Acta Med. Okayama, 2020
Vol. 74, No. 6, pp. 483-493

Acta Medica Okayama
http://escholarship.lib.okayama-u.ac.jp/amo/

Original Article

Potential of Artificial Intelligence for Estimating Japanese Fetal Weights

Yasunari Miyagi[a,b,c]*, and Takahito Miyake[d]

[a]*Medical Data Labo, Okayama 703-8267, Japan,* [b]*Department of Gynecology, Miyake Ofuku Clinic, Okayama 701-0204, Japan,* [c]*Department of Gynecologic Oncology, Saitama Medical University International Medical Center, Hidaka 350-1298, Japan,*
[d]*Department of Obstetrics and Gynecology, Miyake Clinic, Okayama 701-0204, Japan*

We developed an artificial intelligence (AI) method for estimating fetal weights of Japanese fetuses based on the gestational weeks and the bi-parietal diameter, abdominal circumference, and femur length. The AI comprised of neural network architecture was trained by deep learning with a dataset that consists of ±2 standard deviation (SD), ±1.5SD, and ±0SD categories of the approved standard values of ultrasonic measurements of the fetal weights of Japanese fetuses (Japan Society of Ultrasonics in Medicine [JSUM] data). We investigated the residuals and compared 2 other regression formulae for estimating the fetal weights of Japanese fetuses by *t*-test and Bland-Altman analyses, respectively. The residuals of the AI for the test dataset that was 12.5% of the JSUM data were 6.4±2.6, −3.8±8.6, and −0.32±6.3 (g) at −2SD, +2SD, and all categories, respectively. The residuals of another AI method created with all of the JSUM data, of which 20% were randomized validation data, were −1.5±9.4, −2.5±7.3, and −1.1±6.7 (g) for −2SD, +2SD, and all categories, respectively. The residuals of this AI were not different from zero, whereas those of the published formulae differed from zero. Though validation is required, the AI demonstrated potential for generating fetal weights accurately, especially for extreme fetal weights.

Key words: deep learning, artificial intelligence, fetal weight, neural network, ultrasound biometry

Over the past 30 years, the sonographic estimation of fetal weights has been investigated by applying several regression formulae. The precise estimation of fetal weight is important because birth weight is an important predictive parameter for neonatal morbidity and mortality [1]. Problems remain regarding significant discrepancies between the estimated fetal weights and the actual birth weights, however. This is because the multiple regression function for birth weight as a function of fetal biometry measured by ultrasound, which uses the method of finding the straight or non-straight line that most closely fits the data according to a specific mathematical criterion, usually does not cover distant values that are far from the mean or median. This feature of regression has presented a methodological limitation for identifying more accurate formulae for fetal weights.

In 2003, the Japan Society of Ultrasonics in Medicine (JSUM) published a dataset of the standard values of ultrasonic measurements that consisted of −2 standard deviations (SD), −1.5SD, ±0SD, +1.5SD, and +2SD categories of the fetal weight of Japanese fetuses [2-5]. We used this dataset (which has been approved as the standard in Japan) as 'the JSUM data' in the present study. The JSUM (2003) and the Japan Society of Obstetrics and Gynecology (2005) decided that the formula for estimating the fetal weights of Japanese fetuses

Received February 18, 2020; accepted July 16, 2020.
*Corresponding author. Phone: +81-86-281-2020; Fax: +81-86-281-7575
E-mail: ymiyagi@mac.com (Y. Miyagi)

Conflict of Interest Disclosures: No potential conflict of interest relevant to this article was reported.

should be a regression function based on the JSUM data of the values of the bi-parietal diameter (BPD), the abdominal circumference (AC), and the femur length (FL) [2], and this formula has been widely used in Japan since then. The formula is as follows: estimated fetal weight (EFBW) [g] = $1.07 \times \mathrm{BPD}^3$ [cm] + 3.00×10^{-1} AC^2 [cm] × FL [cm]. In the present study, we refer to this formula as the 'JSUM formula'.

In 2018, Miyagi *et al*. proposed a different formula for estimating the fetal weight of Japanese fetuses [6]: EFBW [g] = 8045.1/(1 + Exp (4.747 + 0.2584 × BPD [cm] + 0.1010 × AC [cm] − 1.416 × FL [cm])) [g]. In the present study, we refer to this formula as the 'Miyagi formula'.

Applications of artificial intelligence (AI) in the medical field (including obstetrics and gynecology) have been investigated. Generally, AI is classified as supervised, unsupervised, or reinforcement learning. Supervised deep learning with neural networks is often used as in applications of AI in the medical field. In obstetrics and gynecology, AI has been applied mostly for imaging purposes such as the prognostic prediction of blastocysts in sterility [7-11], estimating the placental volume by 3D ultrasound [12], diagnoses in colposcopy [13-15], and the prediction of local relapse and distant metastasis of cervical cancer [15]. Artificial intelligence has also been used for some non-imaging procedures such as survival analyses [17, 18] and massive hemorrhage during delivery [19]. Since AI consisting of neural networks involves a very large and complex structure of high-dimensional matrices in deep learning, we speculated that a more accurate estimation of fetal weight, not only for mean values but also for distant values from the mean, would be possible if AI is trained properly with a reliable dataset with supervised learning. We hypothesized that neural network architecture trained by supervised deep learning might therefore be feasible for estimating fetal weights.

We conducted the present retrospective study to investigate the potential of AI with the original architecture of a neural network for supervised deep learning, using the published ultrasonic biometric parameters, for generating more precise fetal weights for both the mean and distant values. We did this by evaluating the residuals themselves and by comparing them with those of the JSUM formula and the Miyagi formula. We used the JSUM data to create the AI by supervised deep learning in this study.

Materials and Methods

The datasets published by the JSUM were used [2-5]. The values in this dataset consisted of −2SD, −1.5SD, ±0SD, +1.5SD, and +2SD categories that were obtained with standard ultrasonic measurements of the BPD, AC, and FL of Japanese fetuses from 18 to 41 weeks of gestation.

We speculated that fetal weight could be a function of both the raw values of BPD, AC, and FL — as has often been used in the published regression formulae — and of gestational age in weeks, which is an integer variable. Therefore, we used not z-scores but rather the raw values of BPD, AC, FL, and gestational age in weeks for predicting the fetal weights. We hypothesized that when the standard values of BPD, AC, and FL are used, the estimated fetal weights would be closer to the standard values of fetal weight. The published standard dataset at −2SD, −1.5SD, ±0SD, +1.5SD, and +2SD for each gestational week, all of which follow a normal distribution, is available for use [2-5].

