

Title: Fully automated, real-time 3D-ultrasound segmentation to estimate first trimester placental volume using deep learning

Authors: Pádraig Looney¹

Tel: +44 (0)1865 221004 Email: Padraig.looney@obs-gyn.ox.ac.uk

Gordon N. Stevenson²

Tel: +61 (0)2 9382 6777 Email: gordon.stevenson@gmail.com

Kypros H. Nicolaides³

Tel: +44 (0)20 7924 0894 Email: kypros.nicolaides@kcl.ac.uk

Walter Plasencia⁴

Tel: +34 922 626 240 Email: walterplasencia@icloud.com

Malid Molloholli^{5, 6}

Tel: +44 (0) 1753 633377 Email: malid.molloholli@fhft.nhs.uk

Stavros Natsis⁵

Tel: +44 (0)1865 851165 Email: stavrosnatsis79@gmail.com

Sally L. Collins^{1, 5}

Tel: +44 (0)1865 851165 Email: sally.collins@obs-gyn.ox.ac.uk

¹Nuffield Department of Women's and Reproductive Health, University of Oxford,
Level 3, Women's Centre, John Radcliffe Hospital, Oxford OX3 9DU, UK

²School of Women's and Children's Health, University of New South Wales,
Randwick, New South Wales, Australia.

³Harris Birthright Research Centre of Fetal Medicine, Kings College Hospital, Denmark
Hill London SE5 8RX, UK.

⁴Fetal Medicine Unit. Hospiten Group. Tenerife. Canary Islands. Spain

⁵Fetal Medicine Unit, The Women's Centre, John Radcliffe Hospital Oxford OX3 9DU, UK

⁶Department of Obstetrics and Gynaecology, Wexham Park Hospital, Slough SL3 6AR, UK

Corresponding Author Details:

Email: padraig.looney@obs-gyn.ox.ac.uk

Tel #: +44 (0)1865 222036

Address: Dr Padraig Looney,
Nuffield Department of Women's and Reproductive Health,
University of Oxford,
Level 3, Women's Centre,
John Radcliffe Hospital, Oxford,
OX3 9DU.

Conflict of interest statement:

The authors have declared that no conflict of interest exists.

Abstract:

Objectives: We present a new technique to fully automate the segmentation of an organ from 3D ultrasound (3D-US) volumes, using the placenta as the target organ. Image analysis tools to estimate organ volume do exist but are too time consuming and operator-dependant. Fully automating the segmentation process would potentially allow the use of placental volume to screen for increased risk of pregnancy complications.

Methods: The placenta was segmented from 2393 first trimester 3D-US volumes using a semi-automated technique. This was quality controlled by three operators to produce the 'ground-truth' dataset. A fully convolutional neural network (OxNNet) was trained using this 'ground-truth' dataset to automatically segment the placenta.

Findings: OxNNet delivered state of the art automatic segmentation (median Dice similarity coefficient of 0.84). The effect of training set size on the performance of OxNNet demonstrated the need for large datasets (n=1200, median DSC (inter-quartile range) 0.81 (0.15)). The clinical utility of placental volume was tested by looking at prediction of small-for-gestational-age (SGA) babies at term. The receiver-operating characteristics curves demonstrated almost identical results (OxNNet 0.65 (95% CI; 0.61-0.69) and ‘ground-truth’ 0.65 (95% CI; 0.61-0.69)).

Conclusions: Our results demonstrated good similarity to the ‘ground-truth’ and almost identical clinical results for the prediction of SGA. Our open source software, OxNNet, and trained models are available on request.

Introduction

Researchers have been attempting to ‘teach’ computers to perform complex tasks since the 1970’s. With the falling cost of hardware and availability of open-source software packages, machine learning has experienced something of a renaissance. This has led to the development of several deep learning methods so named as they use neural networks with a complex, layered architecture. Fully convolutional neural networks (fCNNs) have provided state-of-the-art performance for object classification and image segmentation (1). Of 306 papers surveyed in a 2017 review on deep learning in medical imaging (2), 240 were published in the last two years. Of those using 3D data most used relatively small amounts of labelled data, ‘ground-truth’ datasets for training (n = 10; median 66; range 20-1088). The

Dice similarity coefficient (DSC; an index representing similarity between predicted and manually estimated data) in these studies was variable (median 0.84, range 0.72-0.92) and dependent on the difficulty of the segmentation task and imaging modality used. The key advantage of fCNNs is their robust performance when dealing with very heterogeneous input data, a particular challenge in ultrasound imaging. Efforts to segment the fetal skull in 3D ultrasound have obtained a DSC of 0.84 (3). A deep learning method to segment the placenta in MRI with a training set of 50 cases obtained a DSC of 0.72 (4). However, for fCNNs to be trained effectively, large datasets are required that reflect the diversity of organ appearance. Obtaining 'ground-truth' datasets is challenging due to the laborious nature of labelling the data which typically is performed by clinicians experienced with the particular imaging modality. Efforts to segment the placenta using different fCNNs have been recently presented but both used small data sets. A pilot study performed by the authors using a different, simpler architecture and on only 300 cases obtained a DSC = 0.73 (5) whilst another demonstrated a DSC = 0.64 using 104 cases (6). Whilst promising, what remains unclear is whether the DSC value, which is analogous to segmentation performance, is a result of the fCNN used or a reflection of the size the training set used.

