Predicting Birth Outcome Using Cardiotocography and Machine Learning

Josefine Öder

Master's thesis
2022:E13

Lund University

Faculty of Engineering
Centre for Mathematical Sciences
Mathematical Statistics

Centrum Scientiarum Mathematicarum

A Master's thesis

Predicting Birth Outcome Using Cardiotocography and Machine Learning

Josefine Öder
MSc Engineering Mathematics

Supervisors:
Andreas Jakobsson, LTH
Karl Åström, LTH
Ida Arvidsson, LTH
Karel Marsal, MED
Karin Källén, MED

Autumn/Spring, 2021/2022

LUND
UNIVERSITY

Abstract

Cardiotography, CTG, is a monitoring method that is commonly used during childbirth. The method measures the fetal heart rate, FHR, alongside the uterine contractions, TOCO. Clinicians use this tool to evaluate the health of the infant, and to observe changes which might imply hypoxia, lack of oxygen supply, for the fetus. However, abnormalities in the CTG are often non-specific, making the interpretation difficult.

This master's thesis aims to extract features from the CTG and to use these features in different types of machine learning classifiers to predict the child's health after birth. The target is to make the interpretation of the CTG easier for clinicians and to find patterns that could imply hypoxia.

By using the gestational age as the first feature, it was decided to split the data into two cases, preterm pregnancies and full term including postterm pregnancies. For each case, the gestational age was tested as a feature in different classifiers, which provided a benchmark for comparison when testing new features. During the investigation of the FHR signal, a total of eight features were tested. The extracted features were tested with the gestational age separately, followed by a test using all derived features, and lastly a test using all features but the gestational age was made. Some features increased the classifiers' sensitivity, or specificity, but none made the predictions significantly more accurate.

The conclusions of this thesis were that the gestational age should only have been used for stratifying the data into two separate cases, since it for the preterm pregnancies carried much information, therefore derived features did not contribute with further information. For the other case, the gestational age did not carry any information which showed in the results of the last two tests when using all features together. The extracted features should have been tested in different combinations of each other to see if they contributed with correlating information or not. Additional research is required to create an accurate prediction model, e.g., by investigating how the FHR signal correlates with the TOCO signal.

Acknowledgements

Firstly, I would like to thank Andreas Jakobsson, Karl Åström, and Ida Arvidsson, my supervisors at the Faculty of Engineering, for all the interesting discussions, ideas, and help throughout this thesis. Your input and knowledge has been a vital asset to me, and you all have taught me so much.

Secondly, I would like to thank Karin Källén and Karel Marsal, my supervisors at the Faculty of Medicine, for your important input within the obstetrics's field, and for correcting my choice of words to make it medically correct. Your work is truly inspiring.

Thirdly, to all my friends. For the encouragement, for the long study nights, for the joy, and for all the memories. Special thanks to Emma, Erik, Hedvig, and Jessica - I appreciate you so much.

Lastly, I would like to thank Ann Kull, for your guidance, encouragement, and never-ending support.

Contents

Abbreviations

ANN	Artificial Neural Network
bpm	Beats per Minute
CTG	Cardiotocography
DT	Decision Tree
FHR	Fetal Heart Rate
FIGO	The International Federation of Gynecology and Obstetrics
KNN	K Nearest Neighbours
SVM	Support Vector Machine
TOCO	Uterine Contractions

List of Terms

Antepartum	Before Labour
Asphyxia	Physiological Results of Hypoxia
Gestational Age	Measure of the Age of a Pregnancy
Hypoxia	Lack of Oxygen
Intrapartum	During Labour
Postpartum	After Labour

1 Introduction

In Sweden, during the year of 2020, a total of 353 babies were stillborn, and 268 died during their first living year, which were 2.37 per 1000 living births [1, 2]. For comparison, in the beginning of the 1900s approximately 10 percent died within their first living year, and in the middle of the 1960s it was 1 percent [2]. Labour includes different risks, for the mother as well as the child. To reduce these risks, cardiotocography was introduced in the 1960s as a monitoring method [3, 4]. Even if the number of deaths has decreased throughout the decades, it is still high in the year of 2020. This raises the question of what can be done further to reduce the number of deaths. Is there a possibility to strengthen and expand the use of cardiotocography?

Cardiotocography, CTG, is a biophysical method that combines the recording of the fetus heartbeat, cardio, and the uterine contractions, TOCO. Assessment is made by examining the FHR baseline and variability, along with the contractions [3]. Through CTG, clinicians evaluate the state of the fetus and if the fetus is negatively affected [3, 4].

A risk, during childbirth, is mechanic stress which stems from the uterine contractions and affects the brain and/or umbilical cord. This can lead to hypoxia. Another risk is metabolic stress which is when the gas exchange between the mother and fetus is reduced. This is due to the circulation in the placenta being decreased or stopped. The fetus has during the pregnancy developed resilience and would normally manage these stresses during birth. If the resilience is compromised or the stresses are unusually high, the ability to handle the stress will not be enough, this is known as distress. Distress affects the vital functions of the fetus and can lead to brain injuries - or in worst case, death [3, 4].

Deviations and abnormalities in the CTG can be non-specific, they can occur even when there is no present danger to the fetus, making the interpretation difficult. These difficulties have been designated as a reason for the increasing amount of acute cesarean sections in the 1970s [3]. The FIGO guidelines are used when interpreting the CTG. There is however some concern regarding these guidelines, Spilka et al. showed in a study that reaching an agreement on CTG evaluations is hard, even when several clinicians evaluated the CTG, manifesting the complication, and difficulty of interpreting the CTG [5].

1.1 Aim and Research Questions

The aim with this master thesis, which stems from the difficulties interpreting the CTG, is to investigate *if* there is any information in the CTG which indicates that the fetus is not doing well, for example, suffering from asphyxia. If such information can be found, this could be used to aid professionals in the interpretation of CTG, and reduce the uncertainty. The idea is try to find features that distinguishes a birth with a good outcome from a bad outcome, and compare different machine learning techniques to see if one technique can extract information from the features better than others.

The outcome that will be used is the Apgar Score measured at five minutes after birth. The Apgar Score is a metric between 0 and 10, where 10 is an indication that the child is doing well. An Apgar Score below 7 will be considered, and referenced, as a bad outcome, whereas an Apgar Score between

7 to 10 will be considered, and referenced, as a good outcome [6]. The definition of Apgar Score will be further explained in section 2.1.2.

To reach said aim, the following research questions are going to be examined:

- Are there patterns in the CTG that distinguish a good outcome from a bad outcome?
- How do found features affect the results of predictions, when compared to a naive model?
- Which machine learning technique performs the best?

2 Background

This chapter presents background information used as a foundation in this project. Starting with information about obstetrics, specifically cardiotocography and Apgar Score. Then follows a brief explanation of artificial neural network and Bayes' theorem. The last part is the related work section which aims to acquaint the reader with previous work that has been made, and some articles that are referenced to in this report.

2.1 Obstetrics

In this section, the aim is the introduce the vital information about pregnancy, labour and delivery. A pregnancy should last between 37 to 41 weeks, this is called full term. Giving birth before week 37 is called preterm, and passed week 41 is called postterm [7]. During birth the fetal needs to handle periods without normal oxygen supply which happens during labour due to the contractions compressing the umbilical cord or decreasing the blood flow to the placenta, this is called acute hypoxia or asphyxia. Acute asphyxia can lead to fetal acidosis and death [4].

2.1.1 Cardiotocography

Cardiotocography, CTG, is a method used during labour to detect fetal hypoxia, which is lack of oxygen in the blood and in the tissues. The CTG measures the fetal heart rate, alongside with uterine contractions. The FHR signal is measured by using a Doppler sensor, or electrode on the fetal scalp. The TOCO signal is measured by using an external pressure sensor. Figure 1 presents a CTG, where the upper signal is the FHR, and the lower signal is the TOCO. The FHR signal is measured in beats per minute, *bpm*, while the TOCO signal does not have a unit. The signal registration has to be at least 20 minutes before assessment [3].

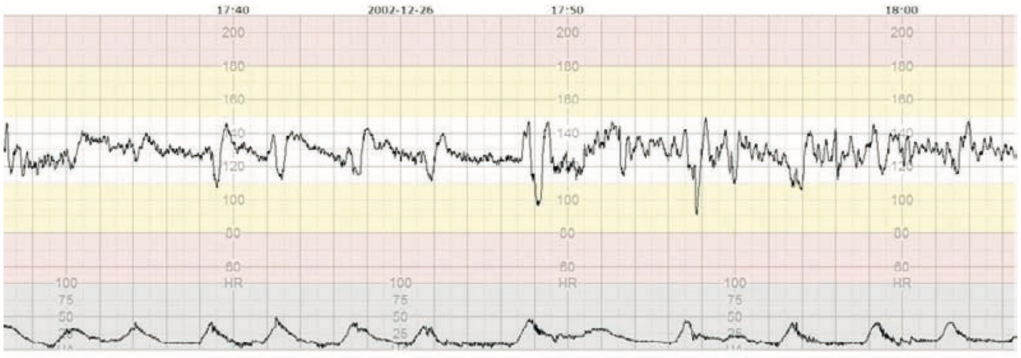


Figure 1: Cardiotocography with FHR signal at the top and TOCO signal at the bottom. The unit on the y-axis is bpm which is the FHR signals unit. The contraction signal does not have a unit. [8]

Based on these signals, the state of the fetus is evaluated throughout the labour. The contractions, which should not have a frequency of more than 4-5 in a period of 10 minutes, affects the FHR signal. As seen in Figure 1, clear decelerations in the FHR are linked to contractions [3]. When the fetus is exposed to contractions, which can lead to increased pressure to the head and compression of the umbilical cord, it has a reflective response and decreases the FHR. The variables that are taken into consideration when evaluating the FHR signal are the following [3]:

- **Baseline Fetal Heart Rate**
  The average FHR, not including accelerations or decelerations, for a period of at least ten minutes. Normal baseline for full term birth is 110-150 bpm, while it can be up to 160 bpm for a preterm birth.
  A baseline FHR of less than 120 bpm is known as bradycardia, where a lesser value is more serious and is a pathological pattern. A baseline FHR of more than 150 bpm is called tachycardia and can be a sign of hypoxia.

- **Baseline Variability**
  The variability of the baseline FHR which should be 5-25 bpm.

- **Acceleration**
  Increase in FHR-signal of at least 15 bpm from baseline, which lasts for at least 15 seconds. There should be at least two or more accelerations in a 20 minute period each hour. If the fetal is asleep accelerations might not be present.

- **Deceleration**
  Temporary decreases in FHR-signal. The decelerations are classified as uniform early, uniform late, variable uncomplicated or variable complicated depending on shape and duration. The most worrying being late decelerations, which is a sign of limited oxygenated maternal blood to the placenta due to a contraction. This leads to hypoxemia for the fetus, who reacts with lowering the FHR.

If all said variables meet their expected value, the CTG is classified as normal. The CTG is classified as suspicious if the baseline FHR is between 100-110 bpm or 150-170 bpm, or if there are no or few accelerations, or the variability is larger than 25 bpm, or the decelerations are lasting longer than 60 seconds or have a big amplitude. If there are numerous variables that deviate, such as, the baseline FHR is lower than 100 bpm or higher than 170 bpm, or there are uniform, i.e., U-shaped, decelerations which can both be seen during and after a contraction, or the FHR looks like a sine curve, the CTG is pathological. The most dangerous case is when the FHR signal does not have any variability or accelerations present, the classification is then preterminal. The fetus has ways of handling the pressure during labour, however, when the pressure is increased or have been ongoing for a long period of time, the risks of brain injuries increases [3].

By using CTG, it is possible to detect and diagnose threatening or manifest hypoxemia. Hypoxia occurs after prolonged hypoxemia, which is low oxygen content in the blood. Hypoxia can lead to organ failure, irreversible tissue damages and death. Abnormalities are more common in the expulsion stage, where the risks of hypoxia increase, due to the contraction intensity increasing [3].

2.1.2 Apgar Score

After birth, clinical professionals evaluates the infant's health by using the Apgar Score. The Apgar Score is the sum of these considered features; breathing, heart rate, skin colour, muscle tone, and response to stimulation. All features are given a score between 0 to 2, where the higher means the better. The Apgar Score is evaluated at 1, 5, and 10 minutes after birth, and is in a range from 0 to 10, where the higher the score, the better the newborn is doing. The Apgar Score determines the degree of possible asphyxia, e.g., a sign of asphyxia is if the infant's skin colour leans more blue

rather than pink [6]. At five minutes, the classification for ranges of the Apgar Score can be defined as: 7-10 reassuring, 4-6 moderately abnormal, and 0-3 abnormal [9].

2.2 Bayes' Theorem

A useful tool in machine learning is Bayes Theorem which gives an idea of the relationship between input data and output classes. In a classification problem, where the probability of a class given an observation should be derived, Bayes Theorem is defined as

$$P(C_i|x) = \frac{P(C_i)P(x|C_i)}{P(x)}, \qquad P(x) \neq 0, \tag{1}$$

where there are i different classes C_i, and x is an observation. Common to all data points is the density function $P(x)$, $P(C_i)$ is the prior distribution for class i, and $P(x|C_i)$ is the density function of the data points belonging to class C_i which is derived from the training data under the assumption there is a certain distribution. The posterior probability $P(C_i|x)$ is derived by using prior knowledge [10, 11].

2.3 Artificial Neural Networks

Artificial Neural Networks, ANNs, or Neural Networks, NNs, are nonlinear statistical models [12]. Figures 2 and 3 depict the structure of neural networks. As seen the neural network consists of nodes, also called units, and connections, known as edges, between the nodes. Figure 2 shows the usage of one neuron, which is the simplest form of a neural network and is referred to as perceptron [13]. The network is fed with the input $\mathbf{X}$. The parameters weight $\mathbf{w}$ and bias b are optimized to minimize the error of the network. The input X_i, for $i = 1, ..., n$, is multiplied with weight w_i, and for all i they are summed together with the bias b. This total sum is fed into function f to produce an output y [12]. The output is defined by the general formula [10]

$$y = f(\mathbf{w}^T\mathbf{X} + b) = f\left(\sum_{i=1}^{n} w_i X_i + b\right) = f(w_1X_1 + ... + w_nX_n + b). \tag{2}$$

Figure 3 shows the usage of multiple neurons and edges, ordered in layers. All layers except the first and last are known as hidden layers. When using more layers the neural network is called a deep neural network. A fully connected layer is when an output is fed as an input to all neurons in the next layer.

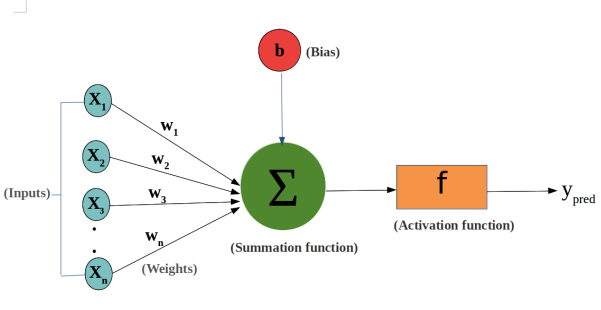


Figure 2: Structure of the smallest neural network, only one perceptron [14].

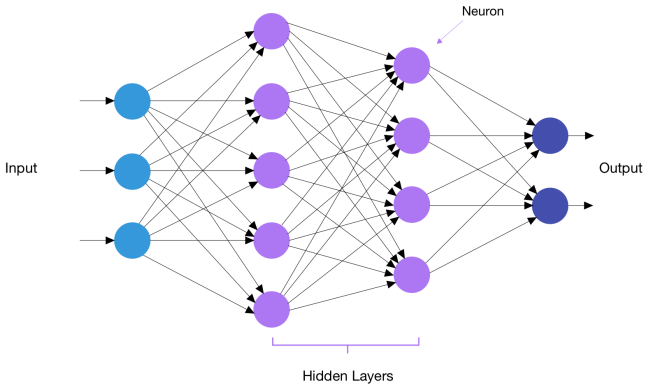


Figure 3: Structure of a neural network with N hidden layers [15].

2.3.1 Activation Functions

Activation functions are used to transform the several inputs into an output within a certain range, and to introduce non-linear complexities in a NN. There are numerous activation functions to chose from when designing a neural network, both in hidden layers and output layer, depending on what is sought after. One can for example use the sigmoid function, also known as the logistic function, which is defined as

$$f(x) = \frac{1}{1 + e^{-x}}. \tag{3}$$

The sigmoid function maps the results within the range 0 to 1. This can be a good pick when the result should be a probability. The softmax function, or the normalized exponential, also produces a probability, but does so for more than one class. The softmax function is defined as

$$f(\mathbf{x}_i) = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}} \qquad \text{for } i = 1, ..., K, \quad \mathbf{x} = (x_1, ..., x_K) \in \mathbb{R}^K. \tag{4}$$

If one wants another range the hyperbolic tangent, tanh, is an option. It maps within the range -1 to 1 using the function

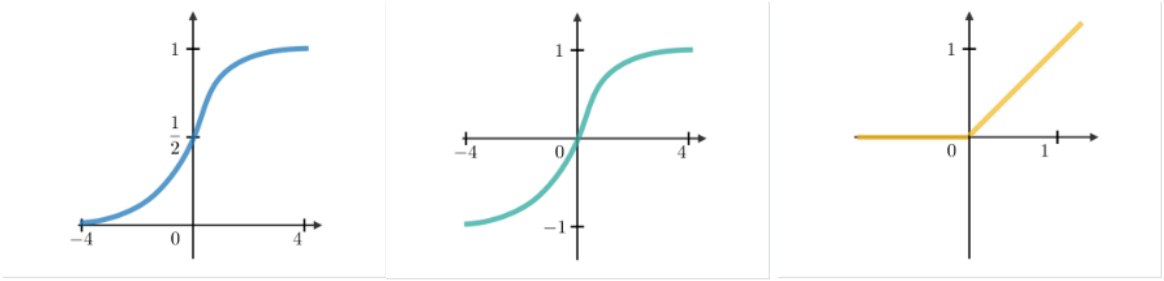
$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}. \tag{5}$$

A benefit of this function is that is maps negative inputs to a negative output. The rectified linear unit (ReLU) is in some literature referred to as the default activation function [10]. It is is defined as $f(x) = \max\{0, x\}$. This function outputs provided input value x if $x \geq 0$ or 0 if $x < 0$. One advantage with the ReLU function is the sparsity it provides with the ability to output a true zero value. The sparsity can make learning less time consuming and also simplifying the model. A down side with ReLU is when the input to the activation function is negative, meaning that the output will always be zero and therefore never activate the unit. This can cause weight swings when optimizing the model [10]. To get around this issue, one can instead use Leaky ReLU which allows negative

inputs. The Leaky ReLU function is defined as

$$f(x) = \begin{cases} \alpha x, & \text{if } x < 0, \\ x & \text{otherwise,} \end{cases} \tag{6}$$

where α is a small scaling factor [16]. Some of the aforementioned activation functions are shown in Figure 4 below.



(a) Sigmoid: $f(x) = \frac{1}{1+e^{-x}}$ **(b)** Tanh: $f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ **(c)** ReLU: $f(x) = \max\{0, x\}$

Figure 4: Different activation functions [17].

2.3.2 Loss Function

To optimize the performance of a network one wants to optimize an objective function. With neural networks it is typically sought after to minimize the error of an objective function, which is also referred to as loss function, L. For example, mean squared error can be used for regression problems, or if one has a classification problem cross-entropy can be used. The cross-entropy is defined as [13]

$$L(y, \hat{y}) = -\sum_{i=1}^{m} y_i \log(\hat{y}_i), \tag{7}$$

where for each class $i = 1, ..., m$, $y_i \in \{0, 1\}$ is binary indicator and $\hat{y}_i$ is the predicted probabilities. The cross-entropy outputs probabilities that are an indication of which class the input belongs to. To reduce the loss, an optimization algorithm, such as gradient descent, or stochastic gradient descent, is used to find suitable updates for the hyperparameters within the network [10].

2.3.3 Weights and Biases

The parameters that are updated in a neural network are the weights and biases. The weights tell the strength of the connections between neurons, and they are updated as

$$\mathbf{w} \leftarrow \mathbf{w} - \alpha \left(\frac{\partial L}{\partial \mathbf{w}} \right), \tag{8}$$

where $\frac{\partial L}{\partial \mathbf{w}} = \left(\frac{\partial L}{\partial w_1}, \ldots, \frac{\partial L}{\partial w_n} \right)$, and α is the learning rate, which can be updated with each iteration step [13]. The invariant part of predictions are captured by the biases which are constant terms. Note that the biases and weights can be updated regardless of one another [12] [13].

To update the weights and biases, a neural network uses *back-propagation*. Forward propagation is shown in Figure 2 and Figure 3, where the input is fed forward until the network produces an

output and a scalar cost. When the scalar cost is sent backward through the network, it is utilizing the back-propagation. This is done such that the scalar cost can influence the calculations of the gradients, and therefore improve the model.

2.3.4 Data Set Augmentation

Large imbalance in collected data creates a problem when classifying, since the accuracy of the classification will be a reflection of the underlying distribution of the classes. The classifier becomes biased towards the majority class. To reduce the imbalance in a set of data, one can create new, fake data of the underlying class by transforming the inputs in the training set, this is known as data set augmentation. Creating synthetic data can be done by adding random noise to the inputs [10].

2.4 Related Work

In the aspiration of making the interpretation of CTG easier numerous studies have been made. Some propose methods to computerize the FIGO guidelines and use these as features in machine learning applications or neural networks, while some present methods to automatically extract patterns in the CTG.

Romano et al. presented a software for automatic analysis of the FHR [18]. Nidhal et al. aimed to derive a computerized baseline of the FHR signal [19]. Agostinelli et al. studied a different approach to the FHR baseline and aimed at finding the correct statistical baseline which they derived as mean(FHR $\pm\Delta$ FHR), where Δ FHR is equal to 8 bpm or 10 bpm. It was found, both statistically and clinically, that Δ FHR = 10 bpm [20]. This was used in a study by Sbrollini et al. who tried to automatically detect decelerations in the FHR signal, and classified these according to the FIGO guidelines, i.e., early or late, V-shaped or U-shaped [21]. Labaj et al. studied how one could extract the correlation between the FHR decelerations and the contractions [22]. Chung et al. used spectral analysis of the FHR variation to predict fetal distress [23].

V. Chudáček et al. studied different features for classification of the FHR signal. It was found that many features correlate and therefore contribute with the same information. The best features were number of accelerations, number of decelerations and the interval index which contributed with uncorrelated information to the classifier [24]. The importance of the decelerations, specifically its depth and duration, was also mentioned by Spilka et al. [25]. They tried to detect fetal acidosis using machine learning, specifically sparse support vector machine. Georgoulas et al. also used support vector machine to predict the risk of metabolic acidosis based on the FHR [26]. Jezewslo et al. used neural networks to classify CTG traces in order to predict the pH-value [27]. Fontenla-Romero et al. used artificial neural networks to create a real time extraction of features from the FHR signal [28].

3 Method

In this section, the data used, examined features, and evaluation are explained. Firstly, the data that was available for this project is introduced, and it is explained how it was preprocessed. Secondly, the selection of features are explained and motivated. Finally, the training and evaluation of classification methods are described. All programming was done in MATLAB.

3.1 Data Set

For this project the gathered data consisted of 98,328 different labours. For each labour, the CTG, and clinical data had been collected. All cases collected were one child cases. The clinical data consisted of the following collected information:

- Study ID
- Parity - number of previous births
- Smoking - yes/no and to what extent
- Ablatio placentae - serious complication
- Pre-eclampsia - serious blood pressure condition
- Gestational age - length of pregnancy in days
- Gender
- Weight
- Weight Deviation
- Mode of Delivery
- Apgar Score at 1 minute
- Apgar Score at 5 minutes
- Vital Status
- pH-value
- Neonatal Care
- IVH - intraventricular bleeding, bleeding into the brain's ventricular system
- HIE - hypoxic ischemic encephalopathy, type of brain dysfunction
- Spasms
- Malformation.

The information of interest in this project was the Apgar Score at 5 minutes. The number of samples having a 5-minute Apgar Score lower than 7 was 1621, making it 1.65% of the data set.

The collected CTG measurements contained FHR and TOCO signals, sensor type for each signal, indication of the quality of the signals, time vector, and patient id. For FHR, the sensor type was either ultrasound or scalp electrode. For TOCO, the sensor type was external pressure. The data points were recorded with a frequency of 4Hz, however for some samples this varied. The time vector sometimes contained negative time stamps or jumps. In Figure 5, a CTG is shown, with the x-axis being time in minutes, where $x = 0$ is where the signal ends. The blue coloured signal is the FHR, the red is the TOCO, yellow is sensor type for FHR, purple is sensor type for TOCO, and khaki is the quality of the signal. The TOCO signal was sometimes not registered, a possible reason is that it was uncomfortable for the patient to have the external pressure sensor on, therefore it was removed.

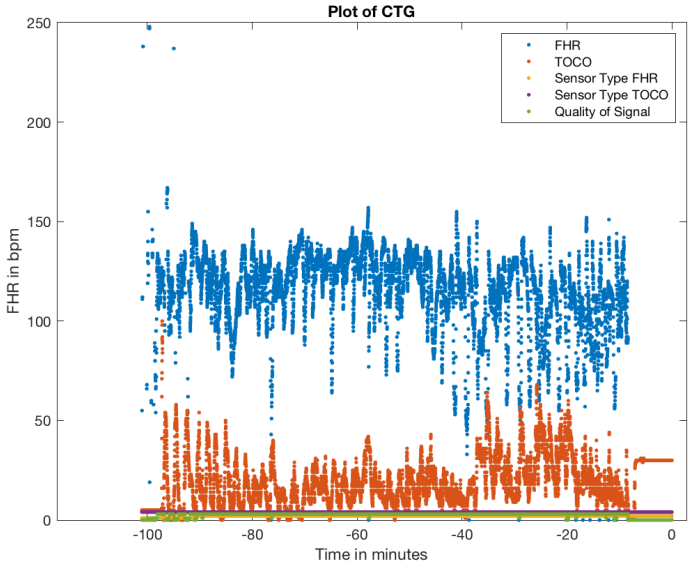


Figure 5: Example of CTG. Blue coloured signal is the FHR, the red is the TOCO, yellow is sensor type for FHR, purple is sensor type for TOCO and khaki is the quality of the signal. The x-axis is the time in minutes in descending order, meaning that the signal ends at $x = 0$. The y-axis is in beats per minute.

