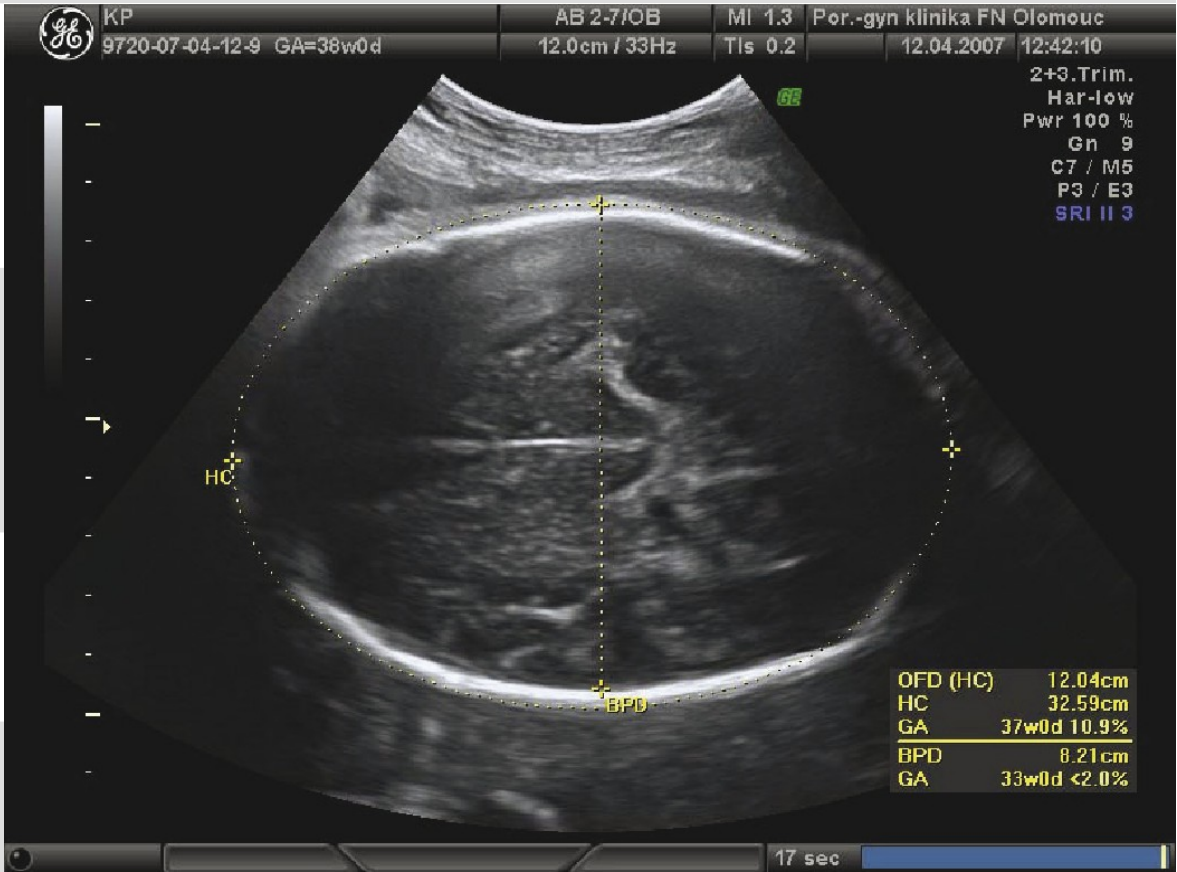
Confident Head Circumference Measurement from Ultrasound with Real-time Feedback for Sonographers

by S. Budd et al. (2019)
[arXiv:1908.02582v1]



MACIEJ MANNA

26.03.2020 r.

Plan prezentacji

1. Wprowadzenie –

– badania prenatalne i pomiar obwodu głowy, automatyzacja pomiaru, cele pracy;

2. Metoda –

– architektura modelu, szacowanie kształtu i obwodu głowy, segmentacja probabilistyczna, oceny wariancji i pewności dokonanego pomiaru;

3. Wyniki eksperymentalne –

– opis zbioru danych i układu eksperymentalnego, wyniki dla pojedynczego i wielokrotnego próbkowania;

4. Wnioski –

– dyskusja nad otrzymanymi wynikami, wnioski i dalsze perspektywy badań

- 1 -

WPROWADZENIE

Badania prenatalne

- Istotny element troski o zdrowie matki i dziecka w czasie ciąży, gdyż wczesne wykrycie anomalii w rozwoju płodu pozwala na skuteczniejsze ich leczenie.
- Standardowo trzy badania referencyjne USG, po jednym w każdym trymestrze ciąży – w tym tzw. **USG połówkowe** w czasie II trymestru (**18-22 tydzień**), którego celem jest pierwsza dokładna ocena rozwiniętej anatomii płodu.
- Badający obrazuje wybrane, „**ustandaryzowane płaszczyzny** *(standardized planes)*“ widoku, na których można dokonać pomiarów biometrycznych wybranych części ciała płodu (głównie głowy, brzucha i kości udowej), na podstawie których szacuje się jego wagę, dokładny wiek oraz przewidywany kierunek dalszego rozwoju (w tym możliwe zabużenia i anomalie).
- **płaszczyzna TV** (*standard transverse brain view at the posterior horn of the ventricle*) – służy do pomiaru obwodu głowy (HC) płodu

