Predicting high risk birth from real large-scale cardiotocographic data using multi-input convolutional neural networks

Alkanan Mohannad¹a), Chihiro Shibata², Kohei Miyata³, Toshiro Imamura³, Shingo Miyamoto³, Hiroaki Fukunishi¹, and Hiroyuki Kameda¹

¹ The Graduate School of Bionics, Computer and Media Science, Tokyo University of Technology Katakura 1404-1, Hachioji, Tokyo 192-0982, Japan
² Hosei University, 2 Chome-17-1 Fujimi, Chiyoda City, Tokyo 102-8160, Japan
³ Department of Obstetrics and Gynecology, Faculty of Medicine, Fukuoka University Nanakuma 7-45-1, Jonan-ku, Fukuoka city, Fukuoka 814-0180, Japan

a) d2119003b2@edu.teu.ac.jp

Received January 3, 2021; Revised March 4, 2021; Published July 1, 2021

Abstract: Apgar score is a test applied 1 minute after birth to check the infant health and can be performed as much as needed. The goal of this paper is to apply a deep learning (DL) method called convolutional neural network (CNN) to predict infants with potentially low Apgar score. Our CNN is a multi-input model that accepts denoised cardiotocography (CTG) images and gestational age. In the first half of the paper, we use basic machine learning (ML) techniques to explore what features and target labels are most effective. In the latter half, we verify to what extent the prediction accuracies can be improved by using our CNN model. Using 5-folds cross validation (CV), the CNN model performance scored an Area Under Curve (AUC) of 0.958 when classifying infants with Apgar score 5 minutes < 7 and AUC of 0.955 if Apgar score 1 or 5 minutes < 6 without using feature extraction algorithms. We conclude that the built model can be utilized as a prognosis tool to predict fetuses with a low Apgar score. Still, we think that a one model isn’t enough as obstetricians could benefit more from multiple models that help predict different risks to fetuses.

Key Words: convolutional neural network, cardiotocography, apgar score, image recognition

1. Introduction

Neonatal death, hypoxic-ischemic encephalopathy (HIE) and respiratory distress are a few examples of risk fetuses may face the moment they born. Such risks can be mitigated or prevented with
knowledge about the fetus condition inside the womb alongside speed and intensive care by medical team. Obstetricians perform different tests to assess the safety of mother and fetus during birth such as umbilical-artery blood pH and Apgar score.

Apgar score is a rapid standardized assessment test that is performed to babies after one minute of birth to assess their health. The test measures five categories: appearance, pulse, grimace, activity and respiratory. An Apgar score of 10 means the newborn scored 2 in all categories, which is the highest. The same test procedures are reapplied after 5 minutes of birth resulting into Apgar score 1 minute and Apgar score 5 minutes. The test may again be repeated, if needed, after 10, 15 or minutes of birth. Generally, an Apgar score of 7 or more is the average and it means the newborn is healthy. Low Apgar score can be an indicator to the risk of fetal compromise. For example, Apgar score 1 or 5 minutes of value between 0 to 6 is associated with higher risk of cerebral palsy and epilepsy. The lower the score is, the higher the possibility can be [24]. This doesn’t necessarily mean the infant will surely develop cerebral palsy, however, based on population studies, Apgar score after 5 and 10 minutes that are lower than 5 signifies higher risk of cerebral palsy [22]. Also, an Apgar score 5 of <7 has shown an association with the occurrence of neurologic disability and low cognitive functions in early adulthood [12], increased the risk of neonatal respiratory distress and hypoxic-ischemic encephalopathy (HIE) [28] and increased the risk of Autistic Disorder [20].

Cardiotocography (CTG) is a monitoring tool used to look for vital information affecting the health safety of mother and fetus during labor. There are multiple guidelines such FIGO, ACOG, JSOG etc. that tried to interpret the meaning of CTG signals and their consequence on mother and fetus. Nevertheless, these guidelines have disparity among them [29] and even obstetricians’ opinions vary in classifying CTG [27]. According to Each Baby Count report of 2019, more than 70% of stillbirths, neonatal death and brain injuries cases would have had different outcome with different care and out of 420 delivery cases, 236 were linked to CTG false interpretation [1]. There has been a debate if visual aided analysis is better than computer analysis or vice versa and some studies tried to compare them in a big scale [13, 21]. Both studies saw no significant advantage of using one over the other. Berglund, S et al. shown that the care for two-third of newborns with low Apgar score was substandard due to misinterpretation of cardiotocography (CTG) or to poor response to abnormal CTG at the right time [6]. Abnormal CTG are usually related to fetal distress. There is an association between low Apgar score and pathological CTG [5] and almost fifth of pathological CTG may have low Apgar score 5 [3].