First trimester placental volume (PIVol) has long been known to correlate with birthweight at term (7-9) and it was suggested as early as 1981 that PIVol measured with B-mode ultrasound could be used to screen for growth restriction (10). Since then many studies have demonstrated that a low PIVol between 11 and 13 weeks' gestation can predict adverse pregnancy outcomes including small for gestational age (SGA) (11) and pre-eclampsia (7). As PIVol has also been demonstrated to be independent of other biomarkers for SGA such as pregnancy associated plasma protein A (PAPP-A) (8, 11) and nuchal translucency (11), a recent systematic review concluded that it could be successfully integrated into a future multivariable screening method for SGA (12) analogous to the 'combined test' currently used

to screen for fetal aneuploidy. As PIVol is measured at the same gestation as this routinely offered ‘combined test’, no extra ultrasound scans would be required making it more economically appealing to healthcare providers worldwide.

Until now, the only way to estimate PIVol is for an operator to examine the 3-dimensional ultrasound (3D-US) image, identify the placenta and manually annotate it. Commercial tools such as VOCAL™ (Virtual Organ Computer-aided AnaLysis; General Electric Healthcare, Milwaukee, WI, USA) and a semi-automated Random Walker derived method have been developed (13) to facilitate this process but they remain too time consuming and operator dependent to be used as anything other than research tools. For PIVol to become a useful imaging biomarker, a reliable, real-time, operator-independent technique for estimation is needed.

This publication has three major contributions to the application of deep learning to medical imaging. Firstly, we apply a novel deep learning fCNN architecture (OxNNet) to a large amount of quality controlled ‘ground-truth’ data to generate a real time, fully automated technique for estimating PIVol from 3D-US scans. Secondly, the relationship of segmentation accuracy to size of the training set is investigated to determine the appropriate amount of training data required to optimise segmentation performance. Finally, the performance of the PIVol estimates generated by the fully automated fCNN method to predict SGA at term was assessed.

Results

The performance of models trained end-to-end on training sets of size 100, 150, 300, 600, 900, 1200 are shown in Figure 1. The mean squared error (MSE) on the validation set decreased monotonically from 0.039 to 0.030 and increased monotonically from 0.01 to 0.025 on the training set. The median (interquartile range) DSC obtained on the validation set

throughout training increased monotonically from 0.73 (0.17) to 0.81 (0.15). Statistical analysis demonstrated a significant improvement in the DSC values with increasing training set size (ANOVA: $p < 0.0001$).

The distributions of the metrics used to evaluate the performance of the automated segmentation are shown in Figure 2. The median (interquartile range) of the DSC, RVD, Hausdorff distance and mean Hausdorff distance were 0.84 (0.09), -0.026 (0.23), 14.6 (9.9) mm and 0.37 (0.46) mm respectively. The correlation coefficients for the four metrics (Table 1), demonstrated a closer correlation between the mean Hausdorff distance and the DSC compared to the Hausdorff distance and DSC or the absolute value of the RVD and DSC.

A visual comparison of the ‘ground-truth’ (RW) segmentation and the OxNNet segmentation with post-processing applied as previously described is shown for a typical case (51st DSC centile in Fig. 2A) in Figure 3 and in Supplemental Digital Content 1 as a video showing the rotation and different slicing through the 3D-US volume.

The median (minimum, maximum) values of PIVol for OxNNet and RW were 59 ml (17, 147) and 60 ml (12, 140) respectively. The PIVol in ml and log PIVol MoMs are shown in Figure 4. There were 157 cases of SGA in the cohort.

The ROC curves for the log PIVol (MoMs) calculated by the fully automated fCNN (OxNNet) and the RW technique to predict SGA are shown in Figure 5. The area under the curve (AUC) for both techniques were almost identical at 0.65 (95% CI; 0.61-0.69) for OxNNet and 0.65 (95% CI; 0.61-0.70) for the RW ‘ground-truth’.

Discussion

In summary, deep learning was used with an exceptionally large ‘ground-truth’ dataset to generate an automatic image analysis tool for segmenting the placenta using 3D-US. To the

best of our knowledge, this study uses the largest 3D medical image data set to date for fCNN training. In a number of data science competitions the best performing models have used similar model architectures to poorer performing models but employed data augmentation to artificially increase the training set (2) suggesting a link between performance and the dataset size. The learning curves presented here demonstrate a key finding of the need for large training sets and/or data augmentation when undertaking end-to-end training. The MSE learning curves of this model architecture the training and validation curves converged towards 0.275 as the training set size is increased. This was reflected in the monotonic increase across training samples where DSC for 1200 training cases = 0.81 and DSC = 0.73 for 100 training samples. These results show that by using approximately an order of magnitude more training data (100 to 1200) segmentation performance measured by DSC increased by 0.08.

