Incorporating Knowledge Base for Deep Classification of Fetal Heart Rate*

Changping Ji[1][0000−0002−6671−8811], Min Fang[2,*][0000−0002−8798−537X], Jie Chen[1][0000−0002−9811−1694], Muhammad Umair Raza[1], and Jianqiang Li[1][0000−0002−2208−962X]

[1] College of Computer Science and Software Engineering, Shenzhen University, 518060 Shenzhen Guangdong, China
jichangping2018@email.szu.edu.cn, {chenjie, lijq}@szu.edu.cn, umair2007pak@gmail.com

[2] Education Center of Experiments and Innovations, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China
fangmin@hit.edu.cn

Abstract. In recent years, remote fetal monitoring has become more and more popular, and it has also brought many challenges. Fetal heart rate records are generally recorded by pregnant women using a fetal monitor at home. Due to the improper operation of the pregnant woman and the surrounding noise, this makes it difficult for the doctor to give an accurate diagnosis. However, the existing methods are difficult to perform well in an environment with noisy data and unbalanced data. To solve the shortcomings of existing methods, we design a novel framework, classification fetal heart rate based on convolutional neural network incorporating knowledge base. In particular, we built a knowledge base for the task of fetal heart rate classification, which can solve the problem of noise and imbalance in the data. To verify the effectiveness of our proposed framework, we conduct extensive experiments on a real-world dataset. the experimental results show that the performance of our framework is better than other methods.

Keywords: Fetal heart rate · Knowledeg base · Classification.

1 Introduction

At present, fetal heart rate (FHR) [13, 10, 6] is an important feature for monitoring the health of the fetus. And electronic fetal monitoring (EFM) [2, 17] for recording FHR has also become an important technology for monitoring fetal health in utero. EFM can detect fetal distress and fetal hypoxia in time and has been widely used in clinical practice. In recent years, the main method of fetal monitoring is that pregnant women use a fetal monitor to record FHR at home [8], and then send it to the doctor for diagnosis through mobile devices, which

* Supported by organization x.

can be found in Figure. 1. Due to the factors such as pregnant women's heartbeat, breathing and improper operation, the recorded FHR is not professional enough, which brings certain challenges to doctors in diagnosing FHR. Hence, improving the method of fetal monitoring has important clinical significance.

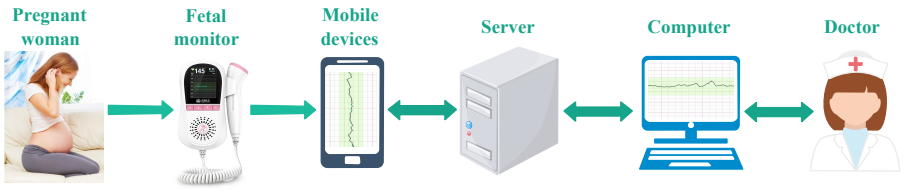


Fig. 1. Remote fetal monitoring system.

In recent years, due to the continuous development of computer science, EFM technology has been continuously improved, and many researchers have made different progress [4]. Georgoulas et al. [9] proposed to use hidden Markov models to classify FHR. Spilka et al. [15] proposed a combination of conventional and nonlinear features to analyze FHR, and later [16] proposed using a multi-scale feature representation method to quantify the variability of FHR, and then used a sparse support vector machine (SVM) to perform supervised classification of FHR. But the model is very dependent on the accuracy of feature extraction. Yu et al. [18] proposed a FHR classification model that uses a non-parametric Bayesian model to distinguish between fetal hypoxia and normal FHR records. This model is mainly based on the hierarchical dirichlet model, which infers a mixed model from healthy FHR and unhealthy FHR, and finally uses the inferred mixed model to classify the new FHR. Dash et al. [7] proposed a FHR classification method based on Bayesian theory and generative models, this framework first extracts the features of fetal status from FHR records, then defines class-specific models for them and uses a set of training data to estimate the parameters of each model. Sbrollini et al. [14] proposed an automatic algorithm for recognition and classification of fetal deceleration, but the model completely relies on the feature of fetal movement records. Based on existing methods, Barnova et al. [3] proposed a hybrid model using multiple independent methods. Li et al. [11] proposed to use the convolutional neural network (CNN) to automatically classify FHR, and vote according to the class of each segment to get the final result, but this voting mechanism is not accurate enough for medical classification.