Trudności pomiaru obwodu głowy

- Pomimo wielkiej wartości, stosunkowo rzadko na tym etapie badań wychwytywane są zaburzenia i anomalie w rozwoju płodu.
- Wynika to z dużego progu niezbędnych umiejętności i ekspertyzy ze strony przeprowadzającego badanie, aby było one skuteczne:
  - pierwsza trudność polega na samym przeprowadzeniu badania i otrzymaniu zdjęć w odpowiednich płaszczyzanch, aby wszelkie pomiary były możliwe i wiarygodne,
  - druga trudność polega na samyym oszacowaniu pomiarów biometrycznych na podstawie tak otrzymanych zdjęć.
- Namacalnym przejawem tego faktu jest zaobserwowane **duże zróżnicowanie w oszacowaniach** między osobami interpretującymi wyniki badań.
- Prowadzi to do prób dokonania automatycznych oszacowań oraz wsparcia osób prowadzących badaniach, aby ograniczyć tę wariancję.

Rozwiązania automatyczne

- Propozycje rozwiązań opartych na metodach **konwencjonalnych**:
  - **2008** - ograniczone probabilistyczne drzewa boostingu (*constrained probabilistic boosting trees*),
  - **2018 –** lasy losowe i szybkie dopasowanie elipsy (*random forests and fast ellipse fitting*).
- Znacznie lepsze wyniki otrzymano przy użyciu metod opartych na **deep learningu**, w szczególności **pełnych, kaskadowych sieciach konwolucyjnych** (oparte na **U-Net**; 2017, 2018).
- Oprócz bezpośredniej estymacji parametrów biometrycznych (w tym HC), wykarzystano również głębokie sieci neuronowe do detekcji i lokalizacji standardowych płaszczyzn widoku (SonoNet; 2017).
- Bezpośrednie metody dają jedynie **punktową predykcję pomiaru, bez oceny poziomu pewności** z jaką została ona dokonana (jak w przypadku ekspertów, kilka zdjęć dokonanych w czasie tej samej sesji może dawać istotnie rozbieżne wyniki ze względu na jakość i płaszczyznę zdjęcia).

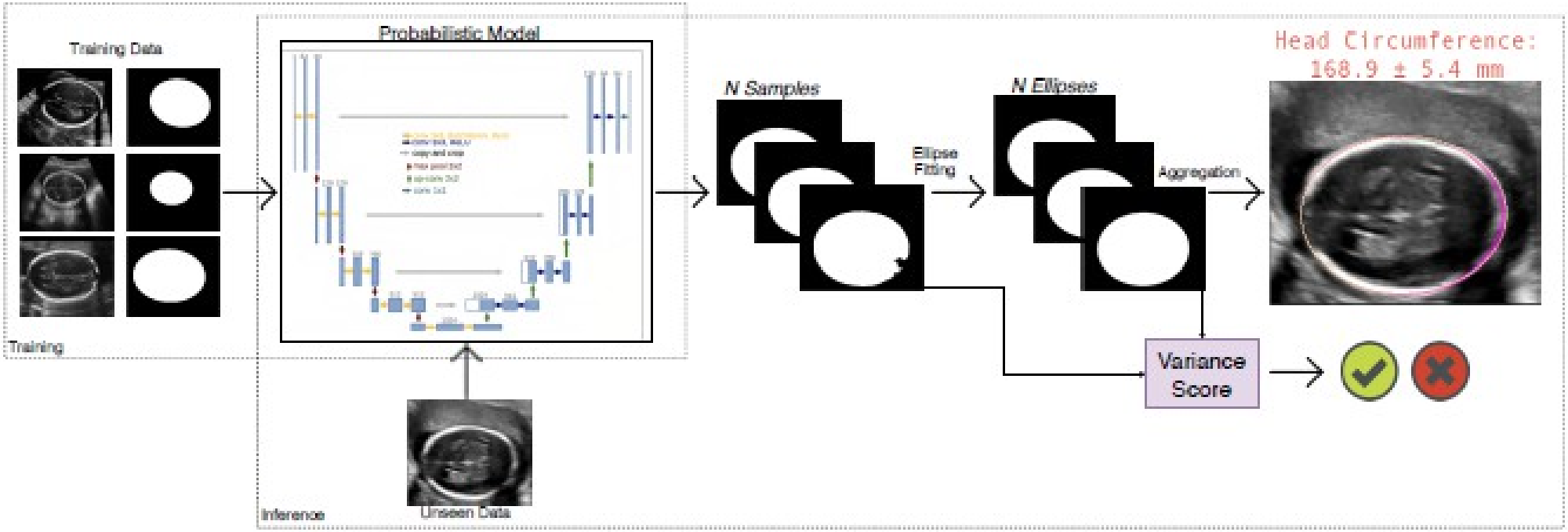
Cele pracy

- Cele przyświecające autorom w tworzeniu ich modelu:
  - **zachowanie skuteczności predykcji** pomiaru porównywalnej z obecnymi rozwiązaniami bazującymi na modelach deep learningowych,
  - wraz z wynikiem punktowym, **podanie zakresu niepewności** z jakim przewidziany został dany pomiar (dolne i górne ograniczenia dopuszczalnych pomiarów),
  - **opracowanie oceny wariancji**, która pozwoliłaby w czasie rzeczywistym stwierdzić, czy dane zdjęcie dobrze nadaje się do przeprowadzenia skutecznego pomiaru (dobra jakość, płaszczyzna) i wspomóc osobę przeprowadzającą badanie, czy należy je zachować, czy odrzucić.