Assessing whether OxNNet can appropriately segment the placenta from a 3D-US volume relies on the benchmark against which it is judged. We have gauged this in two ways; firstly, by how similar it is to the ‘ground-truth’ data set and secondly, how the estimated PIVols perform in the clinically relevant situation of predicting the babies who will be born small. The results are very promising for both. The median DSC of OxNNet was 0.84, a considerable improvement upon previously reported values of 0.64 (6) and 0.73 (5) demonstrating increased similarity between the PIVols estimated by OxNNet and those generated by the ‘ground-truth’ RW algorithm. Previous work to segment the fetus in 3D ultrasound obtained DSC values of 0.84 (3) and to segment of the placenta in MRI images obtained a DSC of 0.71 (4). On assessment of clinical utility, the OxNNet PIVol estimates perform as well for the prediction of SGA as those generated by the previously validated RW technique and outperforms the estimates generated in the original analysis of this data using the proprietorial VOCALTM tool (AUC 0.60 (0.55–0.65)) (14).

In terms of similarity to the 'ground-truth', distributions of the metrics in Figure 2 show that 90% of cases had a DSC > 0.74 and a Hausdorff distance < 28 mm. Discrepancy between the 'ground-truth' segmentation and the prediction by OxNNet must be due either to an error with OxNNet or with the 'ground-truth'. The commonly regarded 'gold standard' for segmentation of a target organ in a ultrasound volume is manual segmentation. This involves painstakingly drawing in the outline of the organ for every slice of the 3D image. This is highly operator dependant. The RW technique has been shown to be comparable to manual segmentation in all aspects of observer reliability (13) and is less time consuming but still remains dependant on the operator's ability to identify the placenta and its boundaries. The major issue here this is that ultrasound images in the first trimester can be very difficult to interpret and ultimately the exact position of the interface between the placenta and the myometrium is often a difficult call even in the hands of a highly experienced sonographer. Any system reliant on human judgement will be open to increased inter and intra-observer variability in these situations and therefore despite considerable efforts to quality control the 'ground-truth' dataset it is highly likely that errors in the segmentation have occurred. It is anticipated that an automatic system working from a voxel level algorithm will be more reproducible when confronted by such difficult boundaries but further investigation is needed to confirm this. Another limitation of this study is that the data was collected several years ago using an ultrasound system which has since been superseded by two newer generations. As B-Mode quality has significantly improved, it is hoped that the image quality will be increased in future studies facilitating easier segmentation.

The RVD metrics demonstrated that generally OxNNet overestimates the volume of the placenta when compared to the 'ground-truth'. Thin mislabelled regions spreading away from the placenta increases the Hausdorff distance without dramatically affecting the DSC values. In some cases, it was evident that OxNNet had labelled some of the myometrium as placental

tissue. This may have been a result of inconsistent identification of the utero-placental interface by the operator in the 'ground-truth' dataset especially when there was difficulty deciding where this interface lay. However, by increasing the context of the fCNN either by using a larger receptive field, employing a secondary recursive neural network (6) or using conditional random fields (15) the mislabelling error may reduce. Further work is required to investigate this.

The PIVol generated by OxNNet demonstrate that for a false positive rate of 10%, the estimated detection rate for SGA is 23% (16-31%) this is an improvement on the previously published (14) detection rate of 18% (12-27%) seen with the same data set but using the operator-dependant VOCAL™ system. This alone is not good enough to provide a clinically useful screening tool. However, much like the improvement in the performance of the nuchal translucency for prediction of aneuploidies, the combination of PIVol with other independent risk factors increases its utility. In the previous study combining the PIVol with maternal characteristics and serum PAPP-A increased the SGA detection rate from 18% to 35% (14). How PIVol performs when combined with other serum markers for SGA such as placental growth factor (PlGF) or ultrasound markers of vascularity such as fractional moving blood volume (FMBV) remains to be seen however with this fully automated tool, large, multicentre studies recruiting many thousands of women can now be undertaken to investigate this. The results should demonstrate the relationship between first trimester PIVol and not only birthweight, but much less common adverse pregnancy outcomes such as pre-eclampsia, placental abruption and stillbirth. If clinical utility is proved, this real-time, operator independent technique makes it possible to use PIVol on a large scale potentially as part of a screening test.

Obtaining large annotated datasets is time consuming and labour intensive. By previous timing of the semi-automated RW method, annotation of the dataset presented represents an

estimated 168.1 hours of segmentation (mean initialisation = 175s; computation = 43.6s; n = 2768) by a single observer. This is usually a major stumbling block when researchers are trying to generate a ‘ground-truth’ training set. However, using transfer learning the limitation of small data sets can circumvent this by using pre-trained networks. Previous work in this area has shown that fine tuning of a pre-trained model based on Google’s Inception v3 architecture on medical data achieved near human expert performance (16, 17). Transfer learning should allow the large dataset and model presented in this work to benefit researchers in other imaging modalities. To enable application of this work to other imaging modalities or different target organs in 3D-US, our source code is freely available (18) and the pre-trained models available on request from the authors. We hope that this will prove beneficial to this rapidly growing area of medical image analysis.