2. Related works
Podda, M. et al. applied neural networks to predict the survival of preterm infants and achieved AUC of 0.91. The study was applied to infants with <30 gestational weeks and needed features that

1This is short for Apgar score after 5 minutes with a score less than 7. In this work, it is shortened to Apgar 5<7.
available within 5 minutes after birth [26]. In our previous work, we built an algorithm to extract
some of the main features obstetricians look for in CTG such as accelerations, decelerations, baseline,
etc. Using the algorithm, we trained a random forest ML model to classify Apgar score 1 or 5 of <6
and achieved AUC of 0.91 [4]. Zhao, Zhidong et al. applied CNN to CTG to predict cord acidemia at
birth, and Petroziello, Alessio et al. are the first to do so [25, 33]. One of their interesting conclusions
is their model will not perform well if cord acidemia is absent and CTG may better be diagnosed with
different models estimating different risks which might cause fetal compromise [25].

This work intentions are to provide the medical team with a prognosis tool to predict infants with
a high risk of low Apgar score so they can be ready to take the required medical procedures before
birth. We use a large dataset that is gathered under clinical conditions. So far, to the best of our
knowledge, no study used deep learning methods to analyze Apgar score classification. No study, we
are aware of, applied convolutional neural network on CTG to classify fetuses with low Apgar score.

3. Goal
This paper aims to build a deep learning convolutional neural network model that acts as a prognosis
tool to predict neonatal with possibly low Apgar score, which is one of the criteria that assess possible
risks on neonatal, so they can be ready to take the required medical procedures. This is done without
using algorithms to extract CTG features such as baseline, accelerations, decelerations, etc.

The CNN model architecture used in this work is a multi-input model of two blocks, a block that
accepts denoised CTG images and a block that accepts gestational age. The two blocks outputs are
concatenated to be trained together to predict low Apgar score as binary classification. Figure 3
shows a detailed visualization of the built model and Table IX describes the applied parameters. This
work, however, doesn’t claim to predict the implication, the meaning, or the condition of infants
that causes Apgar score to be low. It gives the medical team a heads up or foresight of whether the
newborn Apgar score will be high or low and the medical team prepares the necessary procedures
for such concurrence. This is work intent to rely on real data only. The dataset was not augmented
nor synthesized. The chosen metric to measure the model performance was the area under the curve
of receiver operating characteristics (AUC-ROC) because it measures the model’s ability to classify
positive, represented as 1 in this work, from negative classes, represented as 0.

4. Data set and processing
4.1 Description of the collected data
The dataset used in this experiment is a restricted internally available 38,073 CTG records from 2012
to 2017 gathered from several hospitals (Table I). In addition to FHR and UC data, the dataset
metadata include the delivery date, time and type, maternal and gestational age, fertility treatments,
fetal weight and gender, pH, Apgar score 1 and 5. All deliveries are singleton, extracted at the second
stage of labor and nearly all are low-risk pregnancies.

4.2 Denoising process
CTG records are prone to noise and signal drop. This is common to occur because the transducers
attached to the mother’s abdomen may not well fastened or because of the physical activities of both
the mother and the fetus, which affect the readings stability. Our dataset’s CTG recordings were
collected under clinical condition and some records include excessive noise, long signal drop and short
signal recording.

The process of noise removal was inspired by Zhao, Zhidong et al. [33]. However, the noise in the
dataset of this work is complex, so some tweaks were made to [33] approach. If any ≥ 15 seconds in
FHR is completely filled with 0s, then it is dropped. Any other 0s are removed (changed to null). Any
FHR which ≥ 200 bpm or ≤ 50 bpm is removed. If any point within 15 seconds is > |25| bpm than
the mean bpm of that 15 seconds, this point is removed. Finally, linear interpolation is applied to all
removed values and 0s. UC noises were reduced using moving average window of 60. An example of a
CTG denoising process can be seen in Fig. 1. Also, the time length of each record varies as some are
more than 3 hours and some others are less than 30 minutes. We included only the last 60 minutes before birth of every CTG.

4.3 Target tasks for prediction
After the denoising process, the same dataset was replicated to test its performance on two targets of Apgar score:

1. Dataset where Apgar score 1 or 5 < 6 as suggested by our medical researchers.
2. Dataset where Apgar score 5 < 7 as several guidelines and obstetricians label it as low

Creating a dataset where Apgar score 5 < 4 was disregarded because of the number of samples being too small. All datasets share the same size used in our paper [4] and that is 3519. This size was chosen for better comparison with the previous work and to reduce class imbalance. As seen in Table II, in the dataset is limited compared to the negative classes. Almost all positive classes from the original 38,073 records are included. All negative classes were picked at random using Pandas DataFrame.sample, which is a data analysis and manipulation tool in Python. Gestational age is provided by the dataset and it is the only non-image feature used. The mean of gestational age is 38 weeks, max is 42 and min is 22.