There is no gold-standard method for dividing and selecting datasets as the training, validation, and test datasets. We felt that the values of the gestational weeks of relatively smaller datasets such as the validation dataset and the test dataset should not be glomerated. We also thought that the training and validation datasets should be as large as possible for creating better AI. Therefore, because 24 classes of gestational weeks (from 18 to 41 weeks) were used for the preparation for deep learning, we defined the values of BPD, AC, and FL at 18, 19, 20, 22, 23, 25, 27, 28, 29, 31, 33, 34, 35, 36, 37, 38, and 41 weeks of gestation as the training dataset; we defined their values at 21, 26, 30, and 39 weeks of gestation as the validation dataset; and we defined their values at 24, 32, and 40 weeks of gestation as the test dataset. Thus, the ratio of the numbers of the training dataset to the validation dataset to the test dataset was 17 : 4 : 3 (= 0.708 : 0.167 : 0.125), and the datasets did not overlap.

We then weighted the training dataset and the validation dataset of ±2SD, ±1.5SD and ±0SD as 54 : 130 : 399, respectively, which are the integer ratios of the values of the standard normal distribution probability density function at 2, 1.5, and 0. The AI was trained with the training dataset and simultaneously validated with the validation dataset. The AI was then evaluated with the test dataset. No standardizations were applied

to the datasets prior to the training of the AI with deep learning.

The architecture of the present neural network for deep learning was originally created with linear layers [20, 21] catenated with gestational week, BPD, AC, and FL as the scalar, batch normalization layers [22], rectified linear unit layers, [23, 24] and scaled exponential liner unit layers [25] as shown in Fig. 1. The number of maximum training rounds, known as epochs, was 1,500, and the batch size was 64. The optimal trained network was obtained as the best AI.

We compared the predicted values of fetal weight for the test dataset obtained by the best AI with the JSUM data. The predicted values for the test dataset obtained with the JSUM formula and those obtained with the Miyagi formula were also compared with the JSUM data. We then compared the AI, the JSUM formula, and the Miyagi formula with each other by performing a Bland-Altman analysis. Then, after the AI created by using the training dataset seemed feasible for estimating fetal weights, we created a new AI for predicting estimated fetal weights (with the same architecture, training rounds, and batch size) by using all of the JSUM datasets, of which 20% were randomized validation data. We then compared the residuals of this new AI, the JSUM formula, and the Miyagi formula with the JSUM data.

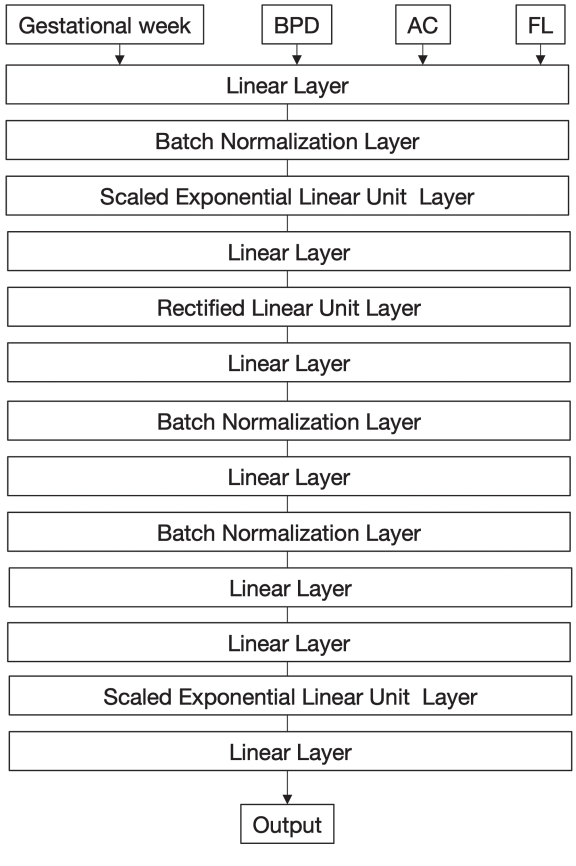

Fig. 1 The architecture of the neural network for estimating fetal weight.

The results obtained were compared with the results of the Miyagi formula and those of the commonly used JSUM formula. The residuals of the AI, the JSUM formula, and the Miyagi formula against the JSUM data were compared by unpaired *t*-test. The relationships among the AI, the JSUM formula, and the Miyagi formula were investigated by a Bland-Altman analysis [26, 27] with a calculation of the exact parametric confidence intervals for limits of agreement [28]. Probability (*p*)-values <0.05 were accepted as significant.

A Mac PC running OS X 10.14.5 (Apple) and Wolfram language 12.0.0.0 (Wolfram Research, Champaign, IL, USA) were used as the development environment.

Results

Figure 2 illustrates the AI-generated fetal weights obtained using the training and validation datasets at the −2SD, −1.5SD, ±0SD, +1.5SD, and +2SD categories for the test dataset, as well as the training and validation datasets themselves. As shown in Fig. 3 and Table 1, the mean ± SD (range) of the residuals of the AI-predicted fetal weight and the JSUM data (rAI) for the −2SD, −1.5SD, ±0SD, +1.5SD, and +2SD categories and all categories of the test dataset were 6.4 ± 2.6 (3.5 to 8.5), −0.8 ± 2.3 (−2.3 to 1.9), −1.4 ± 6.8 (−8.9 to 4.3), −2.0 ± 7.3 (−8.8 to 5.8), −3.8 ± 8.6 (−10.1 to 6.0), and −0.32 ± 6.3 (−10.1 to 8.5) g, respectively. All of the estimated fetal weights generated by the AI created with the training dataset of the JSUM data were not different from the JSUM data. Regarding the test dataset, we also investigated the residuals of the JSUM formula and the JSUM data (rJSUM) and those of the Miyagi formula and the JSUM data (rMiyagi) (Fig. 3, Table 1). The rAI and rJSUM were not different from zero in any of the categories. The rMiyagi values were not different from zero, with the exception of the −2SD and −1.5SD categories. The absolute values of the rJSUM were larger (but not significantly) than those of the rMiyagi in all categories except +1.5SD. The rAI values were smaller

