# 1 Some Key Insight

This study compares **Supervised Learning**, **High-Confidence Pseudo-Labeling**, and **FixMatch** across three models of increasing complexity: **CNN**, **MobileNetV2**, and **EfficientNetB0**. The evaluation includes **accuracy**, **F1 score**, and **recall**, giving a well-rounded view of model generalization, particularly under class imbalance and limited labeled data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Supervised Accuracy | + Pseudo-Labeling Gain | + FixMatch Gain | Final Verdict |
| CNN (Baseline) | 🚫 Low Accuracy & Recall | ⚠️ Minimal F1/Recall gain | 🔻 Performance degradation | ❌ Not viable |
| MobileNetV2 | ✅ Decent across metrics | 👍 Moderate recall/F1 gains | ⚠️ Small extra improvement | ⚠️ Acceptable baseline |
| EfficientNetB0 | 🌟 High Accuracy + F1/Recall | 🚀 Significant F1/Recall boost | ⚠️ Slight gain (high cost) | ✅ Best all-around performer |

**CNN (Baseline): Not Fit for Semi-Supervised Learning**

* CNN underperforms across all metrics—**accuracy, F1 score, and recall**—due to its shallow architecture.
* While **pseudo-labeling** provides more training data, the model’s predictions are too noisy to produce meaningful pseudo-labels.
* **FixMatch makes this worse**, despite theoretical expectations. Why?
  + FixMatch relies on a **confidence threshold (τ)**—only predictions with softmax confidence > τ (e.g., 0.95) are used as pseudo-labels.
  + CNN rarely produces such confident predictions, especially under **strong augmentations**, meaning most pseudo-labels are discarded.
  + Even accepted pseudo-labels may be wrong, leading to **reinforcement of errors** and poor generalization.

🧠 *Mathematical intuition*: FixMatch uses a **binary mask** 𝟙(max(p) > τ) to select pseudo-labels. For CNN, the maximum softmax output max(p) is rarely > 0.95, so 𝟙 = 0 most of the time → the model trains almost entirely on labeled data again.

🔍 *Result*: FixMatch fails with CNNs because the **confidence gating mechanism starves the model of additional data**. It’s like a filter that never opens—making the entire semi-supervised strategy ineffective.

**MobileNetV2: Moderate Learner with Plateauing FixMatch Benefits**

* **Performs reasonably well under supervision, with decent F1 and recall.**
* **Pseudo-labeling provides meaningful improvements, particularly for minority classes, as the model can generate moderately confident labels.**
* **FixMatch adds little extra value, likely due to:**
  + **Inconsistent confidence under augmentation.**
  + **Overly strict thresholds that reject usable pseudo-labels.**
  + **Increased training time without proportional returns.**

**🔍 Takeaway: MobileNetV2 benefits from semi-supervised training, but FixMatch’s stringent pseudo-label selection and heavy augmentation pipeline provide diminishing returns unless confidence calibration is improved or thresholds are adapted.**

**EfficientNetB0: Confident Predictions Power Semi-Supervised Success**

* Delivers the strongest performance in accuracy, F1, and recall under supervised training.
* High-confidence pseudo-labeling yields a major boost, as EfficientNet generates reliable, class-balanced pseudo-labels.
* FixMatch provides only a small additional improvement, despite high theoretical expectations.
  + Most of the pseudo-labels used by FixMatch overlap with those already selected during high-confidence pseudo-labeling.
  + Strong augmentations and stricter filtering bring significant computational overhead, but little extra gain under current conditions.

Mechanism insight: EfficientNet often satisfies max(p) > τ, so FixMatch ends up selecting nearly the same unlabeled examples as pseudo-labeling.

Result: Under the current setup—with limited fine-tuning, frozen layers, and default confidence thresholds—FixMatch’s theoretical strengths appear redundant.  
However, with more extensive fine-tuning, adjusted thresholds, or longer training, it’s possible that FixMatch could still unlock additional gains beyond what was observed here.

**In Summary**

✅ **Higher-performing models tend to benefit more from semi-supervised learning—particularly high-confidence pseudo-labeling**, which enables them to leverage their own confident predictions to extract meaningful signal from unlabeled data.

📈 **FixMatch appears most effective when models can produce confident predictions and maintain consistency under strong augmentations.** In this study, its added value was modest—possibly due to constraints such as frozen layers, limited training time, or lack of augmentation diversity.