![Fig. 1. The process of denoising CTG. The first image represents the original signals and the second image represents the denoised signals.](image)

5. Feature extraction and prediction using machine learning methods
In this section, we introduce an algorithm that follows Japan Society of Obstetricians and Gynecology (JSOG) guidelines for CTG interpretation and apply machine learning methods to test the performance of the algorithm in detecting high risk birth on large data gathered under clinical conditions. The process is tested on umbilical cord pH alone, then on Apgar 1 and 5 only and finally on pH, Apgar score 1 and 5. This is achieved by following three major steps: processing the dataset, build the algorithm, apply SVM, RF, DT and ANN on extracted features. Before applying the algorithm, all CTG records are denoised using the same denoising algorithm explained in the data set and processing section and CTG records with high signal drop are excluded from the experiment. Only the 30 minutes before 20 minutes of giving birth are chosen for the experiment.
Table I. The names of institutes that provided the dataset.

| Name of Institute                  | No. of records |
|-----------------------------------|----------------|
| Fukuoka University Hospital       | 2,193          |
| Kyushu University Hospital Hospital| 3,539          |
| Obstetrics Gynecology             | 5,042          |
| Fukuda Hospital                   | 21,138         |
| Kumamoto University Hospital      | 1,231          |
| Total                             | 38,073         |

Table II. The number of positive and negative samples in training and testing data for the CNN model.

| Label | No. positive samples | No. negative samples |
|-------|----------------------|----------------------|
| Apgar1 | Apgar5 < 6           | 534                  | 2985                  |
| Apgar5 < 7       | 288                  | 3231                 |
| Apgar5 < 4       | 57                   | 3462                 |

5.1 CTG guidelines and feature extraction algorithm

CTG interpretation relies on a set of rules or guidelines to understand them. These guidelines are not universal. There are multiple attempts by experts in the field to standardize the management of CTG patterns [23]. The International Federation of Gynecology and Obstetrics ( FIGO ) is an example of an international guidelines. Still, health institutes usually use guidelines developed regionally or locally but it is not uncommon to find different health institutes in one country use different guidelines. Some examples are American College of Obstetrics and Gynecology ( ACOG ) in the US and Institute for Health and Care Excellence ( NICE ) guidelines in the UK. In Japan, the regularly used guidelines are made by Japan Society of Obstetricians and Gynecology ( JSOG ). Unlike other guidelines, JSOG uses 5-level classification of CTG patterns where 1 normal pattern and 5 as the highest risk variant pattern as shown in Table III. JSOG guidelines proved that it can predict early neonatal outcomes that usually results from low Apgar score, and low pH, thus it is medically acceptable [14].

Table III. JSOG five-tier fetal heart rate classification.

| FHR Pattern Classification Levels | FHR pattern levels & Classification |
|-----------------------------------|------------------------------------|
| Level 1                           | Normal pattern                     |
| Level 2                           | Benign variant pattern             |
| Level 3                           | Mild variant pattern               |
| Level 4                           | Moderate variant pattern           |
| Level 5                           | Severe variant pattern             |

We developed an algorithm that follows JSOG guidelines to detect important CTG features. The features extracted from CTG are: baseline, baseline variability, accelerations, early decelerations, mild/severe variable decelerations, mild/severe late decelerations and mild/severe prolonged decelerations.

The baseline is defined as follows: first, rounding FHR in every quarter second within 10 minutes period to the nearest 5 bpm, and then, the mode (most frequent number) of that 10 minutes is taken and set it as the baseline value. Baseline variability is counted by calculating the difference of FHR peaks per minute. Baseline variability is called “absent” if the difference is 0 bpm, “minimal” if less than 5, “moderate” if between 5 and 25 and “marked” if more than 25.

Regarding the definitions of accelerations and decelerations, in general, acceleration refers to when FHR increase rapidly, and deceleration refers to when FHR decrease rapidly or slowly. If FHR is 15 bpm or more above the baseline for 15-120 seconds, this is counted as acceleration. Variable deceleration is classified as “mild” when FHR is 15 bpm or more below the baseline for 15-120 seconds; and “severe” when the lowest point is less than 70 bpm and the duration is 30 seconds or more, or the lowest point is within 70-80 bpm and the duration is 60 seconds or more. Early deceleration
happens if FHR, which decrease gradually for 30 seconds or more, coincide with the occurrence of a UC where both start and end at the same time. Late deceleration is classified as “mild”, when decrease of FHR continues gradually for more than 30 seconds, begins near the highest point of UC, and ends after UC ends; and “severe”, when FHR lowest point is more than 15 bpm from the baseline. Prolonged deceleration is classified as “mild”, when FHR is more than 15 bpm below the baseline for 2-10 minutes; and “severe”, when FHR lowest point is less than 80 bpm.

within a window of 10 minutes, when accelerations or any type of decelerations occur, they are counted and assigned in their respective column as shown in Fig. 2.

In JSOG guidelines, the extracted features are rated with a corresponding level from the 5-level FHR patterns classification, also called judgment table (Table III), which indicate the health risk condition of the FHR. Instead of using the judgment table, we apply machine learning methods to our collected dataset for more accurate prediction results.