- 2 -

METODA

Architektura modelu



$$X \xrightarrow{g_P(X)=\hat{X}_i} \hat{X}_i \xrightarrow{f(\hat{X}_i)=[a,b,\theta,x_c,y_c]^T} [a,b,\theta,x_c,y_c]^T \xrightarrow{HC=\pi(a+b)(1+\frac{3h}{10+\sqrt{4-3h}})s_{xy}} HC$$

Szkielet architektury

- **Szkielet architektury** – oparty na: Sinclair M. et al., *Human-level Performance on Automatic Head Biometrics In Fetal Ultrasound Using Fully Convolutional Neural Networks* (2018).
- **Elementy** modelu:
  - (1) **segmentacja obrazu** – zwraca maskę segmentacji wyznaczającą kształt głowy – implementowane przez sieć U-Net,
  - (2) **dopasowanie elipsy** – zwraca liczbowe parametry elipsy najlepiej dopasowanej do zwróconej maski – bezpośrednia metoda oparta na błędzie środniokwadratowym dopasowania do zbioru punktów (1996),
  - (3) **oszacowanie obwodu głowy** (elipsy) – zwraca konkretną liczbę – wykorzystuje tzw. aproksymacje Ramanujana:

$$HC = \pi(a+b)(1+\frac{3h}{10+\sqrt{4-3h}})s_{xy} \qquad \text{gdzie} \qquad h = \frac{(a-b)^2}{(a+b)^2}$$

(przybliżenie rzędu O(h^10), zaniedbywalne dla elips zbliżonych do okręgów)

Segmentacja probabilistyczna

- Dla obliczenia zakresów możliwych pomiarów oraz użytecznych miar pewności dla jednego przykładu (obrazu) potrzebujemy, zamiast jednej maski segmentacji, więcej – wybrane z odpowiedniego rozkładu prawdopodobieństwa.
- Musimy **zastąpić zwykłą sieć U-Net jej probabilistyczną alternatywą**, aby móc próbkować odpowiednie maski segmentacji, a następnie:
  - uśredniać ich wyniki, aby otrzymać ostateczną predykcję pomiaru,
  - badać ich współzależności, zakres i wariancje, aby na tej podstawie decydować o stopniu pewności powyższej predykcji.
- Takie podejście symuluje (i zarazem pomoże rozwiązać) problem dużej wariancji w pomiarach dokonywanych przez operatorów USG (a co za tym idzie również w opisach/etykietach zbioru danych).
- Dalej: N – ilość próbek na jeden przykład (obraz).
- Dwie alternatywy probabilistyczne dla deterministycznego U-Netu:
  - wprowadzenie warstwy **dropoutu Monte Carlo** (*MC Dropout*),
  - zastosowanie **sieci probabilistycznej U-Net** (*Probabilistic U-Net*).

Dropout Monte Carlo

- Standardowa warstwa dropout całkowicie wyłącza niektóre neurony (z prawdopodobieństwem *p*) w trakcie procesu uczenia sieci, ale już nie w trakcie predykcji (dzieki temu wynik predykcji jest deterministyczny).
- Dla odmiany, **warstwa dropout Monte Carlo również wyłącza neurony** (z tym samym prawdopodobieństwem) **w fazie predykcji**, dzięki czemu model staje się probabilistyczny – predykcja odpowiada próbkowaniu z pewnego rozkładu prawdopodobieństwa (opartym na mieszance gaussowskiej; *„approximates Bayesian inference in deep Gaussian processes“*).
- Na podstawie eksperymentów dla pojedynczego próbkowania (N = 1) wybrano następujące szczegóły implementacji:
  - $p = 0.6$,
  - Warstwę MC Dropout dodano przed „wąskim gardłem“ (*bottleneck*) sieci U-Net.

Probabilistyczny U-Net

- Probabilistyczny wariant U-Neta (zob.: S. Kohl et al., *A Probabilistic U-Net for Segmentation of Ambiguous Images*, 2019)
- Stanowi **kombinację klasycznego U-Neta z warunkowym wariacyjnym autoenkoderem** (CVAE, *conditional variational autoencoder*).
- Ma mozliwość zwracać dowolną liczbę wiarygodnych hipotez, odtwarzając możliwe warianty segmentacji oraz częstotliwości z jakimi wystepują.

