Practical Work 2: Fetal Head Circumference Measurement using Ultrasound Imaging (HC18)

Le Viet Hoang Lam
Student ID: 22BI13235
ICT Department, USTH

Contents

1 Introduction

1.1 Background

Fetal head circumference (HC) is an important biometric measurement used in prenatal ultrasound examinations to assess fetal growth and detect potential abnormalities. Manual HC measurement is time-consuming and subject to inter- and intra-observer variability. This practical work builds an automated regression model that predicts HC from 2D ultrasound images.

1.2 Objectives

The objectives of this practical work are:

- Explore and analyze the HC18 ultrasound dataset
- Extract HC labels from annotation images
- Develop a regression model to predict HC from ultrasound images
- Evaluate performance using Mean Absolute Error (MAE)
- Experiment with multiple hyperparameter configurations

2 Dataset Description

2.1 Dataset Overview

The HC18 (Head Circumference 18) dataset consists of fetal head ultrasound images with corresponding annotation images in the training set. In our pipeline, HC labels are extracted from the annotation images by estimating the contour perimeter (ellipse-like) in **pixel units**.

Dataset Statistics (from exploration):

- Training images: **806**
- Test images: **335**
- Training annotations: **806**
- Example image dimension: **540** $\times$ **800** pixels (grayscale, uint8)
- HC label range (pixels): **444.56** – **1791.88**
- Mean HC (pixels): **1273.89** $\pm$ **267.91**

2.2 Data Distribution

Figure 1 shows the distribution of HC values and example images/annotations.

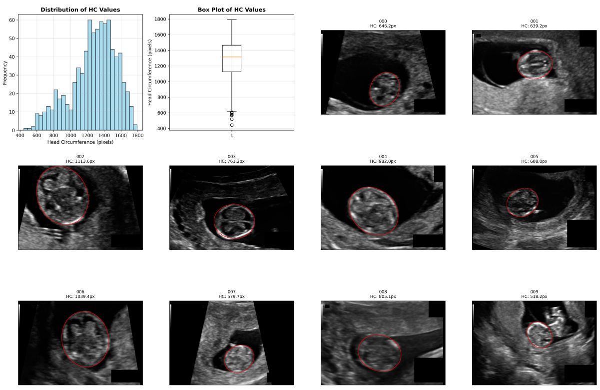

Figure 1: Dataset exploration: HC distribution and example training images with annotations.

2.3 Data Quality

- Missing values: **0**
- Potential outliers (IQR method): **10** samples

3 Methodology

3.1 Label Extraction from Annotations

The HC18 dataset provides training annotations as images. HC is computed from annotation masks by:

1. Loading the annotation image and thresholding to isolate the contour
2. Finding the largest contour (head boundary)

3. Estimating HC using the contour perimeter (and ellipse fit when possible)

This yields HC in **pixels**. Converting to **mm** requires pixel spacing metadata (mm/pixel), which is not used in the current notebook run.

3.2 Data Preprocessing

3.2.1 Train-Validation-Test Split

The dataset was split into:

- Training: **70%** (564 images)
- Validation: **15%** (121 images)
- Test: **15%** (121 images)

3.2.2 Image Preprocessing

All ultrasound images were preprocessed as follows:

1. Resize to **224** $\times$ **224**
2. Normalize pixel intensities to **[0, 1]**
3. Expand channel dimension to shape **(224, 224, 1)**

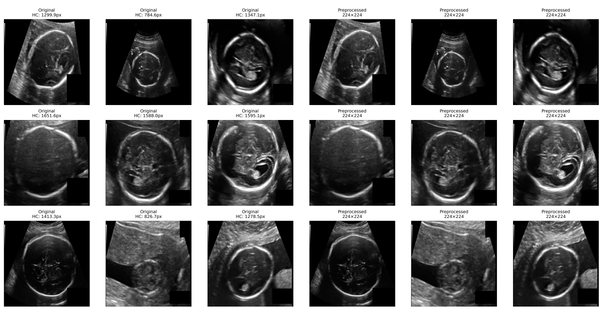

Figure 2: Preprocessing visualization: original images vs. resized/normalized images.

3.2.3 Label Normalization

For stable training, the labels were normalized using training-set statistics:

$$y_{\text{norm}} = \frac{y - \mu}{\sigma},$$

where $\mu = 1275.67$ and $\sigma = 265.04$ (pixels).