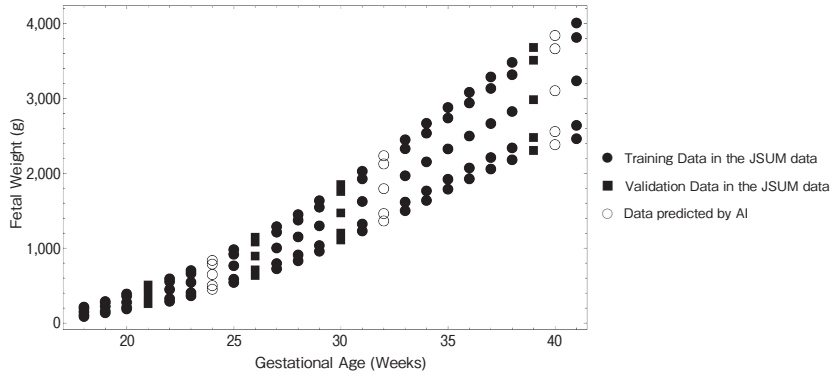

Fig. 2 AI-generated fetal weights. This AI was trained with the published JSUM data [2-5] of −2SD, −1.5SD, ±0SD, +1.5SD, and +2SD at 18, 19, 20, 22, 23, 25, 27, 28, 29, 31, 33, 34, 35, 36, 37, 38, and 41 weeks of gestation as the training dataset, and 21, 26, 30, and 39 weeks of gestation as the validation dataset. After the training, the AI generated fetal weights for the gestational weeks equivalent to the training dataset (●), the validation dataset (■), and the data predicted by the AI for the test dataset (○), which was for 24, 32, and 40 weeks of gestation. AI, artificial intelligence; JSUM, Japan Society of Ultrasonics in Medicine.

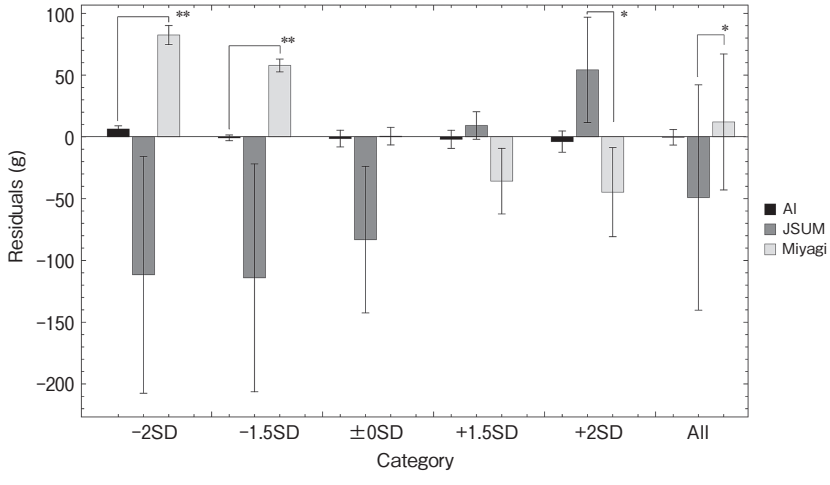

Fig. 3 The mean ± SD of the residuals of estimated fetal weight obtained by the AI and the JSUM data, the JSUM formula and the JSUM data, and the Miyagi formula and the JSUM data at −2SD, −1.5SD, ±0SD, +1.5SD, and +2SD. *$p<0.05$, **$p<0.001$ by *t*-test.

than the rMiyagi values in the −2SD and −1.5SD categories ($p<0.001$). The rMiyagi values were smaller than the rJSUM values in the +2SD category, and the absolute values of the rMiyagi were smaller than those of the rJSUM in all of the categories ($p<0.05$).

Figure 4 and Table 2 provide the results of our comparison of the AI, the JSUM formula, and the Miyagi formula for the test dataset against the JSUM data by Bland-Altman analysis. The differences (95% limits of agreement) between the AI and the JSUM formula, the AI and the Miyagi formula, and the Miyagi formula and the JSUM formula were 48.8 (−133.3 to 230.8), −12.4 (−112.5 to 87.8), and 61.1 (−203.6 to 325.9), respectively. The absolute value of the difference between the AI and the Miyagi formula was the smallest, but not significantly. There were no absolute systematic differences in these three comparisons because all of the *p*-values by *t*-test were not significant. No proportional errors were observed in the three comparisons. The Bland-Altman plots of the JSUM formula and either the AI or the Miyagi formula showed divergence as the mean increased (Fig. 4). The Bland-Altman analysis results thus indicated that the variation of the JSUM formula depended strongly on the magnitude of the

Table 1 Comparison of residuals of the test dataset of the JSUM data and the JSUM formula (rJSUM), the Miyagi formula (rMiyagi), and the AI (rAI) that was created with the training dataset

Statisitc	−2SD	−1.5SD	±0SD	+1.5SD	+2SD	All
Residuals of AI and the JSUM data (rAI)	6.4 ± 2.6 g (N.S.)	−0.8 ± 2.3 g (N.S.)	−1.4 ± 6.8 g (N.S.)	−2.0 ± 7.3 g (N.S.)	−3.8 ± 8.6 g (N.S.)	−0.3 ± 6.3 g (N.S.)
Residuals of the JSUM formula and the JSUM data (rJSUM)	−111.6 ± 95.8 g (N.S.)	−114.1 ± 92.2g (N.S.)	−83.2 ± 59.3 g (N.S.)	9.2 ± 11.1 g (N.S.)	54.2 ± 42.6 g (N.S.)	−49.1 ± 91.2g (N.S.)
Residuals of the Miyagi formula and the JSUM data (rMiyagi)	82.4 ± 7.7 g**	57.8 ± 5.19 g**	0.6 ± 7.1 g (N.S.)	−35.8 ± 26.6 g (N.S.)	−44.8 ± 36.0 g (N.S.)	12.1 ± 55.0 g (N.S.)
rAI vs rJSUM	N.S. ($p=0.167$)	N.S. ($p=0.100$)	N.S. ($p=0.138$)	N.S. ($p=0.221$)	N.S. ($p=0.082$)	N.S. ($p=0.058$)
rAI vs rMiyagi	AI<Miyagi*	AI<Miyagi**	N.S. ($p=0.747$)	N.S. ($p=0.100$)	N.S. ($p=0.128$)	N.S. ($p=0.401$)
rMiyagi vs rJSUM	N.S. ($p=0.07$)	N.S. ($p=0.08$)	N.S. ($p=0.132$)	N.S. ($p=0.054$)	Miyagi<JSUM*	Miyagi<JSUM*

The residuals of each method are shown as the mean ± standard deviation (SD). The rAI and rJSUM values are not different from zero in all of the categories. The rMiyagi values are not different from zero except for the −2SD and −1.5SD categories. The absolute values of the residuals of the rJSUM seem to be larger, but with no significance due to the large number of standard deviations caused by the small size of the test dataset, which was 3 points for each category. The rAI values are significantly smaller than the rMiyagi values in the −2SD and −1.5SD categories ($p<0.001$). The rMiyagi values are significantly smaller than the rJSUM values in +2SD and all of the categories ($p<0.05$). N.S.; not significant, *$p<0.05$, **$p<0.001$ by *t*-test.