3.1.1 Exclusion of Samples

One supplied variable in the clinical data was vital status. It contained information if the infant had died antepartum, intrapartum or postpartum. The antepartum cases were evaluated, since if the child had passed away before labour, there would not be any possible action to prevent this. In total 125 samples had information that the child died antepartum. After further evaluation it was found that 52 of these samples had a CTG that lasted for at least 30 minutes, the other 73 samples were excluded. The 52 cases were checked manually to see if the CTG had some sort of feasible information, i.e., the signal contained a clear FHR signal. If so, it would contradict the information of the infant being dead antepartum, hence those samples could still be used. Out of the 52 samples, 14 were determined to be used, and 38 were excluded.

Some samples had multiple variables missing in the clinical data, for example no specified gender, making the existing variables questionable. One sample had all variables missing. All of these samples were excluded, which was a total of 33.

3.1.2 Segmentation

The collected CTG data rarely contained a continuous signal throughout the labour, as seen in Figure 6. The reason behind this might be that it is common to check the CTG upon arrival, wait for the labour to progress and then check the CTG intermittently. Due to this inconsistency, it was decided to use segments that contained a continuous signal for a 30 minute period, since theoretically at least a 20 minute recording is needed for assessment. A continuous signal was defined as no interruption that lasted longer than 5 seconds, where 5 seconds was selected on account of the sample rate not always being $4Hz$. It was also checked that the FHR signal was not equal to 0 for more than 50% of the continuous signal. The reasoning behind this was having a measurement but no FHR signal would skew the results. In Figure 7, it is visualized how the measurement can be longer than 30 minutes, but not contain a FHR signal throughout.

A 30 minute segment, with a 30 seconds offset, was looked for in each CTG. The 30 second offset was set due to most segments total time never being exactly 30 minutes, but rather 29.9 etc. If the signal was shorter than 30 minutes, or it did not contain any segment of length 30 minutes, it was excluded. These thresholds excluded a total of 14,070, data samples whereof 253 were cases with bad outcome. All of these 253 cases were checked manually to ensure none was excluded by mistake. It was found that 11 out of the 253 could be used. All of them contained negative time jumps so they were extracted manually.

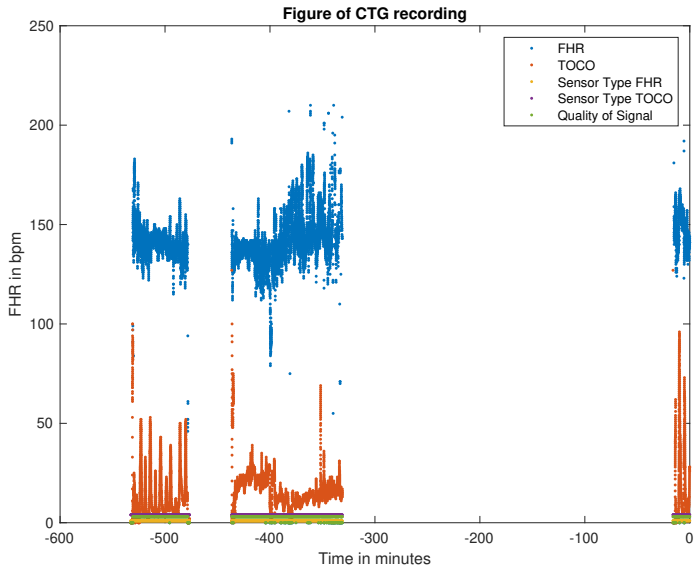


Figure 6: Figure of CTG, visualizing the interruptions in the recording. Blue coloured signal is the FHR, the red is the TOCO, yellow is sensor type for FHR, purple is sensor type for TOCO and khaki is the quality of the signal. The x-axis is the time in minutes in descending order, meaning that the signal ends at $x = 0$. The y-axis is in beats per minute.

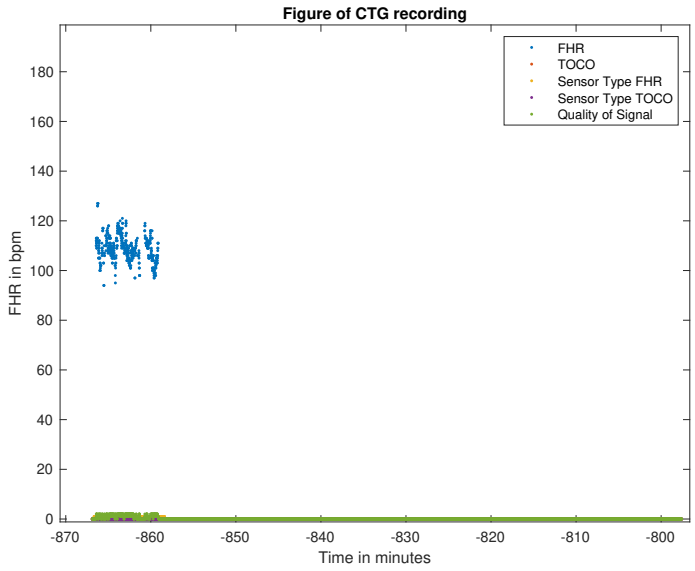
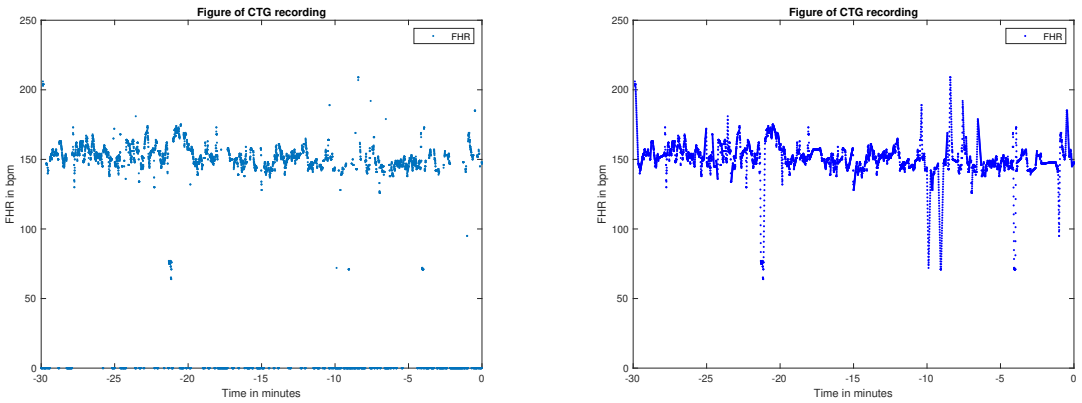


Figure 7: Figure of CTG, visualizing the motivation of having a threshold on the FHR signal when extracting segments. Blue coloured signal is the FHR, the red is the TOCO, yellow is sensor type for FHR, purple is sensor type for TOCO and khaki is the quality of the signal. The x-axis is the time in minutes in descending order. The y-axis is in beats per minute.

If the signal at hand was corresponding to a bad outcome, i.e., 5-minute Apgar Score lower than 7, it was checked if the found segment was longer than 40 minutes, or if the signal contained more segments, of length 30 minutes or more, than the chosen one. If so, it was used in the augmentation step to generate more data. If the segment was originally longer than 40 minutes, a new segment was created by a shift of at least 10 minutes. If there were more segments in original signal, these were used as well. For example, in the Figure 6 above, there are three visible segments. The one to the right is not 30 minutes, hence it was ignored. The other two segments are both longer than 30 minutes, therefore both segments might be used when augmenting.

3.1.3 Linear Interpolation to Remove Missing Data

The FHR signals in the CTG suffers from missing data, some more than others. Figure 8a depicts a segmented FHR signal, where it can be seen that many data points are equal to zero. The main part of the FHR signal is around 150 bpm which indicates that the zero values are missing data, in this particular case it was 34% of the entire signal. Note that this FHR signal still shows a specific pattern, and should therefore still be used in the study. To remove the missing data, linear interpolation was used. The result after using linear interpolation is shown in Figure 8b. If the beginning or the end of the FHR signal consisted of only missing values, these were added as the nearest non-zero value in the FHR signal with random added noise of ± 5 bpm.



(a) Original FHR signal.

(b) Interpolated FHR signal.

Figure 8: Figure of CTG, the left shows original signal, and the right shows the signal after linear interpolation. Blue coloured signal is the FHR, the x-axis is the time in minutes in descending order, meaning that the signal ends at $x = 0$. The y-axis is in beats per minute.

3.2 Extraction of Features

To examine possible features, the very first step made with the data was to divide it into a training set and a test set with the ratio 70:30 using stratification on the different outcomes, i.e., making sure that both sets contained same ratio of cases with bad outcome. The test set was put aside and was **not** used until it was time to test the final model. The training data, with the number of samples being 58,850, was used to find appropriate features and train models. The aim was to find features that would distinguish cases with good outcome and cases with bad outcome. For each feature, histograms are presented, visualizing the normalized proportions for each outcome, where red coloured histogram is the cases with bad outcome, and blue coloured histogram is cases with good outcome. The proportions were normalized, since the number of cases with good outcome is much greater than the number of cases with bad outcome. Given Bayes' Theorem, see Equation 1, the target was to visualize how often the bin corresponding to the features value was present in the data, i.e., P(x|good outcome), blue colour, or P(x|bad outcome), red colour, and see if there was different distributions for the separate outcomes considering their respective histograms. Even if the distributions were overlapping, the examined feature was used to train models to see if it could increase the performance.

The first feature introduced was from the clinical data, and it was used to derive a threshold for the upcoming features derived from the CTG. The next features examined the FIGO guidelines, trying to estimate the FHR baseline, and extract information from this in different manners. Then follows the FHR short term variability, and interval index, two features taken from a related study, followed by the last feature, the mean absolute deviation of the FHR

3.2.1 Gestational Age

A given parameter in the clinical data was the gestational age, i.e., the length of pregnancy in days, was picked as first feature due to being a known fact before birth. In Figure 9, histograms of said parameter are shown to visualize the distribution of cases with good outcome versus bad outcome.

Note that there are samples at number of days equal to 0, which is not a possible gestational age. This is an indication of missing data, though some of these sample had the given information that the infant was taken to neonatal care, i.e., the infant was not doing well after birth and needed intensive care, or the samples had a given Apgar Score. The ones which had an Apgar Score lower than 7 in the fifth minute were kept, and the samples which had an Apgar Score higher than or equal to 7 were excluded, which was 38 samples. This decision was made due to the knowledge of low Apgar Score, or neonatal care are highly correlated with preterm birth [29].

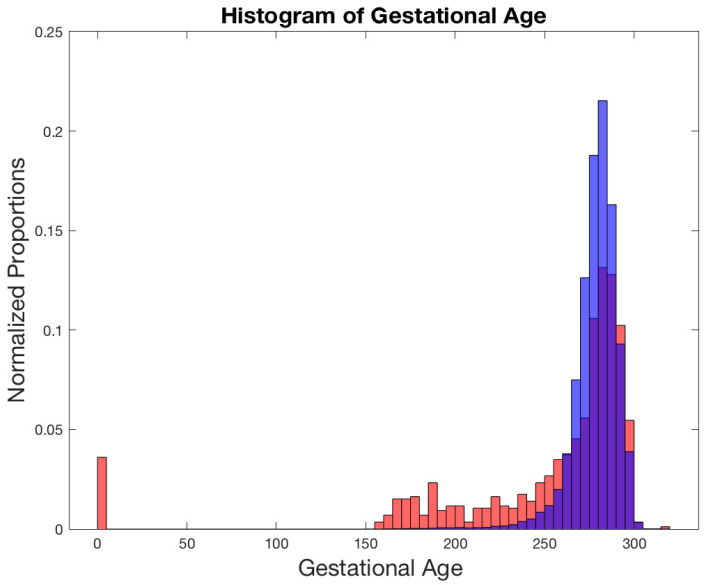


Figure 9: Histograms of gestational age in days. The y-axis is the normalized proportions. The cases with bad outcome are shown in red colour, and the cases with good outcome are shown in blue colour. The proportions of the data adds up to 1 for each colour.

To avoid that this parameter alone is the decisive information, the data was split into two cases:

Case 1 : Preterm, Samples with pregnancy length in days ≤ 258,

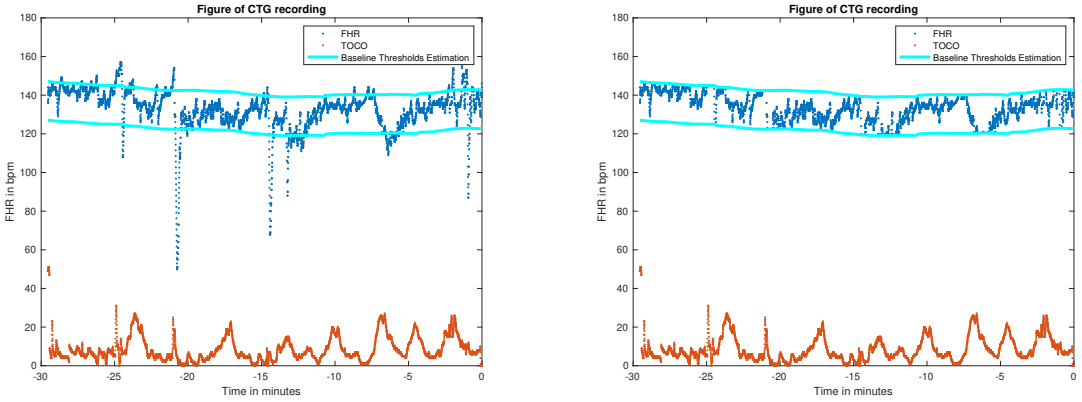
Case 2 : Full term including postterm, Samples with pregnancy length in days ≥ 259.

The split was chosen based on the clinical definition of a full term pregnancy being 37 weeks, i.e, 259 days. There was no overlap in these sets since the cases who are preterm have a bad Apgar Score due to other reasons. Having an overlap would increase the possibility of blurring the results. The number of samples in case 1 was 3,405, and in case 2 there was 55,445 .

3.2.2 Estimation of Fetal Heart Rate Baseline

The FHR baseline is a crucial part when interpreting the CTG, since most patterns clinicians consider are in relation to the baseline. The theory, see [3], defines the baseline as the mean of the FHR signal without accelerations and decelerations, hence it was desired to ”remove” these. Using the equation MA(FHR $\pm$10bpm), see [20], where MA was put to a moving average over a 20 minute sliding window, gave an estimation of an area which the baseline could be within.

Figure 10a shows a segmented CTG signal, including the FHR signals estimated baseline thresholds, in cyan colour. To exclude the accelerations and decelerations, all data points above and below the estimated baseline thresholds were removed, see Figure 10b. The arithmetic mean of the remaining FHR data points were used as baseline when deriving further features that depend on the baseline, the result shown in 11, where the estimated baseline is shown in green colour.



(a) **(b)**

Figure 10: Figure of CTG recordings. In Figure a), the estimated baseline threshold is shown in cyan colour. In Figure b), the FHR values above, and below the baseline threshold have been removed. Blue coloured signal is FHR, and red coloured signal is TOCO. The x-axis is the time in minutes in descending order, meaning that the signal ends at $x = 0$. The y-axis is in beats per minute.

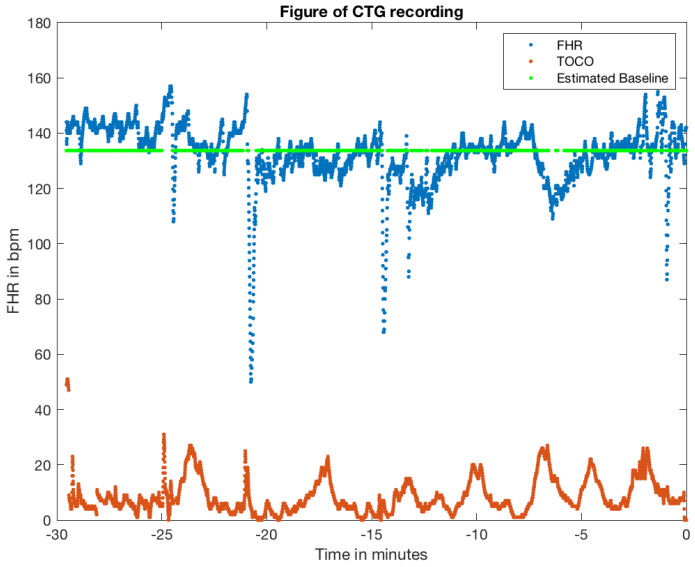
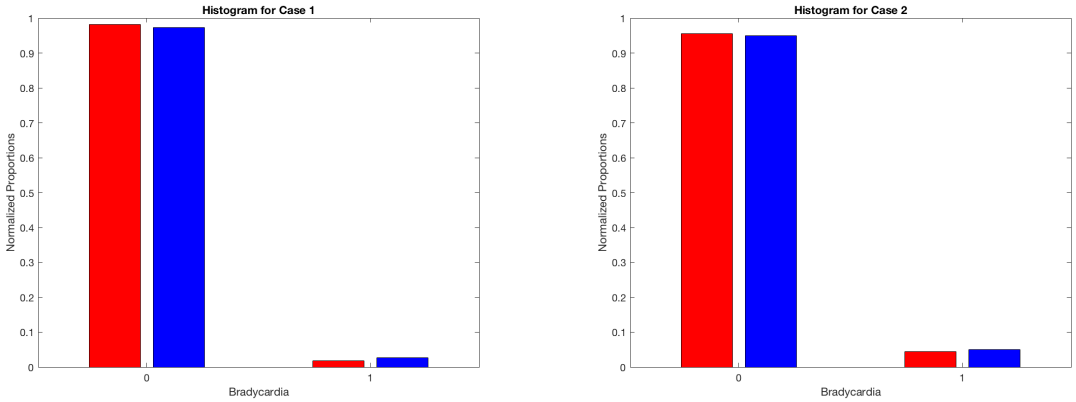


Figure 11: Figure of CTG recording, with the estimated baseline in green colour. Blue coloured signal is FHR, and red coloured signal is TOCO. The x-axis is the time in minutes in descending order, meaning that the signal ends at $x = 0$. The y-axis is in beats per minute.

Bradycardia

Using the estimated baseline, described previously, gave the possibility to investigate if the FHR signal had signs of bradycardia. Bradycardia is when the FHR signal baseline is less than 120 bpm, where the lower, the worse [3], hence a threshold of 120 bpm was used when classifying the presence of bradycardia.

Figure 12a shows the histograms of bradycardia for case 1, and Figure 12b shows the histograms for case 2.



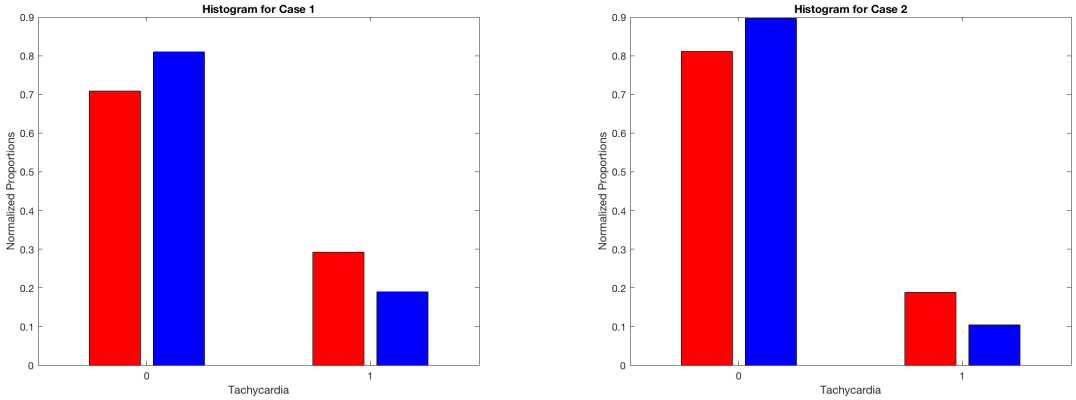
(a) Case 1.

(b) Case 2.

Figure 12: Histograms of bradycardia using a threshold of baseline to be less than 120 bpm. The y-axis is the normalized proportions. The cases with bad outcome are shown in red colour, and the cases with good outcome are shown in blue colour. The proportions of the data adds up to 1 for each colour

Tachycardia

Tachycardia is when the FHR signal baseline is more than 150 bpm [3]. This was derived in the same manner as bradycardia, though instead checking if the estimated baseline was above 150 bpm. Figure 13a shows the histograms of tachycardia for case 1, and Figure 13b shows the histograms for case 2.



(a) Case 1. **(b)** Case 2.

Figure 13: Histograms of tachycardia using a threshold of the baseline to be greater than 150bpm. The y-axis is the normalized proportions. The cases with bad outcome are shown in red colour, and the cases with good outcome are shown in blue colour. The proportions of the data adds up to 1 for each colour.

Estimation of Accelerations

From the baseline estimation, it was also of interest to investigate if the accelerations could be a good feature. An acceleration was defined such that if a data point was 5 bpm higher than the upper estimated baseline threshold. The reasoning behind this is that an acceleration is 15 bpm higher than the FHR baseline, this is depicted in Figure 14, where the red points are classified as accelerations.

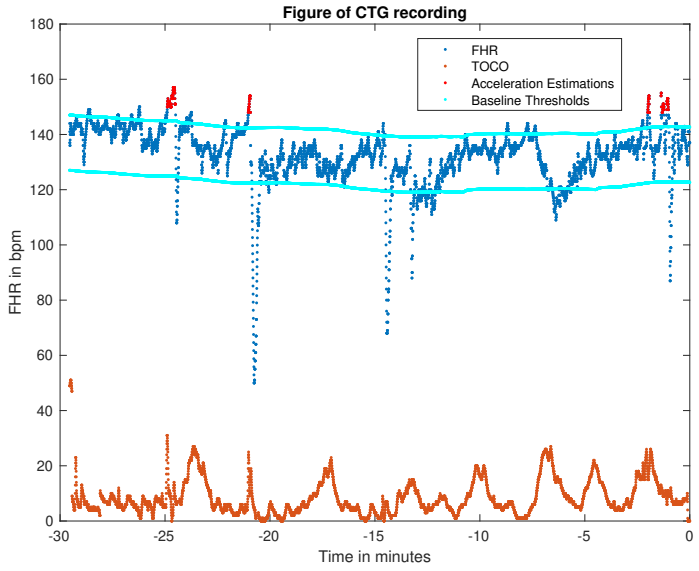
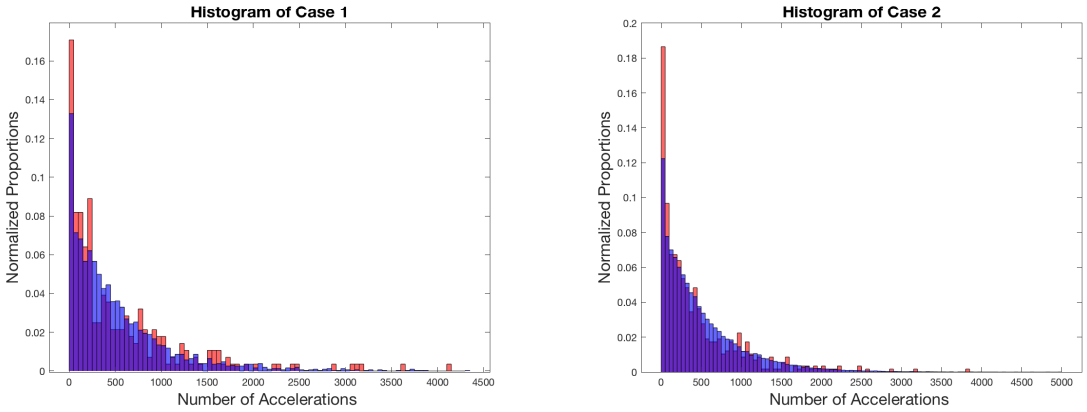


Figure 14: Figure of CTG recording, with the estimated baseline threshold in cyan colour, and the estimated accelerations in red colour. Blue coloured signal is FHR, and red coloured signal is TOCO. The x-axis is the time in minutes in descending order, meaning that the signal ends at $x = 0$. The y-axis is in beats per minute.

The histograms below show the distributions for each outcome having the estimated number of

accelerations as observation. Figure 15a shows case 1, and Figure 15b shows case 2.



(a) Case 1. **(b)** Case 2.

Figure 15: Histograms of estimated accelerations in FHR signal. The cases with bad outcome are shown in red colour, and the cases with good outcome are shown in blue colour. The proportions of the data adds up to 1 for each colour.

Estimation of Decelerations

A deceleration is a temporary decrease in the FHR signal [3]. The number of decelerations was estimated using the lower baseline threshold, where all FHR values below was deemed a deceleration. Figure 16 shows the number of estimated decelerations in red colour.