Although the above studies have made different progress, there are still problems and challenges, which can be summarized in two aspects: 1) Existing models are very dependent on the accuracy of data quality, and do not work well for noisy data; 2) Another prerequisite for the current methods to work well is data balance, but the data in the medical field is usually imbalance. To address the aforementioned challenges, we propose a framework, classification fetal

heart rate based on convolutional neural network incorporating knowledge base (CNN-KB). First, we use the keywords in the pregnant women's question and doctor's answer to build a knowledge base (KB). In the training phase, the keywords in the pregnant woman's question and the doctor's answer are used as an auxiliary input. In the inference phase, because the doctor's answer is not provided, we need to use the keywords in question to query the KB, then use the queried information as an auxiliary input. For the FHR, the training phase and the inference phase are the same, we use CNN to extract its features. Finally, the features of FHR and auxiliary input are merged, and then used for classification. To verify the effectiveness of our framework, we conduct extensive experiments on a real-world dataset. The experimental results show that CNN-KB is much better than baseline methods.

2 Method

2.1 Knowledge Base

We will briefly introduce how the KB in the framework is constructed before introducing the proposed framework. The KB was constructed on data derived from the textual information of the pregnant woman's questions Q and the doctor's answers A. First, we use the word segmentation tool [5] to segment the texts of Q and A, we have also performed synonym substitution [1] processing on keywords because of the different expressions of each person, and then pick out the meaningful words under the guidance of the doctor. The keywords set of Q and A are represented by Q^w and A^w, respectively. Second, traverse the keywords of each question Q_i. If the keyword q of Q_i in the Q^w, then judge whether the keyword a of the corresponding answer A_i is in A^w, if so, create an entity-relationship triples $ert_{q,a}$ between q and a. After traversing, each keyword

Algorithm 1: Construct Knowledge Base

Input: The questions Q and answers A.
Output: Knowledge Base.

perform word segmentation and pick out Q^w and A^w;
foreach $q \in Q_i$ **do**
  if $q \in Q^w$ **then**
    foreach $a \in A_i$ **do**
      if $a \in A^w$ **then**
        create an entity-relationship triple $ert_{q,a}$; $ERT \leftarrow ERT + ert_{q,a}$;
      else
        continue;
  else
    continue;
construct knowledge base based on ERT;

in Q_i may correspond to multiple keywords in A^w, or one, or even none. Finally, build the KB based on the relationship between keywords, the tool we use to build the KB is Protege [12]. The construction details of KB are in Algorithm 1, ERT is the set of entity-relationship triples.

2.2 Classification with Knowledge Base

In order to solve the shortcomings of traditional methods, we propose a fetal heart rate classification framework based on convolutional neural network combined with KB. The model used in the experiment to extract FHR features is a one-dimensional CNN. The CNN-KB framework is shown in the Figure. 2.

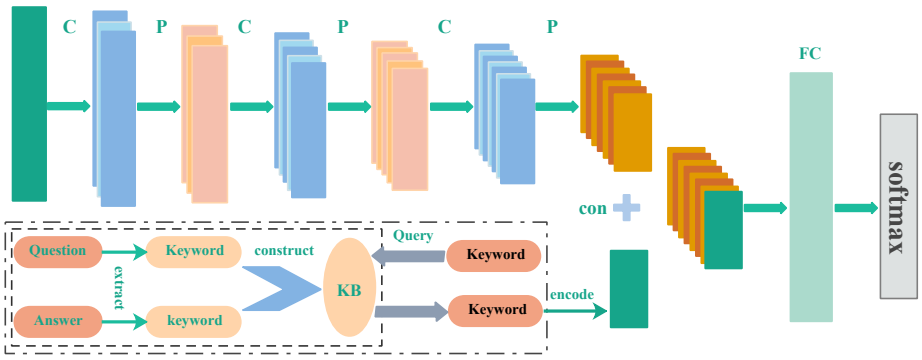


Fig. 2. The framework of CNN-KB.

In the training phase, because the doctor's answer is provided, the real relationship between the keywords in the question and the keywords in the answer is used as an auxiliary input. However, in the inference phase, the doctor's answer is not provided, we need to use the keywords in question to query the KB, then use the queried information as an auxiliary input. The KB can be defined as two parts, entity set $E = \{e_1, e_2, \cdots, e_n\}$ and relationship set $R = \{r_1, r_2, \cdots, r_m\}$. An entity-relationship triple in KB can be defined as $S_{i,j} = \{e_i, r_k, e_j\}$, where e_i, e_j are the entities in E and r_k is the relationship in R. Assuming that an entity e_i in E is related to z entities, then the information set G_{e_i} queried from KB by e_i can be formally defined as follows:

$$G_{e_i} = S_{i,1} \cup S_{i,2} \cup \cdots \cup S_{i,z} \tag{1}$$

An example of querying information in KB through entity is shown in Figure. 3, "anemia" is a keyword entity used for query. The left side of the figure is the set of entities in the KB, and the right side of the figure is the relationship queried by the corresponding entity. Since there may be more than one keyword

in each question, if there are multiple meaningful keywords, the queried information needs to be merged. Therefore, the general form of the KB information set K corresponding to a question can be defined as:

$$K = G_{e_i} \cup G_{e_i+1} \cup \cdots \cup G_{e_j} \tag{2}$$

Then embed and encode the information set K to obtain the feature vector representing T. The calculation formula for T is as follows:

$$T = f(Embedding(K)) \tag{3}$$

where f is LSTM in our framework. On the other hand, for the input FHR,

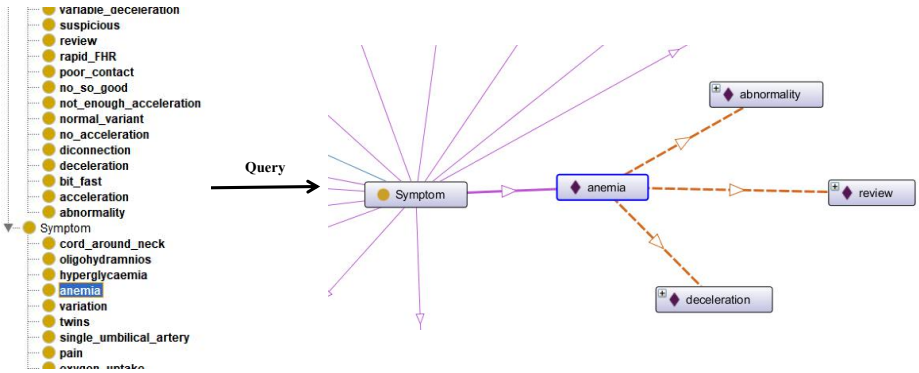


Fig. 3. Example of querying information from KB based on keywords.

the framework uses the CNN model to extract features. As shown in Figure. 2, C represents the convolution operation, P represents the pooling operation, *con* represents the concatenation operation, and FC is the fully connected layer. The model contains three convolutional layers, and the output feature R of each convolutional layer is formally defined as follows:

$$R = P(\sigma(W_c * \hat{R} + b_c)) \tag{4}$$

where $\hat{R}$ is the input of the convolutional layer, the input of the first convolutional layer is FHR; W_c and b_c are the learnable parameters in the convolutional layer; σ is an activation function.

Finally, the features R and T are fused and input into the fully connected layer, and then the $softmax$ activation function is used for classification. The formula is defined as follows:

$$Z = softmax\left(W_{fc}(R \oplus T) + b_{fc}\right) \tag{5}$$

where Z is the classification result, W_{fc} and b_{fc} are the learnable parameters of the fully connected layer, $\oplus$ is concatenation operation.

3 Experiments Setup & Results

3.1 Dataset and Preprocessing

The experimental data for this work was provided by Shenzhen Sdyunban Health Technology Co., Lty. These data are recorded by pregnant women using electronic fetal monitoring at home and then submitted to doctors for diagnosis through the platform. The label of FHR and the answer to the question are all recorded by doctors. The sample of FHR recorded by electronic fetal monitoring is shown in the Figure. 4, which is a segment of data. The time of normal FHR data is at least 20 minutes.

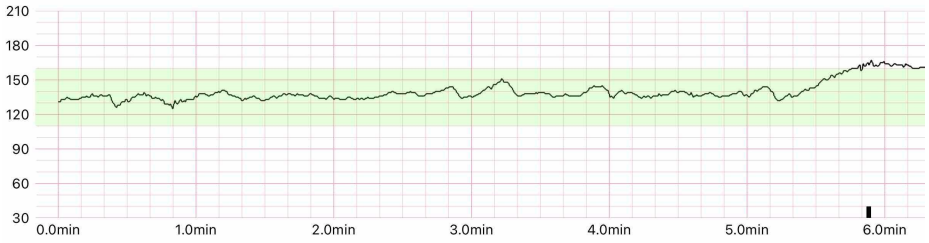


Fig. 4. A segment of FHR data, the ordinate is fetal heart rate, and the abscissa represents the time.