The final form of extracted features for each record is shown in Table IV. The algorithm is applied to every 10 minutes of the chosen 30 minutes where in each 10 minutes, features were extracted and judged independently, resulting in 42 features; 14 features for each 10 minutes. Fetal weight and gestational age which are provided by the dataset are also included.

Table IV. Each row in the dataset should have the below features for each 10 minutes. Since the size of each CTG is limited to 30 minutes, the total features equal (14×3)+2=44.

| Feature name                        | Description                     |
|-------------------------------------|---------------------------------|
| Gestational age                     | At the time of delivery         |
| Fetal Weight                        | At the time of delivery         |
| Variability                         |                                 |
| Absent                              | No. per 10 min                  |
| Minimal                             | No. per 10 min                  |
| Moderate                            | No. per 10 min                  |
| Marked                              | No. per 10 min                  |
| Baseline                            | Of 10 min                       |
| Acceleration                        | No. per 10 min                  |
| Early Deceleration                  | No. per 10 min                  |
| Mild Variable Deceleration          | No. per 10 min                  |
| Severe Variable Deceleration        | No. per 10 min                  |
| Mild Late Deceleration              | No. per 10 min                  |
| Severe Late Deceleration            | No. per 10 min                  |
| Mild Prolonged Deceleration         | No. per 10 min                  |
| Severe Prolonged Deceleration       | No. per 10 min                  |
| Judgment                            | 1-5 rating for features         |

5.2 Applying machine learning methods

Four ML methods were applied followed by their chosen hyperparameters: Support Vector Machine (SVM) (kernel = “linear”, C = 0.01, class_weight = “balanced”), Random Forest (RF) (n-estimators = 100, max_depth = 40), Decision Tree (DT) (max_depth = 6, criterion = “gini”, splitter = “best”), and Artificial Neural Network (ANN) (activation function = “relu”, layers size = (45,45), max iteration = 1000, optimization = “adam”).

All of the aforementioned are applied using Scikit-learn library\(^2\) for Python and the choices of the hyperparameter were based on the use of GridSearchCV in Scikit-learn which is a process for hyperparameter tuning for finding the optimal values for the model. Features in Table IV are used as input data and pH (< 7.1), Apgar score 1 and 5 (<6) as binary target labels where 1 denotes high risk birth and 0 as low risk birth.

The classification accuracy was quantified and compared using ROC-AUC. All of the three processes were tested multiple times using 5-fold cross-validation. To solve class imbalance, random undersampling (for negatives) and SMOTE oversampling (for positives) were applied during training, which

\(^2\)https://scikit-learn.org/
both are well-known methods used for handling data imbalance. Additionally, for better class balance, 3519 records in total were used in the experiment, making sure most of the positive classes in the original 38,073 records are included. This changes the ratio of positive to negative samples to $\sim 11\%$ for pH, $\sim 18\%$ for Apgar score 1 or 5, and $\sim 28\%$ for both (see Table V).

Table V. The number of positive and negative samples in training and testing data for the ML experiment. Positive classes refer to classes that satisfy the conditions of a label where negative classes refer to the ones that don’t.

| Label                  | No. positive samples | No. negative samples |
|------------------------|----------------------|----------------------|
| pH                     | 299                  | 3239                 |
| Apgar1 | Apgar5       | 534                  | 2985                 |
| pH | Apgar1 | Apgar5       | 775                  | 2744                 |

5.3 Experimental results
Both the processes of denoising CTG signals and the algorithm accuracy to predict the classic CTG features such as baseline, baseline variability, acceleration and decelerations have been tested several times under the supervision of expert obstetricians which helped produce satisfying results. ML classification performance outcomes for 5-fold cross validation are summarized in.

As can be seen, adding gestational age to the extracted features significantly improves the performance when Apgar score is the target label as in. A slight improvement also occurs when birthweight is added as a feature. Almost none of the aforementioned is happened when pH (Table VI) is the target label. It is also clear that in Table VIII the performance is less than Table VII because of the addition of pH as a target label.

In all cases, using Apgar score instead of pH produced much better classification performance. This is related to the findings of [14] where they found a significant difference between Apgar score 1 in high-risk birth cases using JSOG guidelines classification compared to low-risk birth cases. On the other hand, Apgar score 1 did not differ significantly between high and low risk birth cases in the subjective classification methods (three level risk classification). Another reason is the overall nature of pH analysis since pH, when obtained by obstetricians, does not discriminate respiratory acidemia from metabolic acidemia and the pH sampling fails with the ratio between 11% to 20% [32].

5.4 Discussions to improve accuracies
So far, we successfully improved the accuracies for detecting high-risk births by introducing various
methods such as denoising, constructing effective features, and choosing appropriate labels. Our method achieves 0.91 in prediction accuracy (see in Table VII, the row of RF with Extracted feature + Gestational age + birthweight). While it is sufficiently high, there is still a room to enhance it as described below.