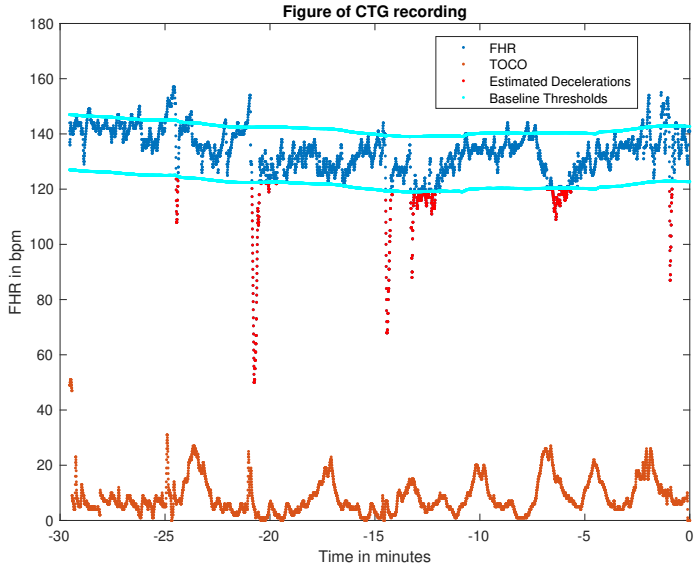
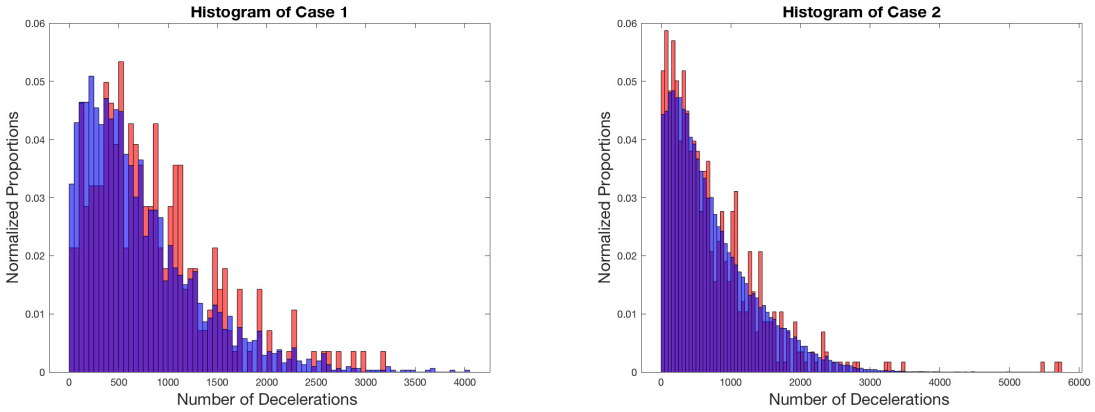


Figure 16: Figure of CTG recording, with the estimated baseline threshold in cyan colour, and the estimated decelerations in red colour. Blue coloured signal is FHR, and red coloured signal is TOCO. The x-axis is the time in minutes in descending order, meaning that the signal ends at $x = 0$. The y-axis is in beats per minute.

The histograms below shows the distributions for the each outcome having the estimated number of

decelerations as observation. Figure 17a shows case 1, and Figure 17b shows case 2.



(a) Case 1.

(b) Case 2.

Figure 17: Histograms of estimated decelerations in FHR signal. The cases with bad outcome are shown in red colour, and the cases with good outcome are shown in blue colour. The proportions of the data adds up to 1 for each colour.

Linear Regression of the Moving Average

The moving average of the FHR signal was also investigated by using linear regression. It was of interest to examine whether or not the slope could be a possible feature, i.e., if the slope being increasing or descending was a distinguishing factor. The moving average was, as before, derived from a 20 minute sliding window. Figure 18 shows the interpolated FHR signal, with its moving average in yellow, and fitted line in black. Figure 19a shows the histogram for derived slopes for case 1, and Figure 19b shows the histogram for derived slopes for case 2.

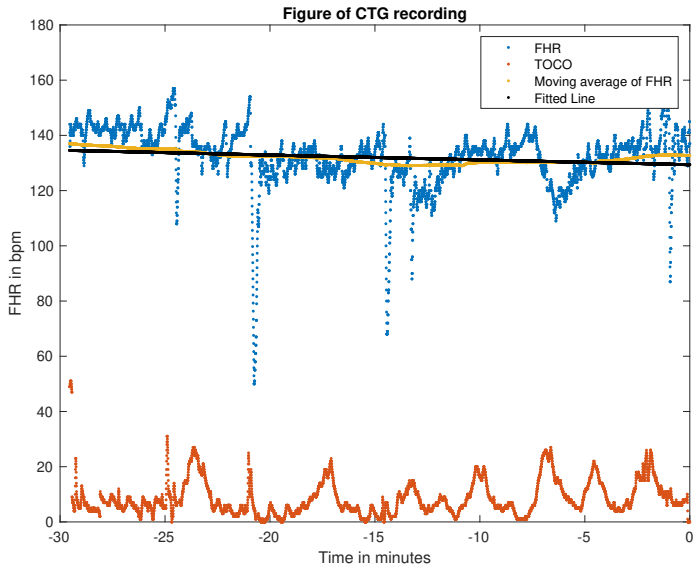
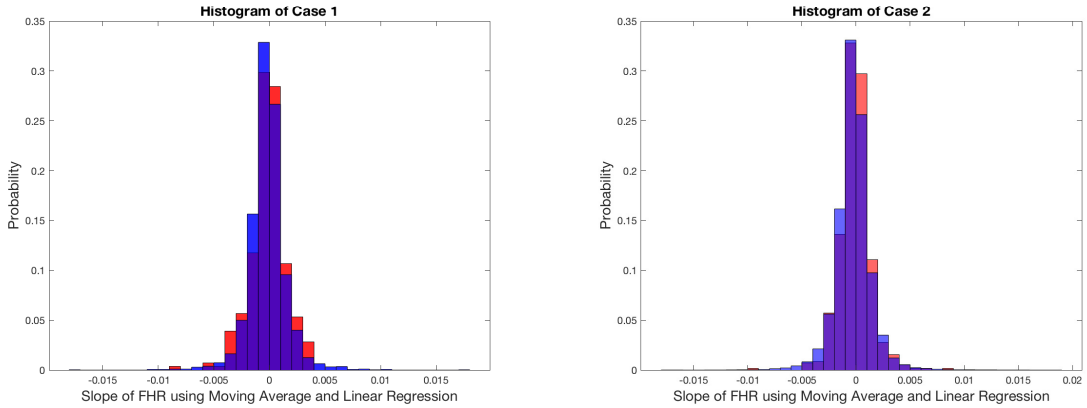


Figure 18: Figure of CTG recording. Blue coloured signal is FHR, red coloured signal is TOCO, yellow line is moving average using 20 minute sliding window of the FHR signal and the black line is the moving average fitted by using linear regression. The x-axis is the time in minutes in descending order, meaning that the signal ends at $x = 0$. The y-axis is in beats per minute.



(a) Case 1. **(b)** Case 2.

Figure 19: Histograms of slope of line derived using linear regression on moving average of the FHR signal. The cases with bad outcome are shown in red colour, and the cases with good outcome are shown in blue colour. The proportions of the data adds up to 1 for each colour.

3.2.3 Short Term Variability and Interval Index

To capture the unseen variability, the short term variability, STV, and its possible influence was investigated. The STV was derived from 60 seconds of the FHR signal, and it was defined as [24]

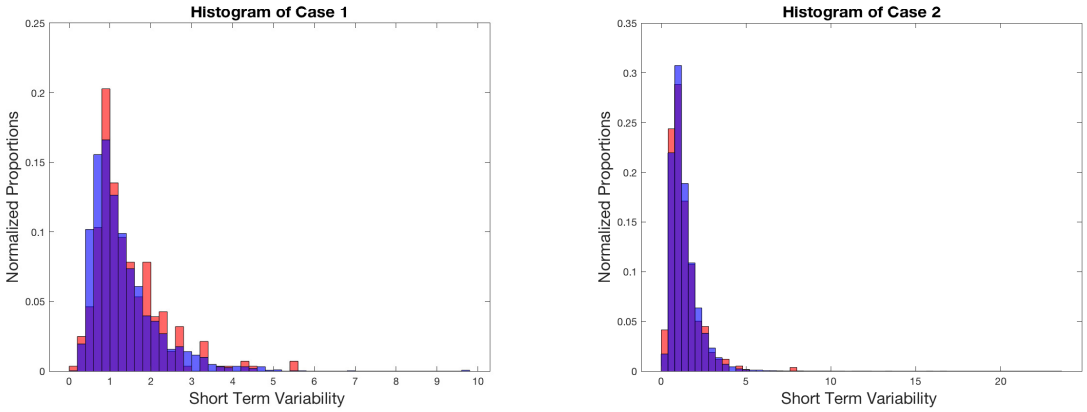
$$\text{STV} = \frac{\sum_{i=1}^{24} |s\text{FHR}(i+1) - s\text{FHR}(i)|}{24}, \tag{9}$$

where $s\text{FHR}(i) = \text{FHR}(10 \times (i-1)+1)$, i.e., the FHR signal taken once every ten samples. From the STV, the interval index can be derived as [24]

$$\text{II} = \frac{\text{STV}}{\text{std}[s\text{FHR}(i)]}, \tag{10}$$

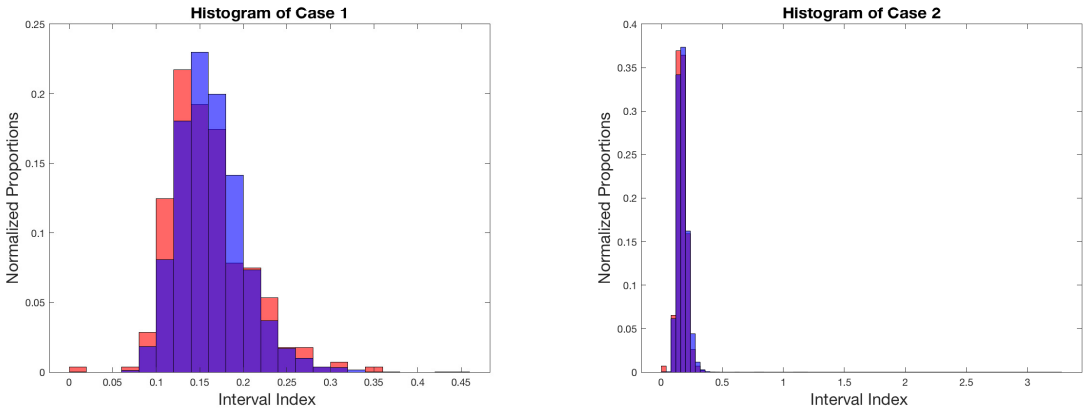
where std is the standard deviation. If the denominator returned 0, i.e., the FHR signal at every ten samples do not change at all, the last sample was added with 1 bpm, forcing a small standard deviation, and getting a large II.
To decrease the sizes of the STV and the II, being 30 values each since being derived from a 30 minute signal, to reduce the dimensions, the mean absolute deviation of each were tested as features.

Figure 20a shows the histograms of STV for case 1, and figure 20b shows the histograms of STV for case 2. Figure 21a shows the histograms of II for case 1, and Figure 21b shows the histograms of II for case 2.



(a) Case 1. **(b)** Case 2.

Figure 20: Histograms of mean absolute deviation of the short term variability. The y-axis is the normalized proportions. The cases with bad outcome are shown in red colour, and the cases with good outcome are shown in blue colour. The proportions of the data adds up to 1 for each colour.

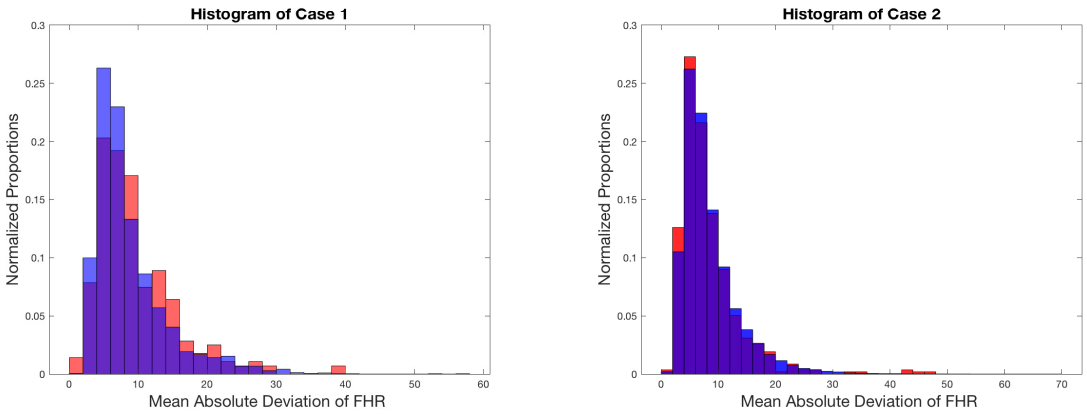


(a) Case 1. **(b)** Case 2.

Figure 21: Histograms of mean absolute deviation of the interval index. The y-axis is the normalized proportions. The cases with bad outcome are shown in red colour, and the cases with good outcome are shown in blue colour. The proportions of the data adds up to 1 for each colour.

3.2.4 Mean Absolute Deviation

To capture the changes over time in the signal, and investigate the spread of the FHR signal, the mean absolute deviation, MAD, of the FHR signal was used. Figure 22a shows histograms for case 1, and Figure 22b shows histograms for case 2.



(a) Case 1. **(b)** Case 2.

Figure 22: Histograms of mean absolute deviation. The y-axis is the normalized proportions. The cases with bad outcome are shown in red colour, and the cases with good outcome are shown in blue colour. The proportions of the data adds up to 1 for each colour.

3.3 Models

With the found features, multiple models were trained. All models used different techniques with the idea to compare the various methods, and not fully rely on one approach. The algorithms used were support vector machine, short SVM, k-nearest neighbours, short KNN, and decision tree, short DT, all of them made by using existing functions in MATLAB. The interested reader can find the theory

behind these methods here [12].

Along with SVM, KNN and DT, a neural network was also implemented, since it allowed tuning the network parameters. The neural network contained input layer, followed by the layers: fully connected of size 10, activation function ReLU, fully connected of size 2, softmax function, and the classification as output. The structure was set after trial and error, tuning the learning rate with initial rate, and drop factor, adding more layers with different output sizes, as well as subtracting layers, changing number of epochs, etc.

To be able to derive a mean and standard deviation of the four classifiers, 10-fold cross validation was used. The mean and standard deviation was sought after to give a better understanding of the results.

3.3.1 Data Set Augmentation

To delete the imbalance of the data set, augmentation was used to reach a 50-50 ratio between cases with good and bad outcomes. For each k-fold, the training sets were augmented, but the validation set was not.

For a case with bad outcome, it was checked if new segments could be extracted, see subsection 3.1.2 Segmentation. For a sample, the number of possible new segments were checked, including shifts. If the number of possibilities were the same as the number of wanted additional samples, all possibilities were used. If there were more possibilities than wanted, the selection was randomized. If there were less possibilities than wanted, all possibilities were evenly added to reach wanted number. Say that it was desired to augment by adding a sample 20 times. If this sample had one new segment, but no shifts, the original sample was used 10 times and the new segment was used 10 times.

When no possible shift or new segment was found in cases with bad outcome, the same segment was used multiple times. To not add the same data, and make the classifiers biased, noise was added to all features. The noise was derived from 10% of the mean of the feature, which was then multiplied with a randomly drawn number from the distribution $\mathcal{N}(0, 0.2)$. The reasoning behind this was to decrease the probability of adding a high valued noise. All samples that had a good outcome also got noise added to their features. This was done because the noise should not influence the classifiers.

Lastly, the features X were normalized to be within the range 0 to 1, using the equation

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}. \tag{11}$$

The scaling was done to keep all features within the same range, since the classifiers are distance based algorithm, and each feature should contribute proportionately.

3.3.2 Evaluation

The models made predictions on the validation data set which was not augmented. This procedure was repeated such that all folds were the validation set once. To compare all results between the different approaches, all classifiers were compared to each other with respect to their F1 score. The

confusion matrix along with accuracy, sensitivity and specificity were also derived, but the F_1-score tells how good the quality of the predictions are which is why it was chosen to be the main comparison parameter. True positive was defined as case with bad outcome that was predicted as such.

Algorithm 1 summarizes the approach previously explained.

Algorithm 1

1. Split training observations into K folds. For each $k = 1, ..., K$:
   (a) Data set augmentation on all but the kth fold of the training data.
   (b) Train models using the augmented folds.
   (c) Predict using the left-out kth fold as validation data.
2. Compute the mean and the standard deviation for accuracy, sensitivity, specificity and F_1-score.

The first step was to derive a threshold to be able to evaluate if a feature would improve the result or not. All models, ANN, DT, KNN, and SVM were trained using the gestational age as single feature. Then each presented feature were used alongside the gestational age to see if it would improve the results for the classifiers. It was then tested to use all presented features in the models, followed by a test using all features but the gestational age. Lastly, the best performing models with features, for each case, were trained using all training data, and then got to predict on the test data.

4 Results

In this section the results are presented. The section is split into cases, i.e., case 1 and case 2. For each case, all described features were used to train different classifiers. A positive prediction is classifying a sample as a case with bad outcome, i.e., Apgar Score will be lower than 7, and therefore a negative prediction is classifying a sample as a case with good outcome, i.e., Apgar Score will be equal to or higher than 7. The results are presented in tables. Each table contain the mean and standard deviation of accuracy, sensitivity, specificity, and F_1-score from all predictions done on the validation set. To summarize and compare the results, the F_1-scores for all features are presented in a common table. The presented results are briefly discussed to make it easier for the reader to compare tables for each classifier. Further discussion of the results will be in the next section.

Confusion matrices for all results can be found in Appendix . They were excluded from this section to keep it comprehensible and simple.

Firstly, the results, when using gestational age as single feature, are presented, followed by results when using gestational age together with another derived feature. Then the models were trained on all derived features, followed by using all, but the gestational age, as features.

At the end of this section, predictions using the test data are made by using the best model, for each case.

4.1 Classifications for Case 1

Case 1 contained all samples with gestational age less than than or equal to 258 days.

4.1.1 Naive Models

In this part, results of the naive classifiers are presented. All classifiers were trained with gestational age as a feature. In Table 1 the mean and standard deviation of accuracy, sensitivity, specificity and F_1 score are presented.

All classifiers got a low mean in sensitivity, i.e., troubles classifying cases with bad outcome. Considering the F_1-score, SVM was the best performing classifier. The KNN was the worst in regards to the F_1-score, though it did have the highest sensitivity. These results were used as thresholds when adding new features to the models.

Table 1: Tables of results from classifiers using different methods (specified in each table), having the gestational age as a feature. Each table shows the mean ± standard deviation of accuracy, sensitivity, specificity, and F_1-score.

a)

SVM	mean ± std
Accuracy	0.8849 ± 0.0113
Sensitivity	0.5164 ± 0.1096
Specificity	0.9181 ± 0.0137
F_1-score	0.4227 ± 0.0621

b)

KNN	mean ± std
Accuracy	0.7786 ± 0.0446
Sensitivity	0.5627 ± 0.1306
Specificity	0.7980 ± 0.0540
F_1-score	0.2962 ± 0.0552

c)

DT	mean ± std
Accuracy	0.8573 ± 0.0303
Sensitivity	0.5308 ± 0.1347
Specificity	0.8867 ± 0.0382
F_1-score	0.3801 ± 0.0710

d)

ANN	mean ± std
Accuracy	0.8072 ± 0.2604
Sensitivity	0.4739 ± 0.1472
Specificity	0.8373 ± 0.2922
F_1-score	0.3686 ± 0.1231

4.1.2 Bradycardia

In addition to the gestational age, all models in this section had presence of bradycardia as a second feature. In Table 2 the mean and standard deviation of accuracy, sensitivity, specificity, and F_1-score are presented.

Looking in table 2, the added feature made small differences for the SVM, KNN, and DT classifiers. Compared to the naive results, all three classifiers either increased the mean of sensitivity, and decreased the mean of specificity, or the other way around. The ANN classifier suffered from many false positives, meaning that the added feature did not provide useful information.

Table 2: Tables of results from classifiers using different methods (specified in each table), having the gestational age and bradycardia as features. Each tables shows the mean ± standard deviation of accuracy, sensitivity, specificity, and F_1-score.

a)

SVM	mean ± std
Accuracy	0.8781 ± 0.0224
Sensitivity	0.5235 ± 0.1014
Specificity	0.9101 ± 0.0262
F_1-score	0.4156 ± 0.0607

b)

KNN	mean ± std
Accuracy	0.7327 ± 0.0856
Sensitivity	0.5659 ± 0.1045
Specificity	0.7478 ± 0.0927
F_1-score	0.2705 ± 0.0681

c)

DT	mean ± std
Accuracy	0.8637 ± 0.0261
Sensitivity	0.5090 ± 0.0847
Specificity	0.8956 ± 0.0279
F_1-score	0.3842 ± 0.0623

d)

ANN	mean ± std
Accuracy	0.4194 ± 0.2464
Sensitivity	0.6107 ± 0.3196
Specificity	0.4026 ± 0.2967
F_1-score	0.1342 ± 0.0514

4.1.3 Tachycardia

In addition to the gestational age, all models in this section also had presence of tachycardia as a second feature. In Table 3 the mean and standard deviation of accuracy, sensitivity, specificity, and F_1-score are presented.

The derived results in Table 3 does not present improvement compared to the naive results in Table 1. Comparing the classifiers between each other, one can see that the KNN classifier had the highest number of false positives, and therefore the lowest specificity which this classifier also had when only using one feature. Decision tree also increased in sensitivity, but lowered its specificity, while the SVM did the opposite. The SVM classifier outperforms the other techniques when comparing the F_1-score, and it also did a minor improvement from the naive classification. The ANN classifier got the lowest F_1-score.

Table 3: Tables of results from classifiers using different methods (specified in each table), having the gestational age and tachycardia as features. Each table shows the mean ± standard deviation of accuracy, sensitivity, specificity, and F_1-score.

a)

SVM	mean ± std
Accuracy	0.8866 ± 0.0191
Sensitivity	0.5055 ± 0.0723
Specificity	0.9209 ± 0.0177
F_1-score	0.4262 ± 0.0683

b)

KNN	mean ± std
Accuracy	0.5962 ± 0.0422
Sensitivity	0.6298 ± 0.0722
Specificity	0.5931 ± 0.0439
F_1-score	0.2059 ± 0.0290

c)

DT	mean ± std
Accuracy	0.7436 ± 0.0494
Sensitivity	0.5483 ± 0.0952
Specificity	0.7612 ± 0.0493
F_1-score	0.2653 ± 0.0603

d)

ANN	mean ± std
Accuracy	0.3592 ± 0.2205
Sensitivity	0.6416 ± 0.3006
Specificity	0.3340 ± 0.2658
F_1-score	0.1374 ± 0.0231

4.1.4 Number of Accelerations

In addition to the gestational age, all models in this section also had the number of accelerations as a feature. In Table 4 the mean and standard deviation of accuracy, sensitivity, specificity and F_1-score are presented.

The ANN classifier did not find any information using this feature, hence it was disregarded. Considering the F_1-score, in Table 4, both KNN and DT made improvements, mostly due to finding true negatives better. The SVM classifier still had a specificity of 91%, as it had for the naive case, making a good separation of true negatives and false positives. The ANN classifier did not improve.

Table 4: Tables of results from classifiers using different methods (specified in each table), having the gestational age and the number of accelerations as features. Each table shows the mean ± standard deviation of accuracy, sensitivity, specificity, and F_1-score.

a)

SVM	mean ± std
Accuracy	0.8813 ± 0.0144
Sensitivity	0.5200 ± 0.0928
Specificity	0.9139 ± 0.0130
F_1-score	0.4191 ± 0.0645

b)

KNN	mean ± std
Accuracy	0.7042 ± 0.0329
Sensitivity	0.5408 ± 0.0936
Specificity	0.7190 ± 0.0345
F_1-score	0.2329 ± 0.0428

c)

DT	mean ± std
Accuracy	0.7284 ± 0.0295
Sensitivity	0.5305 ± 0.0831
Specificity	0.7462 ± 0.0347
F_1-score	0.2439 ± 0.0304

d)

ANN	mean ± std
Accuracy	0.4324 ± 0.1448
Sensitivity	0.6155 ± 0.1943
Specificity	0.4159 ± 0.1714
F_1-score	0.1517 ± 0.0268

4.1.5 Number of Decelerations

In addition to the gestational age, all models in this section also had the number of decelerations as a feature. In Table 5 the mean and standard deviation of accuracy, sensitivity, specificity and F_1-score are presented. Adding this feature did not improve the results for any classifier, and SVM was still the best performing one.

Table 5: Tables of results from classifiers using different methods (specified in each table), having the gestational age and number of decelerations as features. Each table shows the mean ± standard deviation of accuracy, sensitivity, specificity, and F_1-score.

a)

SVM	mean ± std
Accuracy	0.8769 ± 0.0222
Sensitivity	0.5299 ± 0.0688
Specificity	0.9081 ± 0.0220
F_1-score	0.4185 ± 0.0657

b)

KNN	mean ± std
Accuracy	0.7204 ± 0.0212
Sensitivity	0.5619 ± 0.0628
Specificity	0.7346 ± 0.0260
F_1-score	0.2490 ± 0.0207

c)

DT	mean ± std
Accuracy	0.7292 ± 0.0450
Sensitivity	0.5478 ± 0.0993
Specificity	0.7455 ± 0.0510
F_1-score	0.2524 ± 0.0476

d)

ANN	mean ± std
Accuracy	0.3804 ± 0.2516
Sensitivity	0.6549 ± 0.2860
Specificity	0.3557 ± 0.2982
F_1-score	0.1413 ± 0.0400

4.1.6 Slope of Fitted Line

In addition to the gestational age, all models in this section also had the slope of fitted line as a feature. In Table 6 the mean and standard deviation of accuracy, sensitivity, specificity and F_1-score are presented. The tables show a minor improvement in F_1-score for the SVM, but not for the KNN

and DT. The ANN classifier got a higher accuracy than 50% which had not happened since the naive model, but it still had the lowest F_1-score.