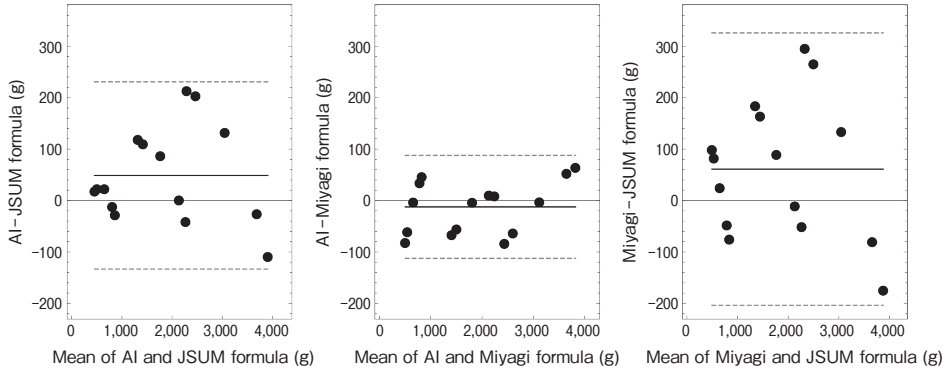

Fig. 4 Bland-Altman plots for comparing pairs of methods. AI and the JSUM formula (*left panel*), AI and the Miyagi formula (*middle panel*), and the Miyagi formula and the JSUM formula (*right panel*) are shown. The mean and 95% limits of agreement of the differences between each pair of methods are shown as a *solid line* and *dashed line*, respectively. The absolute value of the difference between the AI and the Miyagi formula was the smallest, although not significantly. In all three comparisons, no absolute systematic differences and no proportional errors were observed. The JSUM formula and either the AI or the Miyagi formula showed divergence as the mean increased (*left and right panels*). These results suggested that the variation of the JSUM formula depended strongly on the magnitude of the measurements.

measurements.

Table 3 shows the estimated fetal weights that were obtained with the use of a different AI that we created with all of the JSUM data. As shown in Fig. 5 and Table 4, the mean ± SD (range) of the residuals of these AI-predicted fetal weights and the JSUM data (rAI') for −2SD, −1.5SD, ±0SD, +1.5SD, +2SD, and for all categories of the JSUM dataset were −1.5 ± 9.4 (−16.5 to 12.1), 0.2 ± 5.1 (−8.3 to 8.8), 0.6 ± 4.1 (−7.0 to 7.9), −2.1 ± 6.3 (−14.4 to 7.4), −2.5 ± 7.3 (−15.4 to 9.6), and −1.1 ± 6.7 (−16.5 to 12.0) g, respectively. The rAI' values are not different from zero in all of categories. In other words, all of the estimated fetal weights generated by the AI created with all of the JSUM data were not different from the JSUM data. On the other hand, the residuals of the JSUM formula and the JSUM data (rJSUM') and those of the Miyagi formula and the JSUM data (rMiyagi') for all of the JSUM data were different from the JSUM data with the exception of the Miyagi formula at ±0SD. The rAI' values were smaller than the rJSUM' values in all categories and the rMiyagi' in all categories except for the ±0SD category. The rMiyagi' values were smaller than the rJSUM' values in all categories except +1.5SD.

Figure 6 and Table 5 provide the results of the comparison of the AI created with all of the JSUM dataset,

Table 2 Comparison of the AI that was created with the training dataset of the JSUM data, the JSUM formula, and the Miyagi formula for the test dataset by Bland-Altman analysis [26–28]

Statistic	AI vs JSUM formula	AI vs Miyagi formula	Miyagi formula vs JSUM formula
Sample size	15	15	15
Difference (g)	48.8	−12.4	61.1
Lower 95% LoA (g)	−133.3	−112.5	−203.6
Upper 95% LoA (g)	230.8	87.8	325.9
Upper 95% exact CI for lower 95% LoA (g)	−70.3	−77.8	−111.9
Lower 95% exact CI for lower 95% LoA (g)	−254.8	−179.3	−380.3
Upper 95% exact CI for upper 95% LoA (g)	352.3	154.6	502.6
Lower 95% exact CI for upper 95% LoA (g)	167.8	53.1	234.2
The coefficient of Repeatability	200.1	99.7	282.4
p-value by *t*-test for p (H0: Mean = 0)	N.S. ($p = 0.061$)	N.S. ($p = 0.364$)	N.S. ($p = 0.101$)

The absolute value of the difference between the AI and the Miyagi formula is the smallest, although not significantly. LoA, limits of agreement; N.S., not significant.

Table 3 Estimated fetal weights predicted by AI with the use of all of the JSUM data

Gestational Week	−2SD (g)	−1.5SD (g)	± 0SD (g)	+1.5SD (g)	+2SD (g)	5 percentile (g)	95 percentile (g)
18	137.6	149.8	187.9	237.0	256.6	138.7	258.0
19	173.9	192.2	251.0	314.6	337.3	173.5	338.0
20	216.0	240.0	312.9	395.5	422.1	218.0	424.1
21	267.7	297.0	394.1	482.3	519.5	270.7	513
22	328.1	358.5	469.9	587.4	618.7	332.2	619.2
23	392.3	436.9	562.5	695.6	738.1	393.7	741.9
24	473.1	519.8	661.8	803.0	860.9	474.1	855.9
25	552.5	597.5	778.4	937.6	988.7	549.9	983.6
26	643.9	702.5	892.5	1,074.8	1,137.6	645.1	1,132.5
27	744.4	815.8	1,030.9	1,218.6	1,290.2	749.1	1,283.0
28	857.1	934.7	1,165.5	1,382.6	1,462.6	862.5	1,452.2
29	974.3	1,060.2	1,306.8	1,575.0	1,645.7	980.3	1,653.3
30	1,097.9	1,189.9	1,468.0	1,743.5	1,833.7	1,105.4	1,826.3
31	1,218.0	1,325.7	1,632.1	1,929.2	2,023.6	1,225.2	2,017.5
32	1,359.8	1,472.5	1,801.2	2,127.1	2,236.3	1,368.7	2,230.9
33	1,502.0	1,624.6	1,976.4	2,327.4	2,444.2	1,513.3	2,432.6
34	1,647.1	1,776.6	2,151.4	2,529.1	2,659.7	1,656.1	2,649.1
35	1,790.2	1,920.3	2,331.0	2,734.8	2,871.0	1,797.5	2,860.1
36	1,911.1	2,067.9	2,507.9	2,935.9	3,084.7	1,920.6	3,067.0
37	2,042.5	2,204.7	2,669.0	3,134.6	3,286.8	2,058.0	3,271.3
38	2,165.7	2,341.0	2,837.5	3,330.1	3,490.5	2,177.9	3,475.1
39	2,280.5	2,461.8	2,992.0	3,507.5	3,678.8	2,300.2	3,658.7
40	2,372.5	2,566.3	3,125.9	3,679.1	3,860.8	2,390.7	3,853.6
41	2,457.9	2,662.5	3,249.5	3,831.7	4,031.7	2,472.1	4,012.5