While the dataset size provided flexible testing, the actual CTG recordings had plenty of noise to the point that may affect the signals quality and the ability to extract features from it.

One of the causes that limit the performances of the extracted features is related to the length enforced on the signals. Some signals, when checked manually, were found to have significant information that could cause high risk birth. However, the critical information time in the signal is not always consistent. For example, some serious decelerations were found within > 90 minutes before birth in some signals and < 30 minutes in some others. Forcing the length to 30 minutes before 20 minutes of giving birth may miss some serious information.

Still, using CTG guidelines in ML experiment may not be the best method to approach this problem. We are making the machine to form its decision based on guidelines which their accuracy is disputed [29]. Rather than trying to find which guideline is the most accurate, we can input the CTG images directly to CNN model for the classification using minimum features or without the need for interpretation. CNN takes in images and detect (learn) important features that could or couldn’t be perceived by human vision.

Table VI. ML using three feature sets on pH only as target label. In each test, extra features are being added to experiment the performance variation.

| Features                              | method | 5-fold CV AUC |
|---------------------------------------|--------|---------------|
| Extracted features only (#features = 44) | SVM    | .64 ± .06     |
|                                       | RF     | .62 ± .05     |
|                                       | DT     | .57 ± .05     |
|                                       | ANN    | .63 ± .04     |
| Extracted feature + Gestational age   | SVM    | .65 ± .02     |
|                                       | RF     | .64 ± .03     |
|                                       | DT     | .56 ± .04     |
|                                       | ANN    | .63 ± .03     |
| Extracted feature + Gestational age + birthweight | SVM    | .65 ± .05     |
|                                             | RF     | .68 ± .02     |
|                                             | DT     | .61 ± .02     |
|                                             | ANN    | .60 ± .03     |

6. Prediction using deep CNN model

We want our CNN model to be able to accept mixed data where the inputs are not of the same type. In our case, the inputs are two: CTG images and gestational age. Thus, our model consists of two blocks, each of which processes CTG images or embedded gestational-age vectors. This work uses Keras\(^3\), a deep learning API written in Python, and functional API in Keras library which helps manage multi-input models.

6.1 Image processing block

The first block is made solely for handling CTG images as inputs. The block is a pre-trained EfficientNet CNN architecture [30]. The reason to use EfficientNet is because of its high accuracy and being more computationally efficient than the best existing CNN/ConvNets. There are multiple architectures of EfficientNet such as EfficientNetB0 and EfficientNetB1 where the difference among each architecture is the number of layers in it. We applied transfer learning; the model is pre-trained on ‘imagenet’ dataset weights [10]. While the weights in imagenet are trained to classify objects, our model can benefit from the weights of lower layers, for it can detect low-level features such as edges and curves in images, rather than training the model from scratch. We can then edit or add on top layers with our own layers, and fine-tune all the layers.

\(^3\)https://keras.io/
Table VII. ML using three feature sets on both Apgar score 1 and 5 only as target labels. In each test, extra features are being added.

| Features                      | method | 5-fold CV AUC |
|-------------------------------|--------|---------------|
| Extracted features only       | SVM    | .71 ± .03     |
|                               | RF     | .73 ± .03     |
|                               | DT     | .68 ± .01     |
|                               | ANN    | .72 ± .02     |
| Extracted feature + Gestational age | SVM    | .88 ± .03     |
|                               | RF     | .89 ± .03     |
|                               | DT     | .82 ± .03     |
|                               | ANN    | .88 ± .03     |
| Extracted feature + Gestational age + birthweight | SVM    | .89 ± .02     |
|                               | RF     | .91 ± .02     |
|                               | DT     | .82 ± .02     |
|                               | ANN    | .88 ± .02     |

Table VIII. ML using three feature sets on pH, Apgar score 1 and 5 as target labels. In each test, extra features are being added.

| Features                      | method | 5-fold CV AUC |
|-------------------------------|--------|---------------|
| Extracted features only       | SVM    | .65 ± .02     |
|                               | RF     | .67 ± .04     |
|                               | DT     | .64 ± .02     |
|                               | ANN    | .64 ± .02     |
| Extracted feature + Gestational age | SVM    | .78 ± .01     |
|                               | RF     | .80 ± .01     |
|                               | DT     | .76 ± .01     |
|                               | ANN    | .79 ± .02     |
| Extracted feature + Gestational age + birthweight | SVM    | .78 ± .01     |
|                               | RF     | .82 ± .01     |
|                               | DT     | .76 ± .02     |
|                               | ANN    | .77 ± .01     |

On top of EfficientNet, we added two dense layers, which are hidden layers connected to each neuron in the following hidden layer, with one dropout layer after each dense layer. Dropout is used to reduce the likelihood of model overfitting. The extra added dense layers serve as fine-tuning to the model and the number of layers was chosen based on experiments and following best practices in the field.

