Das *et al. Health Inf Sci Syst (2020) 8:16*
https://doi.org/10.1007/s13755-020-00107-7

Health Information Science and Systems

RESEARCH

Cardiotocograph-based labor stage classification from uterine contraction pressure during ante-partum and intra-partum period: a fuzzy theoretic approach

Sahana Das[1], Sk Md Obaidullah[2], K. C. Santosh[3], Kaushik Roy[1*] and Chanchal Kumar Saha[4]

Abstract

Computerized techniques for Cardiotocograph (CTG) based labor stage classification would support obstetrician for advance CTG analysis and would improve their predictive power for fetal heart rate (FHR) monitoring. Intrapartum fetal monitoring is necessary as it can detect the event, which ultimately leads to hypoxic ischemic encephalopathy, cerebral palsy or even fetal demise. To bridge this gap, in this paper, we propose an automated decision support system that will help the obstetrician identify the status of the fetus during ante-partum and intra-partum period. The proposed algorithm takes 30 min of 275 Cardiotocograph data and applies a fuzzy-rule based approach for identification and classification of labor from 'toco' signal. Since there is no gold standard to validate the outcome of the proposed algorithm, the authors used various statistical means to establish the cogency of the proposed algorithm and the degree of agreement with visual estimation were using Bland–Altman plot, Fleiss kappa (0.918 ± 0.0164 at 95% CI) and Kendall's coefficient of concordance ($W = 0.845$). Proposed method was also compared against some standard machine learning classifiers like SVM, Random Forest and Naïve Bayes using weighted kappa (0.909), Bland–Altman plot (Limits of Agreement 0.094 to 0.0155 at 95% CI) and AUC-ROC (0.938). The proposed algorithm was found to be as efficient as visual estimation compared to the standard machine learning algorithms and thus can be incorporated into the automated decision support system.

Keywords: Cardiotocograph, Toco, Stages of labor, Bland–Altman plot, Fleiss kappa, Kendall's coefficient of concordance

Introduction

Continuous electronic fetal monitoring is a part and parcel of modern day obstetric management. In recent years, approximately 3.4 million fetuses in the United Stated were assessed with electronic fetal monitoring (EFM), making it the most common obstetric procedure. Despite widespread belief as an insignificant and sometimes misleading tool, every hospital in all parts of the world is equipped with a CTG monitors which itself speaks about its significance [1]. Fetal heart rate and uterine pressure are two inseparable contents of CTG. Though ignored earlier, uterine contraction monitoring plays a pivotal role in the final interpretation of a CTG trace. The renewed interest in the uterine activity during labor has stemmed up from the evidence, which shows that labor, particularly dysfunctional one can affect the long term health of the offspring. While there has long been suspicion in the clinical literature that abnormal labor can injure the fetus, most of the studies are hampered by small sample sizes or imprecise definitions of dystocia. Specifically, prolonged latent phase was not associated with adverse outcome; but protraction and arrest disorders were, with arrested labor being the major contributor to childhood problems. A major portion of the demonstrable adversity associated with dysfunctional labors was attributable to

*Correspondence: kaushik.mrg@gmail.com
[1] Department of Computer Science, West Bengal State University, Kolkata 700124, West Bengal, India
Full list of author information is available at the end of the article

the mode of delivery or to factors that led to the development of abnormal labor [2]. Though a decline in the cases of neonatal seizures had been noted due to the introduction of fetal monitoring, incidence of perinatal death had not improved in general [3]. During active second stage of labor, the fetus is at the high risk of developing acidosis. Thus, the FHR pattern as well as the uterine activity or toco should be clearly followed. Fetal monitoring system was initially developed to measure FHR. But it was not enough to determine the fetal distress. Hence uterine contraction pressure (UCP) was introduced later as the deceleration pattern of fetal heart rate (FHR) need to be correlated with the duration and amplitude of uterine contraction to determine the status of the fetus. In the early days of the automation of FHR analysis, not much attention was given to the recognition of contraction and the identification of deceleration types. Most modern algorithms used to monitor the status of the fetus were more stable and they were taking into account both the FHR and the UCP. Over the years several commercial systems have shown encouraging results. They are given in the Table 1.

Dawes and Redman worked with FHR signal in 1977 at the Oxford University in 1977. They mainly worked with patients between 26 and 42 weeks of pregnancy. They developed a system called System 8000 which worked on a set of criteria called Dawes/Redman criteria. This system eventually evolved to SonicaidFetalCare. It is based on the crisp logic proposed in the FIGO guidelines [4].

NST-EXPERT was proposed in 1995 by Alonso-Betansos. The system consists of a deterministic and a heuristic unit. The deterministic module is responsible for the acquisition of patient data, whereas, heuristic module is responsible for interpreting these data to help in the diagnostic decision making [5].

In 1995 Betanzos and Berdinas developed NST-EXPERT—an expert system to diagnose the fetal state using non-invasive technique. This system is also capable of proposing a treatment [6, 7].

2CTG2 was proposed by Magenes et al. [8]. This system samples the FHR signal at 2 Hz and extracts the various features of both the signals. Approximate Entropy Analysis was used for the analysis. This algorithm did not produce any significant clinical improvement, however, it reduced inter and intra-observer variability.

In their recent work Comert et al. established the importance of feature selection algorithms in the automated analysis of CTG as this reduces the dimension of the feature set as well as reveals the most significant features without losing the important information [9].

But they all have failed to fulfill the levels of expectation due to the inherent uncertainties associated with the CTG interpretation. The goal of the researchers in this paper is to provide the obgyn experts with an automated decision support system that will help them identify the status of the fetus.

One of the parameters that is needed to determine the fetal wellbeing is to determine the stage of labor. In this paper the authors present a novel method of identifying the stage of labor using soft computing based technique.

Problem statement—why uterine activity assessment is essential

Assessment of uterine activity is an index to predicting not only the stage of labor, but the fetal wellbeing as well. During pregnancy the uterus is stretched progressively. Contractions are the physiological response to this stretch. Frequency and intensity of contraction increases as the pregnancy progresses [10]. Excessive uterine activity may give rise to fetal hypoxia. There are several reasons for monitoring the uterine activity:

i. Fetus may already be compromised even when the mother is not in labor. During active labor such fetus may become further compromised.
ii. Abnormal uterine contraction when the mother is not in labor is ominous.
iii. It shows how the labor is progressing.
iv. In active labor patients if the uterine contraction ceases then it may indicate uterine rupture and subsequent maternal morbidity.

Table 1 Commercialized system for CTG interpretation till date

System name	System details
System 8000	Dawes & Redman (1981). First commercial system.
SonicaidFetalCare	Upgraded version of System 8000
NST-Expert	Alonso-Betanzos (1995). It is an expert system capable of proposing treatment and diagnosis
CAFE (Computer Aided Fetal Evaluator)	Guijarro-Berdiñas, Alonso-Betanzos (2002). Modified version of NST-Expert
2CTG2	Magenes and Signorini (2007)
Omniview SisPorto 3.5	de Campos and his team at the University of Porto, Portugal {2008).

v. Prolonged second stage of labor accompanied by poor cervical dilation may lead to non-progress of labor and eventual fetal compromise.

Although intrauterine pressure estimation is most accurate method of detecting labor, it requires expertise, sophisticated equipment and aseptic technique, which are rarely available in under-developed countries. Thus, a non-invasive method like CTG with accurate interpretation algorithm is the need of the hour.

