# Machine Learning Triage for Handwritten Digits

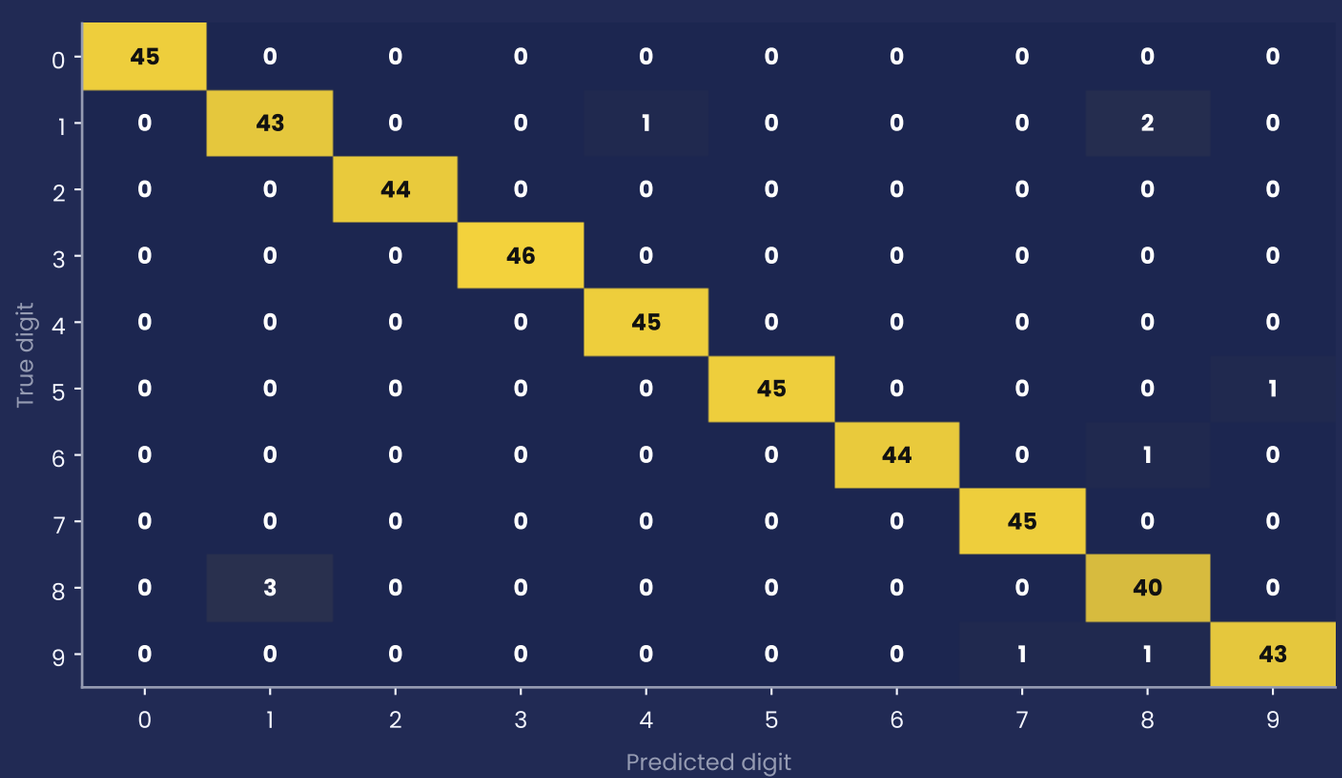
## 94 / 450

held-out digit images sent to a person at a 0.95 confidence cutoff. All 356 predictions accepted automatically were correct.

| 97.8% | 356 of 356 | 20.9% | 0.94 |
|---|---|---|---|
| accuracy before any triage (440 of 450) | correct among auto-accepted at 0.95 | of images referred at 0.95 | lowest cutoff with zero accepted errors |