Methods

Clinical Dataset

The 3D-US data was previously used in a study investigating the predictive value of PIVol, measured using the commercial VOCALTM tool, for the detection of SGA (14). A 3D-US volume containing the placenta was recorded for 3104 unselected singleton pregnancies at 11 + 0 to 13 + 6 weeks’ gestation. This was all the women presenting for their combined

screening for aneuploidies at the Fetal Medicine Centre, London, UK who gave their consent (19, 20). All the women went on to deliver a chromosomally normal baby at term. The 3D-US volume was acquired by trans-abdominal sonography using a GE Voluson 730 Expert system (GE Medical Systems, Milwaukee, Wisc., USA) with a 3D RAB4/8L transducer (21). Of the original 3104 3D-US volumes, 336 had to be discarded as they had been saved using wavelet-compression, which results in significant loss of the underlying raw data thereby preventing further analysis. Another 375 cases were excluded as the volume had been collected with the gain set exceptionally high. This gain setting is inappropriate for imaging the placenta as it removes the subtle variation in the echogenicity of tissues resulting in a ‘stark’, black and white image appearance. It is used in clinical practice as it makes the nuchal translucency more obvious.

The remaining 2393 3D-US volumes were annotated using the Random Walker (RW) algorithm which has been described previously (13). To perform labelling, 3D B-mode data was converted from the pre-scan toroidal geometry GE Voluson format into a 3D Cartesian volume with isotropic 0.6 mm spacing (5). The segmentation was initialised or ‘seeded’ by an operator (SN). These ‘seedings’ were then examined for accuracy by a second, independent, operator (MM) and ‘re-seeded’ where mistakes were evident. Cases where there was uncertainty regarding the boundaries of the placenta were examined by a third operator (SC). The ‘seedings’ were then used to calculate the PIVol with the RW method. The final quality control step for the ‘ground-truth’ dataset involved visually inspecting the segmentation of all the cases seen to be outliers in the distribution of PIVol values. This was performed by three operators (SC, PL and GS), if an error was seen in the segmentation the seeding was checked and the image re-seeded and re-segmented where appropriate. The resulting 2393 quality controlled ‘ground-truth’ segmentations were then used to train, validate and test the models.

3D Deep Learning Segmentation Method

A fCNN, that will be referred to as OxNNet, was created using the framework TensorFlow (version 1.3) using a 3D architecture inspired by a 2D U-net architecture described previously (22). The number of convolutional layers and channels used was customised and max pooling replaced with strided convolutions (23) to accommodate the 12Gb NVIDIA Titan X GPU (24) used for training. Figure 6 shows a full schematic of the architecture. Cross entropy was used as the loss function. The parameters of the Adam optimizer learning rate, β_1 , β_2 and ϵ were set as 0.001, 0.9, 0.999 and 1×10^{-8} respectively. To reduce overfitting, dropout with probability 0.5 was applied to the final layer. A batch size of 30 was used while training the model.

The effect of training set size on the performance of the model was investigated by keeping the validation set fixed and using samples of 100, 150, 300, 600, 900 and 1200 cases trained for 25,000 iterations throughout.

To evaluate the predictive value of PIVol segmentation, 2-fold cross-validation was performed providing training, validation and test partitions of 1097, 100 and 1196 cases respectively. Each volume was normalized to have zero mean intensity and unary variance. Masks of the ultrasound region were input to the fCNN to only consider the field of view. The fCNN was trained for 8 epochs and took 26 hours to run. Validation of the image segments was performed throughout training and a full validation of the whole image was carried out every epoch. Computation of a PIVol following training took on average 11 seconds.

Each predicted segmentation was post-processed to remove disconnected parts of the segmentation less than 40% of the volume of the largest region. The segmentation was binary dilated and eroded using a 3D kernel of radius three voxels and a hole filling filter applied.

These methods removed small regions separated from the largest placental segmented regions, smoothed the boundary of the placenta and filled any holes that were surrounded by placental tissue.

The main evaluation metric was the Dice similarity coefficient (DSC). Relative volume difference (RVD), mean Hausdorff distance and Hausdorff distance were also computed using the Insight Toolkit (ITK; version 4.10). The volumetric metrics for two segmentations, A and B, were defined as:

$$DSC(A, B) = 2 \frac{|A \cap B|}{|A| + |B|}$$

$$RVD(A, B) = \frac{|B| - |A|}{|A|}$$

The Hausdorff distance and mean Hausdorff distance were the maximum and mean of the minimum distances respectively averaged between the surfaces of A and B, and, B and A.