Uterine activity—physiological perspective

Tonnus pressure or the resting tone is the steady contraction of the uterine muscle. It is due to the elastic recoil of the tissues in and around the uterus that causes the pressure to rise to 7.5–15 mmHg. Contraction of the uterus results in intra-uterine pressure (IUP). Blood flow in the uterus is reduced when the IUP is greater than 30 mmHg. Blood flow to the uterus is lowest at the peak of the contraction and slowly returns to normal as the uterine activity ceases. Before, during and after normal contractions blood flow to the uterus and umbilical artery are uninterrupted. But when uterine contractions are long, frequent and with high amplitude, oxygen supply to the fetus is significantly reduced [11]. For the first 30 weeks, uterine activity is comparatively gentle; usually not more than 20 mmHg. Frequency and intensity of uterine activity begin to increase gradually after 30 weeks. These increments become even more prominent in the last weeks of pregnancy. Labor is said to have started when uterine activity (UA) is about three contractions of approximately 40 mmHg in a 10-min window. But it is not possible to differentiate between antepartum and intrapartum period from the UA [12]. In the first stage of labor uterine contraction (UC) gradually increases from 25 mmHg to about 50 mmHg. Frequency of contraction is 3–5 per 10 min and the basal tone is 8–12 mmHg. During the second stage there further increase in the UC and it typically reaches 80–100 mmHg. Frequency of contraction at this stage is 5–6 per 10 min window. The parameters that define uterine activity are:

i. Intensity—it is the degree of uterine systole. It increases as the labor progresses and reaches the maximum value during the second stage of labor.
ii. Duration—usually lasts for 10–15 s and gradually rises to 40–45 s.
iii. Frequency—in the early stage of labor, contraction comes in the interval of 10–15 min, which gradually increases, to maximum in the second stage of labor.

Challenges of uterine activity monitoring

Since the fetus is not directly accessible, diagnosis of its status is based on indirect parameters like presence of certain patterns in FHR and UCP, and the possible correlation. Disparity in diagnosis is quite common because:

i. All observations can be prone to error.
ii. There are widely varying inter and intra observer variations. It has been found that there are 8.86% disagreements among the observers on analysis of existing data set.
iii. Evaluation of fetal status is especially difficult because of the analytic nature of the signals that evolve over a long period of time.

The lowest intra-uterine pressure between contractions is called the resting tone. During labor the resting tone rises to 10–15 mmHg. The parameters needed to identify the stage of labor are number of contractions in a 10 min window and the intensity of contractions. Authors have found that calculation of intensity taking the resting tone as zero gives rise to erroneous identification of contraction which in turn leads to wrong classification of deceleration. Thus before proceeding with finding any other parameter it is necessary to find the stage of labor. Within the definitions provided by NICHD, uterine activity is quantified as a number of contractions present in 10 min window averaged over 30 min [13]. In July 2009, ACOG defined uterine contractions ≥ 6 as tachysystole [14]. These guidelines are suitable for visual interpretation but cannot be effectively used in an automated system. The proposed module was designed using an adaptive process of learning and experimentation. Users can learn from the decision domain by a collaborative process of extracting patterns from the existing data and constructing and refining a predictive model using these patterns.

Novelty of the proposed method

Uterine activity is measured over a fixed time period. The medical literature provides us the parameters of the contraction such as duration, skewness and amplitude. Based on these parameters the progress in labor is measured. Especially, a great emphasis was placed on the skewness, but it was later proved inaccurate [15]. Several methods of measuring the uterine activity that are still in use are shown in Table 2:

MAP was considered the best method for the quantitation of uterine activity, but it could not guarantee that this provided a better insight into the status of the fetus. None of these methods offers a clear idea about how to accurately identify the stage of labor. Hence these methods are not sufficient to be incorporated in

Table 2 Existing methods on uterine pressure monitoring

Method	Researcher	Measurement	Window	Comment
Montevideo unit (MU)	Caldeyro-Barcia, 1957	Mean amplitude × mean frequency	10 min	Did not consider the shape and duration of the contraction.
Alexandria unit (AU)	El-Sahwi, 1967	MU × mean duration	10 min	Did not consider the shape of the contraction. Improvement over MU
Uterine activity unit (UAU)	Han & Paul, 1973	Area under the UCP curve	1 min	Cannot differentiate between active pressure and basal tone.
Uterine activity integral (UAI)	Steer, 1984	Area under the UCP curve	15 min	Independent of the duration of integration period.
Mean active pressure (MAP)	Phillips & Calder, 1987	UAI/900 s	1 s	
Mean contraction activity pressure (MCAP)	Phillips & Calder, 1987	UAI/total duration		

the automated analysis of the uterine activity and classification of the stage of labor.

The authors introduced three parameters, which are evaluated in a thirty minutes window to classify the progress of labor:

i. Maximum value of the amplitude of UCP.
ii. Range of values of UCP within this 30 min window. UCP values are divided into three ranges: less than 35 mmHg, between 35 and 70 mmHg and above 70 mmHg.
iii. Frequency of occurrence of each range.

Materials and methods

Fuzzy logic in medical diagnosis

Medical diagnosis is a field that deals with complexity of the human body, symptoms of disease, overlapping symptoms of different diseases and imprecise knowledge about the symptoms and the diseases. Acquisition of domain knowledge, analysis of human physiology etc. produces large number of cause effect relations. These relations however are a poor approximation of the complex system and there are lots of uncertainty involved [16].

A technique is needed for accurate and fast diagnosis in spite of incomplete or imprecise information. Physicians should be able to classify situations constantly in the same category when comparable cases are encountered. Establishment of consistent classification is possible using intelligent processing by human experts and there are always chances of disagreement in an environment of uncertainty [17]. Fuzzy representations are thus useful for capturing this uncertainty and modeling the acquisition of knowledge by the experts [18].

Proposed work

Dataset description

The data set used in this work was obtained from the CTU-UHB [19], free database of Czech Technical University, Department of cybernetics. Each set of data comprises a CTG trace with corresponding metadata viz. numeric value of uterine pressure and FHR in a timeline. Each CTG trace has a 90 min of recording and signals are sample at 4 Hz. Data sets with missing signal values were excluded from the study. Further, each set of data were classified as 0, 1 and 2 to denote labour stages 1, 2 and 3 respectively. The classification was done by three expert ObGyn clinicians.