the JSUM formula, and the Miyagi formula for all of the JSUM dataset by the Bland-Altman analysis. The differences (95% limits of agreement) between the AI and the JSUM formula, the AI and the Miyagi formula, and the Miyagi formula and the JSUM formula were 40.5 (−111.1 to 192.1), −20.1 (−114.0 to 73.7), and 60.6 (−157.6 to 278.8), respectively. The value of the differences between the AI and the JSUM formula, that between the AI and the Miyagi formula, and that between the Miyagi formula and the JSUM formula

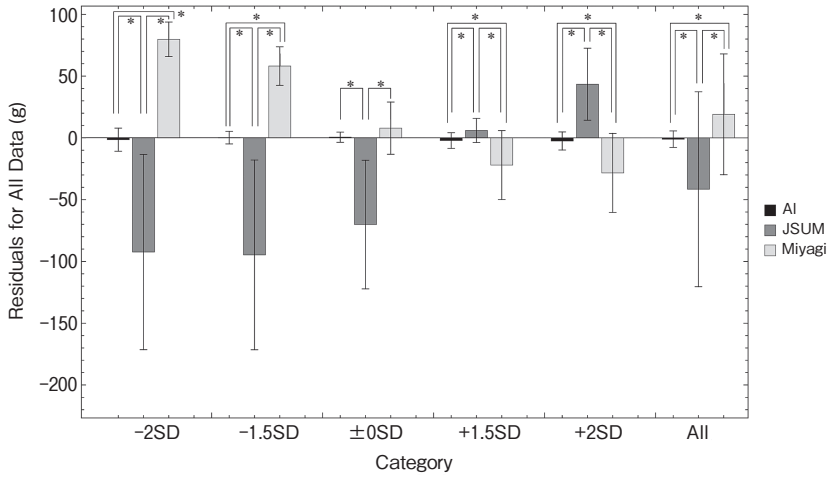

Fig. 5 The mean ± SD of the residuals of estimated fetal weight obtained by the AI and the JSUM data, the JSUM formula and the JSUM data, and the Miyagi formula and the JSUM data at −2SD, −1.5SD, ±0SD, +1.5SD, and +2SD. This AI was trained with all of the JSUM data [2-5] as the training dataset, and 20% of these data were randomly selected as the validation dataset. The estimated fetal weights generated by the AI created with all of the JSUM data were not different from the JSUM data. The residuals of this AI were smaller than those of the JSUM formula in all categories and that of the Miyagi formula in all categories except ±0SD. $^{*}p<0.001$ by *t*-test.

Table 4 Comparison of the residuals of all of the JSUM data and the JSUM formula (rJSUM'), the Miyagi formula (rMiyagi'), and the AI that was created with all of the JSUM data (rAI')

Statistic	−2SD	−1.5SD	±0SD	+1.5SD	+2SD	All
Residuals of AI and the JSUM data (rAI')	−1.5 ± 9.4g (N.S.)	0.23 ± 5.1g (N.S.)	0.6 ± 4.1g (N.S.)	−2.1 ± 6.3g (N.S.)	−2.5 ± 7.3g (N.S.)	−1.1 ± 6.7g (N.S.)
Residuals of the JSUM formula and the JSUM data (rJSUM')	−92.4 ± 79.0g**	−94.7 ± 76.8g**	−70.2 ± 52.1g**	6.0 ± 9.7g*	43.4 ± 29.1g*	−41.6 ± 79.0**
Residuals of the Miyagi formula and the JSUM data (rMiyagi')	79.8 ± 14.0g**	58.1 ± 15.6g**	7.9 ± 21.1g (N.S.)	−22.0 ± 27.9g**	−28.4 ± 32.0g**	19.1 ± 48.9g**
rAI' vs rJSUM'	AI < JSUM**	AI < JSUM**	AI < JSUM**	AI < JSUM**	AI < JSUM**	AI < JSUM**
rAI' vs rMiyagi'	AI < Miyagi**	AI < Miyagi**	N.S. ($P=0.109$)	AI < Miyagi**	AI < Miyagi**	AI < Miyagi**
rMiyagi' vs rJSUM'	Miyagi < JSUM**	Miyagi < JSUM**	Miyagi < JSUM**	JSUM < Miyagi**	Miyagi < JSUM**	Miyagi < JSUM**

The residuals of each method are shown as mean ± SD. The rAI' values are not different from zero in all of the categories. The rAI' values are smaller than the rJSUM' values in all categories. The rAI' values are smaller than the rMiyagi' values in all categories, except ±0SD. $^{*}p<0.01$, $^{**}p<0.001$ by *t*-test.

were 40.5 ± 77.3, −20.1 ± 111.3, and 60.6 ± 47.9 (g), respectively. The absolute value of the difference between the AI and the Miyagi formula was smaller than that between the AI and the JSUM formula ($p=6.68\times10^{-12}$) and that between the Miyagi formula and the JSUM formula ($p=1.23\times10^{-15}$). There were absolute systematic differences in these three comparisons because all of the *p*-values revealed by *t*-test were significant. In other words, the AI created with all of the JSUM data and the AI created with each of the two formulas were different from each other.

In addition, no proportional errors were observed in the three comparisons. The Bland-Altman plots of the JSUM formula and either the AI or the Miyagi formula for all of the JSUM data showed divergence as the mean increased (Fig. 6). The Bland-Altman analysis results thus suggest that the variation of the JSUM formula depended strongly on the magnitude of the measurements.