6.2 Meta-data processing block

The second block is made to take in gestational age. Gestational age is one of the most important fetal factors affecting the heart rate curve of CTG [11]. In fact, Tables VII and VIII show that the prediction accuracies for the Apgar scores are clearly improved by adding gestational age as a feature. Our dataset contains both low and high-risk pregnancies with the majority being low risk. The values of gestational age are indifferent in both cases. In their study on low and high-risk pregnancies, Ahmadpour-Kacho, Mousa et al. [2] found gestational age was the same in low and high-risk mothers. The second block is a simple neural network that includes two dense layers.

6.3 Merging blocks

The output of both blocks is concatenated into one merged input of 64 dimensions vector (32 dim from the first block and 32 from the second). The last layer is a single unit that uses sigmoid activation function, which is suitable for our binary classification. The values of gestational age are scaled by preprocessing, and CTG images are also scaled to [0,1]. The dataset is imbalanced, and this affects the performance. One of the methods to solve the class imbalance is to use a class-weight method which provides each class with a weight. The class with the fewer sample will have heavier weight so it can be given more emphasis when training the model. The chosen optimizer and loss function were Adam and binary cross-entropy since Adam is the most recommend optimizer to be used in CNN and
Fig. 3. The architecture of the proposed model.

Table IX. The chosen parameters for the model.

| Layer Name                  | Parameters                                      |
|-----------------------------|-------------------------------------------------|
| efficientnet-b5: Model      | Max pooling                                     |
| Node_1: Dense               | Relu Activation function, L1 regularization      |
| Node_1.1: Dropout           | 0.3                                              |
| Node_1.2: Dense             | Relu Activation function                        |
| Node_1.3: Batch Normalization| Relu Activation, L1 regularization              |
| Node_1.4: Dense             | Relu Activation function                        |
| Node_1.5: Dropout           | 0.3                                              |
| Node_2: Dense               | Relu Activation function                        |
| Node_2.1: Dense             | Relu Activation function                        |
| Concatenate                 | Output of Both nodes                            |
| Merged_1: Dense             | Relu Activation function                        |
| Merged_1.1: Dense           | Relu Activation function                        |
| Merged_1.2: Dense           | Sigmoid Activation function                      |

the latter is made for binary classification.

Each target label was tested on CTG of 60 minutes and again on CTG of the first 30 minutes from the 60 minutes to test if CTG length has an impact on the performance. The model performance was tested using 5-fold stratified cross validation (CV). CV make multiple model predictions on the dataset. This provides us with more confidence in the performance of the model. Stratified CV is a CV variant that creates folds with a similar percentage of samples for every class. It is useful for our imbalanced dataset because the number of negative samples is more than the number of positive samples.
7. Results, ML vs DL comparison and discussion

Our CNN distinguishes CTG patterns with a low Apgar score from patterns with a high Apgar score more accurately than the extracted features+ML described in the previous section. Moreover, it is possible that the model learns more features from CTG with longer length as shown in Table X. While the change is small, AUC was higher in both labels with CTG with a length of 60 minutes.

Table X. AUC after training the model for different targets.

| Label          | 5-fold CV AUC for 30 minutes CTG | 5-fold CV AUC for up to 60* minutes CTG |
|----------------|---------------------------------|---------------------------------------|
| Apgar5 <6      | 0.949                           | 0.955                                 |
| Apgar5 < 7     | 0.953                           | 0.958                                 |

*Some records length is shorter than 60 minutes either because of their original length or because of the denoising process.

The model trained with a batch size of 8 and 10 epochs of training. More epochs caused the model to start overfitting. For the first block, all versions of EfficientNet were tested. The performance stopped improving after EfficientNetB5, so it is the version we based our results on.

Our ML model in the previous section used a total of 44 features, most of which required an algorithm to be extracted. This required a huge human interference throughout the whole process. Additionally, building an algorithm to extract CTG features can be prone to inaccuracies or may miss critical information and choosing which CTG guidelines to base the algorithm upon makes it an intricate task. The current CNN model relied only on the pre-existing data without the need for feature extraction algorithm. It means that it can bypass one layer of human interference. Relatively, it required less effort than the ML approach and was able to learn distinctive features from CTG images alone paired with gestational age outperforming our previous RF model with AUC 0.91. The training process depended on real data only; no data were augmented or synthesized since we believe that the number of samples is sufficient in our dataset.

Gestational age is an important factor in both the ML and the CNN experiments because of its relationship with Apgar score [2, 11]. The diversity of gestational age in our dataset is one of the reasons our model did well in predicting low Apgar score. The drop in performance could be connected to Cnattingius, Sven & Norman et al. findings where Apgar score prediction performance was better when used in infants with $\leq 31$ gestational weeks than full-term infants [9].