3.2.4 Data Augmentation

To reduce overfitting, data augmentation was applied:

- Rotation: $\pm 15°$
- Width/height shift: $\pm 10\%$
- Zoom: up to 15%
- Horizontal flip

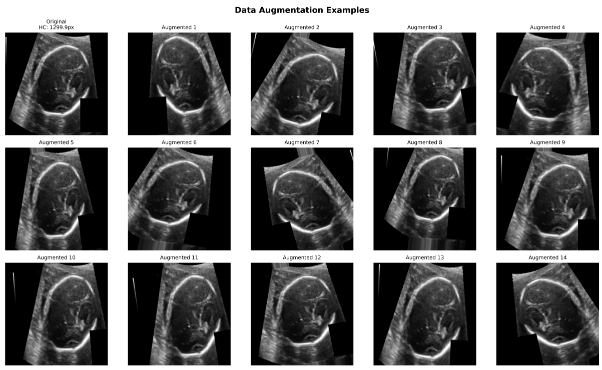

Figure 3: Examples of augmented training images.

3.3 Model Architecture

A CNN regression model was used with the following layers:

- Conv(32, 3×3) + MaxPool
- Conv(64, 3×3) + MaxPool
- Conv(128, 3×3) + MaxPool
- Flatten
- Dense(256) + Dropout(0.5)
- Dense(128) + Dropout(0.3)
- Dense(1) output

Loss: MAE
Optimizer: Adam (default learning rate)
Total parameters: 22,277,121

3.4 Training Strategy

3.4.1 Training Configuration

- Batch size: 32
- Max epochs: 100
- Callbacks:
  - ModelCheckpoint (best model by validation MAE)
  - EarlyStopping (patience = 15, restore best weights)
  - ReduceLROnPlateau (factor = 0.5, patience = 5)

3.4.2 Training Process

The model stopped early after reaching the best validation MAE at epoch **14**. Training curves are shown in Figure 4.

Figure 4: Training history: MAE and loss curves (and LR schedule if tracked).

Best validation MAE:

- Normalized scale: **0.3137**
- Denormalized (pixels): **83.1316 pixels**

4 Experiments

4.1 Hyperparameter Tuning

To satisfy the requirement of experimenting with multiple settings, we evaluated 10 configurations varying:

- Learning rate: {0.001, 0.0001, 0.00001}
- Batch size: {16, 32, 64}
- Dropout: {0.3, 0.5, 0.7}
- Optimizer: {Adam, SGD, RMSprop}
- Architecture: {Simple CNN, Deep CNN}

Note: These experiments used a shorter training schedule (up to 50 epochs, early stopping) and trained directly on **pixel labels** for speed; therefore, the experiment MAE values are reported in **pixels** and are not directly comparable to the final normalized-label training run.

4.2 Experimental Results

Table 1 reports the top 5 configurations (lowest validation MAE in pixels).

Table 1: Hyperparameter experiment results (Top 5 configurations, validation MAE in pixels).

ExpID	LR	Batch	Drop	Opt	Arch	Val MAE
1	0.0010	32	0.5	adam	simple_cnn	120.8008
4	0.0001	16	0.5	adam	simple_cnn	128.9553
8	0.0001	32	0.5	sgd	simple_cnn	129.5413
10	0.0001	32	0.5	adam	deep_cnn	133.0673
6	0.0001	32	0.3	adam	simple_cnn	134.3752

4.3 Effect of Hyperparameters

Figures 5 and 6 summarize the impact of the chosen hyperparameters.

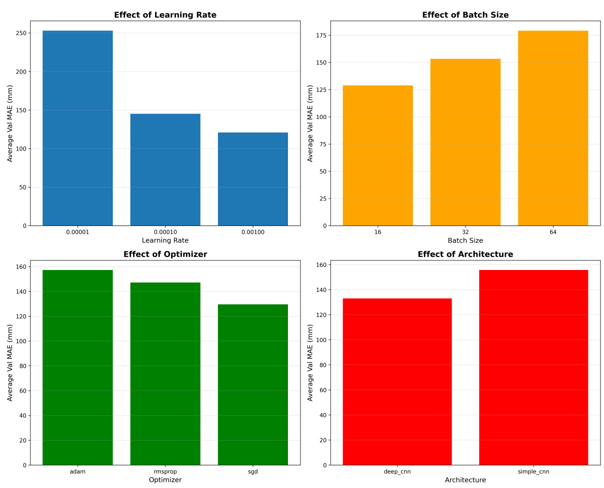

Figure 5: Effect of hyperparameters on validation MAE (experiments).

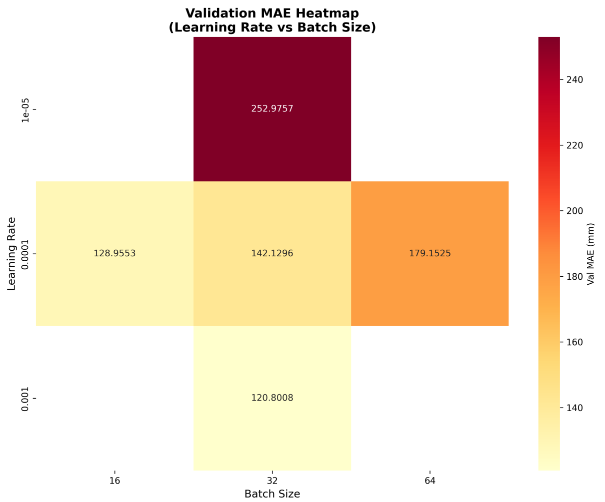

Figure 6: Heatmap visualization of hyperparameter experiments (experiments).