Sample size was tested for adequacy with confidence level 95% and confidence level $\alpha = 0.05$ [20]. A sample FHR and UCP signal is shown in Fig. 1 and an overview of the proposed work is shown schematically in Fig. 2

Fuzzification of parameters Parameters taken into account are highest peak reached by toco in a 30 min window, range of values of toco and frequency of occurrence of each of these range of values. Linguistic variables associated with the parameters are Peak-value, Range and Frequency. Thus the set of parameters is

$$\Pi = \{P_1, P_2, P_3\} = \{Peak\_value, Range, Frequency\}$$

Each of these linguistic variables x is associated with term set $A_{P_i}(x)$ i.e. the set of linguistic terms and a set of discourse X. Thus,

$$A_{P_i}(x) = \{A_1, A_2, A_3\} = \{Low, Medium, High\}$$

The set of toco readings of the patients are given by

$$\mathrm{T} = \{T_1, T_2, \ldots, T_n\}$$

The set of linguistic variables associated with the diagnosis are given by

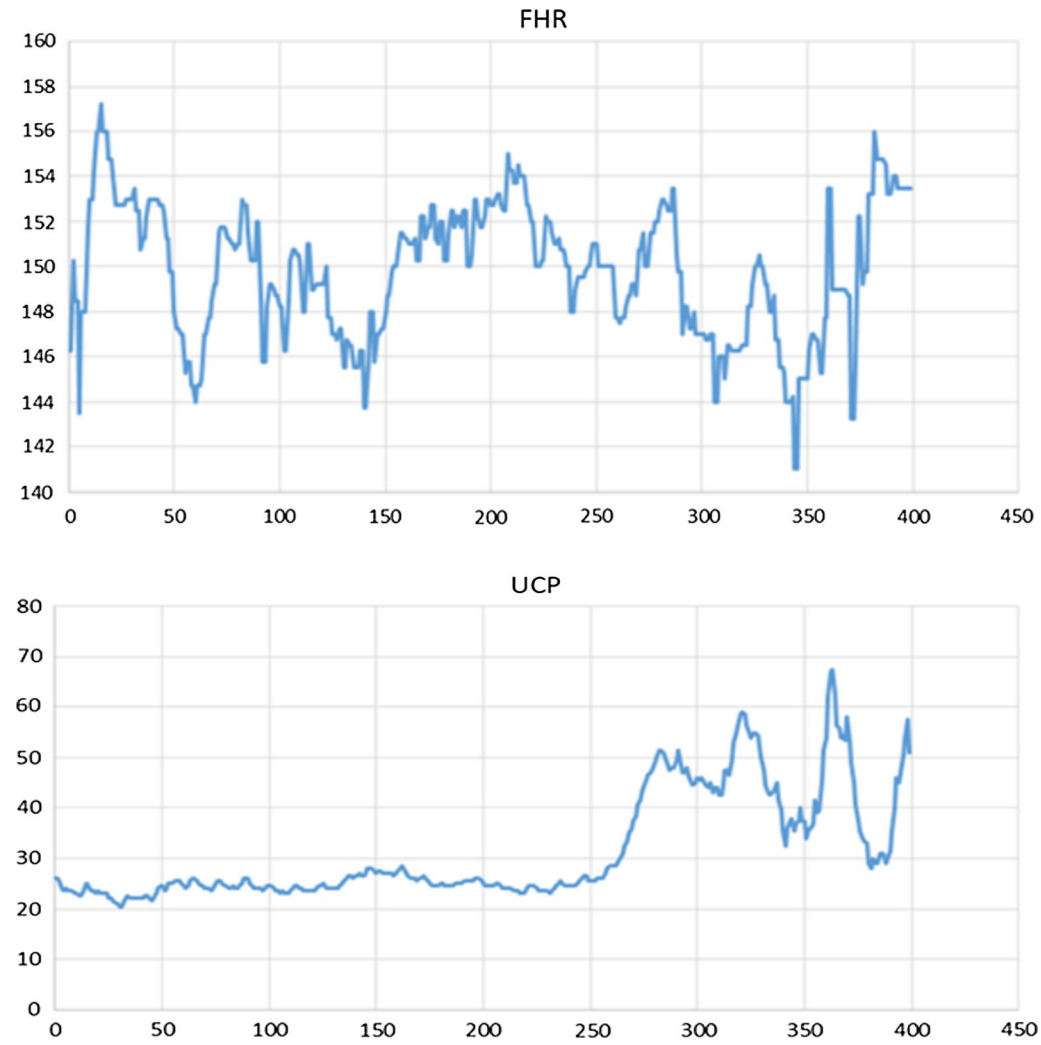

Fig. 1 Fetal heart rate and uterine contraction pressure signal

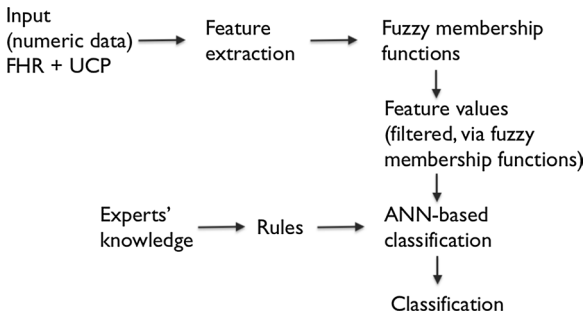

Fig. 2 Schematic diagram of the proposed model

$$\Delta = \{D_1, D_2, D_3\} = \{Normal, Stage1, Stage2\}$$

Main perception of fuzzy logic are belongingness of members to a set in certain degree. This makes fuzzy suitable for modeling observer's perception of vague information. Fuzzifier converts crisp inputs to fuzzy values. Parameter P_i takes the values $\mu_{P_i} \in [0, 1]$, where μ_{P_i} is the degree to which toco exhibits parameters P_i. The relationship between the toco of the patients and the parameters is given by the fuzzy relation.

$$R_{T\Pi} \subset T \times \Pi$$

where $\mu_N(T_i, \Pi_j)$ is the degree to which T_i exhibits parameter Π_j. The relationship between the parameters and diagnosis are

$$R_{\Pi\Delta} \subset \Pi \times \Delta.$$

The universe of discourse for linguistic variable peak value, range and frequency are [0,200], [0, 200] and [0, 20] respectively. Fuzzy membership functions for each of the linguistic variables is calculated as in Eqs. (1)–(6). For the calculation of degree to which the linguistic variables satisfy the linguistic label A_{P_i} authors have use Gaussian, S and Z membership functions [21–23]. Max–min composition of fuzzy relations for patient-diagnosis are given by

$$R_{T\Delta} = R_{T\Pi} \circ R_{\Pi\Delta}$$

Uncertain or imprecise knowledge of the clinicians were captured by interviewing them. They were asked how they interpreted some patterns or the combination of patterns in toco and what diagnosis they had reached. The knowledge of experts is converted into IF....THEN rules which are of the form:

$$\text{IF } (P_1 \text{ is } A_{P_i}) \wedge (P_2 \text{ is } A_{P_j}) \ \wedge \ (P_3 \text{ is } A_{P_k}) \text{ THEN } Dn$$

Classification

Defuzzification is done with the help of neural network. Five-fold cross validation was used to avoid possible bias. Training process is applied to all the folds except one, which is used for testing. The obtained output is the multilevel classification with 0, 1 and 2 indicating the three stages of labor. Fuzzified values of the linguistic variables are the first level input to the neural network. It is a 3-input fuzzy model with 18 rules. Out of the 18 rules some are listed below:

$$R_1 : \textit{If } P_1 \textit{ is } \mu_{P_{1l}} \textit{ and } P_2 \textit{ is } \mu_{P_{2l}} \textit{ and } P_3 \textit{ is } \mu_{P_{3l}} \textit{ Then } f_1 = a_1P_1 + b_1P_2 + c_1P_3 + d_1$$

$$R_2 : \textit{If } P_1 \textit{ is } \mu_{P_{1l}} \textit{ and } P_2 \textit{ is } \mu_{P_{2l}} \textit{ and } P_3 \textit{ is } \mu_{P_{3m}} \textit{ Then } f_2 = a_2P_1 + b_2P_2 + c_2P_3 + d_2$$

$$R_{18} : \textit{If } P_1 \textit{ is } \mu_{P_{1h}} \textit{ and } P_2 \textit{ is } \mu_{P_{2h}} \textit{ and } P_3 \textit{ is } \mu_{P_{3h}} \textit{ Then } f_{18} = a_{18}P_1 + b_{18}P_2 + c_{18}P_3 + d_{18}$$

The general architecture of ANN is given in Fig. 3. The network consists of five layers:

Layer 1 The node function of the ith node in this layer is given by $O_{1,i}$. Thus, the outputs are,

$$O_{1,i} = \mu_{A_i}(P_1),\ O_{1,i} = \mu_{A_i}(P_2),\ O_{1,i} = \mu_{A_i}(P_3).$$

The degree of membership values of fuzzy set that describe the antecedent of the rules R_k are realized in this layer. $O_{1,i}$ is thus the degree to which given input P_i satisfies the linguistic label $_i$. The authors determine the MF for each A_i and these MFs are kept constant throughout the learning process.

