Model Weaknesses and Error Analysis

After training the model for 15 epochs and evaluating it using a classification report and confusion matrix, we analyzed its performance across all Roman numeral classes. While the model achieved a solid overall accuracy of approximately 84%, there were some noticeable weaknesses and error patterns worth highlighting:

1. Underperforming Numerals

Certain Roman numerals consistently showed lower performance:

- **VI (6):** The model often confused “vi” with “iv” and “ix”, possibly due to their structural similarity. It achieved a recall of only 43%, meaning many “vi” samples were misclassified.
- **IV (4):** This numeral had a precision of 72% and recall of 57%, indicating both false positives and false negatives.
- **VII (7):** Many “vii” samples were misclassified as “iii”, which makes sense considering the visual overlap of repeating “i” characters.
- **IX (9):** Although it had a high recall (95%), its precision was lower (62%), meaning many samples from other classes were incorrectly predicted as “ix”.

These numerals are visually similar due to shared characters like “i” and “v”, and their handwritten variations can be difficult to distinguish.

2. Confusion Matrix Observations

By analyzing the confusion matrix:

- Many “vi” samples were incorrectly predicted as “iv” (17 times), “ix” (26 times), or “vii” (7 times).
- “Vii” was frequently misclassified as “iii” (28 times).
- “Iv” was confused with “ix” (30 times), as well as “vi” and “vii”.

These misclassifications indicate the model’s struggle with numerals that have overlapping visual structures.

3. Why These Errors Happen

- Roman numerals inherently share common sub-symbols (e.g., “i”, “v”), and handwritten variations amplify this ambiguity.
- Some classes may be underrepresented in the dataset or harder to recognize due to handwriting inconsistency.
- The model tends to perform better on clearer or more distinct numerals like “I”, “V”, “X” and struggles with mid-range numerals (“IV”, “VI”, “VII”, “IX”).

4. What We Can Do to Improve It

To address these weaknesses, we can:

- Add more diverse training samples, especially for confusing classes like “vi”, “iv”, “vii”, and “ix”.
- Apply stronger data augmentation (rotations, noise, stretching) to help the model generalize better.
- Fine-tune more layers of the MobileNetV2 base model (currently frozen) so it can adapt better to our handwriting dataset.
- Use a class-weighted loss or focal loss to emphasize the hard-to-classify numerals during training.

	precision	recall	f1-score	support
i	0.98	0.94	0.96	100
ii	0.87	0.95	0.91	99
iii	0.73	0.98	0.83	100
iv	0.72	0.57	0.64	98
ix	0.62	0.95	0.75	99
v	0.99	0.94	0.96	100
vi	0.87	0.43	0.57	96
vii	0.83	0.64	0.72	98
viii	0.97	0.97	0.97	95
x	0.95	0.98	0.97	100

```
              accuracy                           0.84       985
             macro avg       0.85      0.83      0.83       985
          weighted avg       0.85      0.84      0.83       985
```

```
Confusion Matrix:
 [[94  2  0  0  0  1  0  0  0  3]
 [ 0 94  5  0  0  0  0  0  0  0]
 [ 0  2 98  0  0  0  0  0  0  0]
 [ 0  2  1 56 30  0  5  4  0  0]
 [ 0  1  0  2 94  0  0  2  0  0]
 [ 2  0  1  1  0 94  0  0  0  2]
 [ 0  5  0 17 26  0 41  7  0  0]
 [ 0  2 28  1  0  0  1 63  3  0]
 [ 0  0  2  1  0  0  0  0 92  0]
 [ 0  0  0  0  2  0  0  0  0 98]]
```