⚠️ **When these conditions aren’t met—as seen with CNN—FixMatch struggles to deliver improvements.** Low-capacity models often fail to produce confident predictions, especially after aggressive augmentations, which limits the number of usable pseudo-labels.

💡 **Mathematically, FixMatch relies on a confidence threshold (τ) to select pseudo-labels, accepting only predictions where max(p) > τ.** In this study, a threshold of **0.85** was chosen—**lower than the more commonly used 0.95**—to ensure more unlabeled examples could contribute to training. This was a **strategic compromise**: while slightly increasing the risk of noisy labels, it substantially improved label coverage and learning signal, especially for models like MobileNet.

🧪 **Even in high-confidence pseudo-labeling, calibration matters**: if confidence doesn’t align with true accuracy, the model may still reinforce incorrect predictions. Future work should explore confidence calibration techniques to further improve label reliability.

# 2 Future Research Directions

Building on the insights of this study, future work could explore several promising paths to further improve semi-supervised image classification performance, especially under real-world constraints.

**1. Extend Training Schedules (Especially for FixMatch)**

Given the **computational constraints**, training epochs were likely kept modest. However, **semi-supervised learning techniques—especially FixMatch—often benefit from longer training durations**, as the model gradually refines pseudo-label quality and improves consistency under augmentation.

**Longer training = better pseudo-labels = stronger unsupervised learning signal.**

**2. Evaluate Confidence Calibration**

Not all high-confidence predictions are accurate. A model’s **confidence calibration**—i.e., how well its predicted probabilities reflect true correctness—plays a major role in the effectiveness of pseudo-labeling and FixMatch.

* **Well-calibrated models** produce **trustworthy pseudo-labels**, enhancing learning.
* Poorly calibrated models may **propagate errors** and **undermine performance**.

Future work could incorporate **temperature scaling**, **expected calibration error (ECE)**, or **reliability diagrams** to evaluate and improve calibration before using confidence thresholds.

**3. Experiment with Confidence Thresholds**

FixMatch uses a fixed threshold (e.g., τ = 0.95) to select pseudo-labels. But this can be too strict—especially for weaker models or hard examples.

* Test **lower or adaptive thresholds** to increase the quantity of usable pseudo-labels.
* Consider threshold annealing over time or per-class thresholds to account for class imbalance.

This could unlock more learning for models like MobileNet or even CNN without compromising quality.

**4. Unfreeze and Fine-Tune More Pretrained Layers**

Due to memory constraints, only top layers were fine-tuned in this study. However, **semi-supervised learning can benefit greatly from unfreezing deeper layers**, especially when adapting pretrained models to new data distributions.

* Consider **gradual unfreezing schedules** (e.g., layer-by-layer).
* If compute allows, fine-tune deeper layers—especially for EfficientNet and MobileNet.

**5. Test on Larger and More Diverse Datasets**

This study used a limited dataset, but **larger and more complex datasets** (e.g., Tiny ImageNet, CIFAR-100, medical or satellite imagery) would stress-test the generalizability of your findings.

* Would FixMatch scale better with more data?
* Would model calibration and pseudo-label quality improve in data-rich environments?

**Scaling the dataset helps assess how semi-supervised techniques generalize and whether conclusions hold under real-world diversity.**

**6. Try Other Semi-Supervised Frameworks**

While FixMatch is state-of-the-art, it's not the only game in town. Future work could compare:

* **UDA (Unsupervised Data Augmentation)**
* **MixMatch**
* **Meta Pseudo-Labels**
* **Self-training with confidence weighting**

These methods offer different trade-offs in consistency, augmentation, and label quality—some may be more robust under low compute settings.

**7. Analyze Error Patterns and Class-Level Performance**

Beyond global accuracy/F1, a **class-wise error breakdown** can reveal:

* Which classes benefit most from semi-supervision?
* Are pseudo-labels more accurate for dominant vs rare classes?
* Are there systematic mistakes being reinforced?

This insight can guide **per-class strategies** (e.g., lower thresholds or oversampling rare-class pseudo-labels).

**8. Optimize Augmentation Pipelines**

FixMatch's strength lies in its use of **strong augmentation** (e.g., RandAugment), but too much distortion can hurt fragile models. Future work could:

* Tune the strength and type of augmentation.
* Explore **automated augmentation search** or **curriculum-based augmentation** (start weak, go strong).