8. Conclusion

The proposed CNN model predicted fetuses with low Apgar score using pre-existing features. The ML approach used an algorithm based on JSOG guidelines to extract important features from CTG signals for interpretation. Different typical ML methods are applied to test the performance of the algorithm in detecting high risk birth on large data gathered under clinical conditions.

CTG signals were denoised for both the CNN and ML methods. Unlike the ML methods, the accuracy of the CNN method achieved higher accuracy even though its features were fewer and available. It learned to distinguish features directly from the CTG images whereas ML relied on the features extracted from the algorithm.

The model acts as a pre-labor tool to guide the medical team during labor. On the other hand, we think one model is not enough to determine fetal safety. Relying on the outcome of only one test such as Apgar score or pH could be misleading [18]. As future work, to progress this work in the right direction, it is desirable to create multiple models and tools for different tests each of which serves a specific purpose. For instance, a combination of models predicts Apgar score and umbilical cord pH can be used as a better indicator to detect complications in the newborn [17, 19]. Apgar score works well as a short-term prognosis tool to assess expected risks on neonatal specifically among preterm infants [8, 15, 31]. CNN and DL in general need extensive time to train and they are computationally demanding. However, after a model is trained, it is quick in analyzing new samples, making them
practical in real life situation.

Acknowledgments

This work was partially supported by JSPS KAKENHI # JP18K11449.