Table 6: Tables of results from classifiers using different methods (specified in each table), and the gestational age and slope of fitted line as features. Each table shows mean ± standard deviation for accuracy, sensitivity, specificity, and F_1-score.

a)

SVM	mean ± std
Accuracy	0.8852 ± 0.0134
Sensitivity	0.5094 ± 0.0616
Specificity	0.9190 ± 0.0134
F_1-score	0.4232 ± 0.0472

b)

KNN	mean ± std
Accuracy	0.7145 ± 0.0275
Sensitivity	0.5198 ± 0.0937
Specificity	0.7321 ± 0.0298
F_1-score	0.2312 ± 0.0364

c)

DT	mean ± std
Accuracy	0.7721 ± 0.0307
Sensitivity	0.5164 ± 0.0928
Specificity	0.7951 ± 0.0381
F_1-score	0.2718 ± 0.0359

d)

ANN	mean ± std
Accuracy	0.5655 ± 0.2285
Sensitivity	0.4357 ± 0.2930
Specificity	0.5768 ± 0.2757
F_1-score	0.1213 ± 0.0540

4.1.7 Short Term Variability

In addition to the gestational age, all models in this section also had the mean absolute deviation, MAD, of the short term variability as a feature. In Table 7 the mean and standard deviation of accuracy, sensitivity, specificity and F_1-score are presented. This feature did not provide any further information when trying to separate the outcomes.

Table 7: Tables of results from classifiers using different methods (specified in each table), having the gestational age and MAD of the short term variability as features. Each table shows the mean ± standard deviation of accuracy, sensitivity, specificity, and F_1-score.

a)

SVM	mean ± std
Accuracy	0.8796 ± 0.0199
Sensitivity	0.5201 ± 0.0869
Specificity	0.9120 ± 0.0227
F_1-score	0.4170 ± 0.0569

b)

KNN	mean ± std
Accuracy	0.6837 ± 0.0382
Sensitivity	0.5020 ± 0.0695
Specificity	0.7001 ± 0.0420
F_1-score	0.2086 ± 0.0307

c)

DT	mean ± std
Accuracy	0.7380 ± 0.0190
Sensitivity	0.4628 ± 0.0545
Specificity	0.7628 ± 0.0206
F_1-score	0.2260 ± 0.0253

d)

ANN	mean ± std
Accuracy	0.4102 ± 0.2430
Sensitivity	0.6195 ± 0.3007
Specificity	0.3914 ± 0.2912
F_1-score	0.1335 ± 0.0490

4.1.8 Interval Index

In addition to the gestational age, all models in this section also had the mean absolute deviation of the interval index as a feature. In Table 8 the mean and standard deviation of accuracy, sensitivity, specificity and F_1-score are presented. Having this as an additional feature, did not improve any result comparing to the naive models.

Table 8: Tables of results from classifiers using different methods (specified in each table), having the gestational age and MAD of the interval index as features. Each table shows mean ± standard deviation for accuracy, sensitivity, specificity, and F_1-score.

a)

SVM	mean ± std
Accuracy	0.8840 ± 0.0130
Sensitivity	0.5160 ± 0.0696
Specificity	0.9171 ± 0.0132
F_1-score	0.4234 ± 0.0513

b)

KNN	mean ± std
Accuracy	0.6978 ± 0.0226
Sensitivity	0.5586 ± 0.0959
Specificity	0.7103 ± 0.0240
F_1-score	0.2336 ± 0.0375

c)

DT	mean ± std
Accuracy	0.7527 ± 0.0216
Sensitivity	0.5473 ± 0.1097
Specificity	0.7711 ± 0.0297
F_1-score	0.2661 ± 0.0352

d)

ANN	mean ± std
Accuracy	0.3233 ± 0.3086
Sensitivity	0.7112 ± 0.3750
Specificity	0.2883 ± 0.3699
F_1-score	0.1312 ± 0.0497

4.1.9 Mean Absolute Deviation

In addition to the gestational age, all models in this section also had the mean absolute deviation of the interpolated FHR signal as a second feature. In Table 9 the mean and standard deviation of accuracy, sensitivity, specificity and F_1-score are presented. No improvements were made from adding this feature. The SVM classifier showed the least deterioration, while both the KNN and DT models had higher sensitivity, however they had lower specificity. The ANN was, yet again, excluded from discussion due to bad results.

Table 9: Tables of results from classifiers using different methods (specified in each table), having the gestational age and MAD of the FHR as features. Each table shows mean ± standard deviation for accuracy, sensitivity, specificity, and F_1-score.

a)

SVM	mean ± std
Accuracy	0.8834 ± 0.0133
Sensitivity	0.5160 ± 0.0696
Specificity	0.9165 ± 0.0129
F_1-score	0.4224 ± 0.0539

b)

KNN	mean ± std
Accuracy	0.6840 ± 0.0310
Sensitivity	0.5377 ± 0.0641
Specificity	0.6972 ± 0.0328
F_1-score	0.2201 ± 0.0291

c)

DT	mean ± std
Accuracy	0.7081 ± 0.0479
Sensitivity	0.5518 ± 0.0560
Specificity	0.7222 ± 0.0540
F_1-score	0.2403 ± 0.0310

d)

ANN	mean ± std
Accuracy	0.3418 ± 0.3493
Sensitivity	0.6823 ± 0.4208
Specificity	0.3111 ± 0.4184
F_1-score	0.1218 ± 0.0558

4.1.10 All Features

In this section all features were used when training the classifiers. In Table 10 the mean and standard deviation of accuracy, sensitivity, specificity and F_1-score are presented. The SVM still outperforms all the other classifiers, even if it got lower results than the naive SVM model.

Table 10: Tables of results from classifiers using different methods (specified in each table), using all derived features. Each table shows mean ± standard deviation for accuracy, sensitivity, specificity, and F_1-score.

a)

SVM	mean ± std
Accuracy	0.8758 ± 0.0111
Sensitivity	0.5337 ± 0.0706
Specificity	0.9065 ± 0.0122
F_1-score	0.4145 ± 0.0435

b)

KNN	mean ± std
Accuracy	0.7903 ± 0.0279
Sensitivity	0.4021 ± 0.0856
Specificity	0.8252 ± 0.0266
F_1-score	0.2423 ± 0.0557

c)

DT	mean ± std
Accuracy	0.7971 ± 0.0237
Sensitivity	0.4879 ± 0.1080
Specificity	0.8249 ± 0.0232
F_1-score	0.2842 ± 0.0600

d)

ANN	mean ± std
Accuracy	0.4871 ± 0.1643
Sensitivity	0.5261 ± 0.2198
Specificity	0.4835 ± 0.1978
F_1-score	0.1420 ± 0.0237

4.1.11 All but One Feature

In this section all features, but the gestational age, were used when training the classifiers. In Table 11 the mean and standard deviation of accuracy, sensitivity, specificity and F_1-score are presented. Comparing these results to the naive models shows that these models performed worse, every single F_1-score decreased, most notably for the SVM classifier. These results, compared to the ones derived when training on all features, also shows a big decrease in the F_1-score. The ANN classifier seems to

be affected the least when not having the gestational age as a feature, though it still had very high standard deviation compared to the other classifiers.

Table 11: Tables of results from classifiers using different methods (specified in each table), using all derived features, but the gestational age. Each table shows the mean ± standard deviation of accuracy, sensitivity, specificity, and F_1-score.

a)

SVM	mean ± std
Accuracy	0.7316 ± 0.0153
Sensitivity	0.3450 ± 0.1273
Specificity	0.7663 ± 0.0214
F_1-score	0.1716 ± 0.0599

b)

KNN	mean ± std
Accuracy	0.7662 ± 0.0262
Sensitivity	0.2064 ± 0.0602
Specificity	0.8166 ± 0.0296
F_1-score	0.1270 ± 0.0368

c)

DT	mean ± std
Accuracy	0.7477 ± 0.0236
Sensitivity	0.2494 ± 0.0615
Specificity	0.7926 ± 0.0259
F_1-score	0.1401 ± 0.0328

d)

ANN	mean ± std
Accuracy	0.4937 ± 0.1653
Sensitivity	0.4990 ± 0.1961
Specificity	0.4933 ± 0.1963
F_1-score	0.1377 ± 0.0229

4.1.12 Comparison: Features vs Features and Classifiers vs Classifiers

Table 12 shows the F_1-score for all different classifier techniques with respective features used. Note the lower mean for all classifiers when excluding the gestational age. The highlighted value was deemed to be the best classifier and best combination of features, due to having the highest mean.

Table 12: Table of F1-score for all classifiers with respective features.

F_1-score	**SVM**	**KNN**	**Decision Tree**	**ANN**
***Gestational Age**	0.4227 ± 0.0621	0.2962 ± 0.0552	0.3801 ± 0.0710	0.3686 ± 0.1231
Bradycardia & *	0.4156 ± 0.0607	0.2705 ± 0.0681	0.3842 ± 0.0623	0.1342 ± 0.0514
Tachycardia & *	0.4262 ± 0.0683	0.2059 ± 0.0290	0.2653 ± 0.0603	0.1374 ± 0.0231
#Accelerations & *	0.4191 ± 0.0645	0.2329 ± 0.0428	0.2439 ± 0.0304	0.1517 ± 0.0268
#Decelerations & *	0.4185 ± 0.0657	0.2490 ± 0.0207	0.2524 ± 0.0476	0.1413 ± 0.0400
Slope & *	0.4232 ± 0.0472	0.2312 ± 0.0364	0.2718 ± 0.0359	0.1213 ± 0.0540
STV & *	0.4170 ± 0.0569	0.2086 ± 0.0307	0.2260 ± 0.0253	0.1335 ± 0.0490
II & *	0.4234 ± 0.0513	0.2336 ± 0.0375	0.2661 ± 0.0352	0.1312 ± 0.0497
MAD & *	0.4224 ± 0.0539	0.2201 ± 0.0291	0.2403 ± 0.0310	0.1218 ± 0.0558
All Features	0.4145 ± 0.0435	0.2423 ± 0.0557	0.2842 ± 0.0600	0.1420 ± 0.0237
All Features excl. *	0.1716 ± 0.0599	0.1270 ± 0.0368	0.1401 ± 0.0328	0.1377 ± 0.0229

4.2 Classifications for Case 2

Case 2 contained all samples with gestational age larger than or equal to 259 days.

4.2.1 Naive Models

In this part results of the naive classifiers are presented. All classifiers were trained with the gestational age as a feature. In Table 13 the mean and standard deviation of accuracy, sensitivity, specificity and F1 score are presented. Table 13 shows that the F_1-scores were very low, implying that the separation between the outcomes were not enough when only using this feature. These results were used as threshold when training new models using added features.

Table 13: Tables of results from classifiers using different methods (specified in each table), having the gestational age as feature. Each table shows the mean ± standard deviation of accuracy, sensitivity, specificity, and F_1-score.

a)

SVM	mean ± std
Accuracy	0.8548 ± 0.0094
Sensitivity	0.2348 ± 0.0457
Specificity	0.8613 ± 0.0094
F_1-score	0.0328 ± 0.0068

b)

KNN	mean ± std
Accuracy	0.5345 ± 0.1084
Sensitivity	0.4786 ± 0.1009
Specificity	0.5351 ± 0.1104
F_1-score	0.0212 ± 0.0032

c)

DT	mean ± std
Accuracy	0.6662 ± 0.0364
Sensitivity	0.3194 ± 0.0719
Specificity	0.6699 ± 0.0372
F_1-score	0.0196 ± 0.0042

d)

ANN	mean ± std
Accuracy	0.7962 ± 0.1400
Sensitivity	0.2209 ± 0.0805
Specificity	0.8023 ± 0.1422
F_1-score	0.0266 ± 0.0100

4.2.2 Bradycardia

In addition to the gestational age, all models in this section also had presence of bradycardia as a feature. In Table 14 the mean and standard deviation of accuracy, sensitivity, specificity and F_1-score are presented. No improvement can be seen for any of the classifiers.

Table 14: Tables of results from classifiers using different methods (specified in each table), having the gestational age and bradycardia as features. Each table shows the mean ± standard deviation of accuracy, sensitivity, specificity, and F_1-score.

a)

SVM	mean ± std
Accuracy	0.8847 ± 0.0258
Sensitivity	0.1882 ± 0.0797
Specificity	0.8921 ± 0.0268
F_1-score	0.0321 ± 0.0079

b)

KNN	mean ± std
Accuracy	0.4508 ± 0.0633
Sensitivity	0.5283 ± 0.1028
Specificity	0.4500 ± 0.0647
F_1-score	0.0197 ± 0.0032

c)

DT	mean ± std
Accuracy	0.6443 ± 0.0473
Sensitivity	0.3421 ± 0.0714
Specificity	0.6474 ± 0.0483
F_1-score	0.0197 ± 0.0032

d)

ANN	mean ± std
Accuracy	0.4489 ± 0.2339
Sensitivity	0.5542 ± 0.2451
Specificity	0.4478 ± 0.2389
F_1-score	0.0204 ± 0.0023

4.2.3 Tachycardia

In addition to the gestational age, all models in this section also had presence of tachycardia as a feature. In Table 15 the mean and standard deviation of accuracy, sensitivity, specificity, and F_1-score are presented. The results shows no overall improvements.

Table 15: Tables of results from classifiers using different methods (specified in each table), and the gestational age and tachycardia as features. Each table shows the mean ± standard deviation of accuracy, sensitivity, specificity, and F_1-score.

a)

SVM	mean ± std
Accuracy	0.8888 ± 0.0038
Sensitivity	0.1883 ± 0.0459
Specificity	0.8962 ± 0.0038
F_1-score	0.0341 ± 0.0082

b)

KNN	mean ± std
Accuracy	0.5107 ± 0.0633
Sensitivity	0.5475 ± 0.1034
Specificity	0.5103 ± 0.0646
F_1-score	0.0229 ± 0.0035

c)

DT	mean ± std
Accuracy	0.6523 ± 0.0201
Sensitivity	0.4111 ± 0.0760
Specificity	0.6548 ± 0.0205
F_1-score	0.0241 ± 0.0042

d)

ANN	mean ± std
Accuracy	0.5068 ± 0.2541
Sensitivity	0.4785 ± 0.2641
Specificity	0.5071 ± 0.2596
F_1-score	0.0194 ± 0.0032

4.2.4 Number of Accelerations

In addition to the gestational age, all models in this section also had the number of accelerations as a feature. In Table 16 the mean and standard deviation of accuracy, sensitivity, specificity, and F_1-score are presented. No improvements can be seen compared to the naive models. Note that for the ANN classifier, the standard deviation of the accuracy, sensitivity and specificity increased,

meaning that the predictions vary much and the classifier could not find information to separate the outcomes.

Table 16: Tables of results from classifiers using different methods (specified in each table), having the gestational age and the number of accelerations as features. Each table shows the mean ± standard deviation of accuracy, sensitivity, specificity, and F_1-score.

a)

SVM	mean ± std
Accuracy	0.7966 ± 0.0591
Sensitivity	0.3057 ± 0.1195
Specificity	0.8018 ± 0.0609
F_1-score	0.0300 ± 0.0047

b)

KNN	mean ± std
Accuracy	0.6351 ± 0.0108
Sensitivity	0.3748 ± 0.0526
Specificity	0.6378 ± 0.0107
F_1-score	0.0210 ± 0.0033

c)

DT	mean ± std
Accuracy	0.6258 ± 0.0082
Sensitivity	0.4128 ± 0.0749
Specificity	0.6281 ± 0.0082
F_1-score	0.0225 ± 0.0041

d)

ANN	mean ± std
Accuracy	0.5183 ± 0.2558
Sensitivity	0.4706 ± 0.2538
Specificity	0.5189 ± 0.2612
F_1-score	0.0196 ± 0.0026

4.2.5 Number of Decelerations

In addition to the gestational age, all models in this section also had the number of decelerations as a feature. In Table 17 the mean and standard deviation of accuracy, sensitivity, specificity and F_1-score are presented. Comparing the classifiers to each other, not including ANN due to bad results, the SVM performed best. Comparing to the naive models, the SVM increased its sensitivity, but lowered in specificity, the same happened for the DT, while the KNN did the opposite.

Table 17: Tables of results from classifiers using different methods (specified in each table), having the gestational age and the number of decelerations as features. Each table shows the mean ± standard deviation of accuracy, sensitivity, specificity, and F_1-score.

a)

SVM	mean ± std
Accuracy	0.7622 ± 0.0266
Sensitivity	0.3504 ± 0.0706
Specificity	0.7666 ± 0.0274
F_1-score	0.0298 ± 0.0045

b)

KNN	mean ± std
Accuracy	0.6327 ± 0.0082
Sensitivity	0.4094 ± 0.0584
Specificity	0.6350 ± 0.0083
F_1-score	0.0227 ± 0.0033

c)

DT	mean ± std
Accuracy	0.6172 ± 0.0065
Sensitivity	0.3886 ± 0.0595
Specificity	0.6196 ± 0.0069
F_1-score	0.0207 ± 0.0030

d)

ANN	mean ± std
Accuracy	0.3801 ± 0.1423
Sensitivity	0.5677 ± 0.1531
Specificity	0.3781 ± 0.1453
F_1-score	0.0187 ± 0.0019

4.2.6 Slope of Fitted Line

In addition to the gestational age, all models in this section also had the slope of fitted line as a feature. In Table 18 the mean and standard deviation of accuracy, sensitivity, specificity and F_1-score are presented. The ANN classifier still had the lowest F_1-score. The other classifiers got better at finding the true negatives than the naive models did which shows in their increased specificity.

Table 18: Tables of results from classifiers using different methods (specified in each table), having the gestational age and slope of fitted line as features. Each table shows mean ± standard deviation for accuracy, sensitivity, specificity, and F_1-score.

a)

SVM	mean ± std
Accuracy	0.9146 ± 0.0170
Sensitivity	0.1554 ± 0.0607
Specificity	0.9226 ± 0.0176
F_1-score	0.0359 ± 0.0095

b)

KNN	mean ± std
Accuracy	0.6686 ± 0.0145
Sensitivity	0.3506 ± 0.0640
Specificity	0.6720 ± 0.0144
F_1-score	0.0217 ± 0.0044

c)

DT	mean ± std
Accuracy	0.7353 ± 0.0146
Sensitivity	0.2782 ± 0.0632
Specificity	0.7401 ± 0.0151
F_1-score	0.0214 ± 0.0043

d)

ANN	mean ± std
Accuracy	0.4764 ± 0.3303
Sensitivity	0.5203 ± 0.3332
Specificity	0.4760 ± 0.3373
F_1-score	0.0196 ± 0.0040

4.2.7 Short Term Variability

In addition to the gestational age, all models in this section also had the mean absolute deviation, MAD, of the short term variability as a feature. In Table 19 the mean and standard deviation of accuracy, sensitivity, specificity, and F_1-score are presented. The SVM, KNN, and DT classifiers improved their results, but all got a lower sensitivity than the naive models. The ANN was excluded from comparison due to low means and high standard deviations.

Table 19: Tables of results from classifiers using different methods (specified in each table), having the gestational age and MAD of the short term variability as features. Each table shows the mean ± standard deviation of accuracy, sensitivity, specificity, and F_1-score.

a)

SVM	mean ± std
Accuracy	0.8587 ± 0.0184
Sensitivity	0.2090 ± 0.0426
Specificity	0.8655 ± 0.0187
F_1-score	0.0302 ± 0.0062

b)

KNN	mean ± std
Accuracy	0.6157 ± 0.0119
Sensitivity	0.4543 ± 0.0498
Specificity	0.6174 ± 0.0118
F_1-score	0.0241 ± 0.0029

c)

DT	mean ± std
Accuracy	0.7497 ± 0.0109
Sensitivity	0.2971 ± 0.0426
Specificity	0.7545 ± 0.0109
F_1-score	0.0242 ± 0.0038

d)

ANN	mean ± std
Accuracy	0.4022 ± 0.2510
Sensitivity	0.5922 ± 0.2381
Specificity	0.4002 ± 0.2561
F_1-score	0.0204 ± 0.0021

4.2.8 Interval Index

In addition to the gestational age, all models in this section also had the mean absolute deviation, MAD, of the interval index as a feature. In Table 20 the mean and standard deviation of accuracy, sensitivity, specificity and F1 score are presented. The SVM classifier got the best F_1-score. The ANN had the highest sensitivity rate, but suffered from many false positives, and high standard deviations.

Table 20: Tables of results from classifiers using different methods (specified in each table), having the gestational age and MAD of the interval index as features. Each table shows the mean ± standard deviation of accuracy, sensitivity, specificity, and F_1-score.

a)

SVM	mean ± std
Accuracy	0.8188 ± 0.0560
Sensitivity	0.2797 ± 0.1094
Specificity	0.8245 ± 0.8245
F_1-score	0.0308 ± 0.0054

b)

KNN	mean ± std
Accuracy	0.5860 ± 0.0133
Sensitivity	0.4266 ± 0.0582
Specificity	0.5877 ± 0.0133
F_1-score	0.0211 ± 0.0030

c)

DT	mean ± std
Accuracy	0.7642 ± 0.0314
Sensitivity	0.2679 ± 0.0735
Specificity	0.7695 ± 0.0322
F_1-score	0.0231 ± 0.0046

d)

ANN	mean ± std
Accuracy	0.5315 ± 0.2453
Sensitivity	0.4830 ± 0.2785
Specificity	0.5320 ± 0.2507
F_1-score	0.0184 ± 0.0075

4.2.9 Mean Absolute Deviation

In addition to the gestational age, all models in this section also had the mean absolute deviation, MAD, of the interpolated FHR signal as a feature. In Table 21 the mean and standard deviation

of accuracy, sensitivity, specificity, and F_1-score are presented. The SVM classifier was still best performing, even if it got lower sensitivity than the naive SVM did. The KNN and DT also improved compared to their naive versions, where in this case they got a higher specificity. The ANN classifier, yet again, got high standard deviations.

Table 21: Tables of results from classifiers using different methods (specified in each table), having the gestational age, and MAD of the FHR as features. Each table shows the mean ± standard deviation of accuracy, sensitivity, specificity, and F_1-score.

a)

SVM	mean ± std
Accuracy	0.8829 ± 0.0293
Sensitivity	0.1849 ± 0.0673
Specificity	0.8902 ± 0.0300
F_1-score	0.0318 ± 0.0094

b)

KNN	mean ± std
Accuracy	0.6059 ± 0.0098
Sensitivity	0.4008 ± 0.0813
Specificity	0.6081 ± 0.0102
F_1-score	0.0208 ± 0.0041

c)

DT	mean ± std
Accuracy	0.7367 ± 0.0125
Sensitivity	0.2866 ± 0.0671
Specificity	0.7415 ± 0.0129
F_1-score	0.0222 ± 0.0048

c)

ANN	mean ± std
Accuracy	0.5844 ± 0.1838
Sensitivity	0.4334 ± 0.2151
Specificity	0.5860 ± 0.1879
F_1-score	0.0191 ± 0.0073

4.2.10 All Features

In this section all features were used when training the classifiers. In Table 22 the mean and standard deviation of accuracy, sensitivity, specificity and F_1-score are presented. The ANN got the best sensitivity, even if it was worse than the naive ANN results. The SVM, KNN and DT increased compared to their naive models, and SVM was the best performing with this combination of features.

Table 22: Tables of results from classifiers using different methods (specified in each table), having all derived features. Each table shows the mean ± standard deviation of accuracy, sensitivity, specificity, and F_1-score.

a)

SVM	mean ± std
Accuracy	0.8887 ± 0.0034
Sensitivity	0.1882 ± 0.0431
Specificity	0.8961 ± 0.0035
F_1-score	0.0341 ± 0.0076

b)

KNN	mean ± std
Accuracy	0.7108 ± 0.0298
Sensitivity	0.3073 ± 0.0506
Specificity	0.7151 ± 0.0301
F_1-score	0.0220 ± 0.0047

c)

DT	mean ± std
Accuracy	0.8622 ± 0.0205
Sensitivity	0.1640 ± 0.0526
Specificity	0.8696 ± 0.0208
F_1-score	0.0246 ± 0.0082

d)

ANN	mean ± std
Accuracy	0.4653 ± 0.1198
Sensitivity	0.5475 ± 0.1142
Specificity	0.4645 ± 0.1222
F_1-score	0.0211 ± 0.0016

4.2.11 All but One Feature

In this section all features, but the gestational age, were used when training the classifiers. In Table 23 the mean and standard deviation of accuracy, sensitivity, specificity and F_1-score are presented. Comparing the classifiers shows that SVM, KNN, and DT got similar accuracy, where SVM had a better sensitivity. The ANN classifiers had the highest sensitivity, but the simultaneously the lowest specificity. When comparing to the results using all features, it should be noted that the SVM classifier did not show any differences at all when classifying the cases with bad outcome.

Table 23: Tables of results from classifiers using different methods (specified in each table), utilizing all derived features, but the gestational age. Each table shows the mean ± standard deviation of accuracy, sensitivity, specificity, and F1-score.

a)

SVM	mean ± std
Accuracy	0.8888 ± 0.0041
Sensitivity	0.1883 ± 0.0418
Specificity	0.8962 ± 0.0043
F1 Score	0.0341 ± 0.0071

b)

KNN	mean ± std
Accuracy	0.8941 ± 0.0069
Sensitivity	0.1175 ± 0.0114
Specificity	0.9023 ± 0.0070
F1 Score	0.0227 ± 0.0027

c)

DT	mean ± std
Accuracy	0.8908 ± 0.0075
Sensitivity	0.1364 ± 0.0348
Specificity	0.8987 ± 0.0076
F1 Score	0.0255 ± 0.0067

c)

ANN	mean ± std
Accuracy	0.5759 ± 0.1510
Sensitivity	0.4282 ± 0.1825
Specificity	0.5775 ± 0.1544
F1 Score	0.0207 ± 0.0044

4.2.12 Comparison: Features vs Features and Classifiers vs Classifiers

In Table 24, the F_1-scores for all different techniques, with varying features, are presented. For every combination of features, the F_1-score had been low throughout. Note the small difference when excluding the gestational age as feature. The result highlighted in yellow, was deemed the best performing model and feature combination, due to having the highest mean.