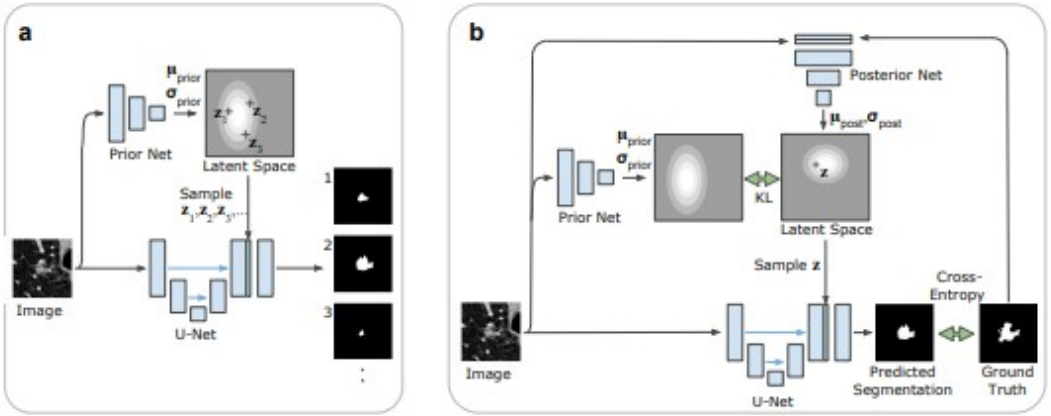


Figure 1: The Probabilistic U-Net. (**a**) Sampling process. Arrows: flow of operations; blue blocks: feature maps. The heatmap represents the probability distribution in the low-dimensional latent space R^N (e.g., $N = 6$ in our experiments). For each execution of the network, one sample $\mathbf{z} \in \mathrm{R}^N$ is drawn to predict one segmentation mask. Green block: N-channel feature map from broadcasting sample $\mathbf{z}$. The number of feature map blocks shown is reduced for clarity of presentation. (**b**) Training process illustrated for one training example. Green arrows: loss functions.

Oceny wariancji

$$g_P(X) = \hat{X}_i$$

$$f(\hat{X}_i) = [a, b, \theta, x_c, y_c]^T$$

- Przy N próbkach masek segmentacji (a następnie także elips oraz obwodów), możemy wyprowadzić i przebadać pewne **miary wariancji**, powiązane z niepewnością predykcji dla danego przykładu (obrazu).
- Zaproponowano cztery takie miary:
  - (1) **wariancja parametrów elipsy**: $\sum_i^5 (\mathrm{Var}(f(\hat{X}_n)_i))$
  - (2) **całkowite pole pierścienia**: $\sum (f(\bigcup_{i=1}^N \hat{X}_i) - f(\bigcap_{i=1}^N \hat{X}_i))$
  - (3) **entropia klasyfikacji maski**: $\sum_{x,y}^K \underline{\hat{X}}(x,y) \log(\underline{\hat{X}}(x,y))$
  - (4) **entropia pewności *Softmax***:

h4) Softmax confidence entropy: given $\dot{X}_i \in \mathbb{R}$ before class assignment, after conversion of the network's final layer's logits with $Softmax(x_i) = \frac{\exp(x_i)}{\sum^t \exp(x_t)}$, the resulting $\hat{X}_i^*$ can be interpreted as two-element prediction confidence $[p_f, p_b]_i = \hat{X}_i^*(x,y)$ for foreground p_f and background p_b. Thus we can estimate class-agnostic prediction entropy by $\sum_i^K p_i \log(p_i)$ where $p_i = \sum_i^N \max([p_f, p_b]_i)$.

- 3 -

WYNIKI EKSPERYMENTALNE

Zbiory danych

- **Zbiór A** – własny zbiór danych autorów artkułu:
  - **2724 zdjęcia** 2d z USG połówkowego (18-22 tydzień) w płaszczyźnie TV (w tym możliwe po kilka zdjęć z jednej sesji),
  - opisane ręcznie przez **45 ekspertów** (poprzez zaznaczenie kształtu głowy na zdjęciu).
- **Zbiór HC18** – publiczny, związany z konkursem w skuteczności pomiaru HC:
  - **1334 zdjęcia** j.w.,
  - opisane przez jednego wyszkolonego operatora USG.

- W obu przypadkach:
  - zdjęcia w rozmiarze **800 x 540 px**,
  - rozmiar jednego piksela: **0.052-0.326 mm**,
  - obrazy downsamplowane do rozmiaru **320 x 384 px**,
  - poddane losowo odbiciom lustrznym oraz niewielkim obrotom (w zakresie +/- 5 stopni).

Pojedyncze próbkowanie – wyniki

- Wyniki dla **N = 1** (a więc zachowanie porównywalne z klasycznym U-Netem dla klasyfikacji).
- Cel – porównanie skuteczności modelu z obecnie uznanym rozwiązaniem (*state-of-the-art*) bez oceny pewności, tj. Sinclair et al. (2018).

	Mean abs difference ± std (mm)	Mean DICE ± std (%)	Mean Hausdorff distance ± std (mm)
Baseline	2.09 ± 1.97	0.982 ± 0.011	1.289 ± 0.880
Dataset A + HC18	1.90 ± 1.90	0.982 ± 0.010	1.292 ± 0.791
Dropout $p = 0.6$	$\mathbf{1.808 \pm 1.65}$	$\mathbf{0.982 \pm 0.008}$	$\mathbf{1.295 \pm 0.664}$

- Wyniki:
  - **osiągnięto rezultaty porównywalne ze *state-of-the-art*** (trening na zbiorze danych A),
  - wyniki poprawia się, kiedy do zbioru treningowego dodamy HC18,
  - kolejną **poprawę** możemy zaobserwować **po zastosowaniu warstwy dropoutu** podczas treningu.

Wielokrotne próbkowanie – wyniki (1)

- Wyniki dla **N > 1** (w tabeli, dla **N = 10**).
- Ostateczna predykcja wartości HC, to średnia lub mediana wartości opartych na wszystkich N segmentacjach.
- Ograniczenie górne i dolne oparte jest na zakresie wyników dla N próbek.
- Cel – porównanie rezultatów dla różnych wartości N oraz skuteczności sieci w zależności od metody próbkowania (probabilistyczny U-Net vs MC dropout).

