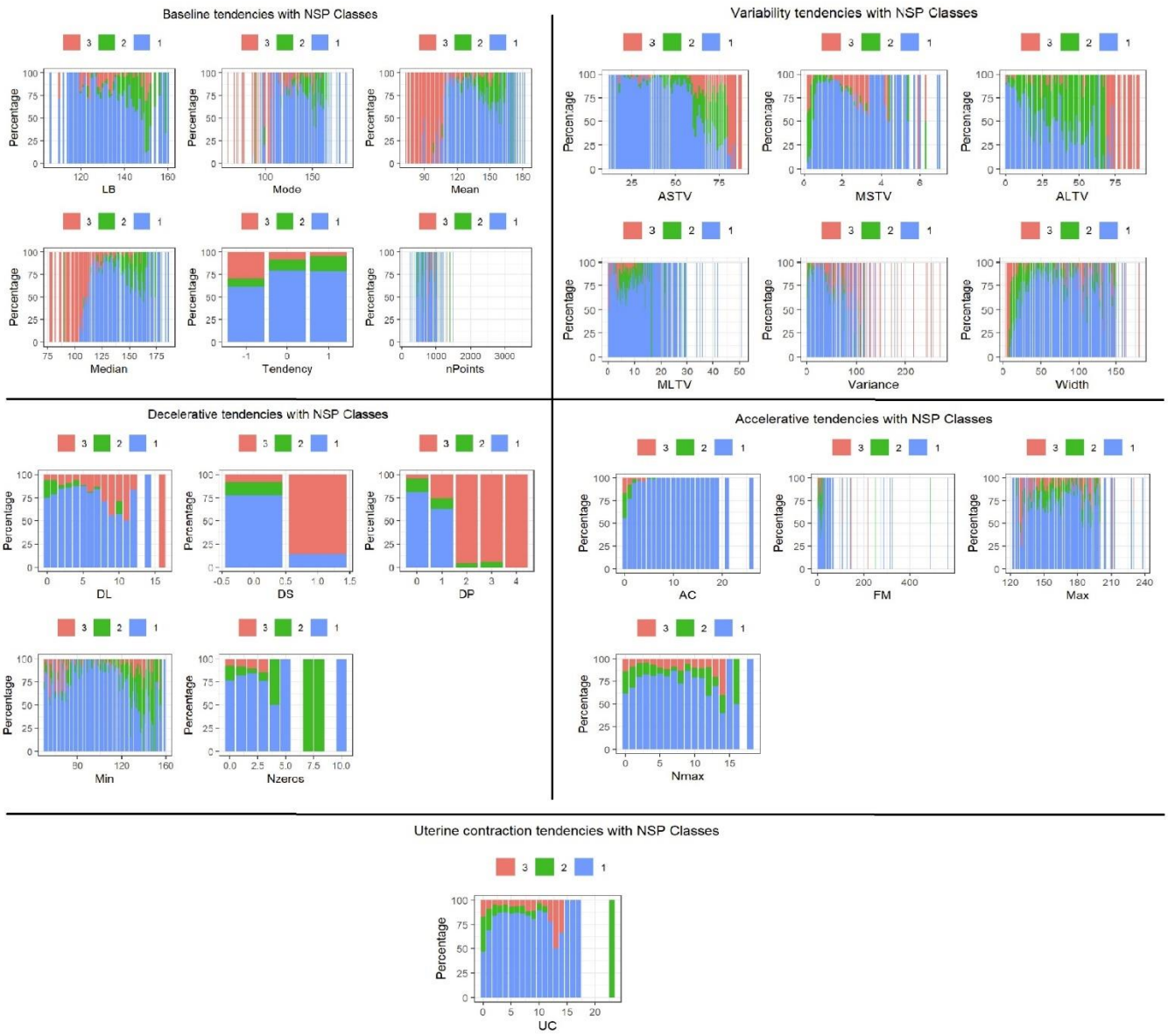


Figure 21: A percentage distribution of NSP class presented in a functionally grouped variable setting.

Figure 21 shows the distribution of the percentage occurrence of NSP in several grouped variables. In baseline and variability setting the pathological class prevails at the low end or after a high threshold. For decelerative group NSP class 3 tends to be more prevalent at the high end. Accelerative and uterine group also display more prevalence at the low end or after a high end

threshold. Some of these cut off ranges and their combinatory presence in related variables would be a great investigation into these features.

Finally, after all the feature investigation, no features were dropped from analysis.

Class Imbalance

Using Equation (1) the data class imbalance for the 3 groups were calculated. The results are reported in Table 6.

```
Class 1:  78 %
Class 2:  14 %
Class 3:  8 %
Imbalance:  0.89
```

Table 6: Percentage distribution of class instances and class imbalance in the data calculated using Equation (1).

The data set is highly imbalanced for NSP class 3 as demonstrated by an imbalance of 0.89 and the fact that only 8% of the instances in the total data represent class 3. Hence, class balancing was perhaps necessary to obtain a good performance. The dataset created in R was exported to Python to make use of the SMOTE-NC in Python imbearn package [36] class (SMOTE-NC was not available in R). SMOTE-NC is an oversampling technique where synthetic data is created from the from the actual data points. SMOTE-NC can handle both numeric and categorical data. A 30-70 stratified test train split produced the following (Figure 22) class distribution in the sets before any over sampling.