Discussion

We developed an AI method that can generate the fetal weight from the gestational age in weeks, the BPD, the AC, and the FL (Fig. 2). The residual obtained by the AI for the test dataset was only −0.32 ± 6.3 g (Table

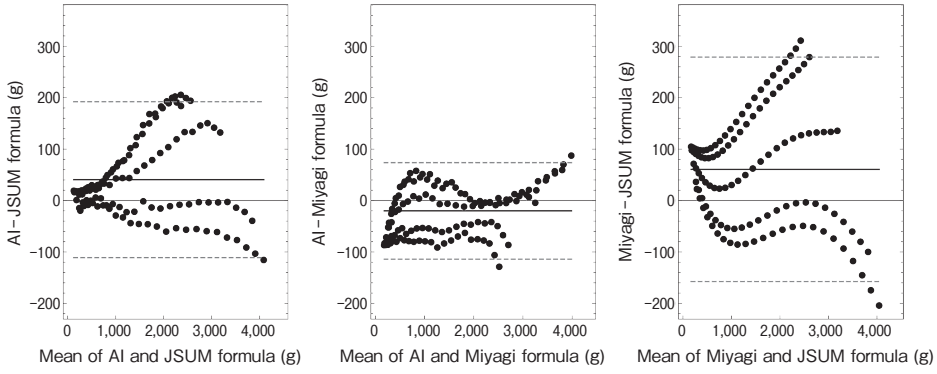

Fig. 6 Bland-Altman plots for comparisons of pairs of methods. The AI and the JSUM formula (*left panel*), the AI and the Miyagi formula (*middle panel*), and the Miyagi formula and the JSUM formula (*right panel*) are shown. This AI was trained with all of the JSUM data [2-5] as the training dataset, and 20% of these data were randomly selected as the validation dataset. The mean and 95% limits of agreement of differences of each pair of methods are shown as a *solid line* and *dashed line*, respectively. The values of the differences between the AI and the JSUM formula, that between the AI and the Miyagi formula, and that between the Miyagi formula and the JSUM formula were 40.5 ± 77.3, -20.1 ± 111.3, and 60.6 ± 47.9 (g), respectively. The absolute value of the difference between the AI and the Miyagi formula was smaller than that of the difference between the AI and the JSUM formula ($p = 6.68 \times 10^{-12}$) and that of the Miyagi formula and the JSUM formula ($p = 1.23 \times 10^{-15}$). In all three comparisons, no absolute systematic differences or proportional errors were observed. The JSUM formula and either the AI or the Miyagi formula showed divergence as the mean increased (*left and right panels*). These results suggested that the variation of the JSUM formula depended strongly on the magnitude of the measurements.

Table 5 Comparison of the AI that was created with all of the JSUM data, the JSUM formula, and the Miyagi formula with all of the JSUM datasets by Bland-Altman analysis [26-28]

Statistic	AI vs JSUM formula	AI vs Miyagi formula	Miyagi formula vs JSUM formula
Sample size	120	120	120
Difference (g)	40.5	−20.1	60.6
Lower 95% LoA (g)	−111.1	−114.0	−157.6
Upper 95% LoA (g)	192.1	73.7	278.8
Upper 95% exact CI for lower 95% LoA (g)	−15.3	−54.7	−19.7
Lower 95% exact CI for lower 95% LoA (g)	−137.6	−130.4	−195.7
Upper 95% exact CI for upper 95% LoA (g)	218.6	90.1	317.0
Lower 95% exact CI for upper 95% LoA (g)	96.3	14.4	141.0
The coefficient of Repeatability	170.5	101.5	247.7
p-value by *t*-test for p (H0: Mean = 0)	$p = 7.45 \times 10^{-8}$	$p = 1.04 \times 10^{-5}$	$p = 2.56 \times 10^{-8}$

LoA: limits of agreement.

1). All of the estimated fetal weights generated by the AI created with the training dataset of the JSUM data were not significantly different from the JSUM data, which are considered the standard dataset of Japanese fetuses. The main advantage of this AI method is its use of the neural network, which may provide good accuracy for extreme fetal weights. The rAI values at −2SD and +2SD were 6.4 ± 2.6 g and -3.8 ± 8.6 g, respectively. On the other hand, the rJSUM values at −2SD and +2SD were -111.6 ± 95.8 g and 54.2 ± 42.6 g, and the rMiyagi values at −2SD and +2SD were 82.4 ± 7.7 g and -44.8 ± 36.0 g, respectively. However, the rAI values were not smaller than the rJSUM values and rMiyagi values except for the Miyagi formula at +2SD, probably because of the large number of standard deviations of the latter two formulae and the small sample size of test dataset which used 3 points for each category. Therefore, the subsequent AI created by using all of the JSUM datasets demonstrated that the rAI' values were significantly smaller than the rJSUM' values and the

rMiyagi' values at not only the −2SD and +2SD categories but also all categories ($p < 0.001$) except for the Miyagi formula at ±0SD (as shown in Table 4). The AI's good accuracy for extreme fetal weights is likely to be very useful. Thus, AI with the neural network seems to have potential for estimating fetal weights.

We applied a Bland-Altman plot analysis to the JSUM formula and the Miyagi formula as a method for comparing two tests. Since this analysis revealed that the differences between the two methods were not significantly different from zero and no divergence was observed, the AI seemed to have features in common with the Miyagi formula. On the other hand, the comparison of the JSUM formula with either the AI or the Miyagi formula showed divergence as the mean increased. We thus consider the AI method and the Miyagi formula to be in agreement, and we suspect that they might be used interchangeably, although using the AI method would be better because of the smaller residuals of AI.

We created an AI method by using all of the datasets after creating an AI method with the training and validation datasets of the JSUM data, and the latter AI method appears to have the ability to estimate fetal weights. The ranges of the 5-95th percentiles as well as the SDs of the estimated fetal weights are given in Table 3. The values of the percentiles may have much more clinical implications in determining small-for-gestational-age and large-for-gestational-age fetuses.

Although the AI method that we created by using all of the JSUM data demonstrated superiority to the JSUM formula and the Miyagi formula compared to the AI method created using the training and validation datasets, the possibility of over-fitting of the AI derived from all of the datasets remains [29-34]. In other words, the AI created by using all of the datasets may not fit well for non-standard fetuses, such as in multiple pregnancies, or for untrained gestational weeks that are not integers, such as 30.4 weeks of gestation. The fetal weights generated by the AI method created by using all of the JSUM datasets were reliable for normal fetuses of integer gestational age, but the AI method created by using the training and validation datasets should be used for patients with a gestational age in a non-integer format or non-standard fetuses, in order to avoid over-fitting.

The existing published formulae use mathematical models that are based on a regression analysis, and the method is to find the line, whether straight or not, that most closely fits the data according to a specific mathematical criterion. However, it is theoretically difficult for a single formula to estimate the values of fetal weight in the −2SD, −1.5SD, +1.5SD, and +2SD categories simultaneously. Since neural networks have complex structures of high-dimensional matrices, it is possible that deep learning could estimate those values as well as the ±0SD values. We speculate that this is why the AI method described herein demonstrated less residuals.