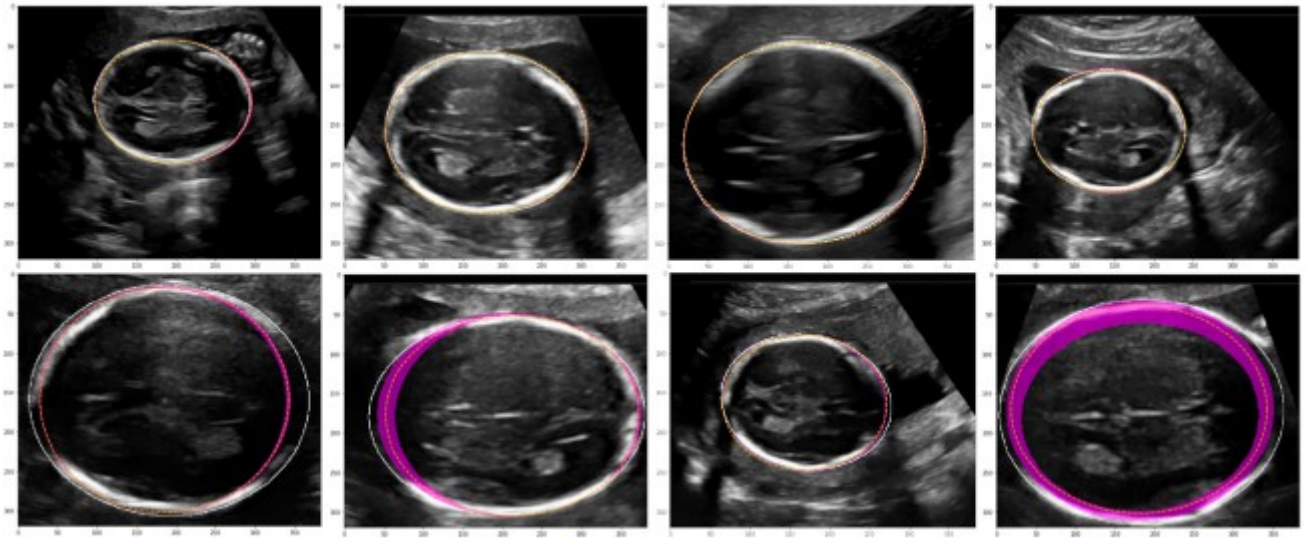
	Mean abs difference ± std (mm)	Mean DICE ± std (%)	Mean Hausdorff distance ± std (mm)	$LB \leq HC_{gt} \leq UB(\%)$
Det.				
MC $p = 0.6$	**1.81 ± 1.65**	**0.982 ± 0.008**	**1.295 ± 0.664**	N/A
Prob. UNet				
Mean	2.22 ± 2.15	0.980 ± 0.011	1.413 ± 0.751	20.4
Median	2.21 ± 2.15	0.980 ± 0.011	1.410 ± 0.748	20.4
MC(inf.)				
Mean	2.15 ± 2.09	0.981 ± 0.010	1.313 ± 0.613	27.8
Median	**2.15 ± 2.07**	**0.981 ± 0.010**	**1.307 ± 0.604**	**27.8**

Wielokrotne próbkowanie – wyniki (2)

	Mean abs difference ± std (mm)	Mean DICE ± std (%)	Mean Hausdorff distance ± std (mm)	$LB \leq HC_{gt} \leq UB(\%)$
Det.				
MC $p = 0.6$	**1.81 ± 1.65**	**0.982 ± 0.008**	**1.295 ± 0.664**	N/A
Prob. UNet				
Mean	2.22 ± 2.15	0.980 ± 0.011	1.413 ± 0.751	20.4
Median	2.21 ± 2.15	0.980 ± 0.011	1.410 ± 0.748	20.4
MC(inf.)				
Mean	2.15 ± 2.09	0.981 ± 0.010	1.313 ± 0.613	27.8
Median	**2.15 ± 2.07**	**0.981 ± 0.010**	**1.307 ± 0.604**	**27.8**

- Wyniki:
  - **słabsza efektywność przy próbkowaniu wielokrotnym, w porównaniu z pojedynczym** (prawdopodobnie wynika z braku zastosowania dropoutu przy inferencji dla pojedynczego próbkowania,
  - wielokrotne próbkowanie **pozwala za to otrzymać górne i dolne ograniczenie dla pomiarów** (średnia różnica: 1.82 +/- 1.78 mm),
  - ilość próbek mieszczących się między dolnym i górnym ograniczeniem silnie zależy od N (wykres na slajdzie 3-8).

Wielokrotne próbkowanie – wyniki (3)