Layer 2 The parameters thus passed to layer 2 are,

$$\begin{bmatrix} P_1 & \mu_{P_{1l}} & \mu_{P_{1m}} & \mu_{P_{1h}} \\ P_2 & \mu_{P_{2l}} & \mu_{P_{2m}} & \mu_{P_{2h}} \\ P_3 & \mu_{P_{3l}} & \mu_{P_{3m}} & \mu_{P_{3h}} \end{bmatrix},$$

where $\{ \mu_{P_{1l}}, \mu_{P_{1m}}, \mu_{P_{1h}} \}$, $\{ \mu_{P_{2l}}, \mu_{P_{2m}}, \mu_{P_{2h}} \}$ and $\{ \mu_{P_{3l}}, \mu_{P_{3m}}, \mu_{P_{3h}} \}$ are the degree of membership of the parameters P_1, P_2 and P_3 respectively to the linguistic label A_{P_i}. Logical AND operation which is defined by the T-norm is performed in this layer. The output from each node is the product of all the incoming signals, i.e. the output from this layer gives information about the fulfillment of rules.

$$O_{2,i} = w_i = T(\mu_{P_{jl}}, \mu_{P_{jm}}, \mu_{P_{jh}}) = \mu_{P_{jl}} \tilde{*} \mu_{P_{jm}} \tilde{*} \mu_{P_{jh}}.$$

Layer 3 The ith node computes the ratio of ith rule's firing strength to the sum of the firing strength of all the rules.

$$O_{3,i} = \bar{w} = \frac{w_i}{w_1 + w_2 + \ldots + w_{18}}.$$

Layer 4 The node function of this layer is,

$$O_{4,i} = \bar{w}_i f_i = \bar{w}_i(a_iP_1 + b_iP_2 + c_iP_3 + d_i)$$

The values of a_i, b_i, c_i, and d_i are set to 1 based on trial.

Layer 5 This layer has a single node to compute the final output by aggregating the incoming signals

$$O_{5,1} = \sum_i \bar{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i}$$

Experiments

Three clinicians marked the stage of labor as stage-1, stage-2 and stage-3 or 0, 1, 2 respectively by observing the uterine contraction pattern from the CTG trace. The authors aim to compute the inter-observer agreement as well as the agreement between the algorithm and the clinicians. For the former we used percentage of agreement and for later we opted for concordance analysis using various statistical methods. A comparison between the clinicians' evaluation and the classification done by the algorithm are given in Table 3 and the inter-evaluator disagreement is shown in Table 4.

Overall agreement among the doctors was nearly 91.3% for all stages of labor. 91.2% of our result was in complete agreement with all the three doctors. This

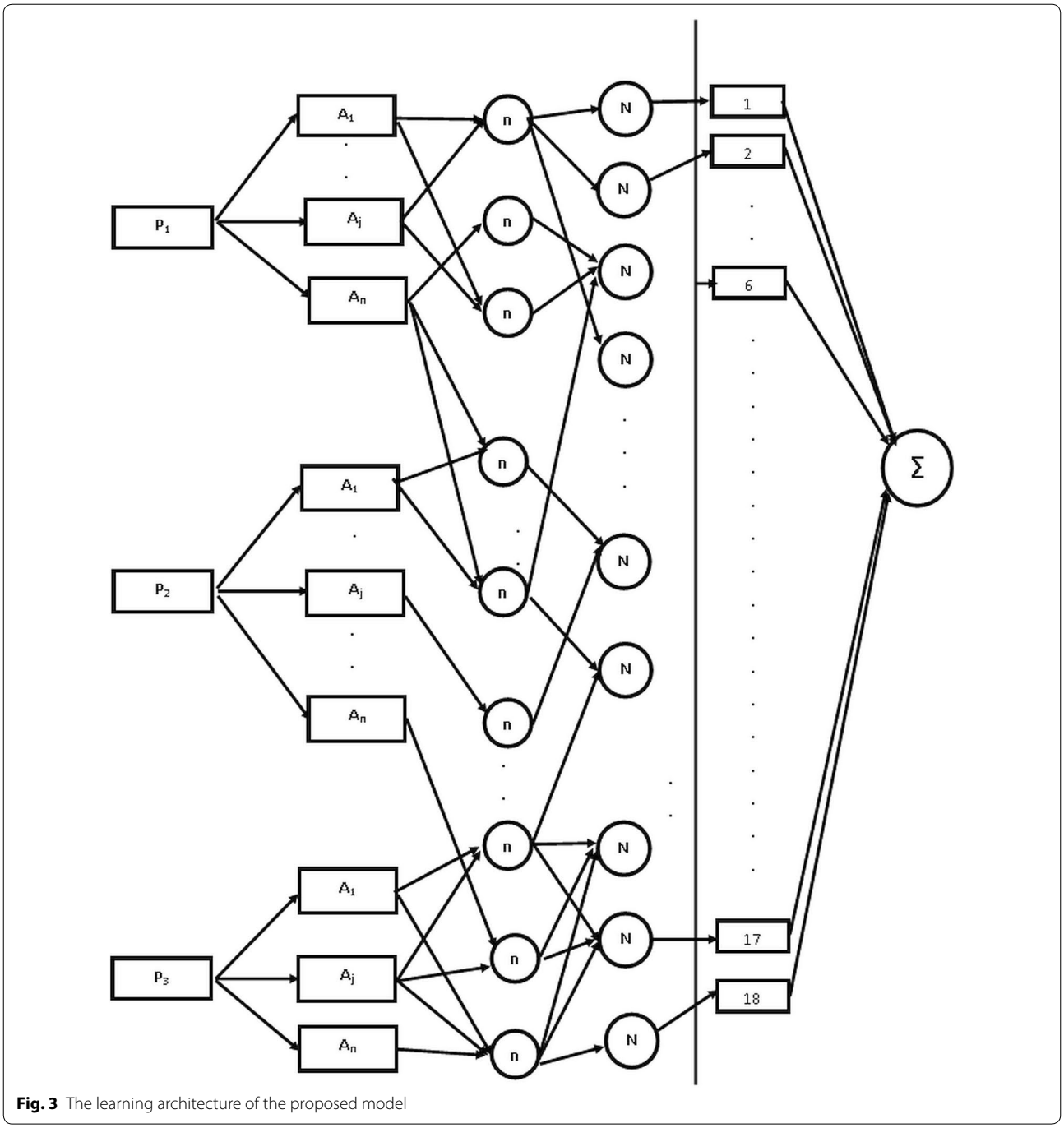

Fig. 3 The learning architecture of the proposed model

was because of the disagreement among doctors and our result in distinguishing stage 2 and stage 3 labor.