Statistics

The difference in the DSC values obtained with different dataset sizes was assessed using one-way analysis of variance (ANOVA). The birthweight percentile for each neonate was taken from a reference range of birthweight for gestation at delivery in the population from which the data was acquired (10). A neonate was considered SGA if it was < 10th percentile birthweight. The distribution of pIVol was made Gaussian by logarithmic transformation (normality was assessed using histograms and probability plots) and differences in gestational age were corrected for by expressing log PIVol as multiples of the median (MoM) of the AGA group. The distribution of log PIVol expressed as multiples of the median (MoM) were calculated for all cases for RW and OxNNet PIVols. Univariate logistic regression was used to build a predictive model to detect SGA for by both techniques (RW and OxNNet). The performance of these models in detecting SGA was assessed by generating receiver-operating

characteristic (ROC) curves. Ultrasound volumes were visualised using 3D Slicer (version 4.6) (25), R (version 3.3.2) (26) was used for data analysis, pROC (version 1.8) (27) was used for the ROC analysis and ggplot2 (version 2.2) (28) for producing the graphs. Statistical significance was assumed at a p value of <0.05.

Study approval

The study had full local ethical approval (ID:02-03-033) and all participants provided written consent.

Author contributions

Guarantors of integrity of entire study, PL, GS, SC; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; literature research, PL, GS, SC; clinical studies, WP; experimental studies, PL, GS, MM, SN, SC; statistical analysis, PL, GS, SC.

Acknowledgments

The authors thank Prof. J. Alison Noble for her valuable input into the original placental imaging analysis that lead to the development of this work. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla GTX Titan X GPU used for this research. PL, SC and research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) Human Placenta Project of the National Institutes of Health under award number U01-HD087209. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. GS is supported by a philanthropic grant from the Leslie Stevens' Fund, Sydney.

References

1. Krizhevsky A, Sutskever I, and Hinton GE. *Advances in neural information processing systems*. 2012:1097-105.
2. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *arXiv preprint arXiv:170205747*. 2017.
3. Cerrolaza JJ, Oktay O, Gomez A, Matthew J, Knight C, Kainz B, et al. *Fetal, Infant and Ophthalmic Medical Image Analysis*. Springer; 2017:25-32.
4. Alansary A, Kamnitsas K, Davidson A, Khlebnikov R, Rajchl M, Malamateniou C, et al. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2016:589-97.
5. Looney P, Stevenson GN, Nicolaides KH, Plasencia W, Molloyholli M, Natsis S, et al. *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*. IEEE; 2017:279-82.
6. Yang X, Yu L, Li S, Wang X, Wang N, Qin J, et al. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2017:711-9.
7. Hafner E, Metzenbauer M, Höfner D, Stonek F, Schuchter K, Waldhör T, et al. Comparison between three-dimensional placental volume at 12 weeks and uterine artery impedance/notching at 22 weeks in screening for pregnancy-induced hypertension, pre-eclampsia and fetal growth restriction in a low-risk population. *Ultrasound in obstetrics & gynecology*. 2006;27(6):652-7.
8. Law L, Leung T, Sahota D, Chan L, Fung T, and Lau T. Which ultrasound or biochemical markers are independent predictors of small-for-gestational age? *Ultrasound in Obstetrics & Gynecology*. 2009;34(3):283-7.
9. Metzenbauer M, Hafner E, Höfner D, Schuchter K, Stangl G, Ogris E, et al. Three-dimensional ultrasound measurement of the placental volume in early pregnancy: method and correlation with biochemical placenta parameters. *Placenta*. 2001;22(6):602-5.
10. Jones TB, Price RR, and Gibbs SJ. Volumetric determination of placental and uterine growth relationships from B-mode ultrasound by serial area-volume determinations. *Investigative radiology*. 1981;16(2):101-6.
11. Collins SL, Stevenson GN, Noble JA, and Impey L. Rapid calculation of standardized placental volume at 11 to 13 weeks and the prediction of small for gestational age babies. *Ultrasound in medicine & biology*. 2013;39(2):253-60.
12. Farina A. Systematic review on first trimester three-dimensional placental volumetry predicting small for gestational age infants. *Prenatal diagnosis*. 2016;36(2):135-41.
13. Stevenson GN, Collins SL, Ding J, Impey L, and Noble JA. 3-D ultrasound segmentation of the placenta using the random walker algorithm: Reliability and agreement. *Ultrasound in medicine & biology*. 2015;41(12):3182-93.
14. Plasencia W, Akolekar R, Dagklis T, Veduta A, and Nicolaides KH. Placental volume at 11–13 weeks' gestation in the prediction of birth weight percentile. *Fetal diagnosis and therapy*. 2011;30(1):23-8.
15. Kamnitsas K, Ledig C, Newcombe VF, Simpson JP, Kane AD, Menon DK, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical image analysis*. 2017;36:61-78.

16. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-8.
17. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*. 2016;316(22):2402-10.
18. Looney P. OxNNet. <https://github.com/plooney/oxnnet>.
19. Kagan K, Wright D, Baker A, Sahota D, and Nicolaides K. Screening for trisomy 21 by maternal age, fetal nuchal translucency thickness, free beta-human chorionic gonadotropin and pregnancy-associated plasma protein-A. *Ultrasound in Obstetrics & Gynecology*. 2008;31(6):618-24.
20. Snijders R, Noble P, Sebire N, Souka A, and Nicolaides K. UK multicentre project on assessment of risk of trisomy 21 by maternal age and fetal nuchal-translucency thickness at 10–14 weeks of gestation. *The Lancet*. 1998;352(9125):343-6.
21. Wegrzyn P, Faro C, Falcon O, Peralta C, and Nicolaides K. Placental volume measured by three-dimensional ultrasound at 11 to 13+ 6 weeks of gestation: relation to chromosomal defects. *Ultrasound in obstetrics & gynecology*. 2005;26(1):28-32.
22. Ronneberger O, Fischer P, and Brox T. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2015:234-41.
23. Milletari F, Navab N, and Ahmadi S-A. *3D Vision (3DV), 2016 Fourth International Conference on*. IEEE; 2016:565-71.
24. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:160304467*. 2016.
25. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin JC, Pujol S, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn Reson Imaging*. 2012;30(9):1323-41.
26. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2016.
27. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics*. 2011;12(1):77.
28. Wickham H. *ggplot2: elegant graphics for data analysis*. Springer; 2016.

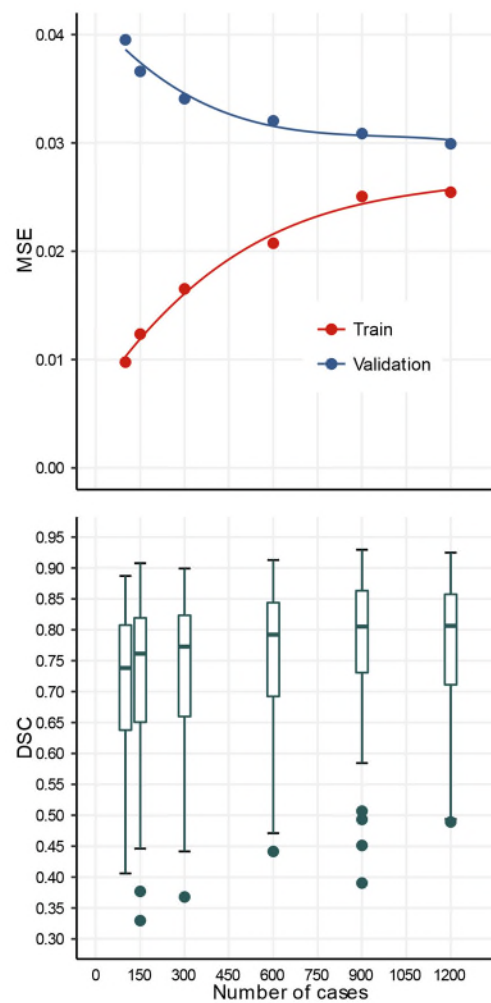


Figure 1. Learning curves of mean squared error (MSE) for both training (red) and validation (blue) datasets for different numbers of cases in training (top). Boxplot of Dice Similarity Coefficients (DSC) for OxNNet using different numbers of cases in training (100-1200; bottom).