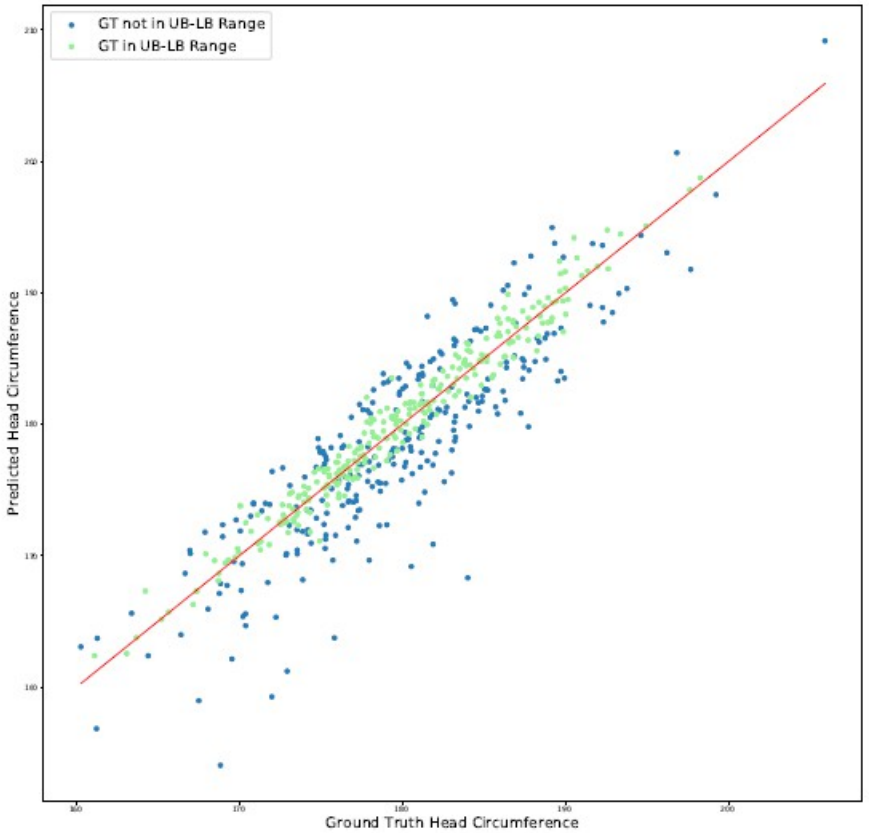


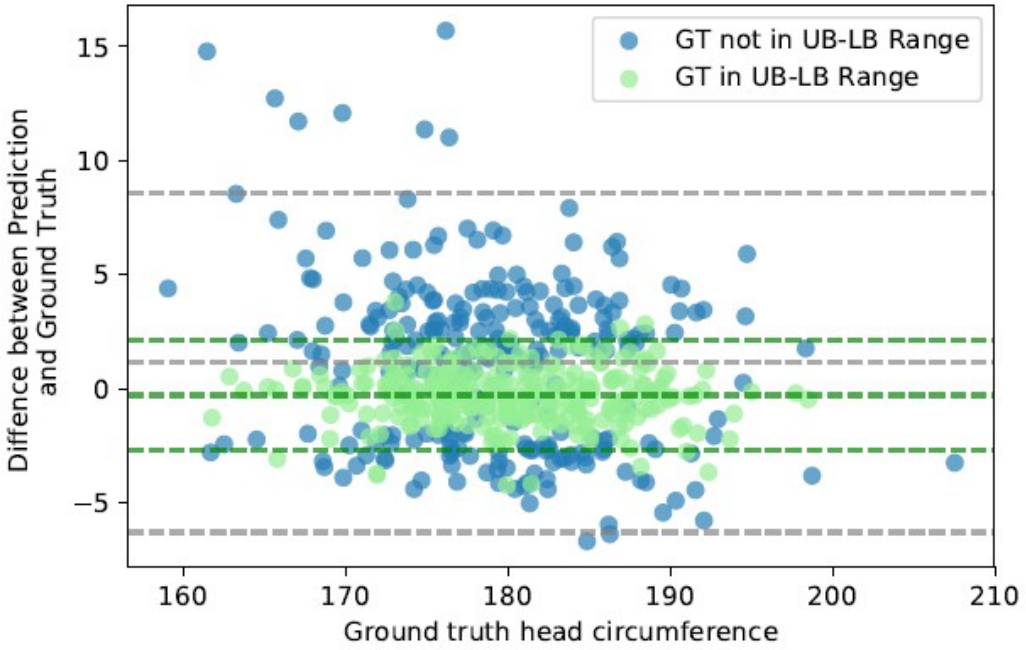
- Wyniki (c.d.):
  - **najlepiej oceniane** (u góry) przykłady dają bardzo wąski zakres między dolnym i górnym ograniczeniem (obie granice są na tych samych pikselach),
  - **najgorzej oceniane** (u dołu) – mają szeroki zakres, który rozmija się z elipsą *ground truth*,
  - Najlepsze – wyraźna biała granica czaszki; najgorsze – rozmyte brzegi wskazujące na słabą jakość lub suboptymalną płaszczyznę widoku.

Zastosowanie ocen wariancji

- Oceny wariancji zastosowano w następujący sposób:
  - Ustalono próg w zakresie od 0 do 1,
  - Jeżeli dane zdjęcie testowe osiągało daną ocenę (po znormalizowaniu) większą niż próg, zostało ono **usuwane ze zbioru testowego**.
- Skuteczność jej zastosowania oceniano poprzez zbadanie jak poprawił się rezultat modelu w stosunku do ilości odrzuconych zdjęć ze zbioru testowego.
- Wyniki:
  - **początkowo**, ustalenie progu powoduje odrzucenie pewnej ilości zdjęć i szybki **wzrost wydajności modelu**,
  - po poczatkowej, szybkiej poprawie, **późniejsze zmiany progu nie powodują istotnych zmian skuteczności**,
  - mechanizm ten nie pozwala odróżnić dobrych zdjęć od słabszych, ale pozwala odrzucić najgorsze, które autorzy utożsamiają z sytuacjami reprezentującymi nieoptymalną płaszczyznę widoku.
- Zaproponowane 4 oceny wariacji wykazują podobne zachowanie w tych okolicznościach.