Evaluation protocol

Concordance analysis

In the field of medicine many times physicians are found to disagree on a diagnosis. Diagnostic medicine is thus prone to error. Hence it is necessary to measure any newly introduced diagnostic technique against a 'gold standard' [24]. As far as the measurement of the stage of labor from the uterine contraction pattern is concerned the only available gold standard is the visual analysis by the physicians. Thus it is necessary to evaluate the degree of agreement both visually and quantitatively. Authors

Table 3 Percentage disagreement among various evaluators and the algorithm

% Disagreement	Doc1 (%)	Doc2 (%)	Doc3 (%)	Algorithm (%)
Doc1	–	5	2.7	3.18
Doc2	5	–	6.36	5.91
Doc3	2.7	6.36	–	4.55
Algorithm	3.18	5.91	4.55	–
Average	3.67	5.67	4.5	4.55

Table 4 Percentage of inter-evaluator disagreement

Among the doctor (%)	With majority of doctors and algorithms (%)	Overall disagreement (%)
8.68	4.95	4.09

thus opted for concordance analysis using various statistical means.

Bland–Altman plot: assessment of agreement on ordinal scale

Mean discrepancy between the estimate provided by the physicians and that provided by the proposed method are assessed using Bland–Altman plot [25]. It assesses the difference between the methods not only quantitatively, but qualitatively as well. Bland–Altman plot is shown with 95% limit of agreement for one of the evaluators is shown in Fig. 4. Bias and its standard deviation and 95% confidence interval (CI) for lower and upper limits of agreement (LoA) are given in Table 5.

Fleiss Kappa: assessment of agreement on nominal scale

It is necessary to verify that the agreement between the obgyn experts and the algorithm is not occurring by chance. As the number of evaluators involved in the classification is more than two the authors used Fleiss' Kappa to measure the inter-evaluators agreement with $\kappa \in [0,1]$ [26]. Here, n (number of subjects), k (number of evaluation categories), m (number of evaluators) are 275, 3 and 4 respectively. For every patient $1 \le i \le n$ the evaluation categories $1 \le j \le k$, x_{ij} = the number of evaluators that assign category j to patient i. Thus,

$$0 \le x_{ij} \le m \qquad \sum_{j=1}^{k} x_{ij} = m \qquad \sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij} = mn$$

The proportion of pair of evaluators that agree in their evaluation of patients is

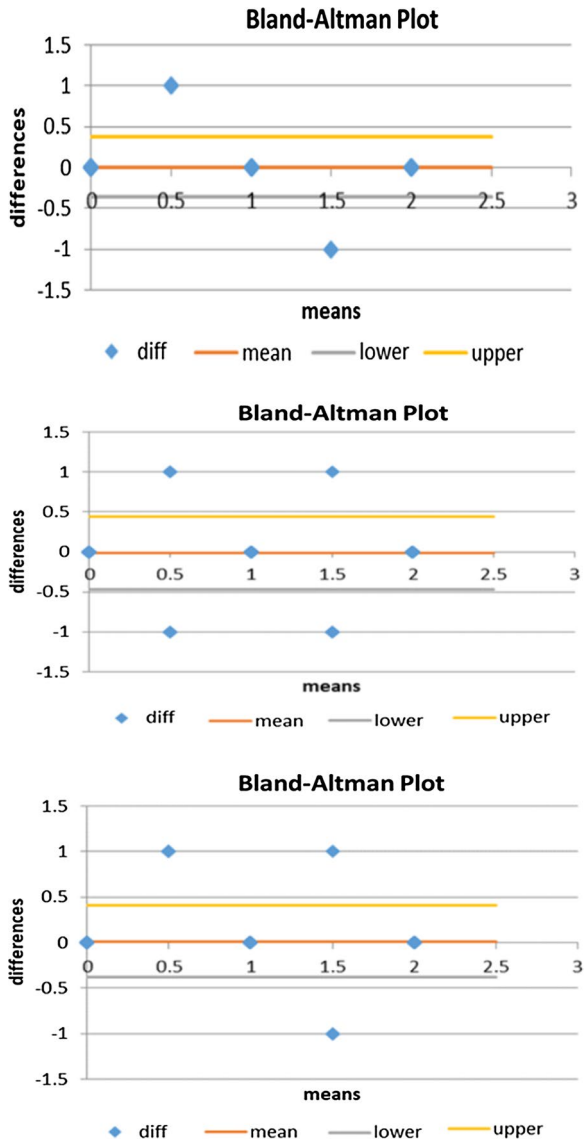

Fig. 4 Bland–Altman plot for Evaluator1 and the Algorithm

Table 5 Mean, standard deviation and LoA for Bland-Altman Plot

	Doc1–Algo	Doc2–Algo	Doc3–Algo
SD of Bias	0.1909	0.2334	0.2023
Bias	0.0091	− 0.0182	0.0136
Std. error	0.0129	0.0157	0.0136
95% CI lower	− 0.3651	− 0.4756	− 0.3828
95% CI upper	0.3833	0.4392	0.4101

$$p_i = \frac{\sum_{j=1}^{k} C(x_{ij}, 2)}{C(m, 2)} = \frac{\sum_{j=1}^{k} x_{ij}(x_{ij} - 1)}{m(m-1)} = \frac{\sum_{j=1}^{k} x_{ij}^2 - m}{m(m-1)}$$

The mean of p_i is given by

$$p_a = \frac{1}{n}\sum_{i=1}^{n} p_i = \frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{j=1}^{k} x_{ij}^2 - m}{m(m-1)}$$

Fleiss' Kappa is defined by

$$\kappa = \frac{p_a - p_\varepsilon}{1 - p_\varepsilon}.$$

Kappa for the jth category of diagnosis is given by

$$\kappa_j = 1 - \frac{\sum_{i=1}^{n} x_{ij}(m - x_{ij})}{mn(m-1)q_j(1-q_j)}.$$

The standard error for kappa for the jth category of diagnosis is

$$SE(\kappa_j) = \sqrt{\frac{2}{mn(m-1)}}.$$

Fleiss kappa was calculated using $\alpha = 0.05$. The results obtained are shown in the Table 6.

Kendell's coefficient of concordance: assessment of agreement among the physicians and the algorithm

It is used to measure the agreement among the clinicians and the algorithm [26]. There are m = 4 evaluators, which include three physicians and the algorithm, to evaluate k = 275 patients. Kendall's W is calculated as

$$W = \frac{12E}{m^2(k^3 - k^2)}.$$

Let c_{ij} be the classification an evaluator j gives to patient i. For each patient i, E_i is the sum of classifications given by m evaluators

$$E_i = \sum_{j=1}^{m} c_{ij}.$$

$$E = \sum_{i=1}^{k} (E_i - \bar{E})^2.$$

The result is W = 0.8448, r = 0.79305, df = 275, p value = 9.23E−58.

Table 6 Result of Fleiss' Kappa calculation

	Total	Stage1	Stage2	Stage3
Kappa(k)	0.918	0.919	0.889	0.948
Std. error (SE)	0.016	0.028	0.027	0.027
CI lower limit	0.879	0.865	0.835	0.894
CI upper limit	0.956	0.973	0.943	1.002

Comparison with some standard classifiers

We have used some standard machine learning algorithms such as Support Vector Machine (SVM), Naïve Bayes (NB) and Random Forest (RF) to classify the UCP and compare the result with the one given by the proposed method as well as the evaluation given by the doctors.