Here, the AI consisted of an original neural network architecture for deep learning with 13 layers, the first layer of which was a linear layer catenated with the gestational week, BPD, AC, and FL as the scalar. We tried different architectures including different types of layers such as dropout layers [35], and different first linear layers catenated with the BPD, AC, and FL but without the gestational week, *etc*. Those architectures resulted in less accuracy (data not shown). Because there is no gold-standard neural network architecture, improvements of the architecture might result in the creation of better AI methods. It is also possible that an improved AI method will be obtained if other parameters are used, such as the head circumference [36-38] and transverse abdominal diameter [39], which are used in published formulae. Because the JSUM has not published the standard values of those parameters for Japanese fetuses, it was not possible to include those parameters in the AI methods in the present study.

An external validation study should be conducted to evaluate the AI methods described herein for their use in clinical practice, although the supervised data we used are approved as the standard data. There seldom are actual fetuses with completely standard fetal biometry measured by ultrasound. The AI methods should be validated on an actual dataset of biometrical evaluations 3-5 days prior to delivery, and the estimated fetal weights should be compared with the birth weights.

Several formulae for estimating fetal weight have been published. There are nine published regression formulae that are functions of at least one of the parameters of BPD, AC, and FL: the JSUM [2], Miyagi [6], Campbell [40], Shepard [41], Mertz [42], Hadlock II and Hadlock III [36], Warsof [43], and Schild (for female fetuses) [37] formulae. Burd *et al.* (2009) reported that among several formulae they examined for 81 fetuses in the U.S., Formula C (described by Hadlock *et al.* [36]) — which is a function of BPD, AC and FL —

had the best performance according to the bias and precision method [44]. In Germany, Siemer *et al*. (2008) reported that among 11 formulae, two Hadlock formulae [36, 38] including abdominal circumference, FL, head circumference, and BPD showed the best levels of accuracy in newborns with a birth weight < 2,500 g (n = 160) [1].

All of these formulae were derived from datasets with completely different racial compositions. Anthropometric differences reflected in fetal biometry may strongly affect the results of the comparisons of formulae. When those formulae are compared, a regression analysis with terms that are similar to the published terms should be performed to see how the regressions coefficients change compared to those reported in other studies. Alternatively, the coefficients in the Japanese dataset being compared with the published coefficients could be obtained in a Bayesian framework. However, a prospective study might be still required for Japanese fetuses. The estimated crude sample sizes that are necessary to validate the significant difference between the JSUM data and the AI method that showed -0.32 ± 6.3 (g) as the residual following a normal distribution for the test dataset are 120,000 for $p < 0.05$ and 200,000 for $p < 0.01$. However, it is possible that the use of neural network architecture with deep learning could be feasible for localized areas considering racial compositions.

The limitations of this study should be considered. We applied only the averaged data of −2SD, −1.5SD, ±0SD, +1.5SD, and +2SD categories of the fetuses to train the AI. We do not know the reliability of this AI method for non-standard fetuses such as in multiple pregnancies, asymmetrically developed fetuses, fetuses with congenital malformations, *etc*. More data or specific data may be necessary to improve the AI method for general use or for specific uses. If some parameters, whether known or unknown, are found to be of importance in estimating fetal weights, and if such parameters are used for AI training, an improved generalized AI method for estimating the fetal weight of Japanese fetuses will be obtained. Such parameters may be values detected by ultrasound or tensors measured by different modalities such as genetic information. Similarly, if specific parameters related to the weight of non-standard fetuses are discovered, a specific AI method can be created for each specific case, such as multiple pregnancies.

In conclusion, AI with deep learning demonstrated potential for accurately estimating fetal weights and might be superior to the commonly used regression formulae. Though improved AI methods for estimating fetal weights can be created by using different neural network architecture, by using other parameters, and by adding data for non-integer gestational ages, prospective studies may be required to validate this. Nevertheless, our present findings revealed that AI could have the ability to accurately estimate fetal weights in conventional clinical practice.

References

1. Siemer J, Egger N, Hart N, Meurer B, Müller A, Dathe O, Goecke T and Schild RL: Fetal weight estimation by ultrasound: comparison of 11 different formulae and examiners with differing skill levels. Ultraschall Med (2008) 29: 159–164.
2. Okai T: Standard values of ultrasonic measurements in Japanese fetuses. J Med Ultrasonic (2003) 30: J415–440 (in Japanese).
3. Okai T: Studies on fetal growth and functional development. Acta Obst Gynaec Jpn (1986) 38: 1209–1217 (in Japanese).
4. Ichijo M: Announcement from perinatal period committee. Acta Obst Gynaec Jpn (1993) 45: 391–394 (in Japanese).
5. Shinozuka N, Masuda H, Kagawa H and Taketani Y: Standard values of ultrasonographic fetal biometry. Jpn J Med Ultrasonics (1996) 23: 879–888 (in Japanese).
6. Miyagi Y, Tada K, Takayoshi R, Oguni N, Sato Y, Shibata M, Kiyokawa M, Hashimoto T, Takada T, Oda T and Miyake T: Formulae based on biomathematics to estimate the standard value of fetal growth of Japanese. Acta Med Okayama (2018) 72: 115–119.
7. Miyagi Y, Habara T, Hirata R and Hayashi N: Feasibility of artificial intelligence for predicting live birth without aneuploidy from a blastocyst image. Reprod Med Biol (2019) 18: 204–211.
8. Miyagi Y, Habara T, Hirata R and Hayashi N: Feasibility of deep learning for predicting live birth from a blastocyst image in patients classified by age. Reprod Med Biol (2019) 18: 190–203.
9. Miyagi Y, Habara T, Hirata R and Hayashi N: Feasibility of predicting live birth by combining conventional embryo evaluation with artificial intelligence applied to a blastocyst image in patients classified by age. Reprod Med Biol (2019) 18: 344–356.
10. Khosravi P, Kazemi E, Zhan Q, Malmsten JE, Toschi M, Zisimopoulos P, Sigaras A, Lavery S, Cooper LAD, Hickman C, Meseguer M, Rosenwaks Z, Elemento O, Zaninovic N and Hajirasouliha I: Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization. NPJ Digit Med (2019) 2: 21.
11. Miyagi Y, Habara T, Hirata R and Hayashi N: Predicting a live birth by artificial intelligence incorporating both the blastocyst image and conventional embryo evaluation parameters. Artif Intell Med Imaging (2020) 1: 94-107.
12. Looney P, Stevenson GN, Nicolaides KH, Plasencia W, Molloholli M, Natsis S and Collins SL: Fully automated, real-time 3D ultrasound segmentation to estimate first trimester placental volume using deep learning. JCI Insight (2018) 3: e120178.
13. Miyagi Y, Takehara K and Mitake T: Application of deep learning