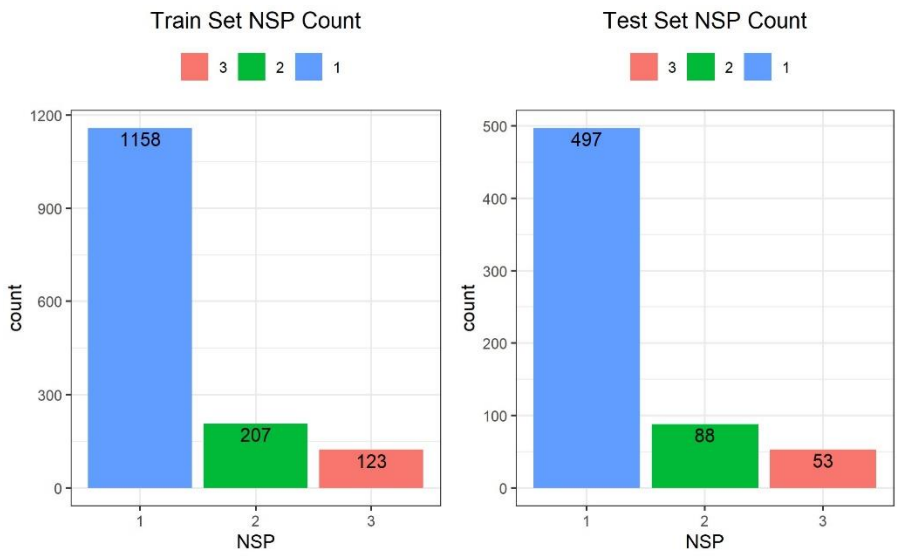


Figure 22: Train test split 30%. (Left) train set, (right) test set.

In addition to SMOTENC, two other standard sampling methods were added to the balancing techniques, they were random minority class up sampling, and random minority class up sampling combined with majority class down sampling. For various levels of desired imbalance, the number of samples created for each class were recorded.

Classification Model Selection

The models investigated were DT, RF classifiers and SVC classifier. Tree classifiers work on principles related to information gain based on splitting data points on features. These choices were made as an introductory learning aid for the author. DTs are easy to implement because they do not require special data preparation and can handle both numerical and categorical variables. They are more robust when no assumption about data distribution can be made. Disadvantages of DT include overfitting and may need pruning and many parameters tuning. RF classifier can

mitigate some of these problems by using ensemble of trees. Finally, SVC model with linear and radial bias kernels were chosen to classify the data.

For DT classifier the hyper parameters chosen were:

Max Depth = [4, 6, 12]

The parameter max depth can be thought of as the maximum allowable longest path from tree root to the leaf. This parameter is very important because a large allowable depth will mean that the tree will capture finer details of the data and as a result may overfit.

For RF classifier the hyper parameters chosen were:

N estimators = [10, 100, 200]

Max Depth = [4, 6, 12]

RF is an ensemble technique of DT, so the additional parameter used here is n estimators which is basically how may trees should be generated and averaged over to extract a low variance model (i.e., not overfit).

For SVC model the hyper parameters chosen were:

C = [0.001, 0.01]

Kernel = [linear, rbs]

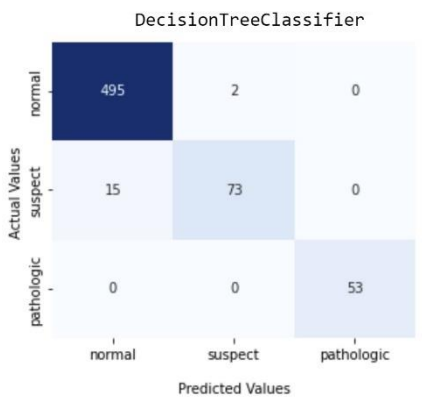
SVM classifier works by drawing a line (with some width) between boundaries by using the data points at the boundaries. The parameter C is the penalty for misclassification, a larger the C means a smaller margin width. A kernel uses the mathematical formula supplied to transform the data into higher dimensions, based on the property of the dataset one kernel might be suitable over the

other. Linear kernel works fine if the data is linearly separable, rbf kernel (radial bias function) may perform better when data is not linearly separable, because it works by creating a higher dimension (height like) based on radial distances between points.

Once the classifiers were selected a simple DT train test with the data was performed to get a baseline.

Examining the Expert Classified Morphological Features

The two origins of the variables in the dataset were the reported parameters from the CTG machines, and the expert analysed morphological classifications (A, B, C, D, E, AD, DE, LD, FS, SUSP, CLASS). Spearman's correlation and Kruskal-Wallis test both show strong influence of the LD, FS, SUSP, and CLASS morphological variables with NSP. From the data source description, it was unclear whether the expert decisions were made with blinded data or not (blinded to extra clinical information or neonatal outcomes). Hence a simple DT model was applied in Python using the data (with 70/30 test train split) to see the predictive value of the morphological variables on the NSP classification. It was initially surprising to see that without any class balancing or any kind of hyperparameter tuning the DT was able to detect with 100% accuracy the NSP class 3, as shown in Figure 23.



	precision	recall	f1-score	support
1	0.97	1.00	0.98	497
2	0.97	0.83	0.90	88
3	1.00	1.00	1.00	53
accuracy			0.97	638
macro avg	0.98	0.94	0.96	638
weighted avg	0.97	0.97	0.97	638

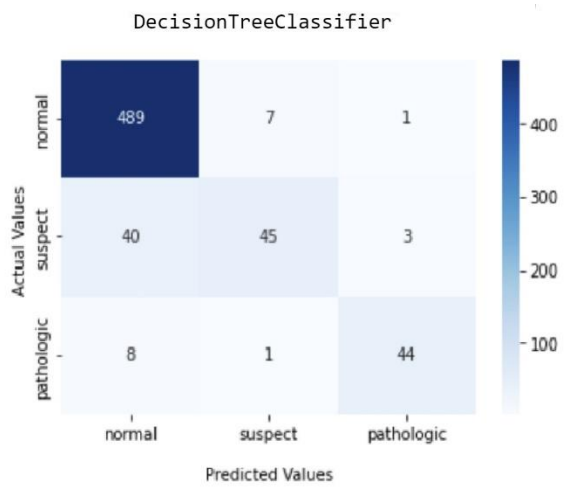
Figure 23: Decision tree predicting NSP class 3 with 100%% accuracy when morphological data is present in the data.