Comparison with proposed algorithm as the reference

Bland–Altman plot with proposed algorithm as reference is shown in Fig. 5 and Table 7 shows the corresponding comparison of mean, standard deviation and 95% confidence interval (CI).

Comparison with doctor's evaluation as the reference

Bland–Altman plot with doctor's evaluation as reference is shown in Fig. 6 and Table 8 shows the comparison of mean, standard deviation and 95% confidence interval (CI).

Inter-rater agreement of each pair of classifiers in terms of weighted kappa and standard error is given in Table 9.

Analysis using ROC curve

In order to visualize the performance of different classification method we have plotted AUC-ROC curve with doctor's evaluation as the classification variable. This is shown in Fig. 7 and the values are given in Table 10.

Result

On the ordinal scale Bland–Altman plot of physicians' evaluation against the algorithmic evaluation was used to check how big the average discrepancy is between the two estimates. In all the three estimates mean and bias line coincide. A priori measurement of accuracy or the difference between the upper and lower limit of agreement (LoA) for all the three assessments is less than one, indicating the closeness of agreement between the proposed method and the experts' evaluation. On the nominal scale kappa values lie between 0.889 and 0.948 for all the three stages of uterine contraction, indicating that the agreement between the proposed method and the visual estimation is not by chance. Standard error of kappa is

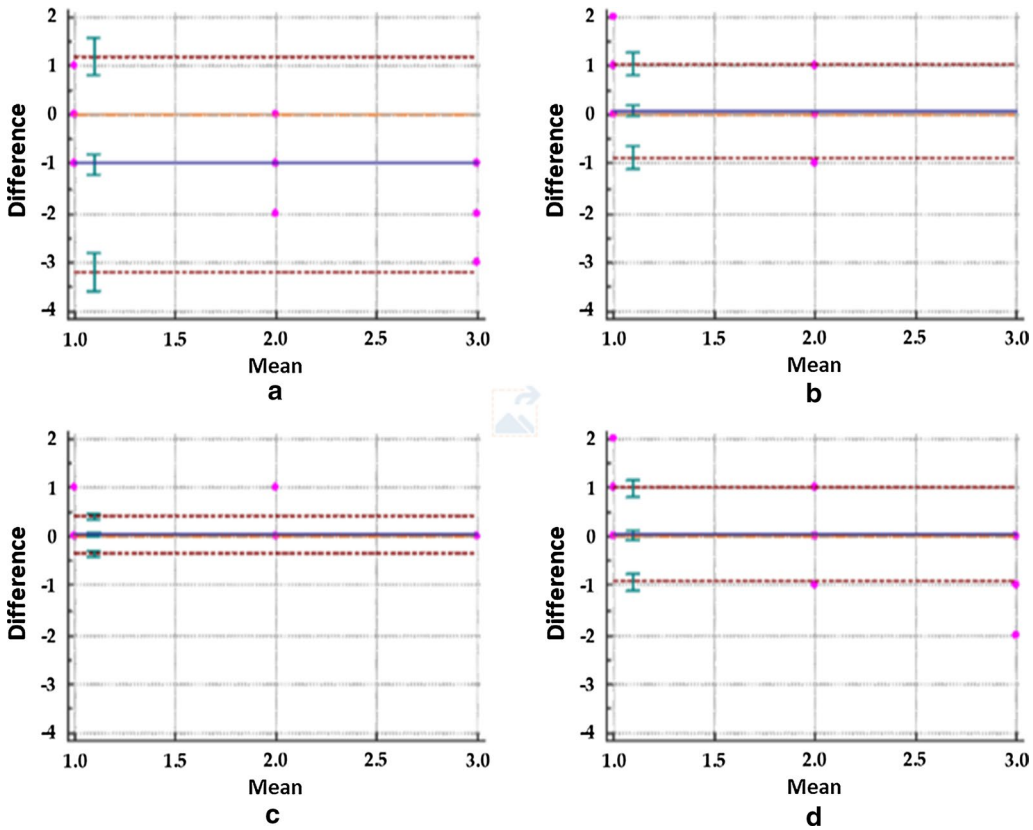

Fig. 5 Bland–Altman plot of **a** majority of doctors evaluation against proposed algorithm, **b** NB against algorithm, **c** RF against proposed algorithm, and **d** SVM against proposed algorithm

0.0275 and overall standard error is 0.019. This indicates that null hypothesis of no agreement is not true.

Kendell's coefficient of concordance $W = 0.84$ indicates good agreement among the evaluators and the algorithm. Also the p-value of $9.2E{-}58 < 0.05 = \alpha$ is consistent with the rejection of null hypothesis that there is no agreement between the evaluators.

When compared with SVM, Naïve Bayes and Random Forest the Bland–Altman has Limits of Agreement (LoA) of 0.094 to 0.0155, weighted kappa for expert-proposed method agreement of 0.909 and AUC-ROC value of 0.938 for the proposed method.

Table 7 Mean, standard deviation and 95% CI for Bland–Altman plot

	Mean	SD	95% of CI
Doc	0.0392	0.2787	−0.0155 to 0.094
Naïve Bayes	0.0392	1.1682	−0.1902 to 0.2687
SVM	0.0784	1.1405	−0.1456 to 0.3024
Random Forest	0.0784	1.1405	−0.1456 to 0.3024

Discussion

Classification of the stage of labor is an integral part of auto-diagnosis of CTG traces. The authors have utilized the CTU-UHB database of Czech Technical University, Department of Cybernetics. The use of fuzzification and final classification by ANN method is very novel in the field of medical diagnosis, which gives an edge over crisp solution in similar data-environment. While evaluating the performance, strength and reliability of the proposed algorithm, the authors have utilized appropriate relevant statistical metrics like concordance analysis, Bland–Altman plot, Fleiss-Kappa and Kendell's coefficient of concordance.

The concordance analysis shows that a significant reduction in the disagreement parameter is achieved by the proposed method (8.68% vs. 4.09%), which can be interpreted as the robustness and positive predictive ability of the algorithm. On ordinal scale using Bland–Altman plot, the algorithm's performance is very encouraging (SE of difference: 0.0129, 0.0157, 0.0136 at 95% CI) when three independent evaluations were compared with the proposed algorithm. This assessment of agreement among the doctors and the algorithm proves

