Week 2: Data Exploration, Baselines, and Calibration Refinements

1 Setup and Notation (Recap)

We keep the notation from Week 1: each input image is $x \in \mathcal{X}$, the binary label is $y \in \{0, 1\}$, and a neural network with parameters θ produces logits

$$\boldsymbol{z}(x;\theta) = \big(z_0(x;\theta), z_1(x;\theta)\big) \in \mathbb{R}^2,$$

which are converted to class probabilities $p_\theta(y = k \mid x)$ by a softmax layer. The predicted label is $\hat{y} = \arg\max_k p_\theta(y = k \mid x)$, and the confidence is $c(x) = \max_k p_\theta(y = k \mid x)$. Training uses cross-entropy, and evaluation uses accuracy, AUROC, and calibration metrics such as ECE.[1]

In Week 2 we focus on:

- understanding basic *dataset statistics* (class balance, splits),
- comparing simple *baseline models*,
- and making a more careful first pass at *calibration analysis* (binning choices and high-confidence errors).

2 Dataset Statistics and Class Balance

Let the dataset be split into three parts:

$$\mathcal{D}_{\text{train}}, \quad \mathcal{D}_{\text{val}}, \quad \mathcal{D}_{\text{test}},$$

with sizes $n_{\text{train}}, n_{\text{val}}, n_{\text{test}}$, respectively. For a given split $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, define the empirical class counts

$$n_y \;=\; \sum_{i=1}^n \mathbf{1}\{y_i = y\}, \qquad y \in \{0, 1\},$$

and the empirical class proportions

$$\hat{\pi}_y \;=\; \frac{n_y}{n}, \qquad \hat{\pi}_0 + \hat{\pi}_1 = 1.$$

These $\hat{\pi}_y$ are simple estimates of the *class priors* $\Pr(y = 0)$ and $\Pr(y = 1)$ for that split.

Class imbalance. If $\hat{\pi}_1$ is very small (rare positives) or very large (rare negatives), the dataset is *imbalanced.* In such cases:

- Accuracy can be misleading (predicting the majority class yields high accuracy but poor detection of the minority class).
- AUROC becomes especially important, because it reflects the ranking quality across classes regardless of any single threshold.
- Per-class metrics (e.g., recall for the positive class) and the confusion matrix are helpful to understand failure modes.

Comparing the empirical class proportions between $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{val}}$, and $\mathcal{D}_{\text{test}}$ also provides a simple check for *distribution shift* between splits.

[1] See the Week 1 notes for precise definitions of cross-entropy, AUROC, ECE, and temperature scaling.

3 Baseline Models and Hypothesis Classes

A classifier is a function $f_\theta : \mathcal{X} \to [0,1]^2$ mapping an input x to class probabilities. Different neural network architectures correspond to different *hypothesis classes*:

$$\mathcal{F}_{\text{smallcnn}} = \{f_\theta \text{ given by a small CNN}\}, \quad \mathcal{F}_{\text{resnet}} = \{f_\theta \text{ given by a ResNet-18}\}, \ \ldots$$

In Week 2 we compare two baselines:

1. **Small CNN.** A relatively low-capacity convolutional network trained from scratch on the target dataset. It is fast and serves as a sanity check that the task is learnable.
2. **ResNet-18 with frozen backbone and learned head.** Let $g_\phi : \mathcal{X} \to \mathbb{R}^d$ be a feature extractor (the ResNet-18 backbone, pre-trained on a large dataset such as ImageNet), and let $h_W : \mathbb{R}^d \to [0,1]^2$ be a linear classifier:

   $$h_W(\boldsymbol{v}) = \operatorname{softmax}(W\boldsymbol{v}), \qquad W \in \mathbb{R}^{2\times d}.$$

   The overall model is

   $$f_{\phi,W}(x) = h_W\big(g_\phi(x)\big).$$

   In *head-only fine-tuning* we fix ϕ and update only W; this is equivalent to training a logistic regression classifier in the learned feature space $g_\phi(x)$.

Comparing these baselines helps to answer questions such as:

- Does a small task-specific model already perform well?
- Does transfer learning (pre-trained features) significantly improve AUROC or calibration?
- Is there evidence of overfitting or underfitting in either model?

4 Confusion Matrix and Error Types

For a fixed threshold decision rule (e.g. argmax), predictions on the test set can be summarized by a *confusion matrix*. In the binary case we define:

$$\text{TP} = \sum_{j=1}^{m} \mathbf{1}\{\hat{y}_j = 1,\, y_j = 1\}, \qquad \text{FP} = \sum_{j=1}^{m} \mathbf{1}\{\hat{y}_j = 1,\, y_j = 0\},$$

$$\text{TN} = \sum_{j=1}^{m} \mathbf{1}\{\hat{y}_j = 0,\, y_j = 0\}, \qquad \text{FN} = \sum_{j=1}^{m} \mathbf{1}\{\hat{y}_j = 0,\, y_j = 1\}.$$

From these we can form:

$$\begin{aligned}\text{Sensitivity / Recall} &= \frac{\text{TP}}{\text{TP}+\text{FN}},\\ \text{Specificity} &= \frac{\text{TN}}{\text{TN}+\text{FP}},\\ \text{Precision (Positive Predictive Value)} &= \frac{\text{TP}}{\text{TP}+\text{FP}}.\end{aligned}$$

While AUROC summarizes ranking quality over all thresholds, the confusion matrix reveals *which type of error* dominates (missed positives vs. false alarms), which is crucial in medical settings.