to the classification of uterine cervical squamous epithelial lesion from colposcopy images. Mol Clin Oncol (2019) 11: 583–589. DOI: 10.3892/mco.2019.1932
14. Miyagi Y, Takehara K, Nagayasu Y and Miyake T: Application of deep learning to the classification of uterine cervical squamous epithelial lesion from colposcopy images combined with HPV types. Oncol Lett (2020) 19: 1602–1610.
15. Sato M, Horie K, Hara A, Miyamoto Y, Kurihara K, Tomio K and Yokota H: Application of deep learning to the classification of images from colposcopy. Oncol Lett (2018) 15: 3518–3523.
16. Shen WC, Chen SW, Wu KC, Hsieh TC, Liang JA, Hung YC, Yeh LS, Chang WC, Lin WC, Yen KY and K CH: Prediction of local relapse and distant metastasis in patients with definitive chemoradiotherapy-treated cervical cancer by deep learning from [18F]-fluorodeoxyglucose positron emission tomography/computed tomography. Eur Radiol (2019) 29: 6741–6749.
17. Miyagi Y, Fujiwara K, Oda T, Miyake T and Coleman RL: Development of new method for the prediction of clinical trial results using compressive sensing of artificial intelligence. J Biostat Biometric App (2018) 3: 202.
18. Matsuo K, Purushotham S, Jiang B, Mandelbaum RS, Takiuchi T, Liu Y and Roman LD: Survival outcome prediction in cervical cancer: Cox models vs deep-learning model. Am J Obstet Gynecol (2019) 220: 381. e1–381. e14.
19. Miyagi Y, Tada K, Yasuhi I, Maekawa Y, Okura N, Kawakami K, Yamaguchi K, Ogawa M, Kodama T, Nomiyama M, Mizunoe T and Miyake T: New method for determining fibrinogen and FDP threshold criteria by artificial intelligence in cases of massive hemorrhage during delivery. J Obstet Gynecol Res (2020) 46: 256–265.
20. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, Petersen S, Beattie C, Sadik A, Antonoglou I, King H, Kumaran D, Wierstra D, Legg S and Hassabis D: Human-level control through deep reinforcement learning. Nature (2015) 518: 529–533.
21. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A: Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition (2015) 1–9.
22. Ioffe S and Szegedy C: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv e-prints (2015) arXiv: 1502.03167v3
23. Glorot X, Bordes A and Bengio Y: Deep sparse rectifier neural networks. Proceedings of the fourteenth International Conference on Artificial Intelligence and Statistics (2011) 315–323.
24. Nair V and Hinton G: Rectified linear units improve restricted Boltzmann machines. Proceedings of International Conference on Machine Learning (2010) 807–814.
25. Klambauer G, Unterthiner T, Mayr A and Hochreiter S: Self-normalizing neural networks. Advances in Neural Information Processing Systems (2017) 971–980.
26. Altman DG and Bland JM: Measurement in medicine-the analysis of method comparison studies. Statistician (1983) 32: 307–317.
27. Bland JM and Altman DG: Statistical methods for assessing agreement between two methods of clinical measurement. Lancet (1986) 1: 307–310.
28. Carkeet, A: Exact parametric confidence intervals for Bland-Altman limits of agreement. Optometry and Vision Science (2015); 92: e71–e80.
29. Yu L, Chen H, Dou Q, Qin J and Heng PA: Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE Trans Med Imaging (2017) 36: 994–1004.
30. Caruana R, Lawrence S and Giles CL: Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. Advances in Neural Information Processing Systems (2001) 402–408.
31. Baum EB and Haussler D: What size net gives valid generalization? Neural Computation (1989) 1: 151–160.
32. Geman S and Bienenstock E: Neural networks and the bias/variance dilemma. Neural Computation (1992) 4: 1–58.
33. Krogh A and Hertz JA: A simple weight decay can improve generalization. Advances in Neural Information Processing Systems (1992) 4: 950–957.
34. Moody JE: The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems. Advances in Neural Information Processing Systems (1992) 4: 847–854.
35. Srivastava N, Hinton G, Krizhevsky A, Sutskever I and Salakhutdinov R: Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res (2014) 15: 1929–1958.
36. Hadlock FP, Harrist RB, Sharman RS, Deter RL and Park SK: Estimation of fetal weight with the use of head, body, and femur measurements — a prospective study. Am J Obstet Gynecol (1985) 151: 333–337.
37. Schild RL, Sachs C, Fimmers R, Gembruch U and Hansmann M: Sex-specific fetal weight prediction by ultrasound. Ultrasound Obstet Gynecol (2004) 23: 30–35.
38. Hadlock FP, Harrist RB, Carpenter RJ, Deter RL and Park SK: Sonographic estimation of fetal weight. The value of femur length in addition to head and abdomen measurements. Radiology (1984) 150: 535–540.
39. Hansmann M, Schumacher H and Voigt U: Mehrparametrische nicht lineare Gewichtsschätzung mittels Ultraschall unter Berücksichtigung des Gestationsalters; in Ultraschalldiagnostik, Kratochwil A and Reinold E eds, Georg Thieme Verlag, Stuttgart (1978) pp 69–71.
40. Campbell S and D Wilkin: Ultrasonic measurement of fetal abdomen circumference in the estimation of fetal weight. Br J Obstet Gynaecol (1975) 82: 689–697.
41. Shepard M, V Richards, R Berkowitz, SL Warsof and JC Hobbins: An evaluation of two equations for predicting fetal weight by ultrasound. Am J Obstet Gynecol (1982) 142: 47–54.
42. Merz E, Lieser H, Schicketanz KH and Härle J: Intrauterine fetal weight assessment using ultrasound. A comparison of several weight assessment methods and development of a new formula for the determination of fetal weight. Ultraschall in Med (1988) 9: 15–24.
43. Warsof SL, Gohari P, Berkowitz RL and Hobbins JC: The estimation of fetal weight by computer-assisted analysis. Am J Obstet Gynecol (1977) 128: 881–892.
44. Burd I, Srinivas S, Paré E, Dharan V and Wang E: Is sonographic assessment of fetal weight influenced by formula selection? J Ultrasound Med (2009) 28: 1019–1024.