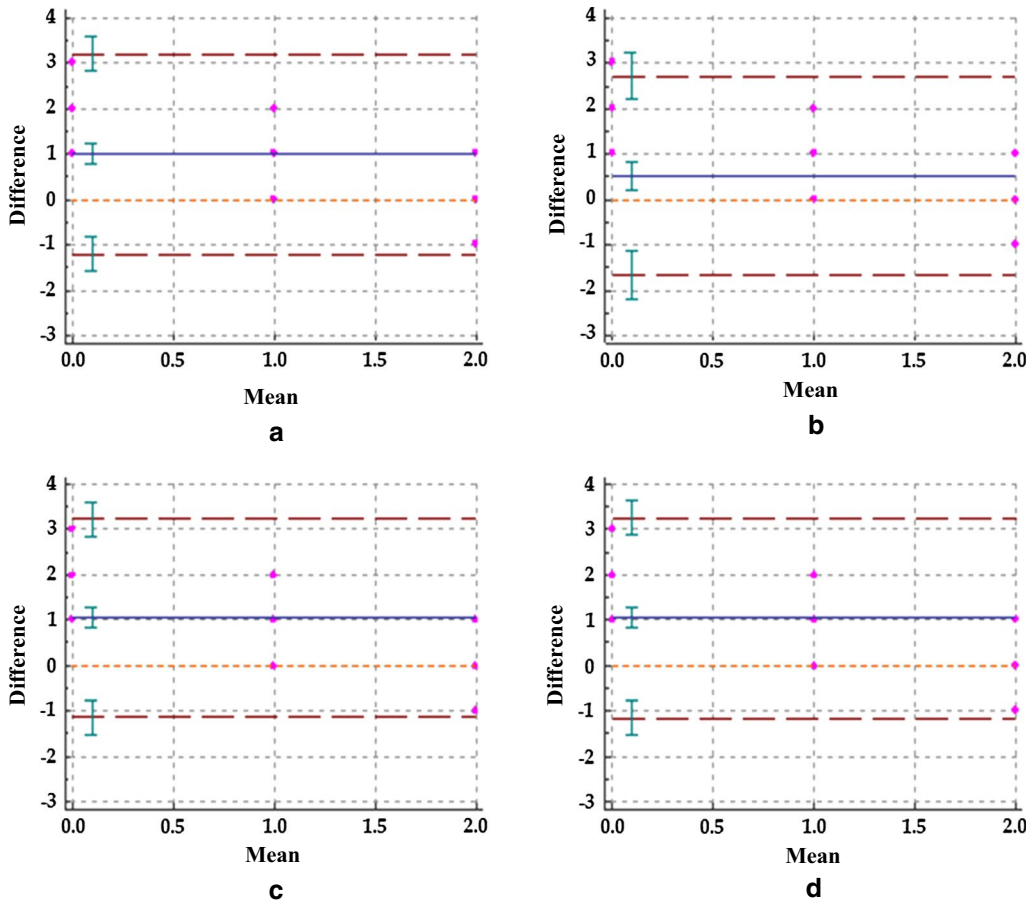

Fig. 6 Bland–Altman plot of majority of doctors evaluation against: **a** proposed algorithm, **b** Naïve Bayes, **c** Random Forest, **d** support vector machine

Table 8 Mean, standard deviation and 95% CI for Bland–Altman plot

	Mean	SD	95% of CI
Algorithm	−0.0392	0.2787	0.094–0.0155
Naïve Bayes	0	1.1432	−0.2245 to 0.2245
SVM	0.0392	1.1251	−0.1818 to 0.2602
Random Forest	0.0392	1.1162	−0.1800 to 0.2585

Table 9 Pairwise comparison of different classification methods using weighted kappa

	Weighted κ	Standard error
Algorithm vs. Doc	0.9094	0.031
Algorithm vs. NB	−0.0252	0.0778
Algorithm vs. RF	−0.0658	0.0744
Algorithm vs. SVM	−0.0353	0.0749
Doc vs. NB	−0.0044	0.0785
Doc vs. RF	−0.0542	0.075
Doc vs. SVM	−0.0214	0.076

the great potential in terms of strength and reliability. Another method of assessment of agreement for various stages of labor employed here is Fleiss Kappa, the result of which is similar to the earlier observation. It gives the algorithm an edge over traditional visual method (Kappa +SE: 0.918 + 0.016 at 95% CI).

Two comparisons are shown graphically for the pairwise comparison of classifiers using Bland–Altman plots in Figs. 5, 6, and 7. One graph uses the proposed algo as the reference and the other uses doctor's evaluation as the reference. Except for the Algo against Doc plot all the others have large interval between upper and lower limits of 95% CI. These cross the maximum allowed difference suggested by the clinicians involved in the study. Also the

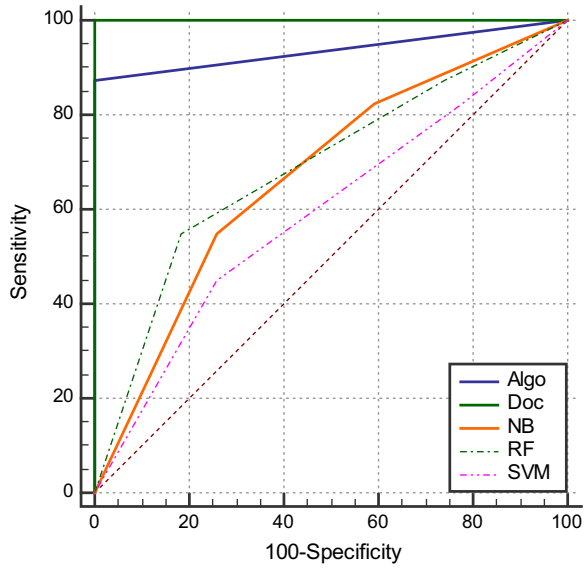

Fig. 7 AUC-ROC curve with doctor's evaluation as classification variable

Table 10 ROC curve results

	Area under the ROC curve	Standard Error	95% CI
Algo	0.938	0.0265	0.850–0.982
Doc	1.000	0	0.946–1.000
NB	0.672	0.0639	0.547–0.782
RF	0.690	0.0610	0.565–0.797
SVM	0.592	0.0657	0.465–0.710

measurement of inter-rater agreement using weighted kappa gives a better value (0.909) for the proposed algo vs. doctor's evaluation.

AUC value approximately equal to one signifies good measure of separability for the classifier. When the doctor's evaluation is taken as standard the proposed method exhibits the best efficiency in classifying the stages of labor while Naïve Bayes and Random Forest show poor performance in classification. Performance of SVM is poorest with AUC value near 0.5 indicating its incapability to distinguish between classes.

Since the proposed algorithm partly relies on the comparable values and performance by the clinicians, it has the inherent chance of fallacy. Absence of a fixed standard pattern of CTG trace is another important reason when the algorithm fails to learn numerous patterns of tracings leading to wrong interpretation. The authors cannot exclude the confounding factors in the datasets, which may arise from an erratic or a novel pattern of UCP trace. Although utmost care is taken to make the algorithm fool proof, data variation and errors, which might crop up in the process of fuzzification may lead to erratic results. Being a novel system using fuzzy algorithm on human data, further testing on a larger dataset is the need of the hour.

Conclusion

Authors in this paper proposed a fuzzy logic based method to estimate the stage of labor. Three obstetricians also evaluated them and the results were compared obtained by the proposed algorithm by various statistical means like Bland–Altman plot, Fleiss Kappa and Kendall's coefficient of concordance. Considering the inter-observer variation which in this case was about 8.68%, was higher compared to the calculated variation which was 4.95%, and also the Kendell's coefficient of concordance $W = 0.845$ showing the unanimity among the different physicians and proving relatively greater strength of the algorithm in detecting labor stage. Concordance analysis by Bland–Altman plot comparing individual evaluator and the proposed algorithm shows encouraging results, which reflects the reliability of this system. Total Fleiss Kappa when calculated was $0.918 + 0.0164$ at 95% confidence interval proves that the greater degree of agreement has not occurred simply by chance. In spite of the small cohort of evaluable CTG traces, this proposed algorithm shows promise in detecting the stage of labor, which is a fundamental part of CTG evaluation. Although this algorithm does not surpass the final overall accuracy to detect the stage of labor, it can be used interchangeably or independently in detecting the stage of labor. This method can be used as a decision support tool both in the presence and absence of clinicians.