We have performed four samples per second on the FHR data to generate sequence data according to the standard [11], that is, the length of a twenty-minute FHR data is 4800. We collected a total of 8,000 pieces of data, divided into three classes: normal, suspicious and abnormal. After deleting some data with missing information, the length distribution of FHR data is shown in the Figure. 5. According to the length distribution of the FHR data, FHR data with a length less than 4800 and a length greater than 8300 are deleted. Since the data was collected by pregnant women at home, some data was not collected due to the improper operation of pregnant women, that is, the value of the sequence data is 0. Therefore, during the data preprocessing, we counted the number of zero values in each data, and then deleted data with more than 500 zero values according to the doctor's opinion. The final number of experimental data is 6,392, of which 80% are used as training data, 10% are used as verification data, and 10% are used as test data.

It can be seen from Figure. 5 that if the length of the data is unified by directly using the padding or cropping method, the short data will be padded too much, and the useful information of the long data may be cropped. Therefore, we use FHR data with a length of 4800 as the standard to down-sample the data with a length of longer than 4800. The down-sampling algorithm is as follows:

$$index = \left\{ \left\lfloor \frac{L_X}{L_X - \hat{L}} \right\rfloor * i \right\}_{i=1}^{L_X - \hat{L}} \tag{6}$$

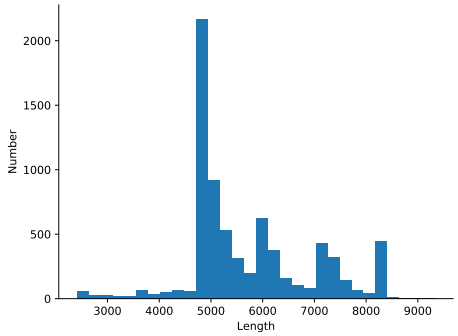


Fig. 5. The length distribution of FHR data.

Where *index* is the set of indexes (starting from 1) of the values to be discarded in data X, L_X is the length of the sequence data X, and $\hat{L}$ is the length of the standard sequence data. Even after using this down-sampling method for the longest sequence (8300) of data, the data sampling frequency is greater than two samples per second. In this case, sufficient feature information is still retained.

3.2 Implementation Details

The configuration and parameter settings of the experiment are mainly as follows: the deep learning framework used in the experiment is Keras whose bottom layer is Tensorflow, and the hardware configuration is two NVIDIA Quadro P5000 GPU. The CNN model includes three convolutional layers, and the number of convolution kernels in each layer is 32, 64, and 128, respectively, the corresponding convolution step size is 5, 5, and 3, respectively. After each convolutional layer, the *ReLU* activation function is used, and the pooling used is the average pooling. When processing the information extracted from the KB, the Embedding layer is used first, and the output dimension is set to 128. Then use LSTM to encode the embedded features, and the output dimension is also set to 128. The learning rate is set to 0.0001, the optimizer uses *Adam*, the final activation function for multi-classification is $softmax$, the activation function for binary classification is *sigmoid*, batch-size is set to 32, and epoch is set to 200.

3.3 Evaluation Metrics

We use three evaluation metrics to evaluate the experimental results. The evaluation metrics are *Accuracy*, *Precision* and *Recall*. The calculation formula of *Accuracy* is as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

where TP, TN, FP, and FN are the number of true positive, true negative, false positive, and false negative, respectively. The difference between $Precision$ and $Accuracy$ is that $Precision$ only pays attention to the number of positive samples predicted to be positive. The calculation formula of $Precision$ is as follows:

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

In the classification of medical diseases, data imbalance is a common phenomenon, so the $Recall$ is necessary. The $Recall$ is defined as follows:

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

3.4 Comparison Methods

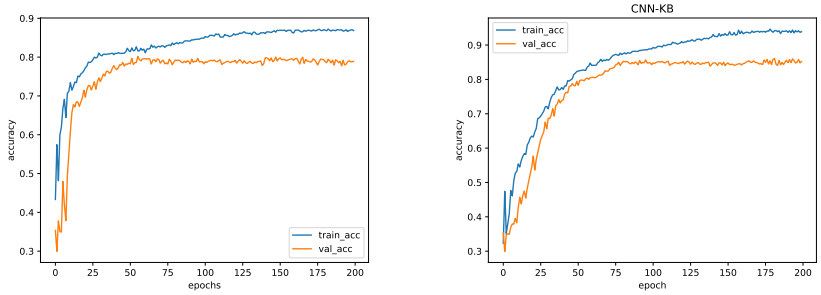
In order to prove the effectiveness of the framework proposed in this paper, we compare with two methods, namely CNN and SVM. Li et al. [11] proposed to use one-dimensional CNN to classify FHR. This method first slices FHR data, then predicts the class of each segment, and finally uses voting to get the class of the entire FHR data. This is inaccurate in the classification of medical diseases. If there is an abnormal segment, it means that the entire FHR is abnormal, not a class with many votes. Therefore, we do not use the voting mechanism in the comparison process, but classify the entire FHR data. SVM is a feature extraction based on the basic statistical method.

