**For Knowledge Distillation (KD)**

1. **Confusion Matrix (like you mentioned):**
   * Great for visual comparison between teacher and student models.
   * Shows how class-wise performance improves post-distillation.
2. **Accuracy / Top-k Accuracy:**
   * Basic but important. Compare:
     + Teacher accuracy
     + Student (pre-KD) accuracy
     + Student (post-KD) accuracy
3. **KL Divergence or Cross-Entropy between Teacher and Student Outputs:**
   * Show how the student's output distribution gets closer to the teacher’s over training.

**For Federated Learning (FL)**

1. **Global Test Accuracy vs. Rounds:**
   * Plot accuracy over communication rounds. Compare:
     + Centralized training
     + Federated training (with and without KD, if you’re doing FL+KD)
2. **Per-Client Accuracy Distribution:**
   * Boxplot or histogram to show accuracy variance across clients.
   * Useful in non-IID settings.
3. **Data Heterogeneity (Non-IIDness Measure):**
   * Use metrics like **Earth Mover’s Distance (EMD)** or **label distribution entropy** to quantify heterogeneity across clients.
   * We use **Earth Mover's Distance (EMD)** to measure how far each client's label distribution is from the global label distribution. This captures the non-IID nature of data quantitatively — the higher the EMD, the more dissimilar the client is from the global view.
4. **Model Drift / Divergence:**
   * Cosine similarity or Euclidean distance between local and global model parameters.
   * Shows how much models drift apart due to data heterogeneity.
   * **Model drift** measures the divergence between each client's local model and the global model. We compute the **cosine similarity** between flattened model weights. A larger drift indicates that local updates have moved far from the global average, often due to data heterogeneity or insufficient communication.
5. **Communication Cost (e.g., Bytes Transferred):**
   * Total or per-round data sent during training.
   * Useful to compare with compressed models or KD models (student smaller than teacher).
6. **Training Time / Computation Load:**
   * Especially relevant with Raspberry Pi.
   * CPU time, energy usage (if you can measure), or FLOPs.