Table 24: Table of F_1-score for all classifiers with respective features.

F_1-score	**SVM**	**KNN**	**Decision Tree**	**ANN**
***Gestational Age**	0.0328 ± 0.0068	0.0212 ± 0.0032	0.0196 ± 0.0042	0.0266 ± 0.0100
Bradycardia & *	0.0321 ± 0.0079	0.0197 ± 0.0032	0.0197 ± 0.0032	0.0204 ± 0.0023
Tachycardia & *	0.0341 ± 0.0082	0.0229 ± 0.0035	0.0241 ± 0.0042	0.0194 ± 0.0032
#Accelerations & *	0.0300 ± 0.0047	0.0210 ± 0.0033	0.0225 ± 0.0041	0.0196 ± 0.0026
#Decelerations & *	0.0298 ± 0.0045	0.0227 ± 0.0033	0.0207 ± 0.0030	0.0187 ± 0.0019
Slope & *	0.0359 ± 0.0095	0.0217 ± 0.0044	0.0214 ± 0.0043	0.0196 ± 0.0040
STV & *	0.0302 ± 0.0062	0.0241 ± 0.0029	0.0242 ± 0.0038	0.0204 ± 0.0021
II & *	0.0308 ± 0.0054	0.0211 ± 0.0030	0.0231 ± 0.0046	0.0184 ± 0.0075
MAD & *	0.0318 ± 0.0094	0.0208 ± 0.0041	0.0222 ± 0.0048	0.0191 ± 0.0073
All Features	0.0341 ± 0.0076	0.0220 ± 0.0047	0.0246 ± 0.0082	0.0211 ± 0.0016
All Features excl. *	0.0341 ± 0.0071	0.0227 ± 0.0027	0.0255 ± 0.0067	0.0207 ± 0.0044

4.3 Final Models

In this section, the final result are presented. For each case, the best performing model are trains using the best performing combination of features.

4.3.1 Case 1

For the antepartum cases, the best performing was SVM using gestational age and tachycardia as features. The results of predicting on the test data is shown in table 25.

Table 25: Results from final SVM model using gestational age and tachycardia as features.

SVM	
Accuracy	0.8873
Sensitivity	0.5758
Specificity	0.9190
F1 Score	0.4856

4.3.2 Case 2

For the full term cases, the best performing was SVM using gestational age and the slope of fitted line from moving average of the FHR as features. The results of predicting on the test data is shown in table 26.

Table 26: Results from final SVM model using gestational age and slope of fitted line as features.

SVM	
Accuracy	0.8130
Sensitivity	0.2980
Specificity	0.8184
F1 Score	0.0318

5 Discussion

This section discusses the approach, derived features and results of the project.

5.1 Segments

The original CTG recordings were cut into 30 minute segments. The selection of using 30 minute was mostly based on the theoretically threshold of *at least* 20 minutes for clinicians when interpreting the CTG. A segment longer than 30 minutes was not selected because it would exclude more samples. However, finding features could be made using shorter or longer segments as well. The theoretical threshold are for clinicians who use medical definitions while non-theoretical features might not be bound to this threshold. No comparison between different length of the segments was made.

Furthermore, the segmentation was selected based on when a signal did not have more than a 5 second time jump between samples. This threshold was selected due to the discontinuity in the time vector, for example the time stamps could be $(0, 0.25, 0.75, 2.5, 2.75, 3)$. There were also cases where the time stamps included a impossible value, for example $(0, 0.25, 0.75, 2500, 1, 1.25)$. Due to this the cases with bad outcomes were checked manually since the middle time stamp is an outlier, and should not lead to a bad case being excluded. However, a 5 seconds threshold could be to small, and increasing this would include more cases with a good outcome. Firstly, a 1 second threshold was chosen, this excluded 13,527 data samples. Comparing this to the 5 seconds threshold which excluded 12,841 samples. The main difference in these exclusions were the number of good cases.

Another threshold for the segments was that the FHR signal had to contain at least 50% non-zero values. This threshold was chosen arbitrarily, and no comparison between different thresholds were made. This may have been too high since the missing data points were afterwards removed with linear interpolation. Though, the cases with bad outcome, that were initially excluded, were all checked manually, which "saved" 11 out of 253, due to having negative time jumps. An increased threshold would include more cases with good outcome, hence a comparison was never made. The negative time jumps were never dealt with for the cases with good outcome either, due to having a much greater number of these than the cases with bad outcome.

5.2 Features

In this subsection, the features are discussed and evaluated. The features, who were derived with the same reasoning, are discussed in the same subsection, since the approach to extract them will be evaluated.

5.2.1 Gestational Age

To have a performance to improve, the gestational age, i.e., the length of the pregnancy in days, was picked as the first feature, due to it being a known factor. The differences in figure 9, showing the histograms, yielded the decision to stratify the data given this feature. Doing so is of utmost importance, since the preterm births have a pathological CTG for other reasons than asphyxia. However, having this as a feature does not add any new information for the medical professionals when evaluating the CTG.

5.2.2 Bradycardia and Tachycardia

To summarize the features bradycardia and tachycardia, none gave the classifiers notable improvement comparing to the naive case. Considering both of these features were based on an estimation of the FHR baseline, with the results at hand, this estimation may not have been good enough. The classification of having bradycardia or tachycardia were checked manually for some CTG. Changing the thresholds could possibly improve the information in these features. For example, increasing the threshold for tachycardia to over 160 bpm instead. Another aspect is that these features are binary, and considering the results, it did not provide enough information to the models, hence it could have been a better idea only using the estimated baseline.

5.2.3 Number of Accelerations and Decelerations

These features were derived in the same manner using the baseline thresholds estimation. Yet again, there might be an underlying problem with the derived baseline thresholds. There is a possibility that the estimated number is far away from the *actual* number of accelerations or decelerations, though it was investigated *if* these features would contribute with any further information to the classifiers, which none of them did.

An attempt to find a more accurate number of accelerations was initially tested. It was made by looking for maximas in the FHR signal. This approach was to arbitrarily since there was no consensus on how close the accelerations could be to one another.

The intention was first to isolate the decelerations in the FHR, and use the derivative to estimate what shape the decelerations had. This would have been applied to the accelerations as well since it *might* have contributed to more information for the classifiers. However, this extraction was never successful. Instead, using the estimated number of accelerations and decelerations were investigated due to observation that for *some* CTG with different outcomes, the cases with bad outcome varied less, i.e., had less accelerations and decelerations.

5.2.4 Slope of Fitted Line

In the FHR signal, deriving the moving average and fitting a line to it using linear regression, aimed to capture if the FHR baseline was increasing or decreasing. This feature did not increase the results much for either classifier. Perhaps it should have been an additional feature when having derived stronger ones.

5.2.5 Short Term Variability and Interval Index

These features were tested due to having a different approach capturing the variability in the FHR signal. Though, the results show that both of them did not contribute to new information. There could be issues taking the mean absolute deviation of the answer to reduce the dimensions. No other technique was tested. It was neither tested to use all 30 values of STV or II. One could also test to take values more or less frequently, compared to taking a value every 10 seconds, to see if this would have made improvement.

5.2.6 Mean Absolute Deviation

Multiple time domain features from the FHR were derived, e.g., RMS, power, kurtosis, and skewness. These are not presented in the report, due to the histograms of P(x|bad outcome) and P(x|good outcome) did not show any particular differences. The mean absolute deviation had somewhat more ”interesting” histograms. The hope was to see a difference in the variability for the different outcomes, but this did not contribute to valuable information to the classifiers. The intention was to also investigate the frequency domain, and use a combination of features from time and frequency domain.

5.2.7 Combination of Features

All Features
Firstly, all mentioned features were used when training the models. For both cases this did not improve the results, implying that some features are contributing with the same, or contradicting information and could possibly only confuse the classifiers.

All but One Feature
This test was made to see how much the information the gestational age carries. Excluding this as a feature showed a big decrease in the results for case 1, i.e., the gestational age carries vital information. Though, it should be questioned using this as a feature for the preterm labours. The clinicians are already aware of the low gestational age and that it is a high risks labour. The CTG might then be abnormal due to other reasons than hypoxia.

For case 2, however, the results improved for the SVM, KNN, and DT classifiers. This case is interesting, since an abnormal CTG would not be connected to the gestational age, but probably the fetus experiencing hypoxia. This concludes that the gestational age should not have been selected as first feature in this case, but only yielding a stratification of the data. Having this as a feature skewed the results, and therefore the found features did not seem to improve anything when testing these alongside the gestational age, while together they were better without the gestational age. Not including the gestational age in the classifiers would also show if the CTG features did yield any information or if some of them have correlating information.

5.2.8 Feature Extraction

This thesis mainly targeted the FHR signal for feature extraction. The features were created using the information on how clinicians interpret CTG, while others were extracted due to differences in the signals for *some* samples with different outcomes. For example, the number of accelerations and decelerations, came from the visual that some cases with good outcome had a variability in the CTG, i.e., more accelerations and decelerations as they were defined in this project. Though even if many different CTG signals were reviewed, these features were not able to distinguish the different outcomes. However, there was no check point to see if these features were somewhat correct, i.e., no comparison to clinical evaluations, something most studies rely on. This raises the question if the features were somewhat similar to what clinicians interpret when looking at the CTG, and if patterns who were considered normal by the models might be considered pathological by clinicians. To increase the certainty of the feature extraction, there could be a comparison made to statements from medical professionals on how they evaluate the CTG.

To only use the FHR signal for feature extraction had the reasoning that sometimes the sensor, that measures the contraction, was removed, perhaps due to being uncomfortable for the mother. The contraction signal was sometimes not present, while there were FHR signal. However, it was never checked if adding a requirement of having the TOCO present excluded a much higher number of samples.

The idea with the features were to test if a machine learning method could extract more information than the human eye can, though the results were disappointing. There are many other ways to extract further features (to be mentioned in the next section), but a main issue is to find those who do not contribute with correlated information.

Another aspect to discuss is that when adding a new feature, mostly it contributed to an increase in sensitivity while lowering the specificity, or the other way around. With the main target to find the true positives, i.e., the cases with bad outcome, should one then allow an increase in the false positives? The introduction of the CTG as a monitoring method already led to an increase of cesarean sections, something that preferably should not be increased when using new technique to extract further information [3].

5.3 Different Models

Testing different techniques had the aim to capture the topology of the distribution in the data, and not rely on one technique.

The SVM classifier ended up being the chosen technique for each case. It was tested due to its robustness and it handled the extracted features best. The KNN and DT classifiers were tested to see if the outcomes were separable, which deeming the overall results, they were not. The ANN was selected to be able to tune the network parameters, though this proved to be unsuccessful. It needed a higher number of features to make better generalization, and therefore predict better.

In order to reach good classification results, select the right model, with parameters that are correctly tuned, and appropriately extracted features, where the last mentioned was the main issue with this thesis. The features did not distinguish the different outcomes enough.

6 Conclusions

6.1 Conclusions

The objective of this master thesis, was to find features that would distinguish cases with good versus bad outcome from one another, and to investigate if said features would increase the correct predictions when compared to a naive model. The found features did not improve the results. Most of the derived features gave a decreasing sensitivity, while for some, simultaneously, increasing the specificity. There was also almost no improvement in the F_1-score, hence the derived features were deemed to not be carrying any valuable information that distinguishes the outcomes.

This project used known clinical fact to create a ground truth to have a reference point. This approach should be changed such that the gestational age could be used as a feature in naive models, but not use this in addition to features derived from the CTG. This would have created a benchmark on how good predictions could be made from known information, and then trying to find features that would beat this benchmark on their own or in combination of each other. This should especially be done for case 2, where the exclusion of the gestational age showed an improvement in the F_1-score.

It was investigated if the best performing machine learning algorithm could be found. From this thesis, it was found to be SVM, though, this could probably change when using better and possibly more features.

To conclude, there needs to be better features that distinguishes the outcomes from one another, and finding such features is difficult.

6.2 Future Outlook

This thesis touches the surface of feature extraction from the cardiotocography, and there are numerous things to investigate further. A natural step forward could be the continuation of feature extraction from the FHR. The extraction of the FHR signal could be altered, meaning that chosen thresholds could be changed. The tested features were only examined along with the gestational age and in combination of each other, though it was not examined if any features contribute with correlated information to the classifiers. The frequency domain could also be investigated to see if there is any variance in how the FHR lies within each given frequency band. One could also try applying time series analysis on the FHR signal, to check if the parameters of a model would be interesting.

The features based on background information could be re-made with clinical statements to have some sort of correctness in the feature extraction, and increase the understanding of deviating patterns in the CTG.

The approach in this project focused much on trying to find information in the FHR signal, therefore it would be of interest to focus more on the contraction signal, TOCO. Features using only the TOCO signal could be time domain features or frequency domain. Another aspect to investigate could be how the FHR and TOCO signals relate to each other, since it is an important how the fetus manages a contraction, specifically how decelerations appear in the FHR connected to the TOCO.

An approach might be to isolate the contractions, and see if there is a corresponding deceleration.

Furthermore, the use of neural network could be expanded. This thesis derived the best resulting ANN by trial and error, and changing the internal parameters can be tested more. Another idea is to use of convolutional neural network on the time series data to find fitting features, or use long short-term memory to have feedback connections, and capture important traits of the CTG signals. Regarding the different techniques, it could be interesting combining them, i.e., doing ensemble learning, and making the prediction from the ensemble aggregation.

These steps could prove to distinguish the respective outcomes from one another, and therefore improve the results of the classifiers.

References

[1] SCB. *Dödfödda Efter År*. Available at: `https://www.scb.se/dodfodda`. (Accessed: 26 January 2022).

[2] SCB. *Döda i Sverige*. Available at: `https://www.scb.se/doda-i-sverige`. (Accessed: 26 January 2022).

[3] A. Herbst and I. Amer-Wåhlin. *Fosterövervakning under förlossningen*. In K. Marsal, H. Hagberg, and M. Westgren, editors, *Obstetrik*, chapter 19. Studentlitteratur, 2008.

[4] H. Hagberg and M. Blennow. *Fosterasfyxi*. In K. Marsal, H. Hagberg, and M. Westgren, editors, *Obstetrik*, chapter 17. Studentlitteratur, 2008.

[5] J. Spilka, V. Chudácek, P. Janku, L. Hruban, M. Burša, M. Huptych, L. Zach, and L. Lhotská. *Analysis of Obstetricians' Decision Making on CTG Recordings*. *Journal of Biomedical Informatics*, 2014.

[6] R. Persson and P.Johansson (ed). (2020). *Undersökningar av det nyfödda barnet*. Available at: `https://www.1177.se/forlossning/undersokningar-av-det-nyfodda-barnet`. (Accessed: 1 december 2021).

[7] K. Bengtsson and C. Lilliecreutz (ed). *Graviditet efter vecka 41 - att gå över tiden* (2021). Available at: `https://www.1177.se/graviditet/att-ga-over-tiden`. (Accessed: 15 december 2021).

[8] K. Marsal, H. Hagberg, and M. Westgren. *Fosterövervakning under förlossningen*. In *Obstetrik*, chapter 19. Studentlitteratur, 2008.

[9] American College of Obstetricians and Gynecologists. *Apgar Score (2021)*. Available at: `https://www.acog.org/clinical/apgarscore`. (Accessed: 30 december 2021).

[10] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[11] T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[12] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2008.

[13] C. C. Aggarwal. *Neural Networks and Deep Learning*. Springer, 2018.

[14] S. Ganesh. *What's the Role of Weights and Bias in a Neural Network* (2020). Available at: `https://towardsdatascience.com/neural-network`. (Accessed: 30 december 2021).

[15] Towards Data Science. *A Layman's Guide to Deep Neural Networks*. Available at: `https://towardsdatascience.com/deepNN`. (Accessed: 25 January 2022).

[16] Andrew L. Maas. *Rectifier Nonlinearities Improve Neural Network Acoustic Models*. 2013.

[17] B. Kanani. *Activation Functions in Neural Network*. Available at: `https://ML.com/activation`. (Accessed: 26 January 2022).

[18] M. Romano, P. Bifulco, M. Ruffo, G. Improta, F. Clemente, and M. Cesarelli. *Software for Computerised Analysis of Cardiotocographic Traces*. 2015.

[19] S. Nidhal, M.A. Mohd Ali, A.A. Zaidan, B.B. Zaidan, and H. Najah. *Computerized Algorithm for Fetal Heart Rate Baseline and Baseline Variability Estimation based on Distance Between Signal Average and α Value. International Journal of Pharmacology*, 2011.

[20] A. Agostinelli, E. Braccili, R. Rosati, A. Sbrollini, L. Burattini, M. Morettini, F. Di Nardo, S. Fioretti, and L. Burattini. *Statistical Baseline Assesment in Cardiotocography. IEEE*, 2017.

[21] A. Sbrollini, A. Carnicelli, A. Massacci, L. Tomaiuolo, T. Zara, I. Marcantoni, L. Burattini, M. Morettini, S. Fioretti, and L. Burattini. *Automatic Identification and Classification of Fetal Heart-Rate Decelerations from Cardiotocography Recordings. IEEE*, 2018.

[22] P. Labaj, M. Jezewski, A. Matonia, T. Kupka, J. Jezewski, and A. Gacek. New approach to quantitative description of deceleration of fetal heart rate for the patterns classification. 2007.

[23] D.Y. Chung, Y.B. Sim, K.T. Park, S. H. Yi, J.C. Shin, and S.P. Kim. *Spectral Analysis of Fetal Heart Rate Variability as a Predictor of Intrapartum Fetal Distress.* 2000.

[24] V. Chudáček, J. Spilka, M. Koucký, L. Lhotská, and M. Huptych. *Automatic Evaluation of Intrapartum Fetal Heart Rate Recordings: A Comprehensive Analysis of Useful Features. Physiol. Meas. **32***, 2011.

[25] J. Spilka, J. Freon, R. Leonarduzzi, N. Pustelnik, P. Abry, and M. Doret. *Sparse Support Vector Machine for Intrapartum Fetal Heart Rate Classification. IEEE*, 2017.

[26] G. Georgoulas, C. D. Stylios, and P. P. Groumpos. *Predicting the Risk of Metabolic Acidosis for Newborn Based on Fetal Heart Rate Signal Classification Using Support Vector Machines. IEEE*, 2006.

[27] M. Jezewski, J. Wrobel, P. Labaj, J. Leski, N. Henzel, K. Horoba, and J. Jezewski. *Some Pracitcal Remarks on Neural Networks Approach to Fetal Cardiotocograms Classification. IEEE EMBS*, 2007.

[28] O. Fontenla-Romero, A. Alonso-Betanzos, and B. Guijarro-Berdiñas. *Adaptive Pattern Recognition in the Analysis of Cardiotocographic Records.* 2015.

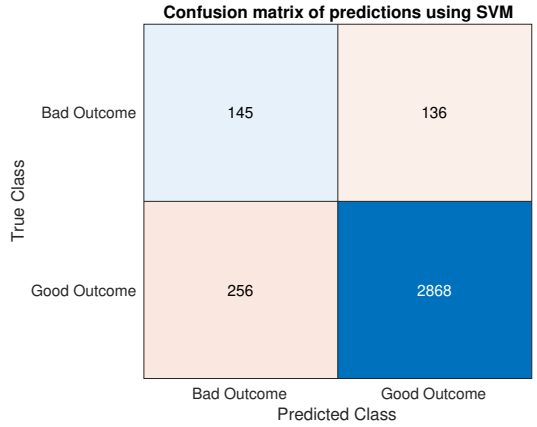
[29] J. Magnusson Österberg and P. Johansson (ed). *När barn föds för tidigt.* Available at: `https://www.1177.se/barn-gravid/tidigt`. (Accessed: 14 February 2022).

A Confusion Matrices

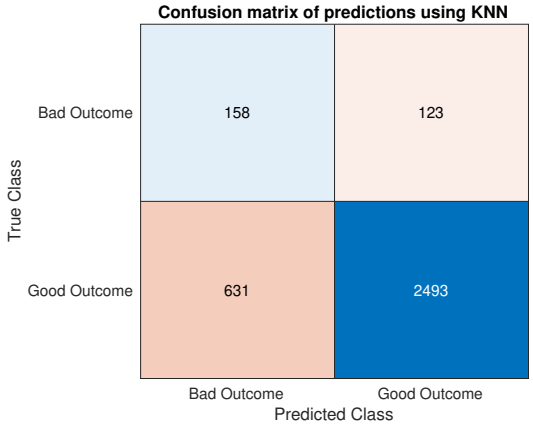
Here the confusion matrices for each case and each classifier a confusion matrix are presented. The values are the sum of predictions for each validation set. The appendix is split into the cases and each feature. Lastly the confusion matrices of the final models are shown.

A.1 Case 1

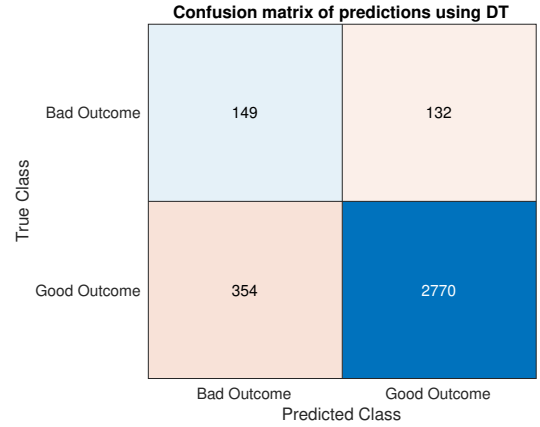
A.1.1 Naive models



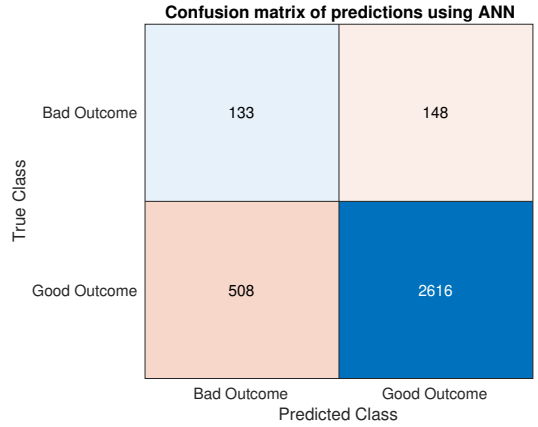
(a) SVM



(b) KNN



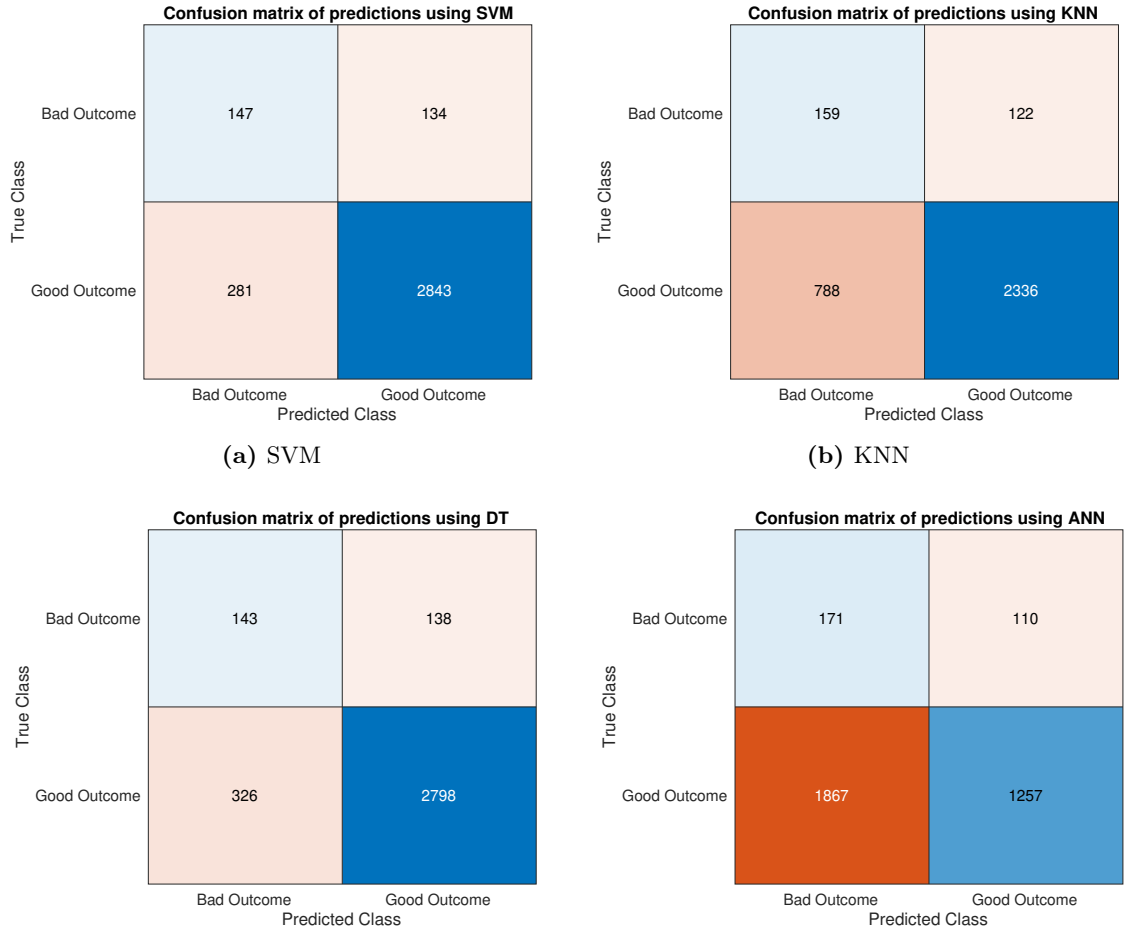
(c) DT



(d) ANN

Figure 23: Confusion matrices of predictions using different classifiers. The classifier is specified in the sub caption of each figure. The classifiers were trained using the gestational age as a feature. The x-axis shows the predicted class, and the y-axis shows the true class.