Further inspection revealed that only the LD, FS, SUSP patterns were sufficient to classify NSP class 3 with 100% accuracy. Since the NSP classification was also done by the same experts who made the morphological classifications, it is conceivable that only the morphological patterns can classify the NSP 3 category with 100% accuracy. While observer classification and intra observer variance is an interesting topic, it was deemed that the morphological data needs to be examined in other manners to improve predictive power feasibly. Therefore, at this stage all expert classified morphological variables were removed and focus was shifted to the machine reported parameters. This is an acceptable, and perhaps more practical, change because the purpose of the ML model/s is to predict fetal risk without any human interpretation of data.

CTG only Features: Baseline Predictive Model Results

After removal of the expert classified parameters the remaining 22 machine reported variables (LB, AC, FM, UC, DL, DS, DP, ASTV, MSTV, ALTV, MLTV, Width, Min, Max, Nmax, Nzeros, Mode, Mean, Median, Variance, Tendency, nPoints) were used to build a DT model to scope the baseline predictive results. Figure 24 shows that a DT model with the 22 variables produce an accuracy, recall, and f1 score of 0.91, 0.83 and 0.87. Taking these values as the baseline, further

data and model optimizations were carried out to improve the recall and f1 of NSP class 3 (pathological) classification.



	precision	recall	f1-score	support
1	0.91	0.98	0.95	497
2	0.85	0.51	0.64	88
3	0.92	0.83	0.87	53
accuracy			0.91	638
macro avg	0.89	0.78	0.82	638
weighted avg	0.90	0.91	0.90	638

Figure 24: Baseline predictive results using a decision tree model and CTG machine data only, without upsampling or hyperparameter tuning.

Training Validation and Hyper Parameter Tuning

After the 70/30 stratified test train split of the whole set, the training set was used for a 10-fold cross validation with stratified up sampling of the train set in each fold and no up sampling of the validation folds. Since identifying NSP class 3 correctly was the main objective, f1, recall and precision for class 3 were recorded alongside overall accuracy. The mean validation accuracy score for all folds was the metric used to optimize hyper parameter combinations for each classifier. Figure 25 shows the data split, cross validation, up sampling of training folds and test strategy steps.

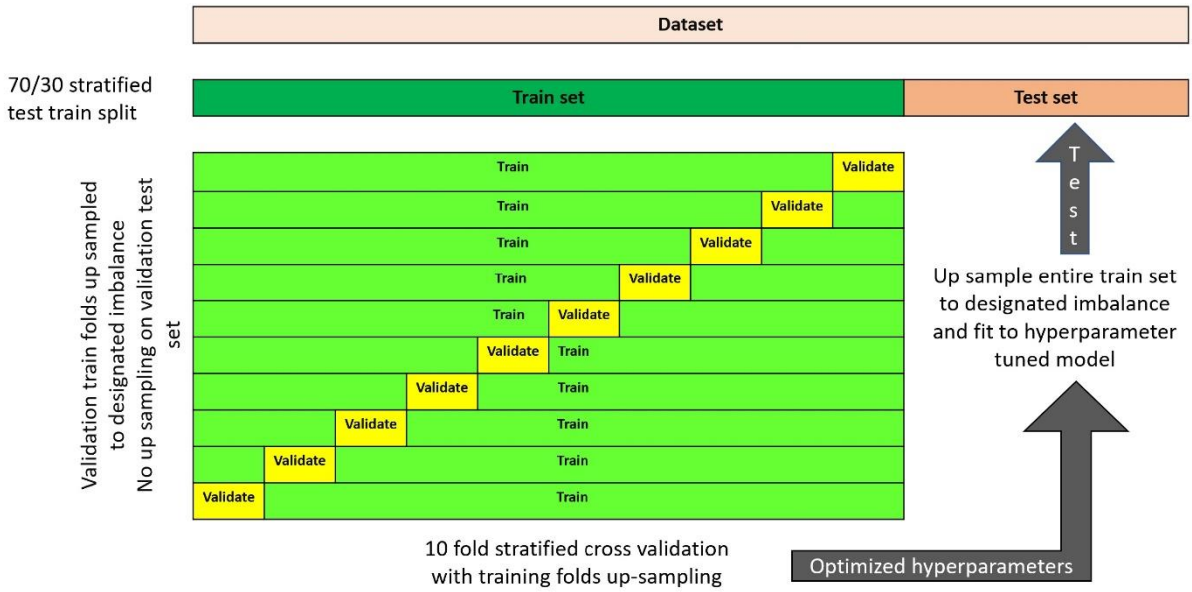


Figure 25: Data processing diagram for model fitting, all runs were done on the same fold of the data.

The function for upsampling and cross validation was modified form [37]. The following modifications were made to the code: change random seed integers to keep the model and data random states consistent for each run, stratified cross validation with 10 folds, upsample folds using SMOTE-NC, upsampling, upsampling and down sampling combo and optimizing model by mean accuracy score and reporting values of validation f1, precision and recall score for class 3.

Once the optimized model (hyper parameter tuned) was obtained. The whole train data set was upsampled to the desired imbalance and used to train the model.

In order to determine how much over sampling to create, Equation (1) was reversed to calculate the number of class 3 items (n_3) needed to meet an investigative *imbalance* value [39]. This imbalance calculation was done in Python (Appendix A). For ease of calculation the number of class 2 and class 3 items needed were assigned to be the same (i.e. $n_2 = n_3$). Several values of

imbalance [0.89, 0.69, 0.49, 0.29, 0.09] were examined to compare and choose the best one that produces the scores when the trained model is applied on the test set.

Figure 26 below shows the upsampled test set NSP class level distribution to achieve an imbalance of 0.39.