Wyniki vs *ground truth*





Przykłady GT w ograniczeniach a N

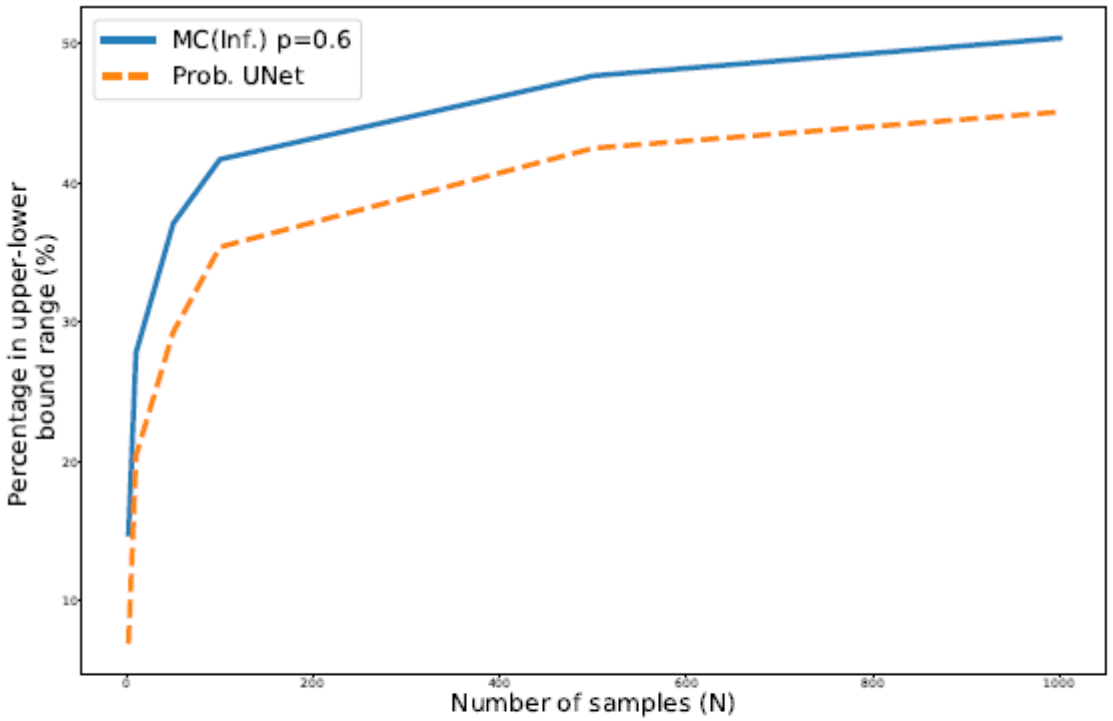
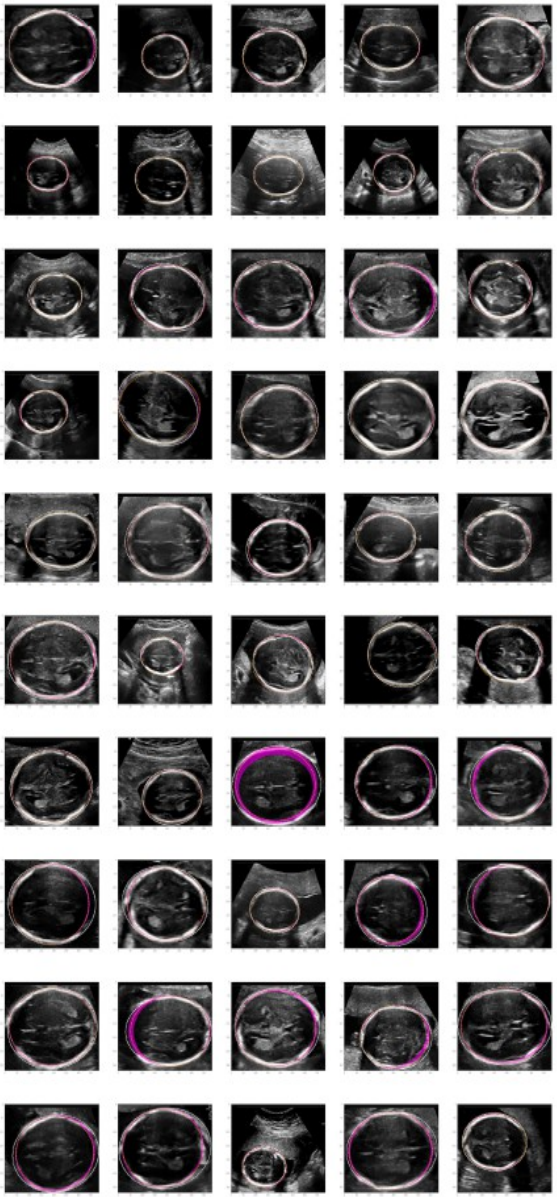
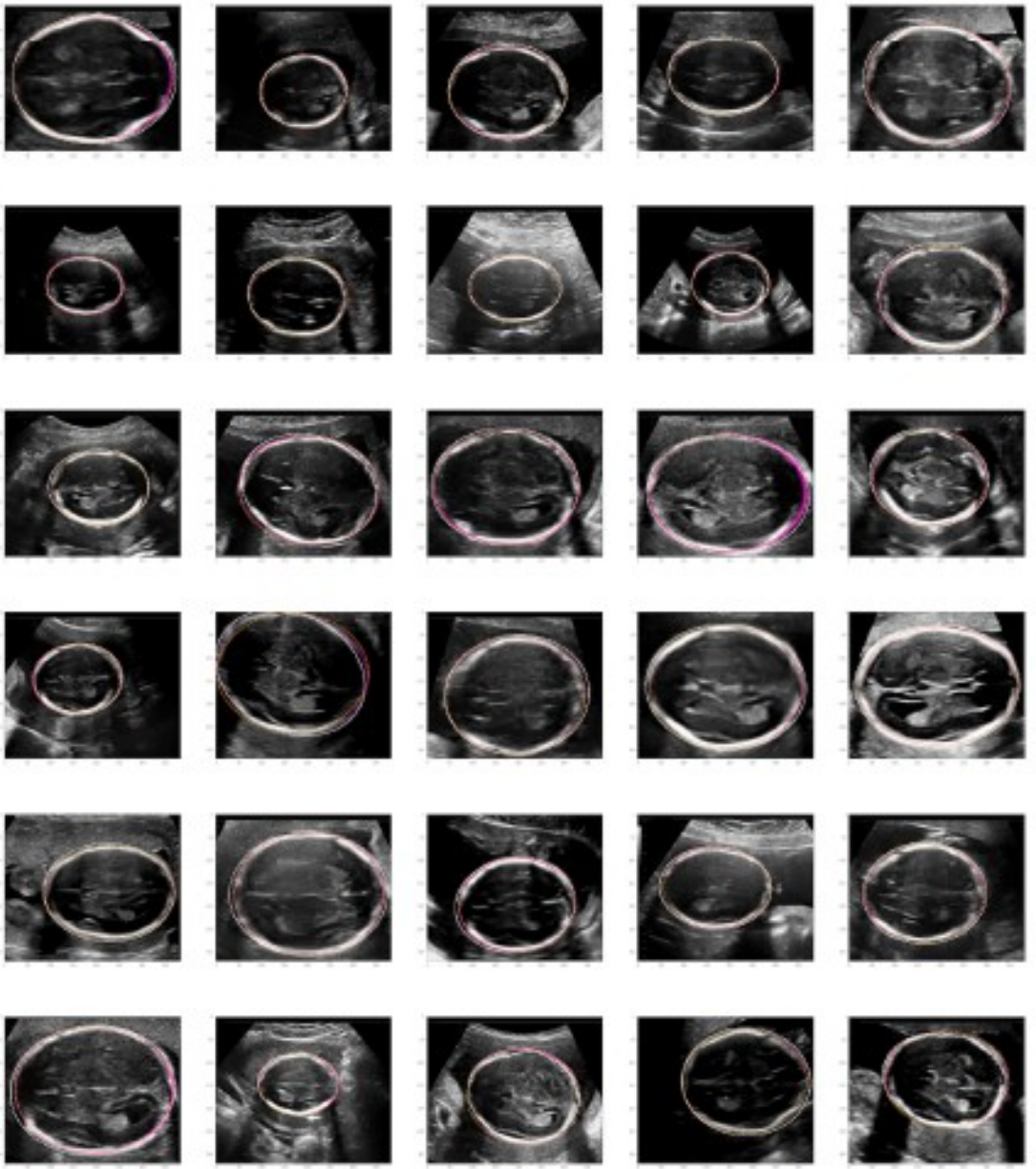