A.1.2 Bradycardia

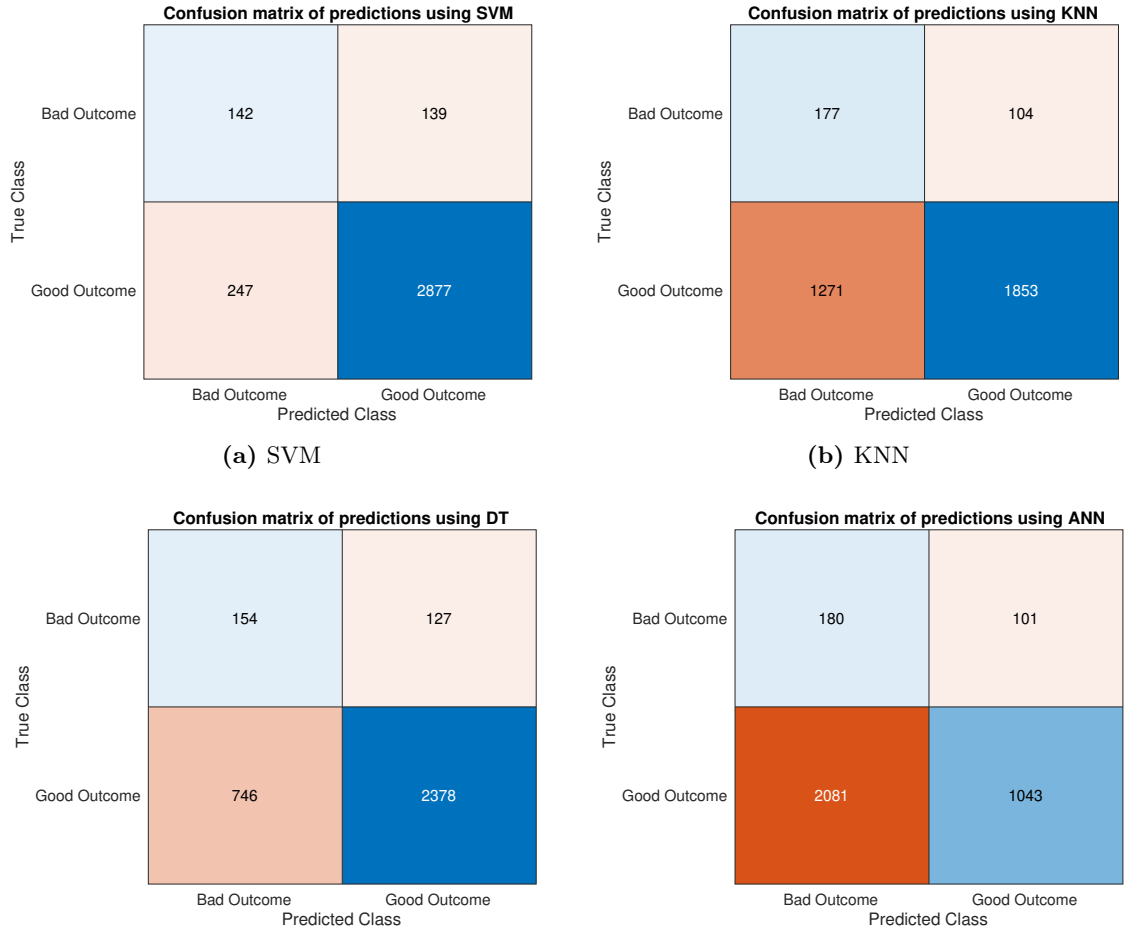


(a) SVM **(b)** KNN

(c) DT **(d)** ANN

Figure 24: Confusion matrices of predictions using different classifiers. The classifier is specified in the sub caption of the figures. The classifiers were trained using the gestational age and bradycardia as features. The x-axis shows the predicted class, and the y-axis shows the true class.

A.1.3 Tachycardia

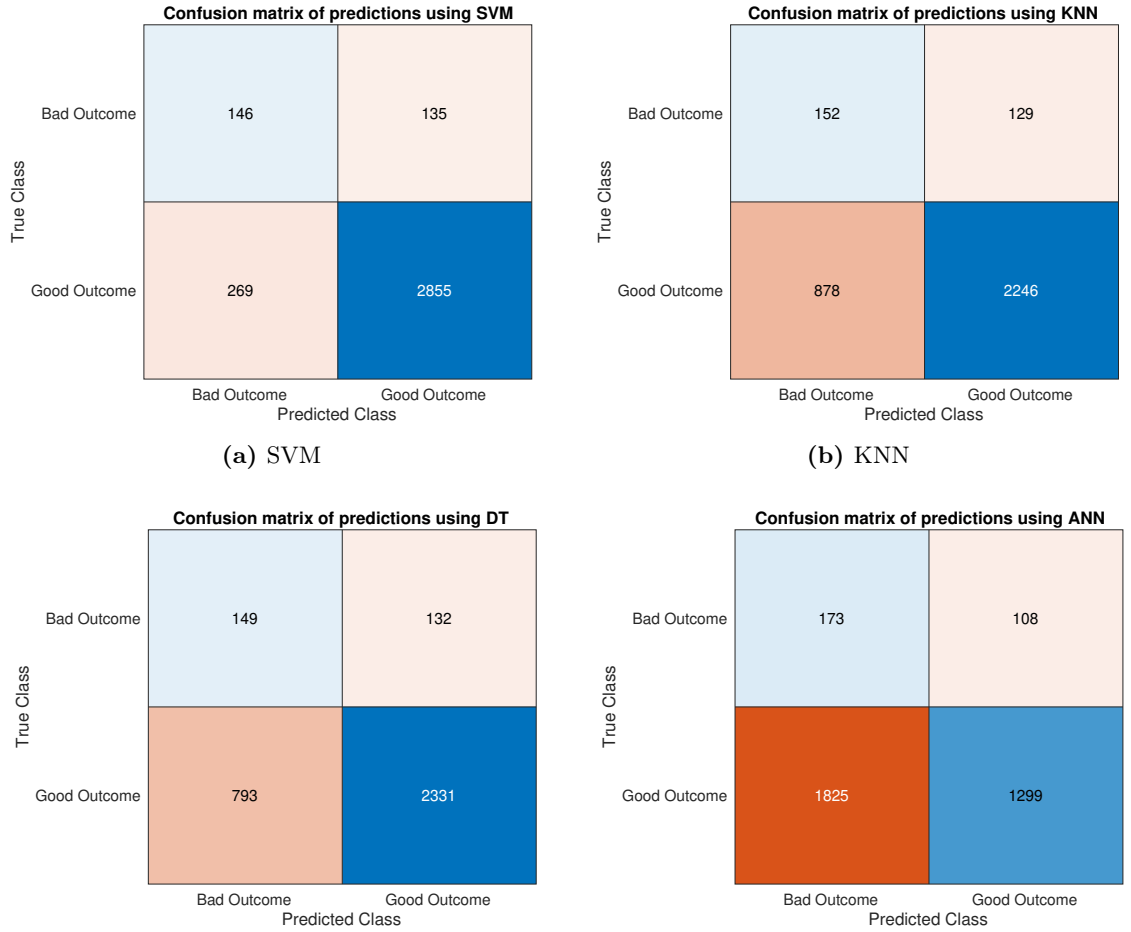


(a) SVM **(b)** KNN

(c) DT **(d)** ANN

Figure 25: Confusion matrices of predictions using different classifiers. The classifier is specified in the sub caption of the figures. The classifiers were trained using the gestational age and tachycardia as features. The x-axis shows the predicted class, and the y-axis shows the true class.

A.1.4 Number of Accelerations

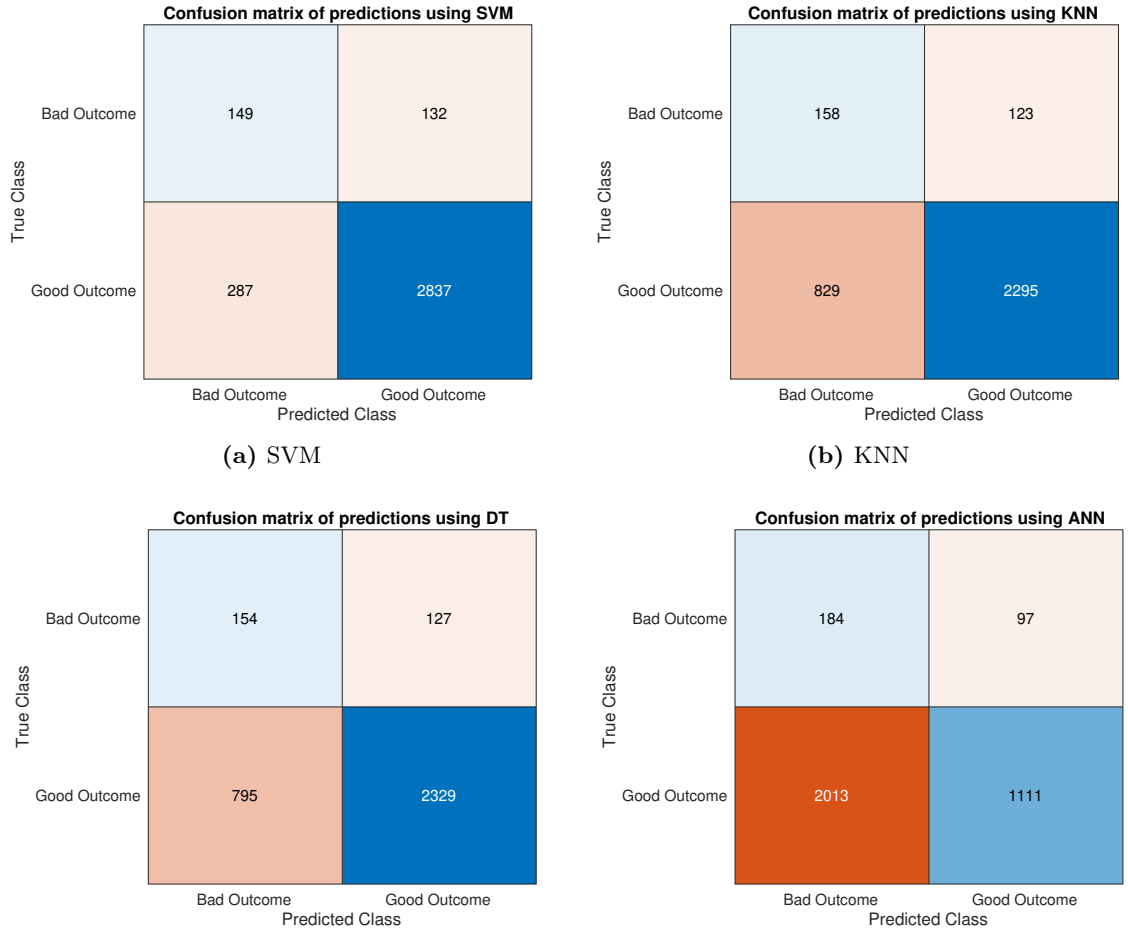


(a) SVM **(b)** KNN

(c) DT **(d)** ANN

Figure 26: Confusion matrices of predictions using different classifiers. The classifier is specified in the sub caption of the figures. The classifiers were trained using the gestational age and number of accelerations as feature. The x-axis shows the predicted class, and the y-axis shows the true class.

A.1.5 Number of Decelerations

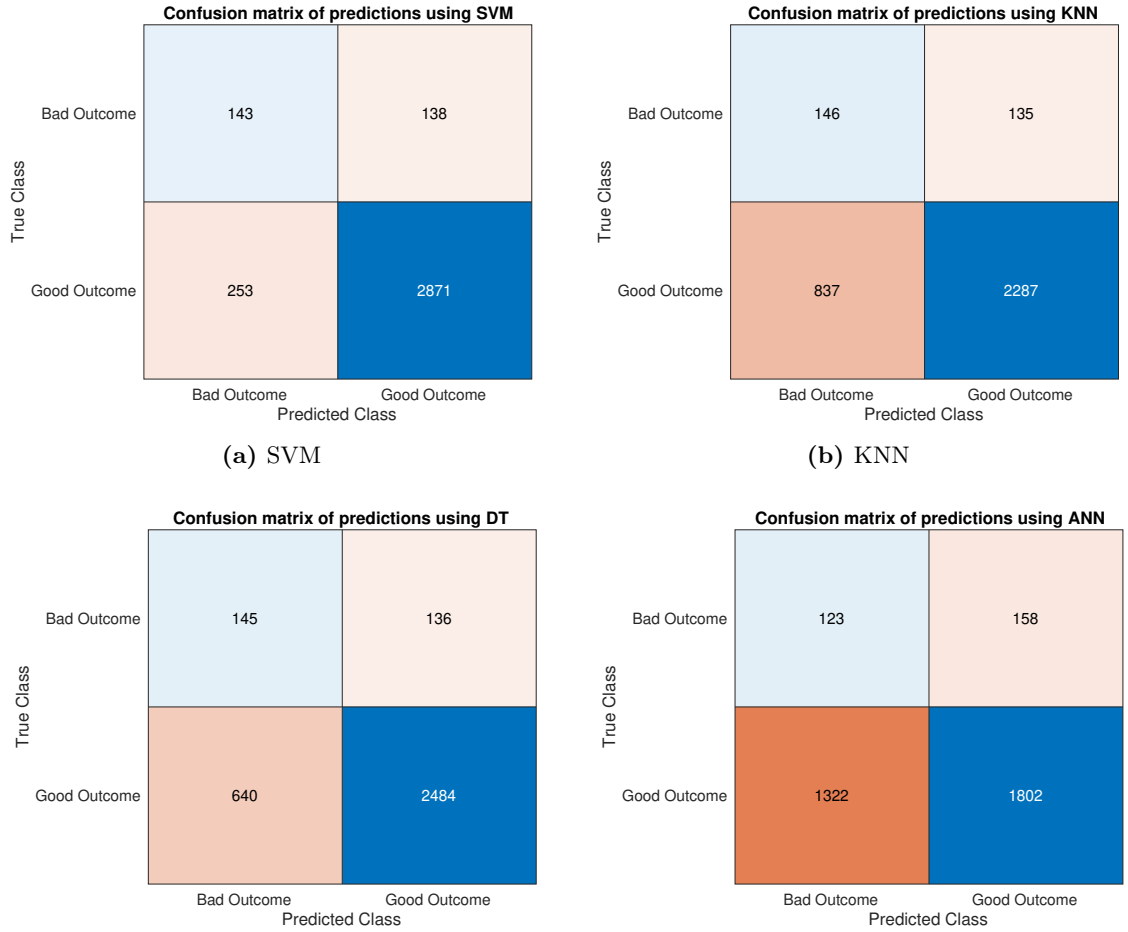


(a) SVM **(b)** KNN

(c) DT **(d)** ANN

Figure 27: Confusion matrices of predictions using different classifiers. The classifier is specified in the sub caption of the figures. The classifiers were trained using the gestational age and the number of decelerations as feature. The x-axis shows the predicted class, and the y-axis shows the true class.

A.1.6 Slope of Fitted Line

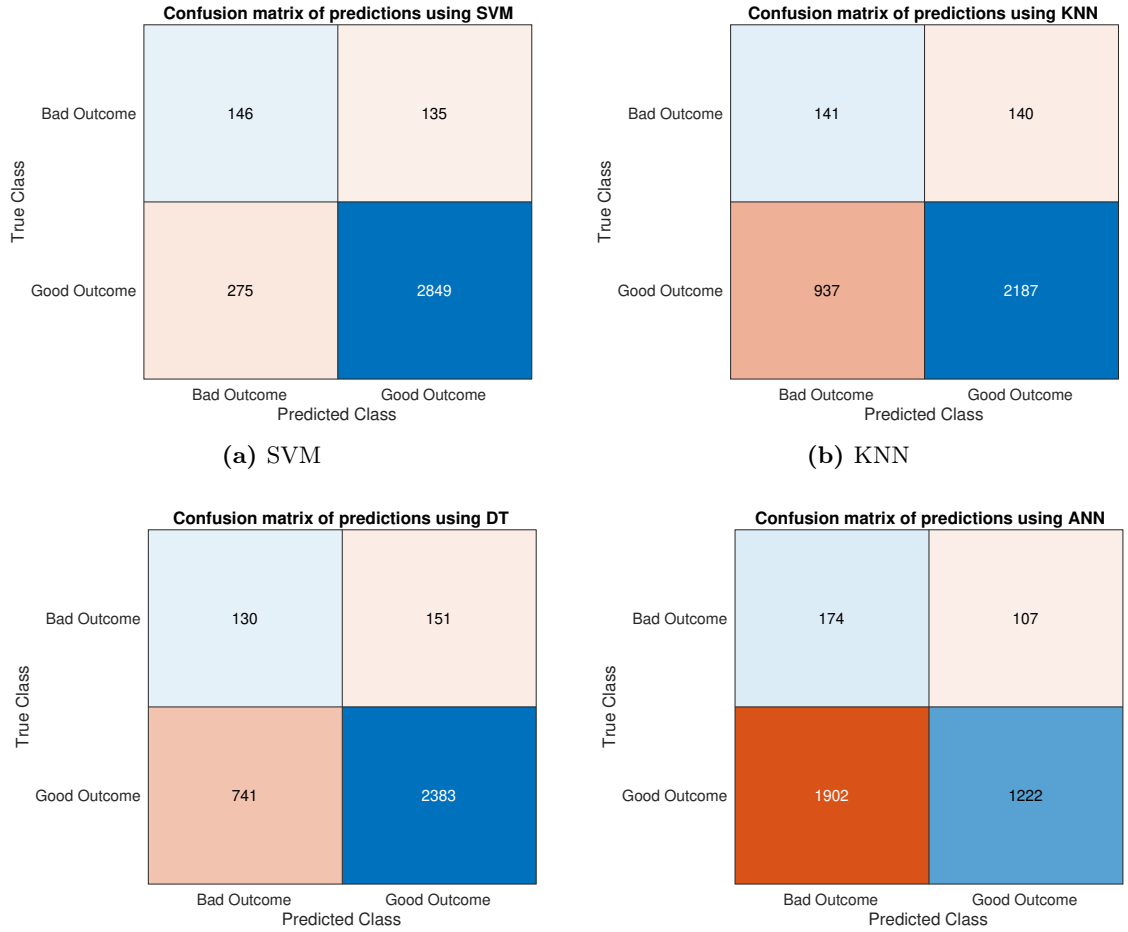


(a) SVM **(b)** KNN

(c) DT **(d)** ANN

Figure 28: Confusion matrices of predictions using different classifiers. The classifier is specified in the sub caption of the figures. The classifiers were trained using the gestational age and slope of fitted line as features. The x-axis shows the predicted class, and the y-axis shows the true class.

A.1.7 Short Term Variability

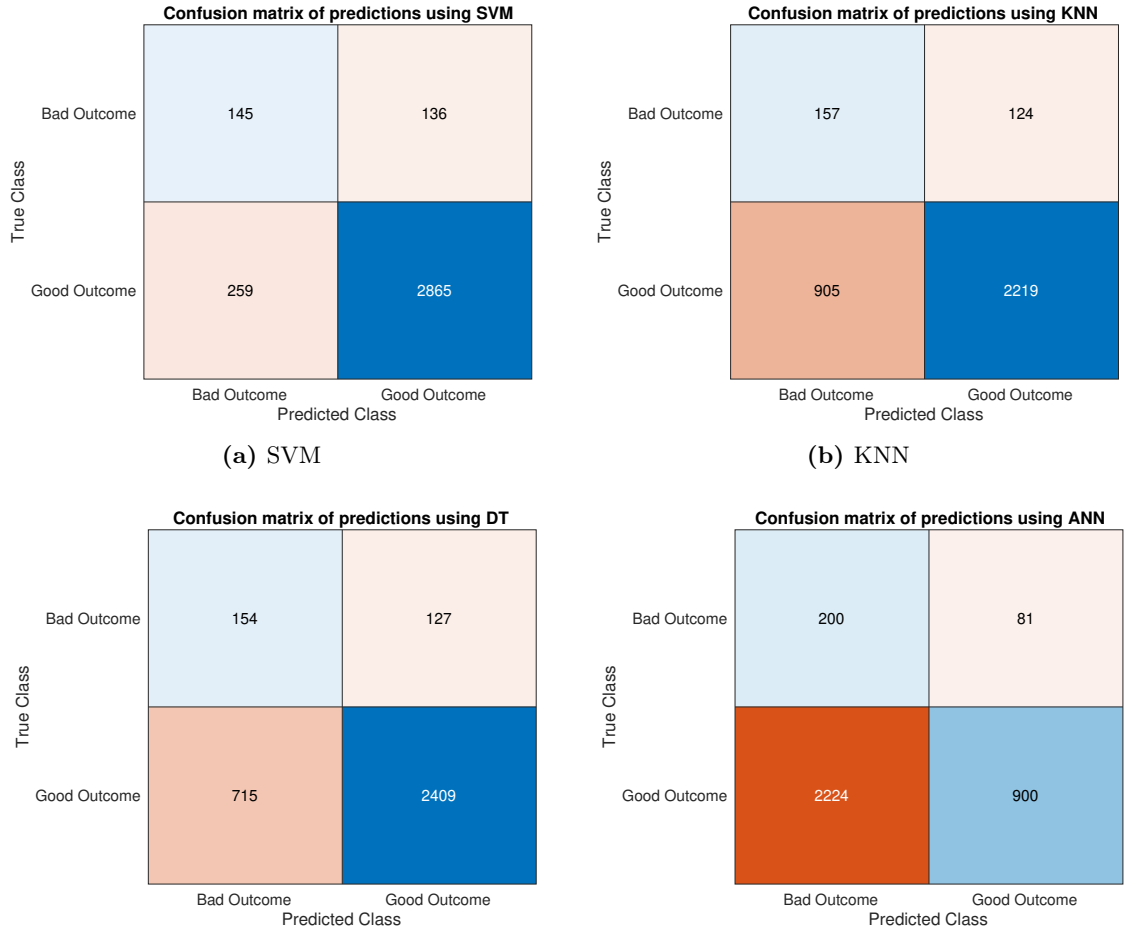


(a) SVM **(b)** KNN

(c) DT **(d)** ANN

Figure 29: Confusion matrices of predictions using different classifiers. The classifier is specified in the sub caption of the figures. The classifiers were trained using the gestational age and MAD of the short term variability as features. The x-axis shows the predicted class, and the y-axis shows the true class.

A.1.8 Interval Index

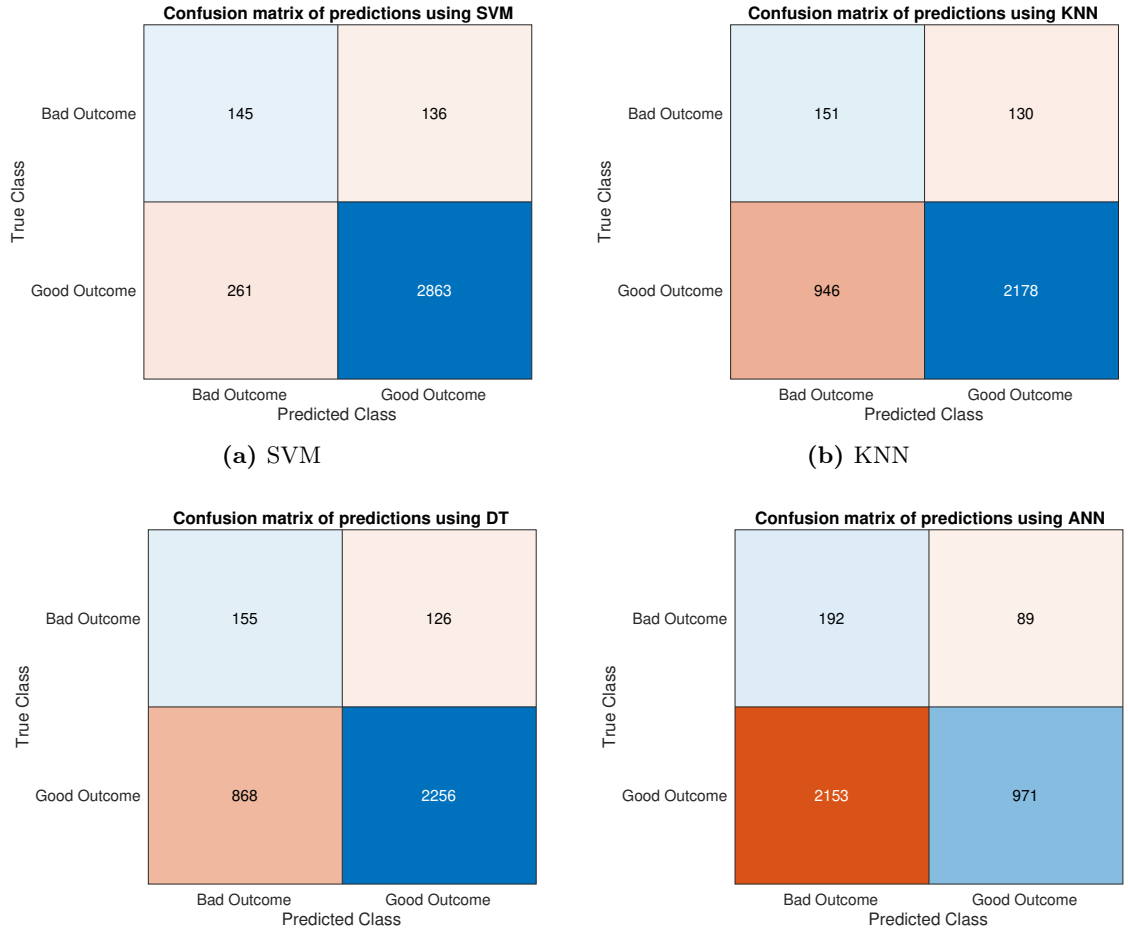


(a) SVM **(b)** KNN

(c) DT **(d)** ANN

Figure 30: Confusion matrices of predictions using different classifiers. The classifier is specified in the sub caption of the figures. The classifiers were trained using the gestational age and MAD of the interval index as features. The x-axis shows the predicted class, and the y-axis shows the true class.