5 Results

5.1 Final Model Performance

Table 2 reports MAE/RMSE/MSE/R^2 on train/validation/test for the best saved model (denormalized to pixels).

Table 2: Final model performance metrics (pixels).

Dataset	MAE (px)	RMSE (px)	MSE	R^2
Training	36.2442	55.3751	3066.4031	0.9563
Validation	83.1316	131.3256	17246.4131	0.7763
Test	81.0071	130.2845	16974.0568	0.7670

5.2 Prediction Visualization

Figure 7 shows predicted vs. true HC and residuals on the test set.

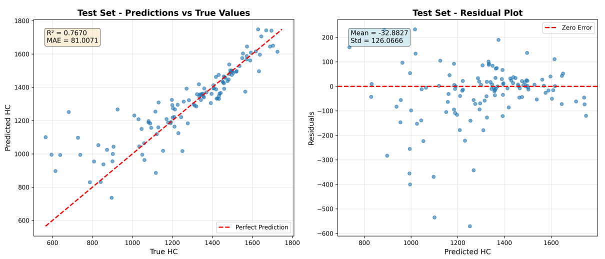

Figure 7: Test set predictions: predicted vs. true HC (pixels) and residual plot.

5.3 Error Analysis

The error distribution is shown in Figure 8. Summary statistics on the test set:

- Mean absolute error: **81.0071 px**
- Standard deviation of absolute error: **102.0388**
- Maximum error: **570.7864**
- Predictions with error > 5 px: **111** (**91.74%**)

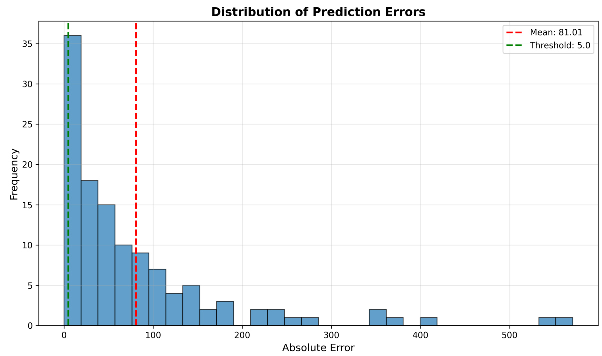

Figure 8: Distribution of absolute prediction errors on the test set (pixels).

6 Leaderboard

The HC18 challenge leaderboard typically reports error in **mm**. In this notebook run, HC labels were extracted and modeled in **pixels**. A reliable conversion to mm requires per-image pixel spacing (mm/pixel), which is not incorporated here. Therefore, a direct numerical comparison to the leaderboard is not provided in this report.

To enable leaderboard comparison in future work:

- Retrieve pixel spacing metadata (mm/pixel) for each image (if available).
- Convert labels and predictions from pixels to mm: HC(mm) = HC(px) $\times$ spacing(mm/px).

7 Discussion

7.1 Performance Interpretation

The model achieved a test MAE of **81.01 pixels**. The R^2 score of **0.7670** indicates a moderate-to-strong correlation between predictions and true HC values, but there is still significant error variability across difficult cases.

7.2 Strengths

- Fully automatic pipeline: label extraction $\rightarrow$ preprocessing $\rightarrow$ training $\rightarrow$ evaluation
- Simple CNN architecture with reasonable predictive performance
- Reproducible outputs saved as figures/tables for reporting

7.3 Limitations

- Model predicts HC from the entire image without explicitly localizing the fetal head
- Large errors occur on challenging images (noise, partial views, artifacts)
- Reported units are pixels; missing mm conversion limits clinical interpretability and leaderboard comparison

7.4 Future Improvements

- Add head localization/segmentation to guide regression (multi-task learning)
- Try transfer learning (e.g., ResNet/EfficientNet) adapted for grayscale inputs
- Use attention mechanisms to focus on the head boundary
- Incorporate metadata to convert HC to mm for proper leaderboard comparison

8 Conclusion

This practical work implemented an automated CNN regression system for fetal head circumference estimation from ultrasound images. Using extracted pixel-based HC labels, the final model achieved **81.01 px** MAE on the test split and demonstrated reasonable predictive capability. Hyperparameter experiments were conducted to evaluate training trade-offs and guide model selection.

References

[1] MICCAI 2018 Grand Challenge: HC18 - 2D Fetal Head Circumference.