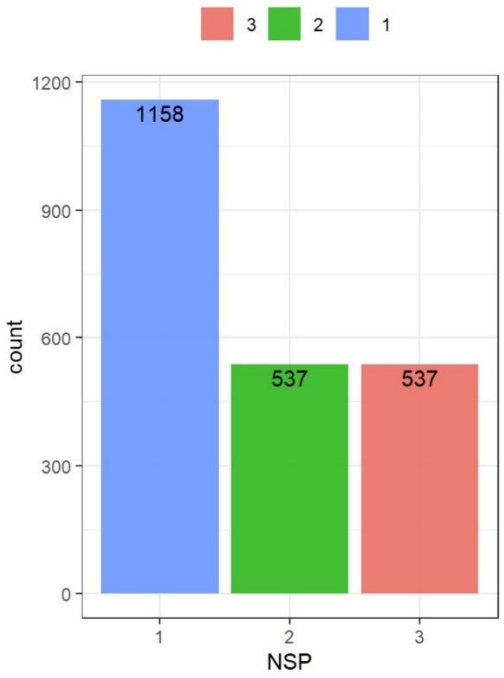


Figure 26: Train set after SMOTE-NC NSP class 2 and 3 up sampling based on imbalance of 0.39.

Predictive Model Results and Comparison

In total 96 models were created and tested. The models were formed as follows: (2 normalization state x 3 classifiers x 5 imbalances x 3 sampling techniques) + (2 normalization states x 3 classifiers with no tuning, or data sampling).

Figure 27 shows that the test accuracy of the DT model is the highest. DT with SMOTENC sampling also had the least overfitting, as demonstrated by smaller gap between test (dashed) and

validation accuracies (solid). Effect of normalizing data was also least pronounced for DT and RF. Test and validation accuracy remained similar or decreased slightly for DT and RF was data was more balanced. The impact of data normalization (red vs. blue) was most pronounced for SVM classifier.

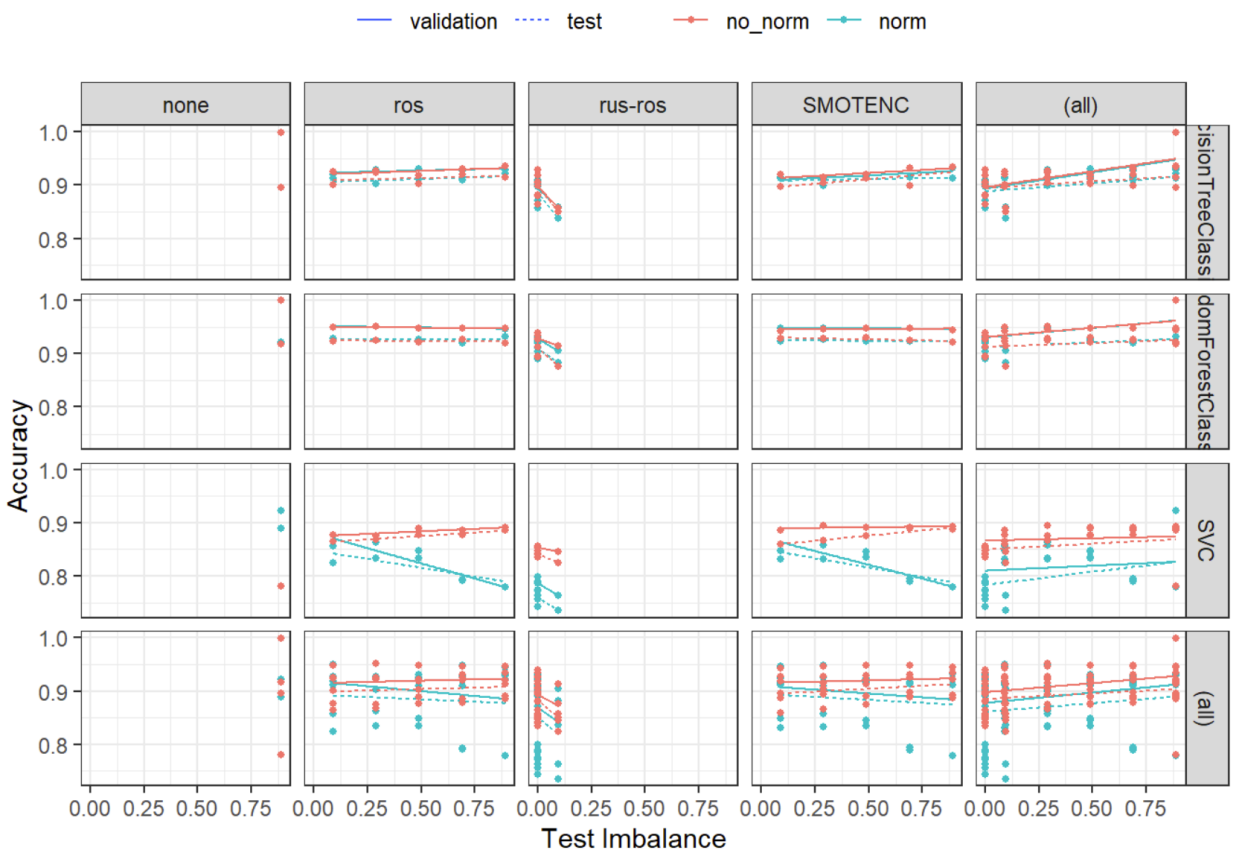


Figure 27: Test accuracy of each model as a function of several parameters.