EXHIBIT 1

### Confusion matrix on 450 held-out images



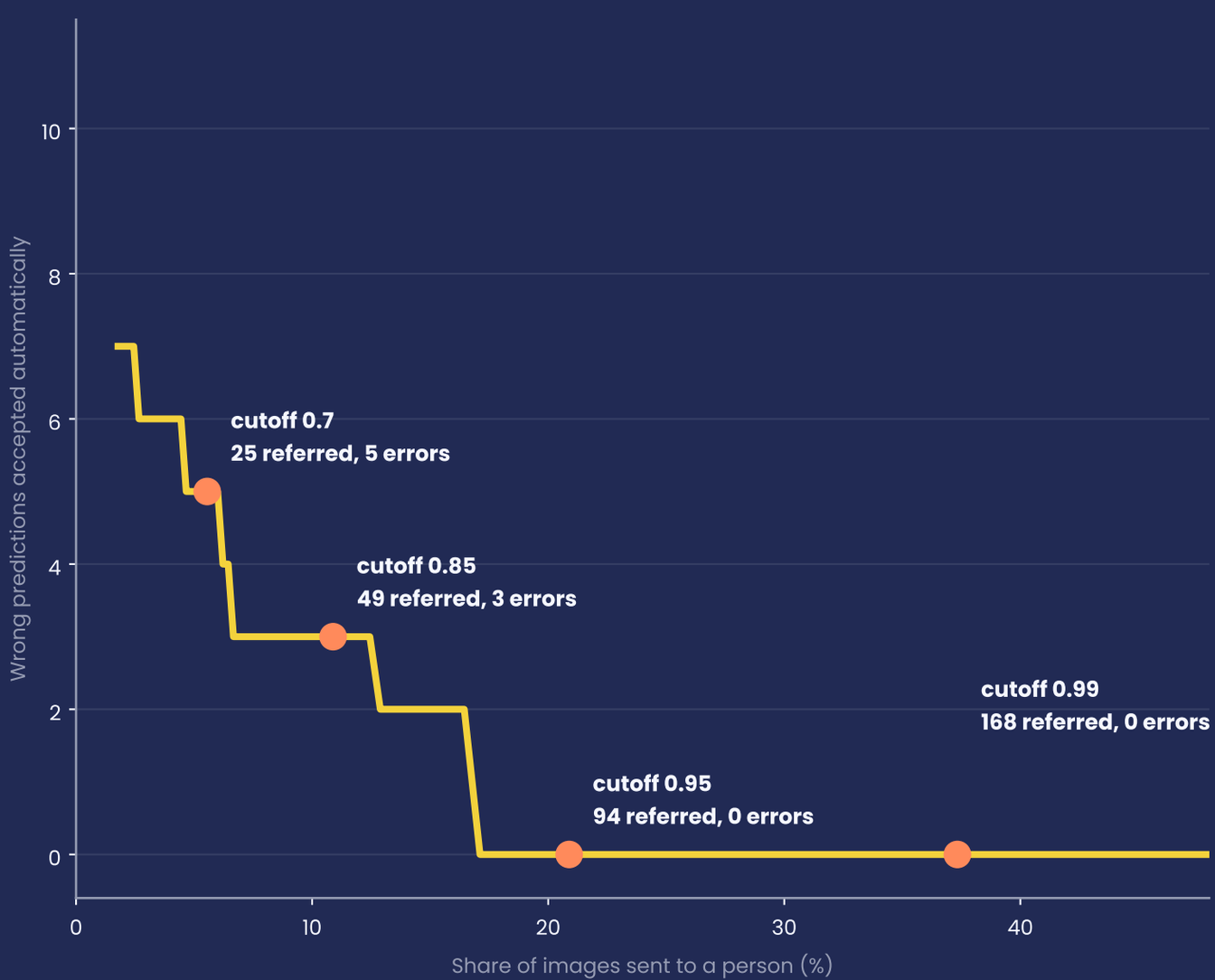

440 on the diagonal. The ten misreads are scattered across pairs of similar shapes.