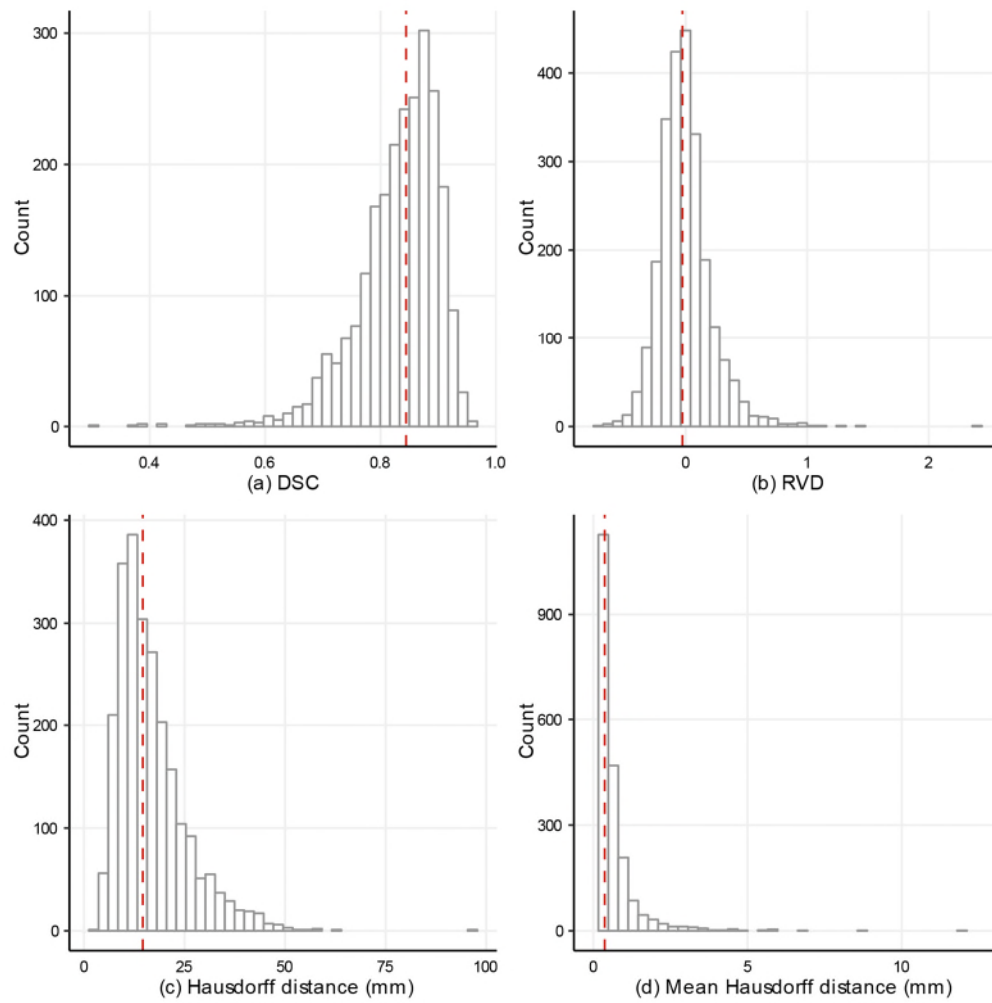


Figure 2. Histograms showing the distribution of the Dice similarity coefficient (DSC), relative volume difference (RVD) and Hausdorff distance (actual (c) and mean values (d)) for the cross validated test sets of 2393 cases. The median is shown by the red dashed line in each figure.



Figure 3. Placental segmentations with 2D B-mode plane (left). The RW segmentation (centre; red) and the OxNNNet prediction (right; blue). The values of the Dice similarity coefficient and Hausdorff distance metrics for this case were 0.838 and 12.6 mm, respectively.

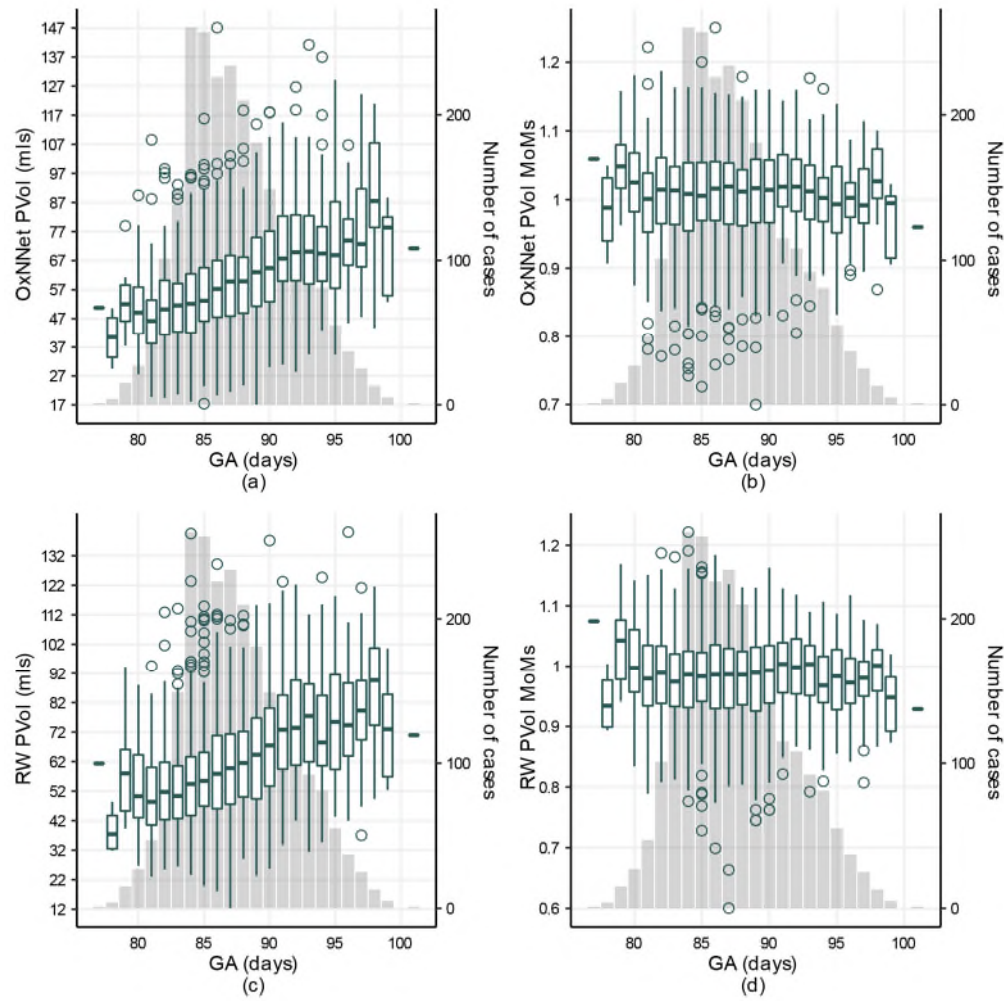


Figure 4. Box plots showing the distribution of actual placental volumes (PIVol) for OxNNet (a), and logarithm of the multiples of the medians (MoMs) for OxNNet (b), actual placental volumes (PIVol) for Random Walker (c) and logarithm of the multiples of the medians (MoMs) for Random Walker (d) versus the gestational age (GA). The number of cases for each GA (vertical axis on the right) is plotted as a column chart in the background.