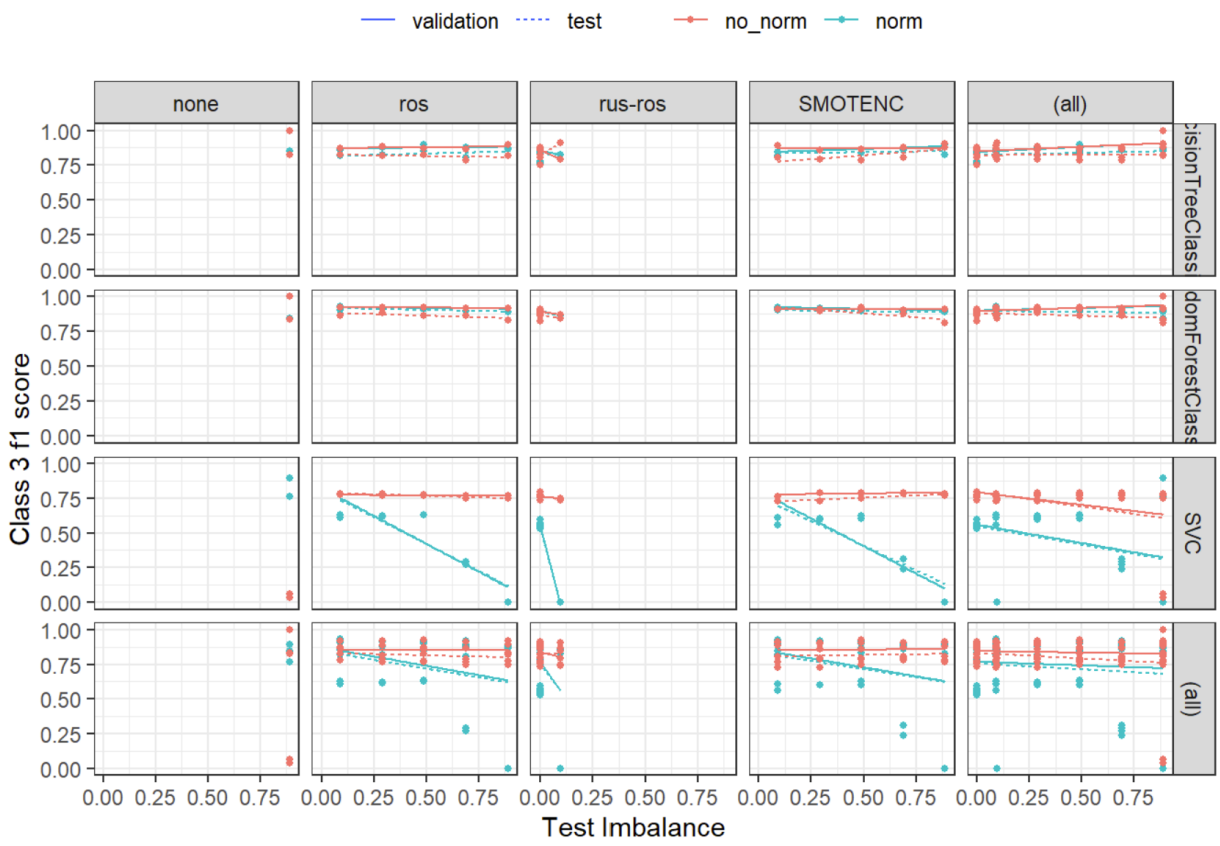


Figure 28: NSP class 3 f1 score as a function of several parameters.

Figure 28 shows that the class 3 NSP f1 score obtained using DT and RF remained invariant under normalization, class balancing and sampling technique. Random over sampling for DT has some overfitting. SVC performed the worst and was very dependent on the normalization state and imbalance state.

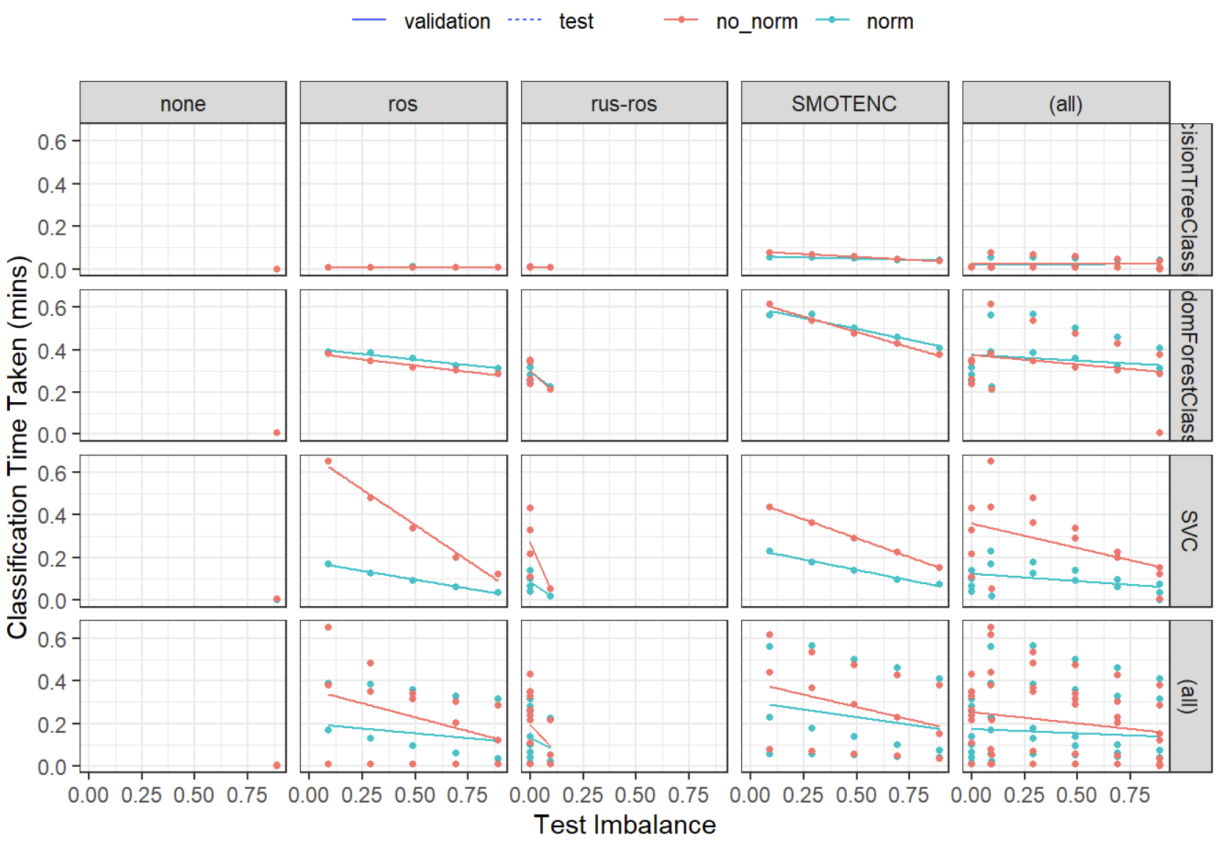


Figure 29: Time taken to validate train and test each model as a function of several parameters.