# What the analysis shows

- The classifier reads 440 of 450 held-out digits correctly on its own. All ten misreads fall below 0.95 confidence, so a 0.95 cutoff sends every one of them to a person.
- That safety costs a queue of 94 images (20.9%). The lowest cutoff that still accepts no errors is about 0.94, which would refer 77 images instead.
- At 0.85 the queue halves to 49 but three wrong predictions are accepted; at 0.70 it is 25 with five accepted errors.
- Review load is uneven by digit: 8s are referred most often (49%) and 0s least (9%).
- Confidence is informative but not perfectly calibrated in the middle bins, where counts are small; it should be treated as a ranking score.

EXHIBIT 2

## Queue size against accepted errors

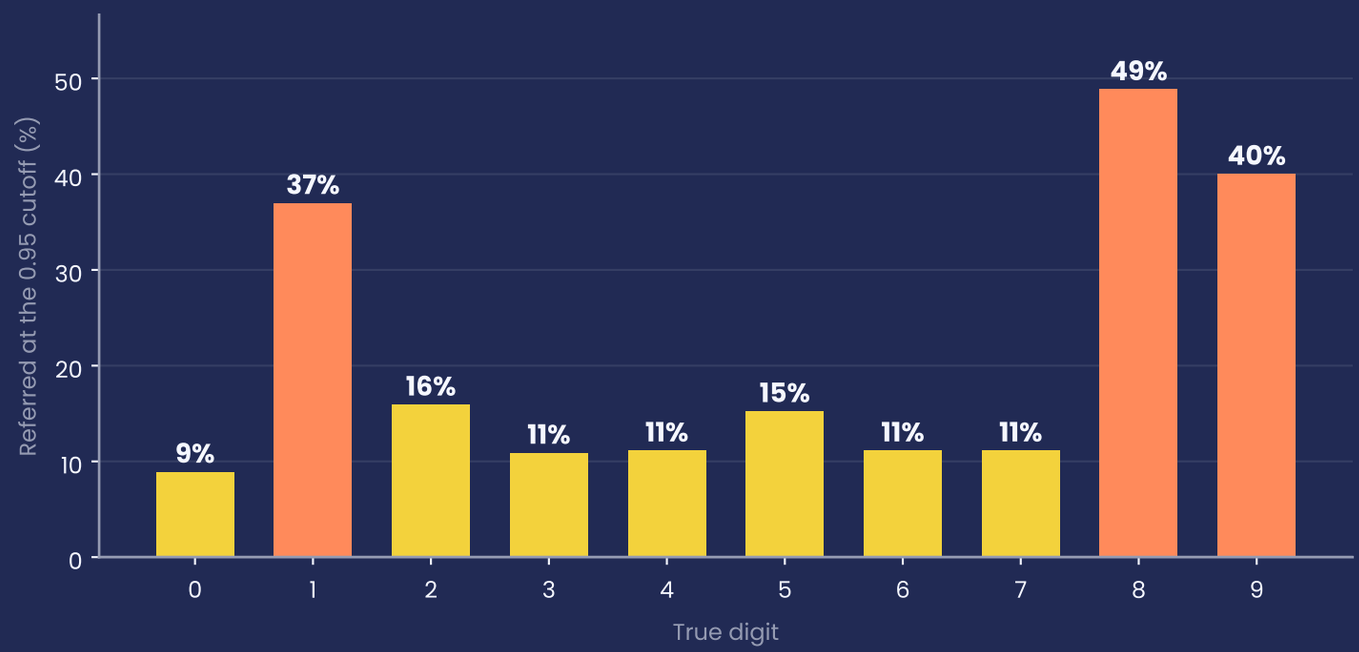



Each point on the curve is a cutoff. Moving right sends more images to people and lets fewer errors through.