A.1.9 Mean Absolute Deviation

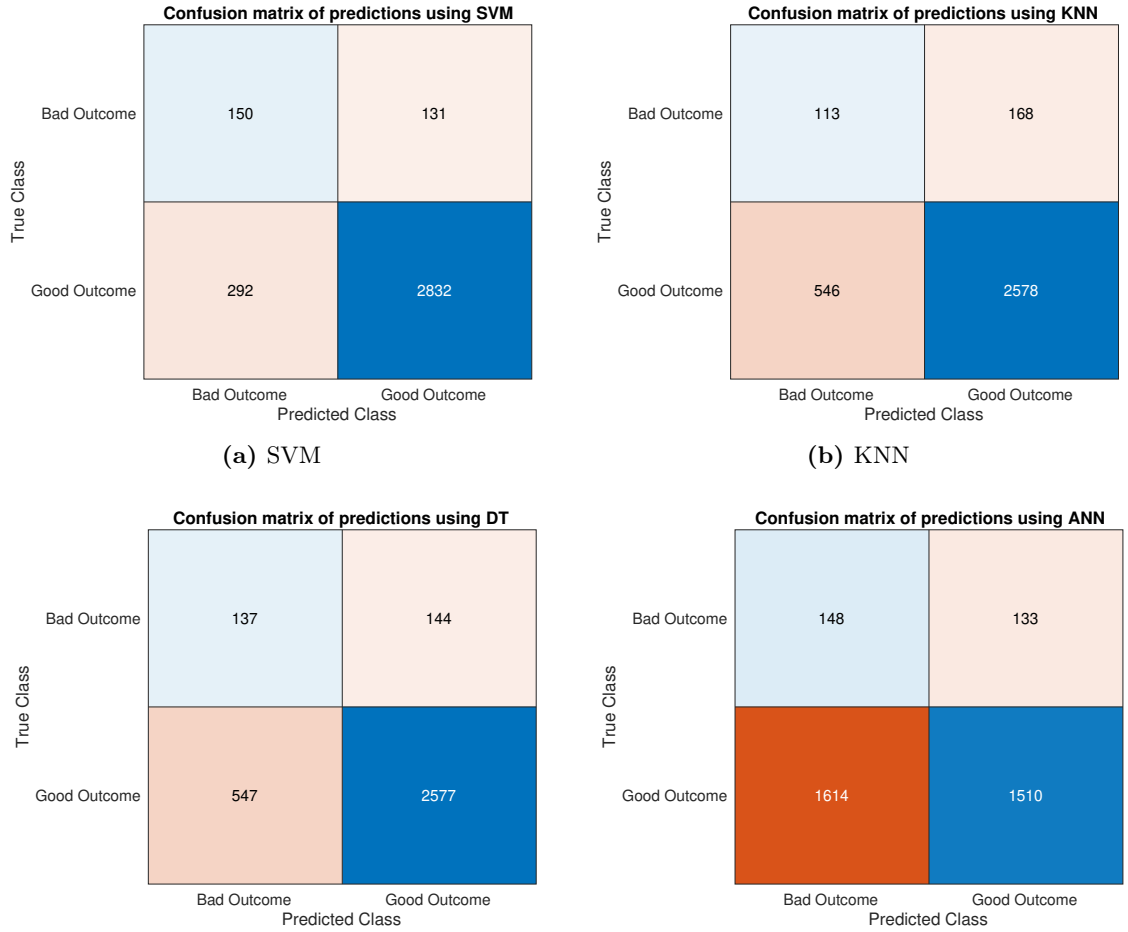


(a) SVM **(b)** KNN

(c) DT **(d)** ANN

Figure 31: Confusion matrices of predictions using different classifiers. The classifier is specified in the sub caption of the figures. The classifiers were trained using the gestational age and MAD of the FHR as features. The x-axis shows the predicted class, and the y-axis shows the true class.

A.1.10 All Features

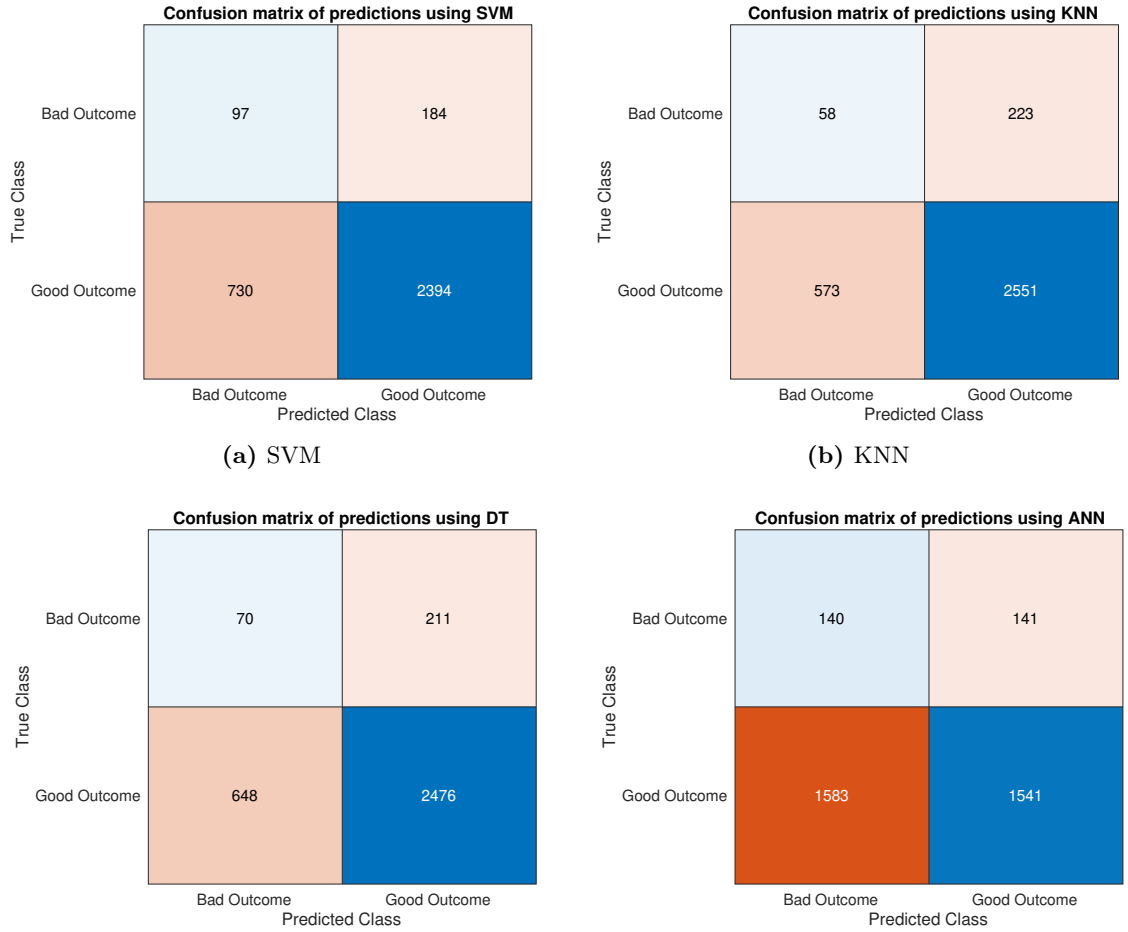


(a) SVM **(b)** KNN

(c) Decision Tree **(d)** ANN

Figure 32: Confusion matrices of predictions using different classifiers. The classifier is specified in the sub caption of the figures. The classifiers were trained using all derived features. The x-axis shows the predicted class, and the y-axis shows the true class.

A.1.11 All but One Feature



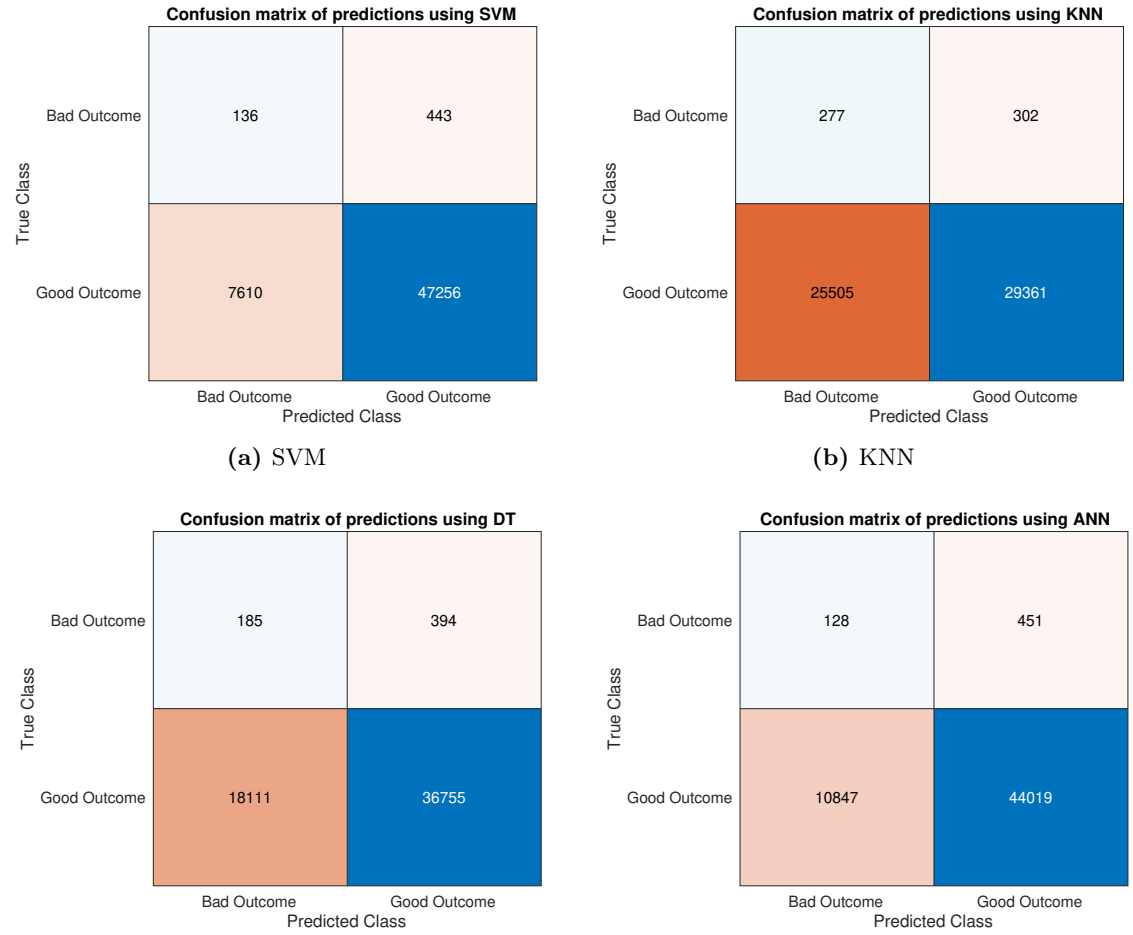
(a) SVM **(b)** KNN

(c) Decision Tree **(d)** ANN

Figure 33: Confusion matrices of predictions using different classifiers. The classifier is specified in the sub caption of the figures. The classifiers were trained using all derived features, but the gestational age. The x-axis shows the predicted class, and the y-axis shows the true class.

A.2 Case 2

A.2.1 Naive models

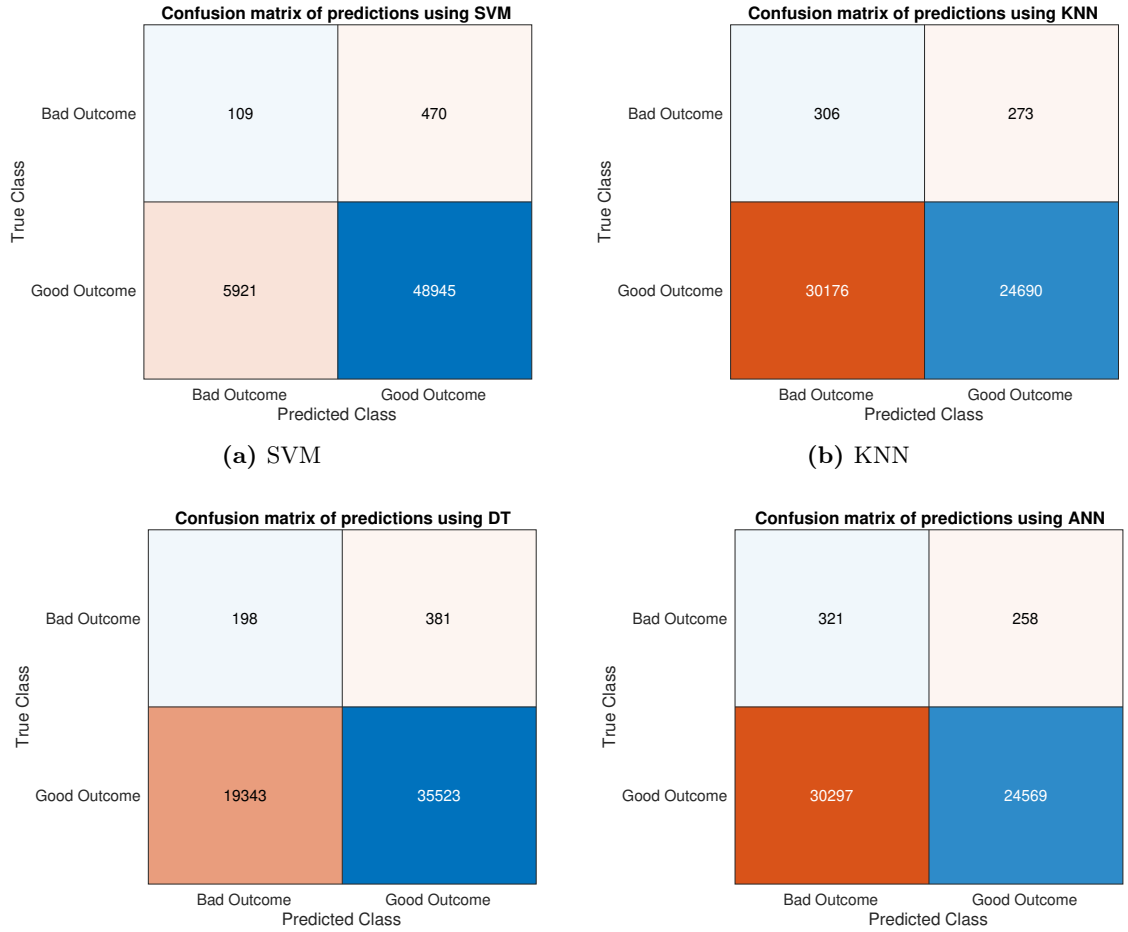


(a) SVM **(b)** KNN

(c) DT **(d)** ANN

Figure 34: Confusion matrices of predictions using different classifiers. The classifier is specified in the sub caption of the figures. The classifiers were trained using the gestational age as a feature. The x-axis shows the predicted class, and the y-axis shows the true class.

A.2.2 Bradycardia

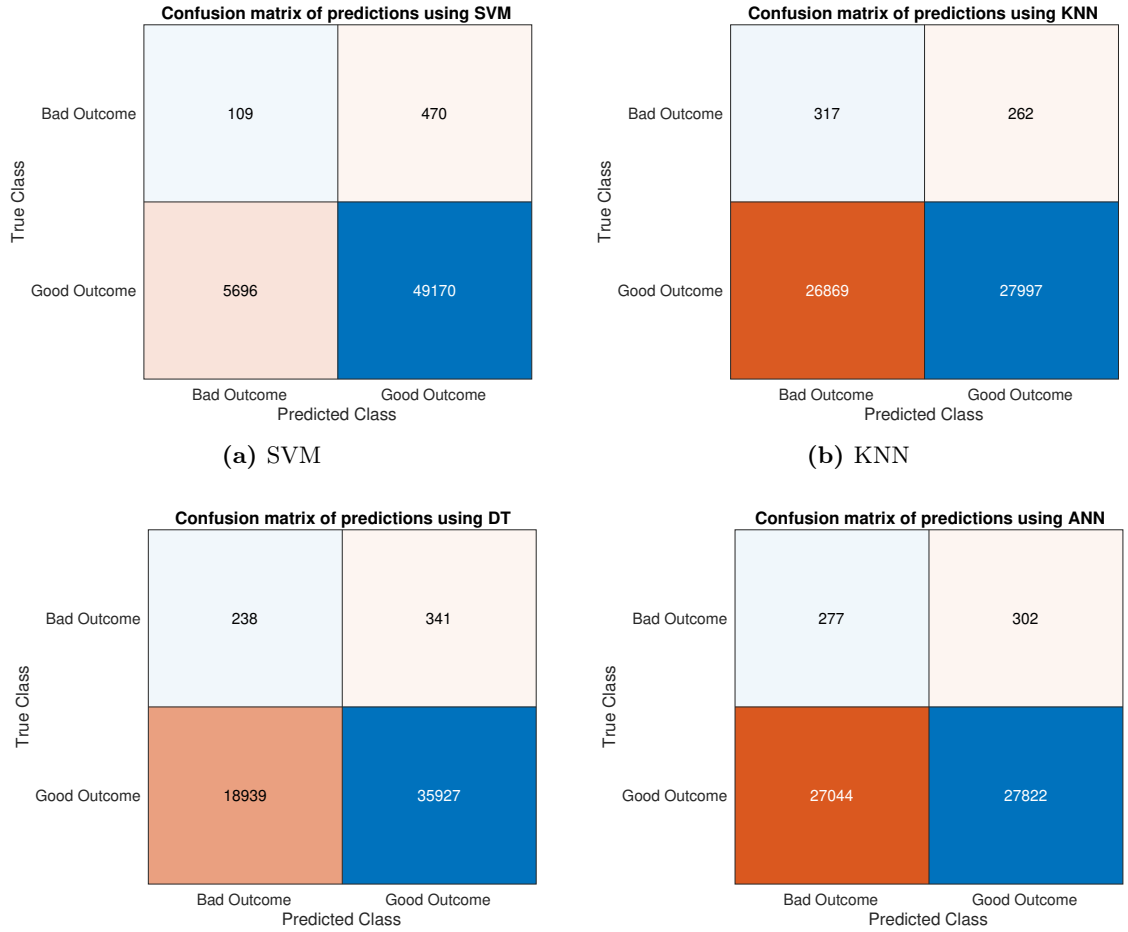


(a) SVM **(b)** KNN

(c) DT **(d)** ANN

Figure 35: Confusion matrices of predictions using different classifiers. The classifier is specified in the sub caption of the figures. The classifiers were trained using the gestational age and bradycardia as feature. The x-axis shows the predicted class, and the y-axis shows the true class.

A.2.3 Tachycardia

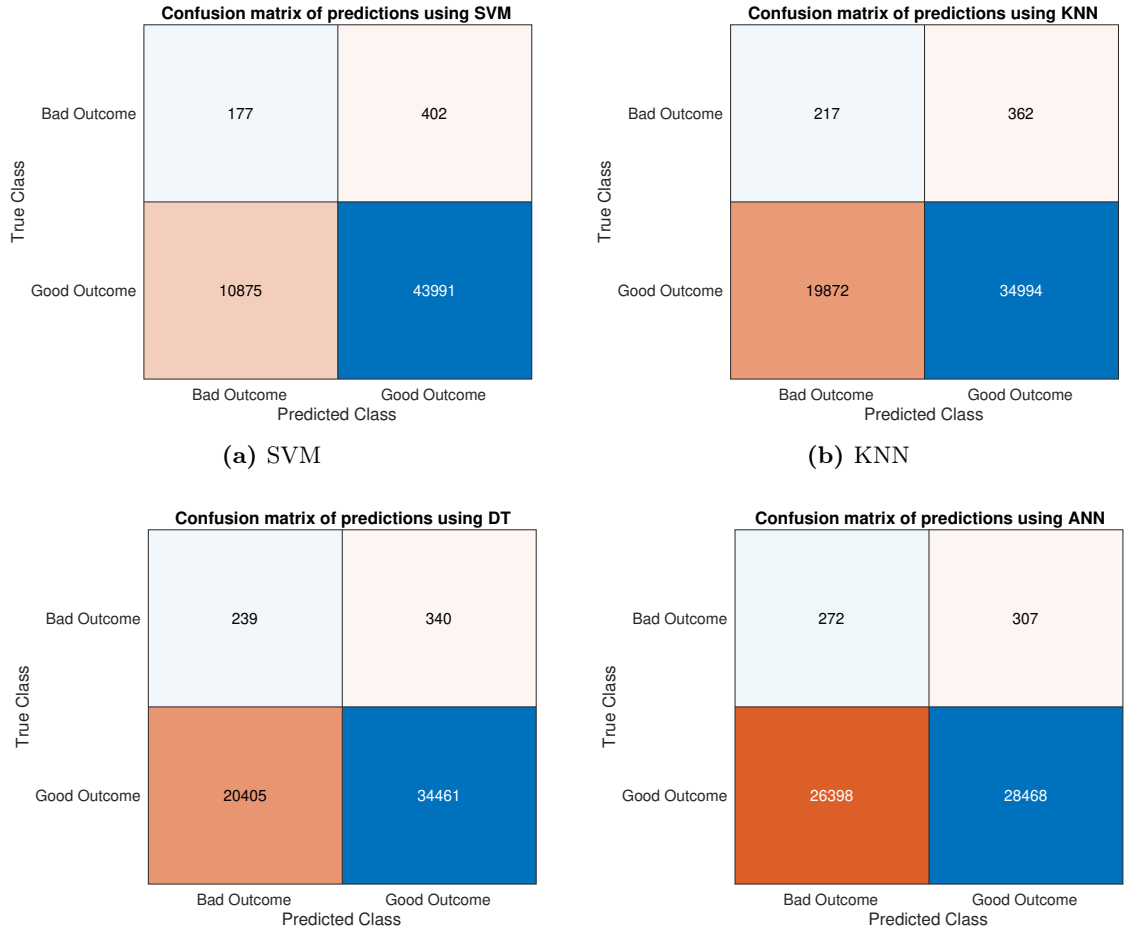


(a) SVM **(b)** KNN

(c) DT **(d)** ANN

Figure 36: Confusion matrices of predictions using different classifiers. The classifier is specified in the sub caption of the figures. The classifiers were trained using the gestational age and tachycardia as features. The x-axis shows the predicted class, and the y-axis shows the true class.

A.2.4 Number of Accelerations

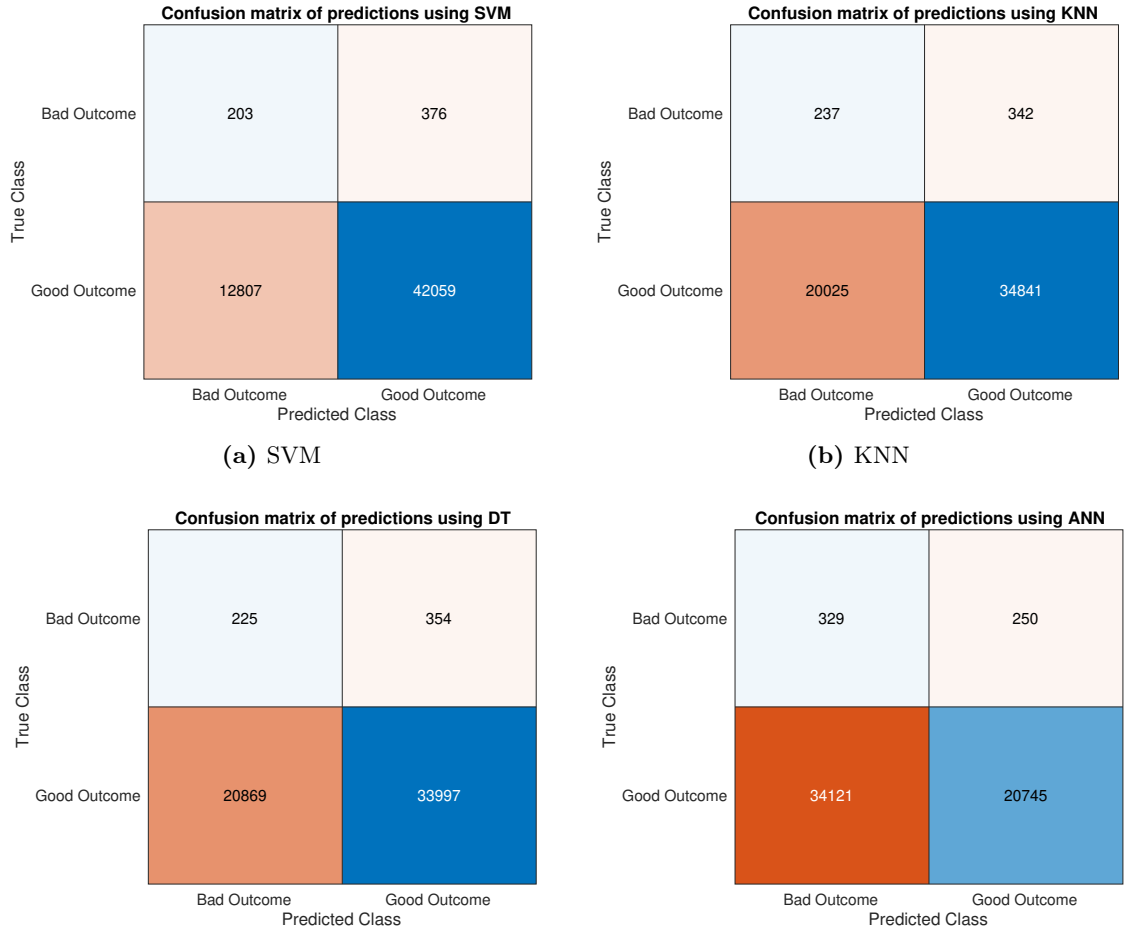


(a) SVM **(b)** KNN

(c) DT **(d)** ANN

Figure 37: Confusion matrices of predictions using different classifiers. The classifier is specified in the sub caption of the figures. The classifiers were trained using the gestational age and the number of accelerations as features. The x-axis shows the predicted class, and the y-axis shows the true class.