Figure 29 shows that SVC took the longest to train and test. And it took longer to train the no normalisation data. While RF time increased with more upsampling (smaller imbalance), the effect was least pronounced for DT and random over sampling. SVC classification time also had the highest variability.

Figure 30 shows that the DT classifier with a max depth of 12 has the highest test accuracy of 0.934 with no normalization of data and with 0.89 imbalance maintained using SMOTENC over sampling.

```
hyper_param_tuning      {'random_state': 42, 'max_depth': 12, 'criteri...
validation precision                                            0.887981
validation recall                                               0.885897
validation f1                                                    0.88051
validation accuracy                                             0.933466
test precision                                                  0.890909
test recall                                                     0.924528
test f1                                                         0.907407
test accuracy                                                   0.934169
sampling                                                         SMOTENC
sampling_strategy                          [{1: 1158, 2: 207, 3: 165}]
test imbalance                                                  0.889785
model name                                        DecisionTreeClassifier
y_pred                  [3, 1, 1, 1, 1, 1, 1, 1, 2, 1, 3, 2, 1, 1, 1, ...
normalization                                                    no_norm
time                                                            2.126208
```

Figure 30: Best test accuracy model DT (max depth 12) with 0.89 imbalance, SMOTENC over sampling, and test accuracy of 0.934.

Table 7 shows classifier the score improvements from baseline were, test accuracy +0.039, test target class recall +0.113 and test target class f1 score +0.080.

model name	normalization	sampling	sampling_strategy	validation accuracy	test accuracy	test f1	test recall	test precision	time
DecisionTreeClassifier	no_norm	none	[{1: 1158, 2: 207, 3: 123}]	0.999328	0.894984	0.826923	0.811321	0.843137	0.034532
DecisionTreeClassifier	no_norm	SMOTENC	[{1: 1158, 2: 207, 3: 165}]	0.933466	0.934169	0.907407	0.924528	0.890909	2.126208
RandomForestClassifier	norm	ros	[{1: 1158, 2: 207, 3: 165}]	0.945551	0.931034	0.886792	0.886792	0.886792	18.812940
SVC	no_norm	SMOTENC	[{1: 1158, 2: 314, 3: 314}]	0.891116	0.888715	0.780952	0.773585	0.788462	13.584096
SVC	norm	none	[{1: 1158, 2: 207, 3: 123}]	0.922043	0.888715	0.764045	0.641509	0.944444	0.129701

Table 7: Comparisons of scores of all the models with highest intergroup test accuracy. In red the model with best test accuracy.

Comparing class 3 f1 values, decision tree had the best value (0.907) over RF (0.887) and SVC (0.781).

Model Comparison Statistical Analysis

Since the tuned model was only tested once on the train set the scores of the test were not sufficient to do any high-powered statistical tests. A simple McNamar's test was done to see whether the models and the difference in their prediction results were significant. This test captures the errors made by both models specifically the relative difference in the proportion of error between the models. McNamar's test was chosen for its simplicity of application and no assumption of any data distribution. Even though it has low probability of type I error it also has low power. The result of the pairwise test between the models was insignificant (P=0.978). Which means that the best accuracy may have been obtained from DT and not RF or SVC by chance.

Discussion

The undertaking was to improve detection of class 3 NSP condition of fetal risk, which is pathological and requires action. However, in clinical practice there are many false positives and that leads to a high cost and health implications of operative delivery. This undertaking was to see whether there can be improvements made to decreasing the numbers of FP while not impacting the TP rate. The one combined parameter used to make that assessment was the f1 statistic of NSP class 3. A baseline value for this statistic was found to be f1 = 0.827 using a DT model and no data resampling. Then further techniques were applied to improve upon this value.

Models were created and hyperparameters were tuned and model validated using a 10-fold cross validation technique, on a train set that was created using a 70/30 test train data split. The hyper parameters investigated in a grid search were max_depth = [4, 6, 12] for DT and RF, n_estimators = [10, 100, 200] for RF and C = [001, 0.01], kernel = [linear, rbf] for SVM. The hyperparameter combination that produced the highest mean cross validation accuracy was chosen to be the best

tuned model. Finally, the models were evaluated using data that was min max normalized and not min max normalized. Upon comparing the outputs of all the modeling sessions, the best test scores obtained were, test accuracy = 0.934, (validation accuracy = 0.933), target class test f1 score = 0.907, target class test precision = 0.891, and target class recall = 0.925, using a DT classifier with data imbalance of 0.890, which was sampled using the SMOTENC technique on data with no min max normalization. When compared to a baseline DT with no up sampling and hyperparameter tuning, the improvement in scoring was as follows, test accuracy +0.039, test target class recall +0.113 and test target class f1 score +0.080. The model was trained in 2 minutes. However, there was no statistically significant difference found in the maximum test accuracies of the DT, RF, SVM model when compared pairwise using McNamar's test (P=0.978), or all together using Kruskal-Wallis test (p=1). Both DT and RF models were robust in performing well under high class imbalance, but SVM accuracy (for both test and validation) improved as the class imbalance was reduced. In terms of resampling, a minimal resampling using SMOTENC performed the best and the random minority sampling combined with majority down sampling performed the worst. The highest accuracy reported, using this dataset in the literature was found to be 0.990 using bagging and RF model [11] [12] in WEKA [13] which is much higher than the reported values in this study. This study was undertaken under manual coding with some manually chosen hyperparameters as the starting point of an investigation, and hence while they provided some insight, the models presented here, still have room for further development.

On all aspects, performance scores, low training time, and robustness under change of imbalance and normalization, DT and RF were first and close second performers as classifiers. SVM model was the least stable, accurate and required the most time to train.