# Inside the numbers

EXHIBIT 3

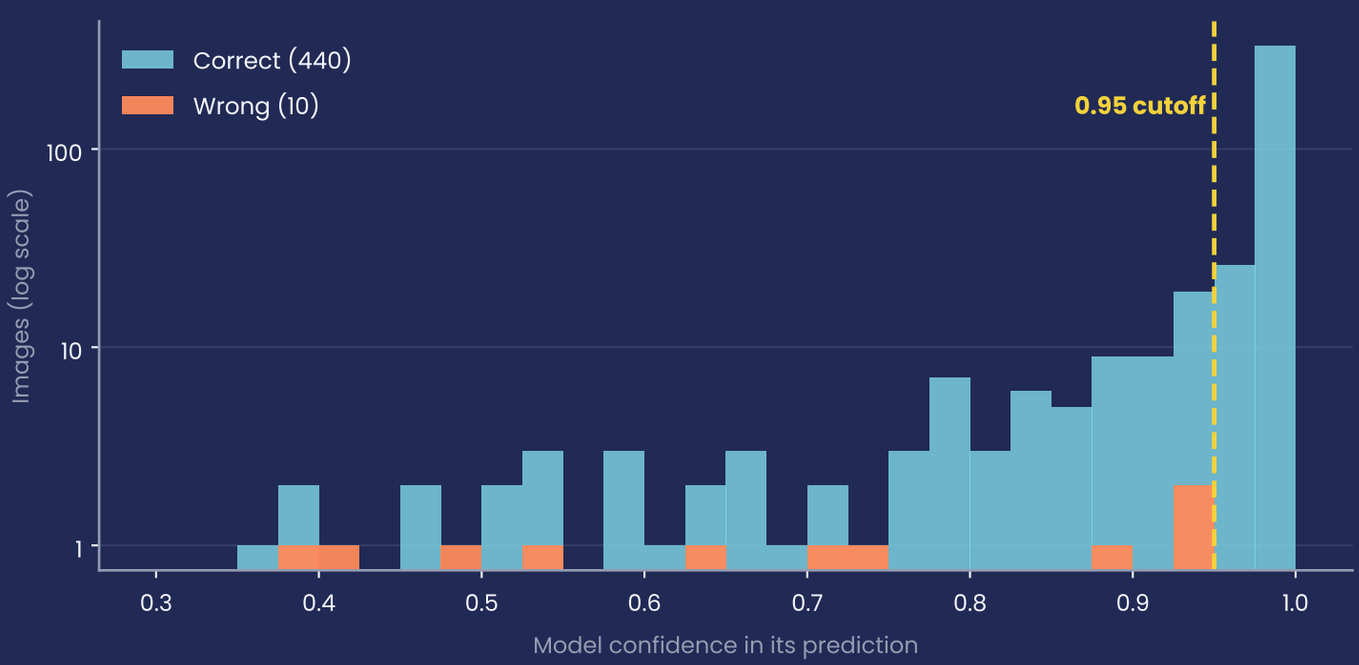
## Who ends up in the queue




Share of each true digit referred at 0.95.