References

[1] Royal college of obstetricians and gynaecologists, each baby counts: 2019 Progress Report, 2020.
[2] M. Ahmadpour-Kacho, N. Asnafi, M. Javadian, M. Hajiahmadi, and N. Taleghani, “Correlation between umbilical cord ph and apgar score in high-risk pregnancy,” Iranian journal of pediatrics, vol. 20, no. 4, p. 401, 2010.
[3] L. Ali, R. Mushtaq, and N. Ahmed, “Frequency of pathological ctg in low risk women and its outcomes,” Pak J Surg, vol. 30, no. 4, pp. 340–345, 2014.
[4] K. Miyata, T. Imamura, S. Miyamoto, H. Fukunishi, A. Mohammad, and C. Shibata, “Application of machine learning techniques to classify fetal hypoxia using japan society of obstetricians and gynecology guidelines,” 2020. (Accepted but not published yet).
[5] S. Banu, “Relationship between abnormal cardiotocography and fetal outcome,” Nepal Journal of Obstetrics and Gynaecology, vol. 10, no. 2, pp. 36–39, 2015.
[6] S. Berglund, H. Pettersson, S. Cnattingius, and C. Grunewald, “How often is a low apgar score the result of substandard care during labour?,” BJOG: An International Journal of Obstetrics & Gynaecology, vol. 117, no. 8, pp. 968–978, 2010.
[7] B.M. Casey, D.D. McIntire, and K.J. Leveno, “The continuing value of the apgar score for the assessment of newborn infants,” New England Journal of Medicine, vol. 344, no. 7, pp. 467–471, 2001.
[8] S. Cnattingius, S. Johansson, and N. Razaz, “Apgar score and risk of neonatal death among preterm infants,” New England Journal of Medicine, vol. 383, no. 1, pp. 49–57, 2020.
[9] S. Cnattingius, M. Norman, F. Granath, G. Petersson, O. Stephansson, and T. Frisell, “Apgar score components at 5 minutes: risks and prediction of neonatal mortality,” Paediatric and perinatal epidemiology, vol. 31, no. 4, pp. 328–337, 2017.
[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” In 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255, Ieee, 2009.
[11] A.P. Drogtrop, R. Ubels, and J.G. Nijhuis, “The association between fetal body movements, eye movements and heart rate patterns in pregnancies between 25 and 30 weeks of gestation,” Early human development, vol. 23, no. 1, pp. 67–73, 1990.
[12] V. Ehrenstein, L. Pedersen, M. Grijota, G.L. Nielsen, K.J. Rothman, and H.T. Sørensen, “Association of apgar score at five minutes with long-term neurologic disability and cognitive function in a prevalence study of danish conscripts,” BMC pregnancy and childbirth, vol. 9, no. 1, pp. 1–7, 2009.
[13] M.G. Frasch, G.B. Boylan, H. Wu, and D. Devane, “Commentary: Computerised interpretation of fetal heart rate during labour (infant): a randomised controlled trial,” Frontiers in physiology, vol. 8, p. 721, 2017.
[14] M. Hayashi, A. Nakai, A. Sekiguchi, and T. Takeshita, “Fetal heart rate classification proposed by the perinatology committee of the japan society of obstetrics and gynecology: reproducibility and clinical usefulness,” Journal of Nippon Medical School, vol. 79, no. 1, pp. 60–68, 2012.
[15] H.C. Lee, M. Subeh, and J.B. Gould, “Low apgar score and mortality in extremely preterm neonates born in the united states,” Acta Paediatrica, vol. 99, no. 12, pp. 1785–1789, 2010.
[16] F. Li, T. Wu, X. Lei, H. Zhang, M. Mao, and J. Zhang, “The apgar score and infant mortality,” PloS one, vol. 8, no. 7, p. e69072, 2013.
[17] P.-A. Lönnqvist, “Smith’s anesthesia for infants and children,” 2011.
[18] G.C. Martin, R.S. Green, and I.R. Holzman, “Acidosis in newborns with nuchal cords and normal apgar scores,” Journal of Perinatology, vol. 25, no. 3, pp. 162–165, 2005.
[19] P. Meena, M. Meena, and M. Gunawat, “Correlation of apgar score and cord blood ph with severity of birth asphyxia and short-term outcome,” Int J Contemp Pediatr, vol. 4, no. 4, pp. 1325–1328, 2017.
[20] A. Modabbernia, S. Sandin, R. Gross, H. Leonard, M. Gissler, E.T. Parner, R. Francis, K. Carter, M. Bresnahan, D. Schendel, et al., “Apgar score and risk of autism,” European journal of epidemiology, vol. 34, no. 2, pp. 105–114, 2019.
[21] I. Nunes, D.A. Campos, A. Ugwumadu, P. Amin, P. Banfield, A. Nicoll, S. Cunningham, P. Sousa, C. Costa-Santos, and J. Bernardes, “Central fetal monitoring with and without computer analysis,” Obstetrics & Gynecology, vol. 129, no. 1, pp. 83–90, 2017.
[22] Committee on Obstetric Practice American Academy of Pediatrics-Committee on Fetus and Newborn, Committee opinion no. 644: the apgar score, Obstet Gynecol, vol. 126, no. 4, 2015.
[23] J.T. Parer, “Standardization of fetal heart rate pattern management: is international consensus possible?,” Hypertension Research in Pregnancy, vol. 2, no. 2, pp. 51–58, 2014.
[24] M. Persson, N. Razaz, K. Tedroff, K.S. Joseph, and S. Cnattingius, “Five and 10 minute apgar scores and risks of cerebral palsy and epilepsy: population based cohort study in sweden,” Bmj, vol. 360, 2018.
[25] A. Petrozzionello, C.W.G. Redman, A.T. Papageorghiou, I. Jordanov, and A. Georgieva, “Multimodal convolutional neural networks to detect fetal compromise during labor and delivery,” IEEE Access, vol. 7, pp. 112026–112036, 2019.
[26] M. Podda, D. Bacci, A. Micheli, R. Bellù, G. Placidi, and L. Gagliardi, “A machine learning approach to estimating preterm infants survival: development of the preterm infants survival assessment (pisa) predictor,” Scientific reports, vol. 8, no. 1, pp. 1–9, 2018.
[27] M. Rei, S. Tavares, P. Pinto, A.P. Machado, S. Monteiro, A. Costa, C. Costa-Santos, J. Bernardes, and D. Ayres-De-Campos, “Interobserver agreement in ctg interpretation using the 2015 figo guidelines for intrapartum fetal monitoring,” European Journal of Obstetrics & Gynecology and Reproductive Biology, vol. 205, pp. 27–31, 2016.
[28] E.M.A. Salustiano, J.A.D.B. Campos, S.M. Ibidi, R. Ruano, and M. Zugaib, “Low apgar scores at 5 minutes in a low risk population: maternal and obstetrical factors and postnatal outcome,” Revista da Associação Médica Brasileira, vol. 58, no. 5, pp. 587–593, 2012.
[29] S. Santo, D.A. Campos, C. Costa-Santos, W. Schnettler, A. Ugwumadu, L.M.D. Graça, and FM-Compare Collaboration, “Agreement and accuracy using the figo, acog and nice cardiotocography interpretation guidelines,” Acta obstetricia et gynecologica Scandinavica, vol. 96, no. 2, pp. 166–175, 2017.
[30] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” In International Conference on Machine Learning, pp. 6105–6114, PMLR, 2019.
[31] B. Weinberger, M. Anwar, T. Hegyi, M. Hiatt, A. Koons, and N. Paneth, “Antecedents and neonatal consequences of low apgar scores in preterm newborns: a population study,” Archives of pediatrics & adolescent medicine, vol. 154, no. 3, pp. 294–300, 2000.
[32] E. Wiberg-Itzel, C. Lipponer, M. Norman, A. Herbst, D. Prebenesen, A. Hansson, A.L. Brygelsson, M. Christoffersson, M. Sennström, U.B. Wennerholm, et al., “Determination of ph or lactate in fetal scalp blood in management of intrapartum fetal distress: randomised controlled multicentre trial,” Bmj, vol. 336, no. 7656, pp. 1284–1287, 2008.
[33] Z. Zhao, Y. Zhang, Z. Comert, and Y. Deng, “Computer-aided diagnosis system of fetal hypoxia incorporating recurrence plot with convolutional neural network,” Frontiers in physiology, vol. 10, p. 255, 2019.