A.2.5 Number of Decelerations

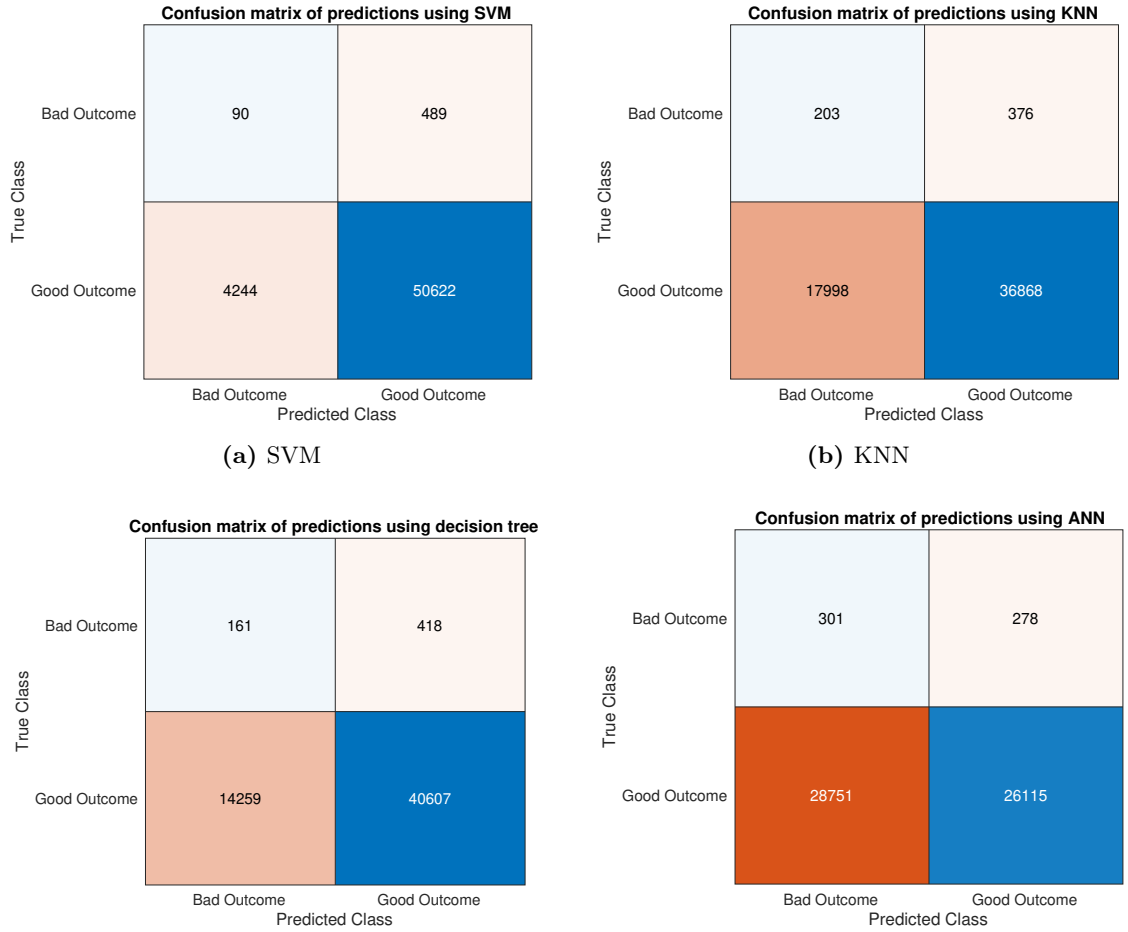


(a) SVM **(b)** KNN

(c) DT **(d)** ANN

Figure 38: Confusion matrices of predictions using different classifiers. The classifier is specified in the sub caption of the figures. The classifiers were trained using the gestational age and the number of decelerations as features. The x-axis shows the predicted class, and the y-axis shows the true class.

A.2.6 Slope of Fitted Line

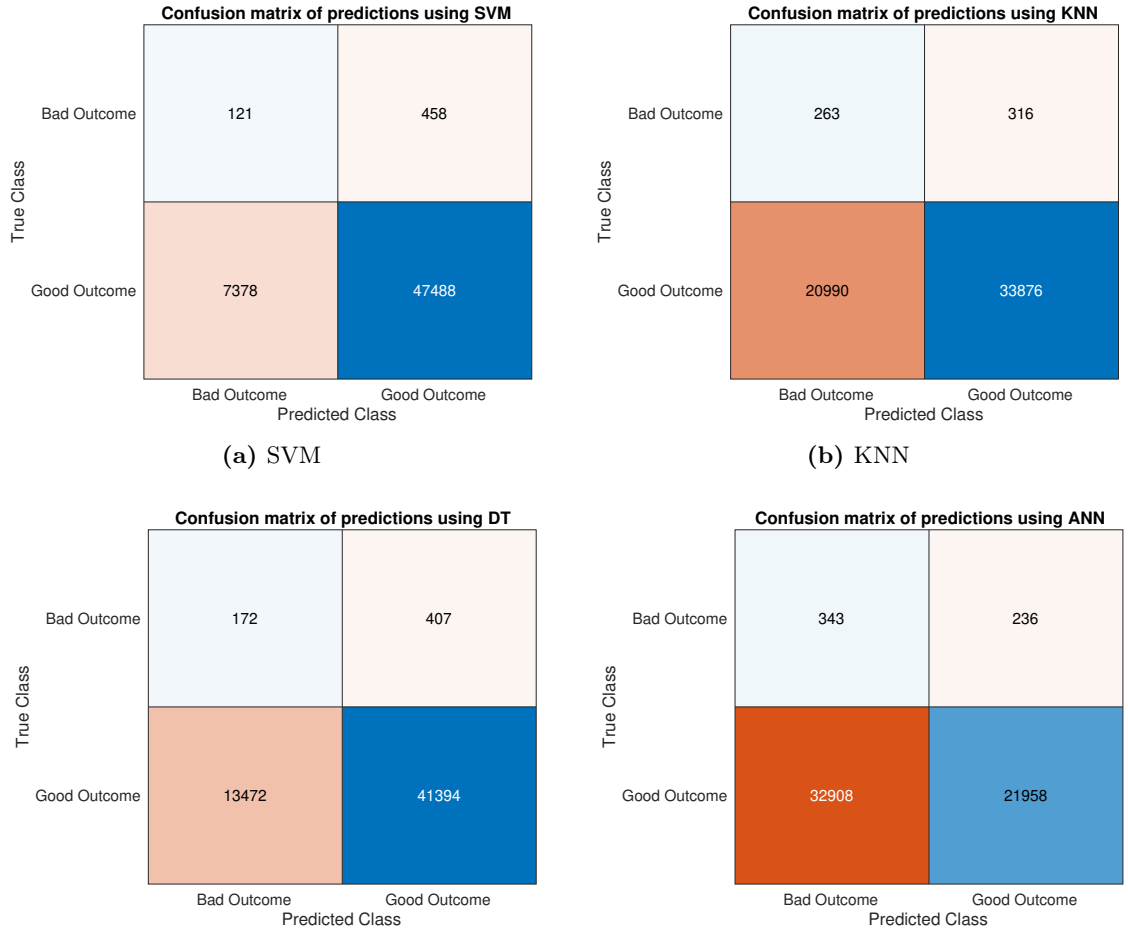


(a) SVM **(b)** KNN

(c) DT **(d)** ANN

Figure 39: Confusion matrices of predictions using different classifiers. The classifier is specified in the sub caption of the figures. The classifiers were trained using the gestational age and slope of fitted line as features. The x-axis shows the predicted class, and the y-axis shows the true class.

A.2.7 Short Term Variability

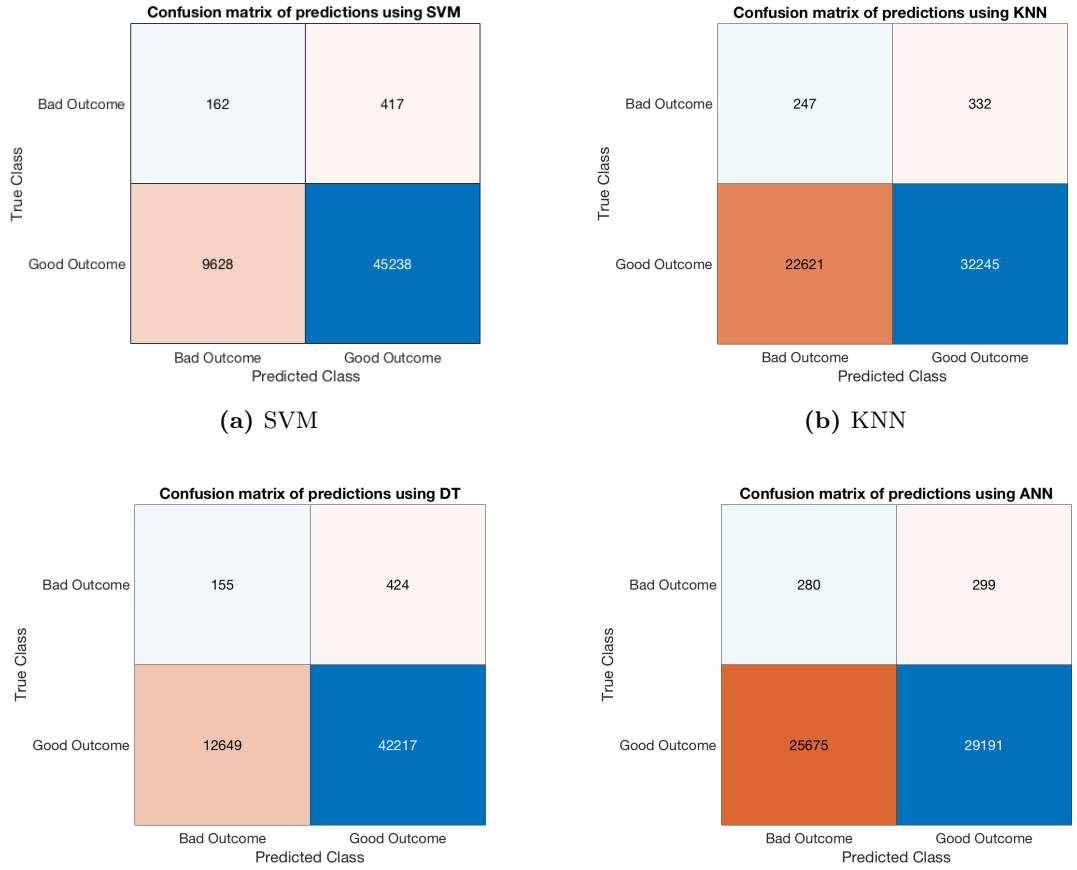


(a) SVM **(b)** KNN

(c) DT **(d)** ANN

Figure 40: Confusion matrices of predictions using different classifiers. The classifier is specified in the sub caption of the figures. The classifiers were trained using the gestational age and MAD of the short term variability as features. The x-axis shows the predicted class, and the y-axis shows the true class.

A.2.8 Interval Index

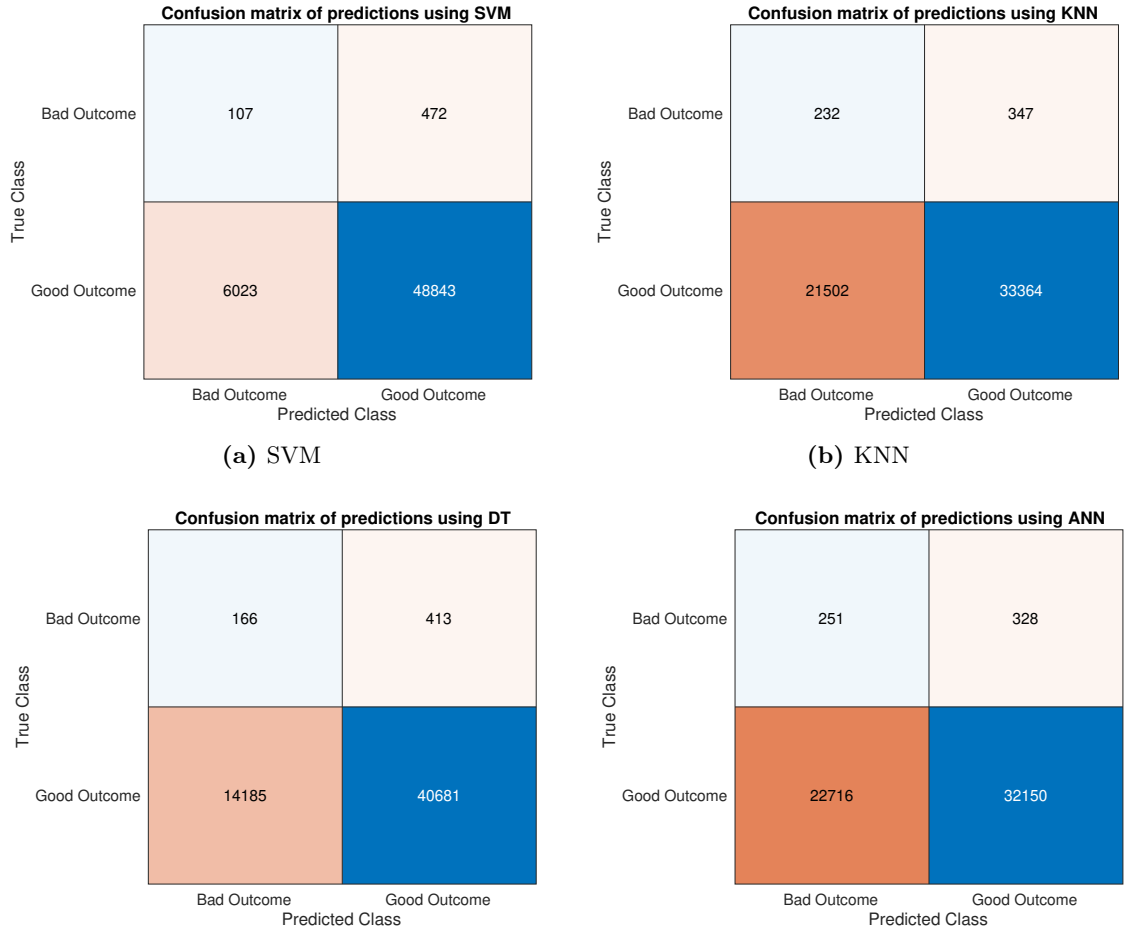


(a) SVM **(b)** KNN

(c) DT **(d)** ANN

Figure 41: Confusion matrices of predictions using different classifiers. The classifier is specified in the sub caption of the figures. The classifiers were trained using the gestational age and MAD of the interval index as features. The x-axis shows the predicted class, and the y-axis shows the true class.

A.2.9 Mean Absolute Deviation

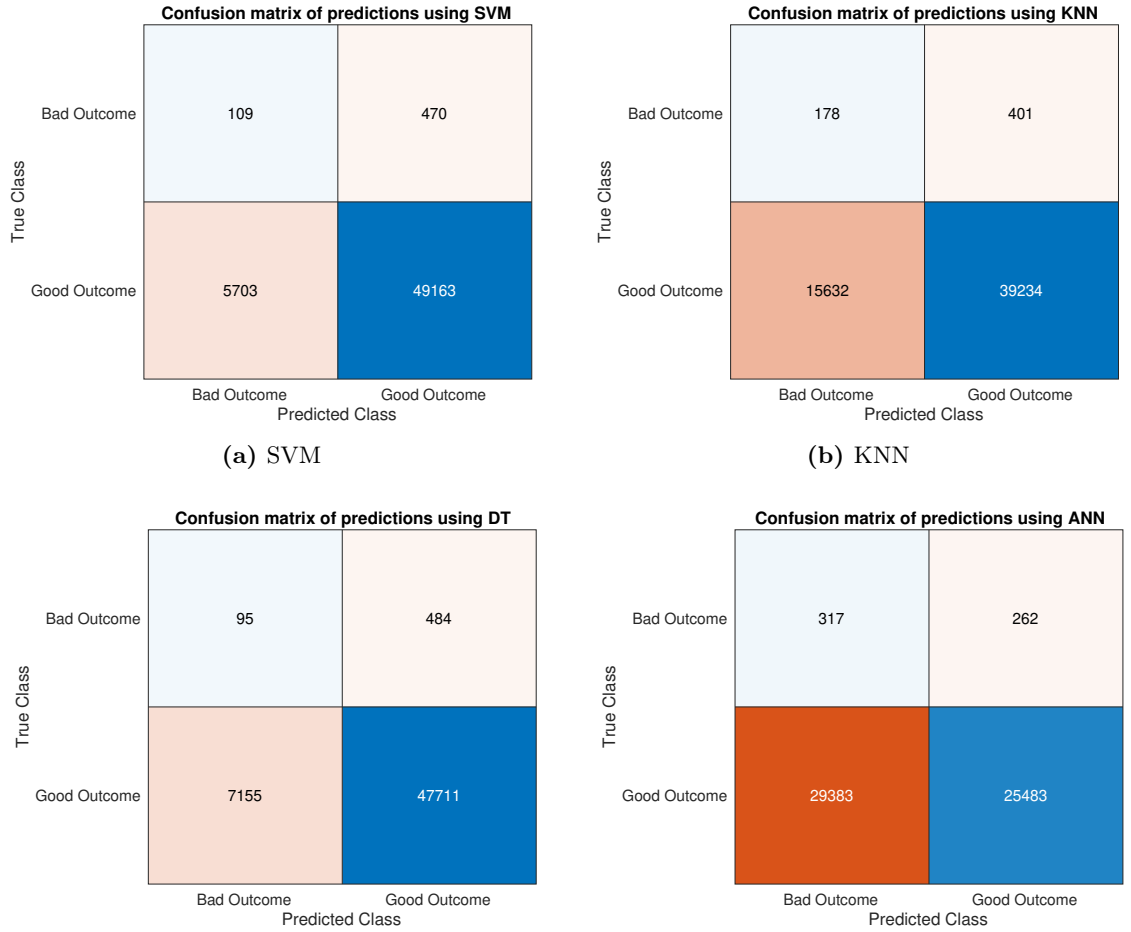


(a) SVM **(b)** KNN

(c) DT **(d)** ANN

Figure 42: Confusion matrices of predictions using different classifiers, specified in title and sub captions. The classifiers were trained using the gestational age and MAD of the FHR as features. The x-axis shows the predicted class, and the y-axis shows the true class.

A.2.10 All Features

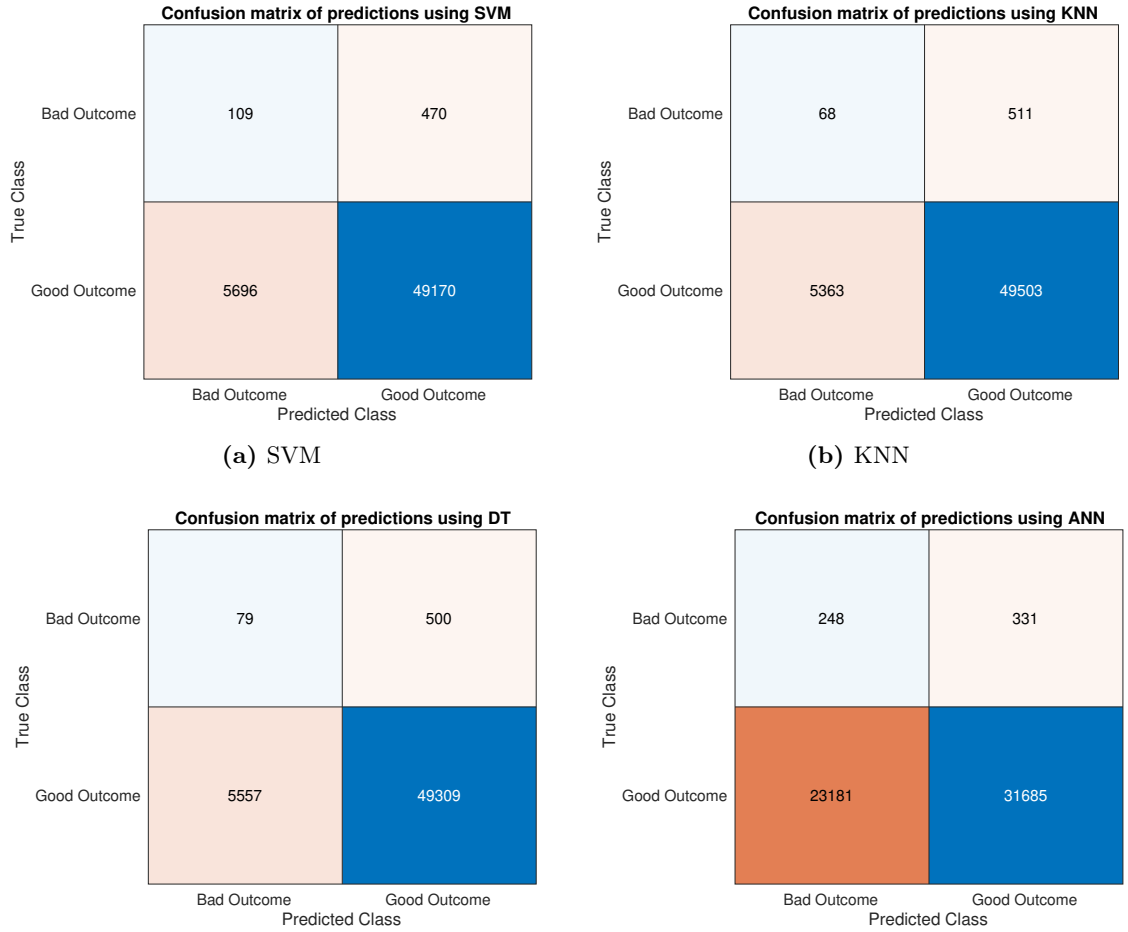


(a) SVM **(b)** KNN

(c) DT **(d)** ANN

Figure 43: Confusion matrices of predictions using different classifiers. The classifier is specified in the sub caption of the figures. The classifiers were trained using all derived features. The x-axis shows the predicted class, and the y-axis shows the true class.

A.2.11 All but One Feature



(a) SVM **(b)** KNN

(c) DT **(d)** ANN

Figure 44: Confusion matrices of predictions using different classifiers. The classifier is specified in the sub caption of the figures. The classifiers were trained using all derived features, not including the gestational age. The x-axis shows the predicted class, and the y-axis shows the true class.

A.3 Final Models

A.3.1 Case 1

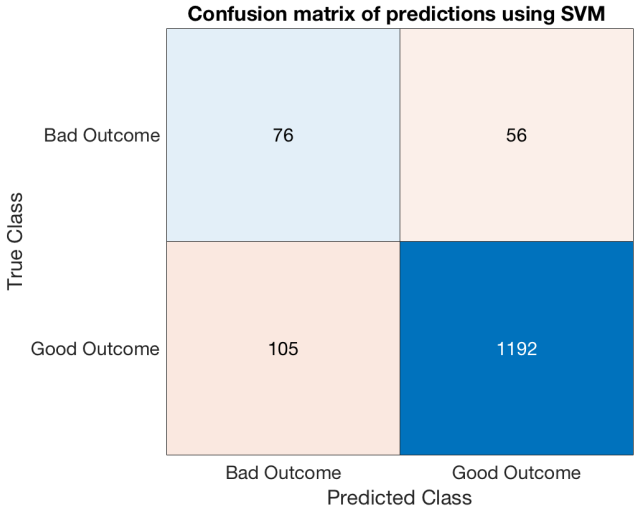


Figure 45: Caption

A.3.2 Case 2

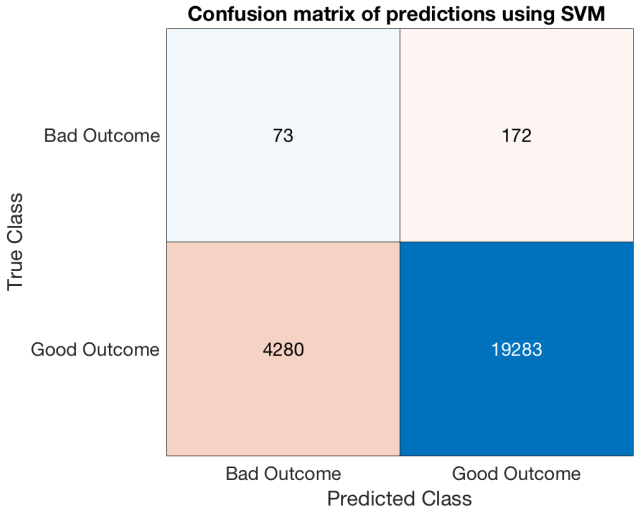


Figure 46: Caption

Master's Theses in Mathematical Sciences 2022:E13
ISSN 1404-6342

LUTFMS-3440-2022

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lu.se/