EXHIBIT 4

## Confidence of correct and wrong predictions

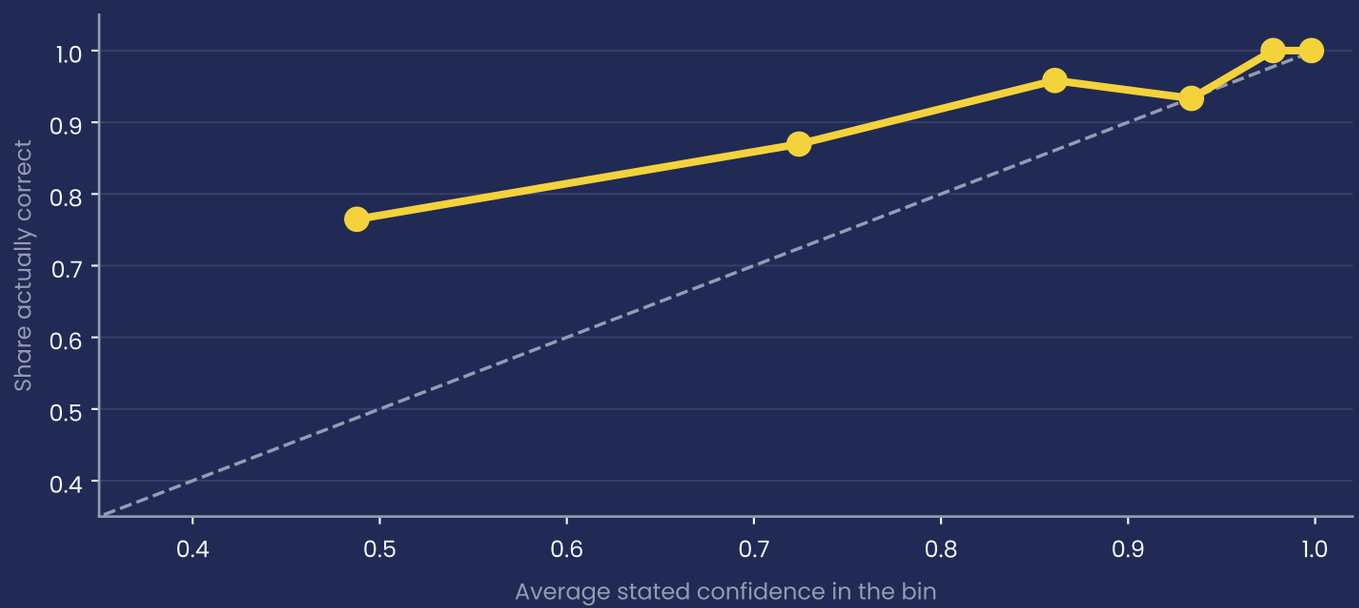



Every wrong prediction sits left of the 0.95 line; most correct ones sit at the far right.

# What changes the answer

EXHIBIT 5

## Does stated confidence match accuracy?




Points on the dashed line would be perfectly calibrated. High-confidence bins are reliable; sparse middle bins are noisy. Sample sizes, lowest to highest confidence: 17, 23, 24, 30, 74 and 282.

EXHIBIT 6

## The ten images the model misread



**true 1, read 8**
**confidence 0.54**



**true 8, read 1**
**confidence 0.71**



**true 9, read 7**
**confidence 0.40**



**true 6, read 8**
**confidence 0.75**



**true 8, read 1**
**confidence 0.65**



**true 9, read 8**
**confidence 0.42**



**true 5, read 9**
**confidence 0.89**



**true 8, read 1**
**confidence 0.93**



**true 1, read 4**
**confidence 0.94**



**true 1, read 8**
**confidence 0.48**

Each would be caught by the 0.95 cutoff. Several are ambiguous to a human reader too.

# How it was built and how far it goes

### RECOMMENDED USE

- Start at 0.95: no accepted errors on this holdout, for a one-in-five review load.
- If the queue is too long, test cutoffs between 0.94 and 0.95 on a second holdout before lowering it.
- Monitor referral rate by digit; a jump signals drift in the incoming images.

### METHOD

- UCI optical recognition digits (1,797 8×8 images); stratified 75/25 split with seed 42, giving 450 held-out images.
- Standard scaling fitted on the training set, then multinomial logistic regression.
- Confidence is the highest predicted class probability. The model was re-run for this report and reproduces the project's holdout counts exactly.

### LIMITS

- One split of a small, clean benchmark; zero accepted errors in 356 does not guarantee zero in production.
- Real handwriting varies more than this dataset.
- Human reviewers are assumed to be correct.

### NUMBERS RECONCILED

| Quantity | Value | How it was checked |
|---|---|---|
| **Held-out images** | 450 | 25% stratified split, seed 42 |
| **Correct** | 440 | re-run; matches project file |
| **Referred at 0.95** | 94 | re-run; matches holdout table |
| **Accepted at 0.95** | 356, all correct | re-run; matches holdout table |
| **Referred at 0.85** | 49 (3 errors accepted) | re-run; 398 of 401 correct |
| **Referred at 0.70** | 25 (5 errors accepted) | re-run; 420 of 425 correct |