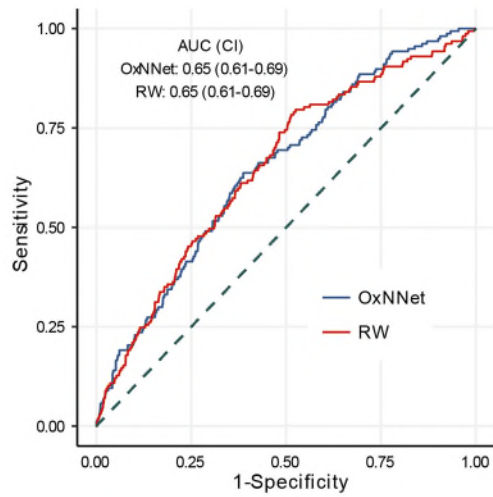


Figure 5. Receiver-operating characteristics (ROC) curves of placental volume calculated by both the fully automated fCNN (OxNNet) and the Random-Walker (RW) technique to predict small for gestational age (SGA: <10th percentile birth weight). Area under the curve (AUC) and 95% confidence intervals are shown for each model.

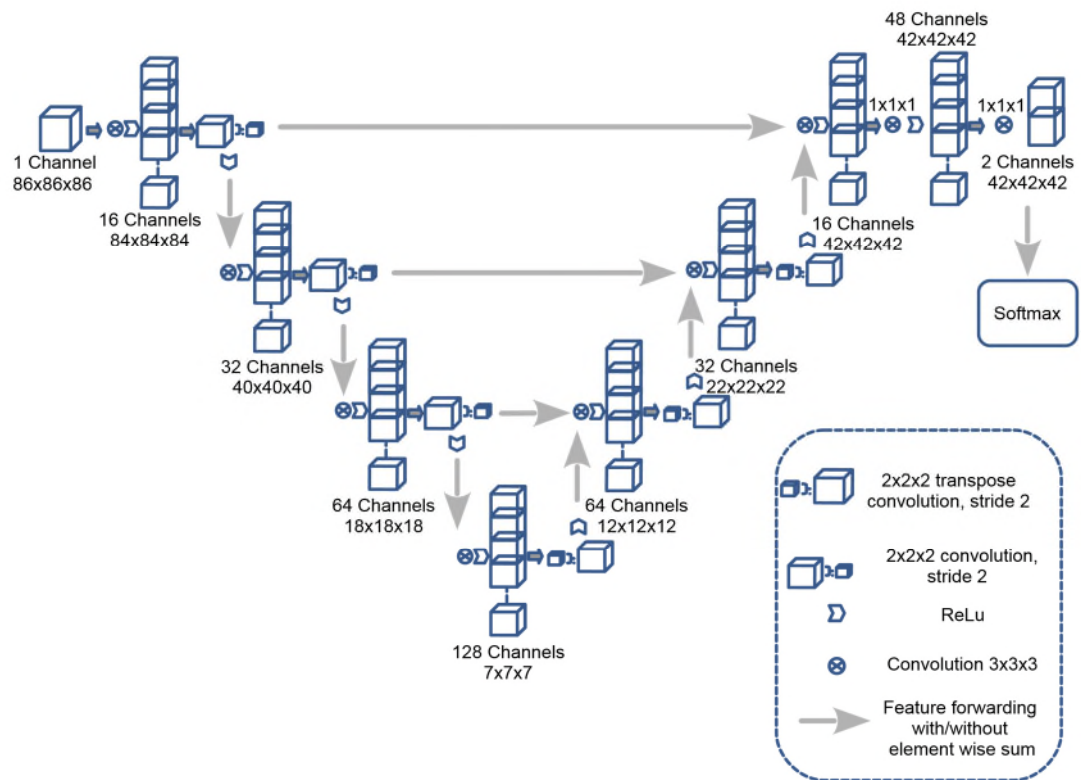


Figure 6. The architecture of the OxNNet fully convolutional neural network (fCNN).

Table 1: Pearson's correlation coefficient (r) and 95% confidence intervals for the metrics.

Dice Similarity Coefficient = DSC; Relative Volume Difference = RVD.

	DSC	Mean Hausdorff	Hausdorff	Absolute RVD
DSC	1	-0.81 (-0.82, -0.79)	-0.59 (-0.62, -0.57)	-0.66 (-0.68, -0.64)
Mean Hausdorff	-0.81 (-0.82, -0.79)	1	0.70 (0.68, 0.72)	0.51 (0.49, 0.55)
Hausdorff	-0.59 (-0.62, -0.57)	0.70 (0.68, 0.72)	1	0.38 (0.35, 0.42)
Absolute RVD	-0.66 (-0.68, -0.64)	0.51 (0.49, 0.55)	0.38 (0.35, 0.42)	1