5 Refined Calibration: Binning Strategies

In Week 1, calibration and the Expected Calibration Error (ECE) were defined using B fixed, equal-width bins on $[0,1]$.[2] Here we make this more explicit and introduce an alternative *equal-frequency* binning strategy.

Let $\{c_j\}_{j=1}^m$ be the confidences on the test set, and let $b(j) \in \{1,\ldots,B\}$ denote the bin index for example j.

[2] See Section 5 of the Week 1 notes for the original definitions of $\text{conf}(b)$, $\text{acc}(b)$ and ECE.

5.1 Equal-Width Binning

In equal-width binning, we split $[0,1]$ into intervals

$$I_b = \left[\frac{b-1}{B}, \frac{b}{B}\right), \qquad b = 1, \dots, B-1, \quad \text{and} \quad I_B = \left[\frac{B-1}{B}, 1\right],$$

and assign

$$b(j) = b \quad \text{if} \quad c_j \in I_b.$$

For each bin b, define:

$$\text{conf}(b) = \frac{1}{|b|} \sum_{j:b(j)=b} c_j, \qquad \text{acc}(b) = \frac{1}{|b|} \sum_{j:b(j)=b} \mathbf{1}\{\hat{y}_j = y_j\},$$

with $|b|$ the number of points in bin b. The ECE is the weighted average

$$\text{ECE} = \sum_{b=1}^{B} \frac{|b|}{m} \left|\text{acc}(b) - \text{conf}(b)\right|.$$

Equal-width bins are simple, but in regions where few samples fall (e.g. very high confidences) some bins may have very small $|b|$, making $\text{acc}(b)$ noisy.

5.2 Equal-Frequency (Quantile) Binning

In equal-frequency binning, we choose bin boundaries so that each bin contains approximately the same number of samples. Let $q_0 = 0$ and $q_B = 1$, and choose quantiles

$$0 = q_0 < q_1 < \cdots < q_B = 1$$

such that roughly $\frac{m}{B}$ confidences lie between q_{b-1} and q_b. We then define

$$I_b = [q_{b-1}, q_b), \qquad b = 1, \dots, B-1, \quad \text{and} \quad I_B = [q_{B-1}, q_B].$$

As before, we assign $b(j)$ based on which interval I_b contains c_j and compute $\text{conf}(b)$, $\text{acc}(b)$, and ECE in exactly the same way.

Equal-frequency bins reduce the chance of empty or extremely small bins, so the reliability diagram and ECE estimate are often more stable in regions with many points (e.g. high-confidence predictions). The trade-off is that bin widths in confidence space are no longer uniform, so the horizontal axis of the reliability diagram is slightly harder to interpret.

Reporting ECE. In practice, it is useful to:

- show the reliability curve (accuracy vs. confidence per bin),
- overlay the histogram (or bar plot) of $\frac{|b|}{m}$,
- and report the final ECE value in the title or caption (e.g. with three decimal places).

This makes it easy to compare different models or calibration schemes.

6 High-Confidence Errors

Calibration analysis becomes especially important for *high-confidence errors*. Define the error set

$$E = \{j \in \{1, \dots, m\} : \hat{y}_j \neq y_j\}.$$

For a fixed confidence threshold $\tau \in (0,1)$ (e.g. $\tau = 0.9$), we define the high-confidence error set

$$E_{\text{high}}(\tau) = \{\, j \in E : c_j \geq \tau \,\}.$$

Each $j \in E_{\text{high}}(\tau)$ corresponds to a test example where the model is both *confident* and *wrong*. Inspecting the images $\{x_j : j \in E_{\text{high}}(\tau)\}$ helps answer questions such as:

- Are these mistakes visually ambiguous even for humans?
- Do they contain artifacts (e.g. text overlays, unusual cropping, severe noise)?
- Are they concentrated in a particular subclass or acquisition pattern?

Such qualitative analysis complements quantitative metrics and guides model improvements.

7 Optional: Post-hoc Temperature Scaling

As in Week 1, we may apply a single scalar *temperature* $T > 0$ to the logits to improve calibration without changing the ranking:

$$\tilde{\boldsymbol{z}}(x;\theta,T) = \frac{\boldsymbol{z}(x;\theta)}{T}, \qquad \tilde{p}_\theta(y = k \mid x; T) = \frac{\exp\big(\tilde{z}_k(x;\theta,T)\big)}{\sum_\ell \exp\big(\tilde{z}_\ell(x;\theta,T)\big)}.$$

The temperature $T^\star$ is chosen by minimizing validation NLL. Empirically, this often reduces ECE (especially for models that are over-confident) while leaving AUROC nearly unchanged, since dividing logits by a positive constant does not change their order.

In Week 2, temperature scaling can be used as a simple, optional experiment to compare:

- ECE before vs. after calibration,
- and to verify that AUROC (discrimination) remains effectively the same.

8 Summary

Week 2 connects three pieces:

- *Dataset understanding* through class counts and basic EDA.
- *Model comparison* via simple baselines (small CNN vs. transfer learning) using accuracy and AUROC.
- *Calibration refinement* using more careful binning strategies and an explicit focus on high-confidence errors.

These ingredients provide a more complete picture of how reliable a model is, beyond a single scalar metric, and prepare the ground for the later self-supervised learning experiments.