3.5 Results Analysis

We first performed multi-classification (normal, suspicious, and abnormal) experiments on the three methods. The experimental results of the CNN and CNN-KB methods are shown in the Figure. 6. It can be seen from the figure that the fitting speed of the CNN model is faster than the CNN-KB model. This is because the CNN-KB model integrates KB information, which leads to slower model convergence. At the same time, it can also be seen that although the CNN-KB model has a slower fitting speed, the final performance of the model is higher than the CNN model, which also verifies the effectiveness of our proposed framework. It proves that the text information has a certain restraint ability on the classification results.

Table 1. Evaluation results of multi-class CNN and CNN-KB models on the test set.

	Accuracy	Precision	Recall
SVM	0.7301	0.7413	0.7001
CNN	0.7905	0.8172	0.7814
CNN-KB	0.8218	0.8216	0.8131

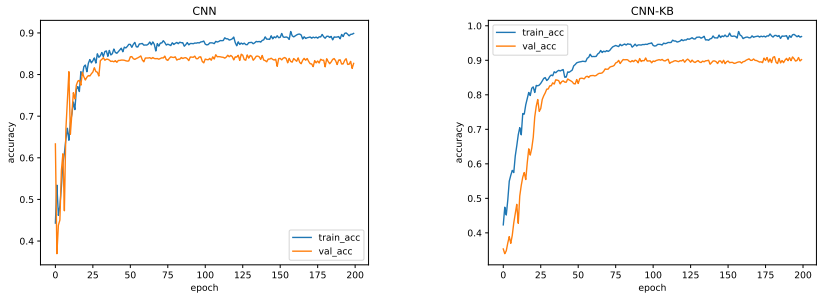


(a) Training and verification results of one-dimensional CNN.

(b) Training and verification results of the CNN-KB model.

Fig. 6. The performance comparison of CNN and CNN-KB models in multi-classification.

We evaluated the performance of the three models of SVM, CNN and CNN-KB on the test set, and the final evaluation results are shown in Table 1. It can be seen from the table that the SVM model has achieved the worst results on three metrics. This is because the experimental data has insufficient professionalism, which leads to poor quality of the features extracted for SVM. The performance of the CNN model and the CNN-KB model on the *Precision* metric is relatively close, while in the other two metrics, the performance of CNN-KB is higher than the CNN model. In general, although the proposed framework CNN-KB is slower in model fitting, its performance is higher than other methods.



(a) Training and verification results of one-dimensional CNN.

(b) Training and verification results of the CNN-KB model.

Fig. 7. The performance comparison of CNN and CNN-KB models in binary classification.

In the FHR multi-classification task, the proportion of abnormal data is too small, about 3%, which will make the model perform poorly in abnormal classi-

fication. From the perspective of auxiliary diagnosis, we consider suspicious and abnormal as one class under the advice of doctors, and only need to distinguish whether it is a normal FHR. Therefore, we conducted binary classification experiments on three methods. The training results of the CNN model and the CNN-KB model are shown in Figure. 7. It can be seen from the figure that the performance of the binary classification task is better than the multi-class classification task, this is because the data imbalance becomes lower.

Finally, as with the multi-classification task, we also evaluated the performance of the three models of SVM, CNN and CNN-KB on the test set, and the final evaluation results are shown in Table 2. It can be seen from the results in the table that SVM still achieved the lowest score on the three metrics, and the reason is the same as in the multi-classification task. The performance of the CNN-KB model on the three evaluation metrics is still higher than the CNN model. The experimental results once again verify the effectiveness of the proposed framework.

Table 2. Evaluation results of binary classification CNN and CNN-KB models on the test set.