Comparison of the proposed method with some standard machine learning algorithms like Naïve Bayes, SVM and Random Forest using various statistical means have shown that the proposed method has the best agreement with the experts' evaluation. The weighted kappa for the proposed method was 0.909 and area under the ROC curve was 0.908. Also the limits of agreement in the Bland–Altman's plot was the best (0.094 to 0.0155 at 95% CI) when the expert's classification was plotted against the proposed method.

We have tested the algorithm with noisy data and still obtained encouraging result. The authors plan to elaborate on this in subsequent studies. Later, we also plan to extract other parameters of uterine activity and check if they can be correlated with the health of the fetus. For this purpose we are going to build a database of our own by collecting live data from the patients. We also plan to involve clinicians with different levels of experience to evaluate the CTG.

Compliance with ethical standards

Conflict of interest
Authors declare that they have no conflicts of interest.

Research involving human participants or animals
This article does not contain any studies with human participants or animals performed by any of the authors.

Author details
[1] Department of Computer Science, West Bengal State University, Kolkata 700124, West Bengal, India. [2] Department of Computer Science & Engineering, Aliah University, Kolkata 700156, India. [3] Department of Computer Science, The University of South Dakota, Vermillion, SD, USA. [4] Department of Obstetrics & Gynecology, Biraj Mohini Matri-Sadan & Hospital, Kolkata, West Bengal 700122, India.

Received: 1 October 2019 Accepted: 13 March 2020
Published online: 30 March 2020

References

1. Teegala B, Kalyansundar A, Ramesh B. Uterine contraction measurement device and a fetal monitoring system. 2018. WO 2011/023521 Al.
2. Ricci SS, Kyle T. Maternity and pediatric nursing. New York: Lippincott William and Wilkins; 2009.
3. Das S, Roy K, Saha CK. A novel step towards machine diagnosis of fetal status in utero: Calculation of baseline variability. In Proceedings of 2015 IEEE International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN). Kolkata: IEEE Press; 2015, p. 230–234. https://doi.org/10.1109/icrcicn.2015.743424.
4. Dawes GS, Redman CWG. Numerical analysis of the human fetal heart rate: the quality of ultrasound records. Am J Obstet Gynecol. 1981;141(1):43–52.
5. Alonso Betanzos A, Moret Bonillo V, Devoe LD, Searle JR, Boveda Alvarez C. NST EXPERT: an Intelligent Program For NST Interpretation. Artif Intell Med. 1995;7(4):297–313.
6. Das S, Roy K, Saha CK. A linear time series analysis of fetal heart rate to detect the variability: measures using cardiotocography. In Bhattacharyya S, Das N, Bhattacharjee D, Mukherjee A, editors. Handbook of research on recent developments in intelligent communication application (pp. 471-495). Hershey, PA: IGI Global. 2017. https://doi.org/10.4018/978-1-5225-1785-6.ch018.
7. Guijarro-Berdiñas B, Alonso-Betanzos A, Prados-Méndez S, Fernández-Chaves O, Alvarez-Seoane M, Ucieda-Pardinas F. A hybrid intelligent system for the pre-processing of Fetal Heart rate signals in antenatal testing. In: Mira J, Moreno-Díaz R, Cabestany J, editors. Biological and artificial computation: from neuroscience to technology. IWANN 1997. Lecture Notes in Computer Science, vol 1240. Berlin, Heidelberg: Springer; 1997.
8. Magenes G, Signorini M, Ferrario M, Lunghi F. 2CTG2: A new system for the antepartum analysis of fetal heart rate. In: Jarm T, Kramar P, Zupanic A, editors. IFMBE Proceedings of 11th Mediterranean Conference on Medical and Biomedical Engineering and Computing 2007, vol 16. Berlin, Heidelberg: Springer; 2007.
9. Cömert Z, Şengür A, Budak Ü, Kocamaz AF. Prediction of intrapartum fetal hypoxia considering feature selection algorithms and machine learning models. Health Inf Sci Syst. 2019;7(1):17.
10. Winn HC, Hobbins JC. Clinical maternal-fetal medicine. New York: The Parthenon Publishing Group; 2000.
11. Cunnigham FG, Leveno KJ, Bloom SL, Spong CY, Dashe JS, Hoffman BL, Casey BM. Williams obstetrics. New York: McGraw Hill; 2001.
12. Macones GA, Hankins GD, Spong CY, Hauth J, Moore T. The 2008 National Institute of Child Health and Human Development workshop report on electronic fetal monitoring: update on definitions, interpretation, and research guidelines. J Obstet Gynecol Neonatal Nurs. 2008;37(5):510–5. https://doi.org/10.1111/j.1552-6909.2008.00284.x **PMID:18761565**.
13. ACOG Practice Bulletin No. 106: Intrapartum fetal heart rate monitoring: nomenclature, interpretation, and general management principles (2009), Obstet. Gynecol. pp. 192–202.
14. Bakkar PC, Kurver PH, Kuik DJ, Van Geijn HP. Elevated uterine activity increases the risk of fetal acidosis at birth. Am J Obstet Gynecol. 2007;196(4):331.
15. Maojo V, Sannanders J, Billhartdt H, Crespo J. Computational intelligence techniques in medical decision making: the data mining perspective. In: Schmitt M, Teodorescu HN, Jain A, Jain A, Jain S, editors. Studies in fuzziness and soft computing, vol. 96. Heidelberg: Physica; 2002. p. 13–44.
16. Konoenko I. Machine learning for medical diagnosis: history, state of the art and perspective. Artif Intell Med. 2001;23(1):89–109.
17. Das S, Roy K, Saha CK. Fuzzy membership estimation using ANN: a case study in CTG analysis. In: Satapathy S, Biswal B, Udgata S, Mandal J, editors. Advances in intelligent systems and computing, vol. 327. Cham: Springer; 2015. p. 221–8.
18. Czech Technical University (CTU) in Prague and University Hospital in Brno (UHB) database. 2010. http://physionet.nlm.nih.gov/pn3/ctu-uhb-ctgdb/. Accessed 8 Aug 2018
19. Sakpal PV. Sample size estimation in clinical trial. Perspect Clin Res. 2010;1(2):67–9.
20. Lawson AE, Daniel ES. Inferences of clinical diagnostic reasoning and diagnostic error. J Biomed Inform. 2011;44(3):402–12.
21. Das S, Guha D, Dutta B. Medical diagnosis with the aid of using fuzzy logic and intuitionistic fuzzy logic. Appl Intell. 2016;44(3):850–7.
22. Jang JSR, Sun CT, Mizutani E. Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence. Upper Saddle River: Prentice Hall; 1997.
23. Kwiecien R, Kopp-Schneider A, Blettner M. Concordance analysis: Part 16 of a series on evaluation of scientific publications. Dtsch Arztebl Int. 2011;108(30):515–21.
24. Altman DG, Bland JM. Design, analysis, and interpretation of method comparison-studies. AACN Adv Crit Care. 2008;19(2):223–34 **PMID:18560291**.
25. Zainotz C. Fleiss' Kappa. 2014. http://www.real-statistics.com/reliability/fleiss-kappa/ Accessed 8 Aug 2018.
26. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med. 2012;22(3):276–82.