Fig. 4. Graph showing the effect of taking more samples from the network on the number of ground truth measurements lying between the generated upper and lower bounds. We can see a sharp increase in this percentage as samples increasing, plateauing around 50% for MC Dropout during inference.

Przykłady (mała pewność)





Przykłady (duża pewność)

- 4 -

WNIOSKI

Wnioski – podsumowanie

- Dla pojedynczego próbkowania – wyniki porównywalne ze *state-of-the-art*, dla wielokrotnego – mniejsza skuteczność, ale uzyskujemy użyteczne w ocenie pewności dolne i górne ograniczenia.
- Oceny wariancji:
  - zdolność odrzucenia zdjęć, które dają bardzo niepewne pomiary,
  - możliwość wykorzystania w systemie pomagającym w czasie rzeczywistym ocenić operatorowi USG, czy zdjęcie nadaje sie do pomiaru – zmniejszając zróżnicowanie między ocenami różnych operatorów,
- Uwaga: oceny wariancji nie reprezentują bezpośrednio niepewności modelu, ale pokazują pewność danej predykcji modelu względem jego możliwości i zadanych mu przykładów treningowych

Dalsze perspektywy badań

- Zmiany w modelu:
  - alternatywne metody wielokrotnego próbkowania (inne niż probabilistyczny U-Net i MC dropout),
  - alternatywne metody łączenia wielu próbek segmentacji,
  - alternatywne oceny wariancji.

- Dalsza ewaluacja modelu:
  - zbadanie działania modelu na innych zestawach danych (sprawdzenie, czy nie ma jakichś ukrytych *bias*ów),
  - zbadanie działanie modelu dla przypadków anomalnych, które są najistotniejsze do wykrycia z perspektywy diagnostycznej i profilaktycznej.

Dziękuję za uwagę!

[kontakt: **maciejmanna@gmail@com**]
[slajdy: **github.com/xann16/talks**/ml/hc-meas-conf/hc-meas-conf.pdf]