	Accuracy	Precision	Recall
SVM	0.7526	0.7612	0.7413
CNN	0.8201	0.8331	0.8051
CNN-KB	0.8454	0.8512	0.8269

4 Conclusion

In order to solve the problems of imbalance and lack of professionalism in FHR data. we propose a framework, classification fetal heart rate based on convolutional neural network incorporating knowledge base. We first use the keywords of the pregnant women's questions and the doctor's answers to build a KB. And then in the inference phase, we use the keywords in the pregnant women's questions to query the KB. Finally, the FHR features and the query information are merged and classified. To verify the effectiveness of our proposed framework, we conduct extensive experiments on a real-world dataset. the experimental results show that the performance of our framework is better than other methods. In the future, we will study how to automatically build a KB so that it can be applied to more scenarios.

References

1. Abrahamsson, E., Forni, T., Skeppstedt, M., Kvist, M.: Medical text simplification using synonym replacement: Adapting assessment of word difficulty to a

compounding language. In: Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations. pp. 57–65 (2014)
2. Alfirevic, Z., Gyte, G.M., Cuthbert, A., Devane, D.: Continuous cardiotocography (ctg) as a form of electronic fetal monitoring (efm) for fetal assessment during labour. Cochrane Database of Systematic Reviews (2) (2017)
3. Barnova, K., Martinek, R., Jaros, R., Kahankova, R.: Hybrid methods based on empirical mode decomposition for non-invasive fetal heart rate monitoring. IEEE Access **8**, 51200–51218 (2020)
4. Cahill, A.G., Tuuli, M.G., Stout, M.J., López, J.D., Macones, G.A.: A prospective cohort study of fetal heart rate monitoring: deceleration area is predictive of fetal acidemia. American Journal of Obstetrics and Gynecology **218**(5), 523–e1 (2018)
5. Chang, P.C., Galley, M., Manning, C.D.: Optimizing chinese word segmentation for machine translation performance. In: Proceedings of The Third Workshop on Statistical Machine Translation. pp. 224–232 (2008)
6. Cömert, Z., Kocamaz, A.F.: Open-access software for analysis of fetal heart rate signals. Biomedical Signal Processing and Control **45**, 98–108 (2018)
7. Dash, S., Quirk, J.G., Djurić, P.M.: Fetal heart rate classification using generative models. IEEE Transactions on Biomedical Engineering **61**(11), 2796–2805 (2014)
8. Fanelli, A., Ferrario, M., Piccini, L., Andreoni, G., Matrone, G., Magenes, G., Signorini, M.G.: Prototype of a wearable system for remote fetal monitoring during pregnancy. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology. pp. 5815–5818. IEEE (2010)
9. Georgoulas, G.G., Stylios, C.D., Nokas, G., Groumpos, P.P.: Classification of fetal heart rate during labour using hidden markov models. In: 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541). vol. 3, pp. 2471–2475. IEEE (2004)
10. Ibrahimy, M.I., Ahmed, F., Ali, M.M., Zahedi, E.: Real-time signal processing for fetal heart rate monitoring. IEEE Transactions on Biomedical Engineering **50**(2)
11. Li, J., Chen, Z.Z., Huang, L., Fang, M., Li, B., Fu, X., Wang, H., Zhao, Q.: Automatic classification of fetal heart rate based on convolutional neural network. IEEE Internet of Things Journal **6**(2), 1394–1401 (2018)
12. Musen, M.A.: The protégé project: a look back and a look forward. AI matters **1**(4), 4–12 (2015)
13. Nageotte, M.P.: Fetal heart rate monitoring. In: Seminars in Fetal and Neonatal Medicine. vol. 20, pp. 144–148. Elsevier (2015)
14. Sbrollini, A., Carnicelli, A., Massacci, A., Tomaiuolo, L., Zara, T., Marcantoni, I., Burattini, L., Morettini, M., Fioretti, S., Burattini, L.: Automatic identification and classification of fetal heart-rate decelerations from cardiotocographic recordings. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 474–477. IEEE (2018)
15. Spilka, J., Chudacek, V., Koucky, M., Lhotska, L., Huptych, M., Janku, P., Georgoulas, G., Stylios, C.: Using nonlinear features for fetal heart rate classification. Biomedical signal processing and control pp. 350–357 (2012)
16. Spilka, J., Frecon, J., Leonarduzzi, R., Pustelnik, N., Abry, P., Doret, M.: Intrapartum fetal heart rate classification from trajectory in sparse svm feature space pp. 2335–2338 (2015)
17. Stout, M.J., Cahill, A.G.: Electronic fetal monitoring: past, present, and future. Clinics in Perinatology **38**(1), 127–142 (2011)
18. Yu, K., Quirk, J.G., Djurić, P.M.: Fetal heart rate classification by non-parametric bayesian methods. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 876–880. IEEE (2017)