The validation sets were trained on a 10-fold cross validation of the train set and was used to tune the hyperparameters of the models. The choice of the hyper parameters was not substantiated by any analysis other than some average and extreme values were used.

The contribution of this work is more towards the authors understanding of basic ML models and data handling techniques. The outcome of this project certainly did not improve upon the best reported results found in the literature.

Conclusion

The analysis presented in this work improved the classification outcome of NSP class 3 pathological state using a DT model which was hyperparameter tuned using a 10-fold cross validation, and upsampled to an imbalance of 0.89 using SMOTENC. The tuned hyperparameter for DT was max depth of 12, and the time to train was 2 minutes. The validation and test accuracies were 0.933 and 0.934 respectively. The target statistic of this work, the f1 score for class 3 improved from 0.827 at baseline, to 0.907 after hyperparameter tuning. While the preliminary results show an improvement, further work needs to be done to promote it to a statistically significant result that is close to the high values reported in the literature.

Future Work

Some simple improvements to this work can be employment of a random search instead of a grid search of parameters, to hyper tune to a better model. Employing an outer validation loop that performs test on the test split more than once will also enable a better confidence establishment around the model test scores. There are several other ML models that can also be employed to test this data. While the previous recommendations revolve around the ML techniques. Another approach would be to investigate the data more deeply. Exploring causal relationship and threshold

for the features and creating latent variables should be also investigated. The morphological classifications while not useful in predicting this dataset outcomes directly can be used in a reverse engineering way where some of the variables highly correlated with NSP can be investigated to uncover patterns in the machine data that connect to them. This kind of human machine hybrid training technique is gaining popularity currently. Lastly, a more detailed dataset is publicly available [39]. This set includes the actual waveform recordings and can be explored further to extract whole new parameters that are not conceivable using the current dataset.

References

[1] A. M. Ponsiglione, C. Cosentino, G. Cesarelli, F. Amato, and M. Romano, "A comprehensive review of techniques for processing and analyzing fetal heart rate signals," *Sensors*, vol. 21, no. 18, p. 6136, 2021.

[2] "S1-Guideline on the Use of CTG During Pregnancy and Labor," *Geburtshilfe Frauenheilkd*, vol. 74, no. 8, pp. 721–732, Aug. 2014, doi: 10.1055/s-0034-1382874.

[3] J. J. Arnold and B. L. Gawrys, "Intrapartum Fetal Monitoring," *AFP*, vol. 102, no. 3, pp. 158–167, Aug. 2020.

[4] D. Ayres-de-Campos, C. Y. Spong, E. Chandraharan, and FIGO Intrapartum Fetal Monitoring Expert Consensus Panel, "FIGO consensus guidelines on intrapartum fetal monitoring: Cardiotocography," *Int J Gynaecol Obstet*, vol. 131, no. 1, pp. 13–24, Oct. 2015, doi: 10.1016/j.ijgo.2015.06.020.

[5] Z. Hoodbhoy, M. Noman, A. Shafique, A. Nasim, D. Chowdhury, and B. Hasan, "Use of Machine Learning Algorithms for Prediction of Fetal Risk using Cardiotocographic Data," *Int J Appl Basic Med Res*, vol. 9, no. 4, pp. 226–230, 2019, doi: 10.4103/ijabmr.IJABMR_370_18.

[6] C. Antoine and B. K. Young, "Cesarean section one hundred years 1920–2020: the Good, the Bad and the Ugly," *Journal of Perinatal Medicine*, vol. 49, no. 1, pp. 5–16, Jan. 2021, doi: 10.1515/jpm-2020-0305.

[7] D. Ayres-de-Campos, J. Bernardes, A. Garrido, J. Marques-de-Sa, and L. Pereira-Leite, "SisPorto 2.0: a program for automated analysis of cardiotocograms," *Journal of Maternal-Fetal Medicine*, vol. 9, no. 5, pp. 311–318, 2000.

[8] "1.10. Decision Trees," *scikit-learn*. https://scikit-learn/stable/modules/tree.html (accessed Mar. 28, 2022).

[9] "sklearn.ensemble.RandomForestClassifier," *scikit-learn*. https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (accessed Feb. 19, 2022).

[10] "1.4. Support Vector Machines," *scikit-learn*. https://scikit-learn/stable/modules/svm.html (accessed Feb. 19, 2022).

[11] H. Sahin and A. Subasi, "Classification of the cardiotocogram data for anticipation of fetal risks using machine learning techniques," *Applied Soft Computing*, vol. 33, pp. 231–238, Aug. 2015, doi: 10.1016/j.asoc.2015.04.038.

[12] L. Huang *et al.*, "Investigating the interpretability of fetal status assessment using antepartum cardiotocographic records," *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, p. 355, Dec. 2021, doi: 10.1186/s12911-021-01714-4.

[13] "Citing Weka - Weka Wiki." https://waikato.github.io/weka-wiki/citing_weka/ (accessed Apr. 04, 2022).

[14] D. Ayres-de-Campos, "Electronic fetal monitoring or cardiotocography, 50 years later: what's in a name?," *American Journal of Obstetrics & Gynecology*, vol. 218, no. 6, pp. 545–546, Jun. 2018, doi: 10.1016/j.ajog.2018.03.011.

[15] N. Lisenbee and J. A. Tyndall, "Fetal Heart Rate Monitoring," in *Atlas of Emergency Medicine Procedures*, L. Ganti, Ed. New York, NY: Springer, 2016, pp. 639–642. doi: 10.1007/978-1-4939-2507-0_109.

[16] D. Ayres-de-Campos, M. Rei, I. Nunes, P. Sousa, and J. Bernardes, "SisPorto 4.0 - computer analysis following the 2015 FIGO Guidelines for intrapartum fetal monitoring," *J Matern Fetal Neonatal Med*, vol. 30, no. 1, pp. 62–67, Jan. 2017, doi: 10.3109/14767058.2016.1161750.

[17] L. S. Cahill *et al.*, "Determination of fetal heart rate short-term variation from umbilical artery Doppler waveforms," *Ultrasound in Obstetrics & Gynecology*, vol. 57, no. 1, pp. 70–74, 2021, doi: 10.1002/uog.23145.

[18] "Recommendations | Intrapartum care for healthy women and babies | Guidance | NICE." https://www.nice.org.uk/guidance/cg190/chapter/Recommendations#initial-assessment (accessed Feb. 19, 2022).

[19] M. I. Evans, D. W. Britt, S. M. Evans, and L. D. Devoe, "Changing Perspectives of Electronic Fetal Monitoring," *Reproductive Sciences*, p. 1, doi: 10.1007/s43032-021-00749-2.

[20] M. E. O'Sullivan, E. C. Considine, M. O'Riordan, W. P. Marnane, J. M. Rennie, and G. B. Boylan, "Challenges of Developing Robust AI for Intrapartum Fetal Heart Rate Monitoring," *Front Artif Intell*, vol. 4, p. 765210, Oct. 2021, doi: 10.3389/frai.2021.765210.

[21] INFANT Collaborative Group, "Computerised interpretation of fetal heart rate during labour (INFANT): a randomised controlled trial," *Lancet*, vol. 389, no. 10080, pp. 1719–1729, Apr. 2017, doi: 10.1016/S0140-6736(17)30568-8.

[22] A. Georgieva *et al.*, "Computer-based intrapartum fetal monitoring and beyond: A review of the 2nd Workshop on Signal Processing and Monitoring in Labor (October 2017, Oxford, UK)," *Acta Obstetricia et Gynecologica Scandinavica*, vol. 98, no. 9, pp. 1207–1217, 2019, doi: 10.1111/aogs.13639.

[23] Z. Zhao, Y. Deng, Y. Zhang, Y. Zhang, X. Zhang, and L. Shao, "DeepFHR: intelligent prediction of fetal Acidemia using fetal heart rate signals based on convolutional neural network," *BMC Med Inform Decis Mak*, vol. 19, no. 1, p. 286, Dec. 2019, doi: 10.1186/s12911-019-1007-5.

[24] V. Chudáček, J. Spilka, P. Janků, M. Koucký, L. Lhotská, and M. Huptych, "Automatic evaluation of intrapartum fetal heart rate recordings: a comprehensive analysis of useful features," *Physiol. Meas.*, vol. 32, no. 8, pp. 1347–1360, Jul. 2011, doi: 10.1088/0967-3334/32/8/022.

[25] J. Balayla and G. Shrem, "Use of artificial intelligence (AI) in the interpretation of intrapartum fetal heart rate (FHR) tracings: a systematic review and meta-analysis," *Arch Gynecol Obstet*, vol. 300, no. 1, pp. 7–14, Jul. 2019, doi: 10.1007/s00404-019-05151-7.

[26] M. G. da Silva Neto, J. P. do Vale Madeiro, and D. G. Gomes, "On designing a biosignal-based fetal state assessment system: A systematic mapping study," *Computer Methods and Programs in Biomedicine*, vol. 216, p. 106671, Apr. 2022, doi: 10.1016/j.cmpb.2022.106671.

[27] "UCI Machine Learning Repository." https://archive-beta.ics.uci.edu/ml/datasets/cardiotocography (accessed Jan. 30, 2022).

[28] D. Jagannathan and M. Phil, "Cardiotocography-a comparative study between support vector machine and decision tree algorithms," *International Journal of Trend in Research and Development*, vol. 4, no. 1, 2017.

[29] M. T. Rayhan, A. S. Arefin, and S. A. Chowdhury, "Automatic detection of fetal health status from cardiotocography data using machine learning algorithms," *Journal of Bangladesh Academy of Sciences*, vol. 45, no. 2, Art. no. 2, 2021, doi: 10.3329/jbas.v45i2.57206.

[30] D. Bhatnagar and P. Maheshwari, "Classification of Cardiotocography Data with WEKA," *International Journal of Computer Science and Network - IJCSN*, Apr. 30, 2016. http://eprints.rclis.org/29886/ (accessed Feb. 19, 2022).

[31] "Technology Central Fetal Monitoring." http://www.omniview.eu/ing/technology/technology (accessed Feb. 19, 2022).

[32] “Fetal Monitoring,” *elearning for healthcare*. https://www.e-lfh.org.uk/programmes/electronic-fetal-monitoring/ (accessed Feb. 19, 2022).

[33] E. de Jonge and M. van der Loo, *An Introduction to Data Cleaning with R*. Statistics Netherlands, 2013.

[34] E. Collins, N. Rozanov, and B. Zhang, “Evolutionary data measures: Understanding the difficulty of text classification tasks,” *arXiv preprint arXiv:1811.01910*, 2018.

[35] “RStudio | Open source & professional software for data science teams.” https://www.rstudio.com/ (accessed Feb. 18, 2022).

[36] “SMOTENC — Version 0.10.0.dev0.” https://imbalanced-learn.org/dev/references/generated/imblearn.over_sampling.SMOTENC.html (accessed Feb. 18, 2022).

[37] 262588213843476, “Example of cross-validation with unbalanced data,” *Gist*. https://gist.github.com/kiwidamien/bcbe8e527a5f0cc9f28c4fe692f70cbc (accessed Mar. 08, 2022).

[38] J. Wang, A. Bhowmick, M. Cevik, and A. Basar, “Deep learning approaches to classify the relevance and sentiment of news articles to the economy,” in *Proceedings of the 30th Annual International Conference on Computer Science and Software Engineering*, 2020, pp. 207–216.

[39] V. Chudáček *et al.*, “Open access intrapartum CTG database,” *BMC Pregnancy and Childbirth*, vol. 14, no. 1, p. 16, Jan. 2014, doi: 10.1186/1471-2393-14-16.

Appendix A

Code GitHub Repository link:

https://github.com/fayruzkibria/Cardiotocography-Fetal-Risk-Classification