A. Appendix

A.1. Distribution

In Fig. A.1, we present the visual representation of the frequency of each pathology in each task. The blue bars correspond to the pathologies associated with the current task, while the light blue bars correspond to the other pathologies, and the blue contour represents the frequency of each disease in the original dataset. During each task, we keep all the images in the dataset containing at least one of the pathologies associated to the task; however, other diseases may be present even though the information on the presence of such pathologies is not available, hence they're hidden pathologies. As can be noticed, there's some intersection between tasks; however, some diseases such as those of task 1 are very rare, hence they rarely appear in subsequent tasks.

A.2. Evaluation metrics

The primary metric used to evaluate the methods is the F1 score, which is the harmonic mean of precision and recall. It provides a balanced measure by accounting for both false positives and false negatives, making it suitable for imbalanced datasets. The formula for the F1 score is:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{6}$$

where:

$$precision = \frac{TP}{TP + FP} \tag{7}$$

$$recall = \frac{TP}{TP + FN} \tag{8}$$

For the CL scenario, we calculate the percentual forgetting and percentual relative gap. Percentual forgetting is the average, across all tasks except the last one, of the difference between the initial F1 and final F1 for each task, divided by the initial F1. It is given by:

$$forgetting = \frac{\sum_{i=1}^{n-1} \frac{F1_{i,i} - F1_{i,n}}{F1_{i,i}}}{n-1} \tag{9}$$

where n is the number of tasks and $F1_{i,j}$ is the F1 score relative to che classes of task i of the model trained on task j. A positive forgetting indicates a degradation in F1 performance.

The relative gap is a measure of the difference between the final F1 of a CL method and the F1 of Joint Training, the upper baseline. It is computed as:

$$relativeGap = \frac{F1_{JointTraining} - \frac{\sum_{i=1}^{n} F1_{i,n}}{n}}{F1_{JointTraining}} \tag{10}$$

A.3. Forgetting Results

In Fig. A.2 the average F1 across all tasks after training on each task is presented, for our proposed method and three comparative methods: LwF, LwF Replay, and Pseudo-Label. Fig. A.2a and A.2b illustrate that both LwF and LwF Replay exhibit significant performance degradation on previous tasks after training on new tasks. The Pseudo-Label method demonstrates a substantial reduction in forgetting, which is further minimized with our RCLP method, as can be seen from Fig A.2c and A.2d. Indeed, the curves of Fig. A.2d are remarkably stable, demonstrating an overall forgetting rate of just 2.4%.

B. Ablation study

To assess the contribution of each of the four components in our method, we conducted a series of ablation experiments, systematically excluding one block at a time. The results, including final average F1 score, percentage forgetting, and relative performance gap, are summarized in Table 2. For comparison, we also report the performance of three baseline methods: RCLP, Pseudo-Label, and Replay.

As anticipated, both the Masked Loss and Distillation Loss components had the lowest impact on performance. When either of these blocks was excluded from the modified RCLP, the final F1 score was 0.26, compared to 0.27 for the full RCLP. Therefore, the combined contribution of Distillation Loss and Masking Loss amounts to 0.02, improving RCLP's performance from 0.25 to 0.27. Additionally, the Forgetting rate increased to 9.3% without the Masked Loss and 11% without the Distillation Loss, in contrast to the 2.4% observed with the complete RCLP method.

The removal of the forward step in RCLP significantly degrades the method's performance, with the final F1 score dropping to 0.17—considerably lower than both RCLP and Pseudo-Label, and only slightly above the final F1 of Replay. This result is expected, as the forward step is crucial for mitigating interference, which cannot be addressed by the backward step alone. Specifically, the backward step adjusts the memory buffer so that each sample contains information about all labels from tasks $i, i+1, \ldots, n$, where i is the task of origin and n is the most recent task. However, without the forward step, samples from task $n+1$ only retain information about task $n+1$ labels, leading to interference between these new samples and those replayed from earlier tasks.

The removal of the backward step also leads to a decline in performance, though not as pronounced as in the case of the forward step. The final F1 score decreases to 0.25, still slightly higher than that of Pseudo-Label, with a forgetting rate of 18%. The key difference is that while the forward step incorporates information about past labels into both the memory buffer and current task samples, the backward step

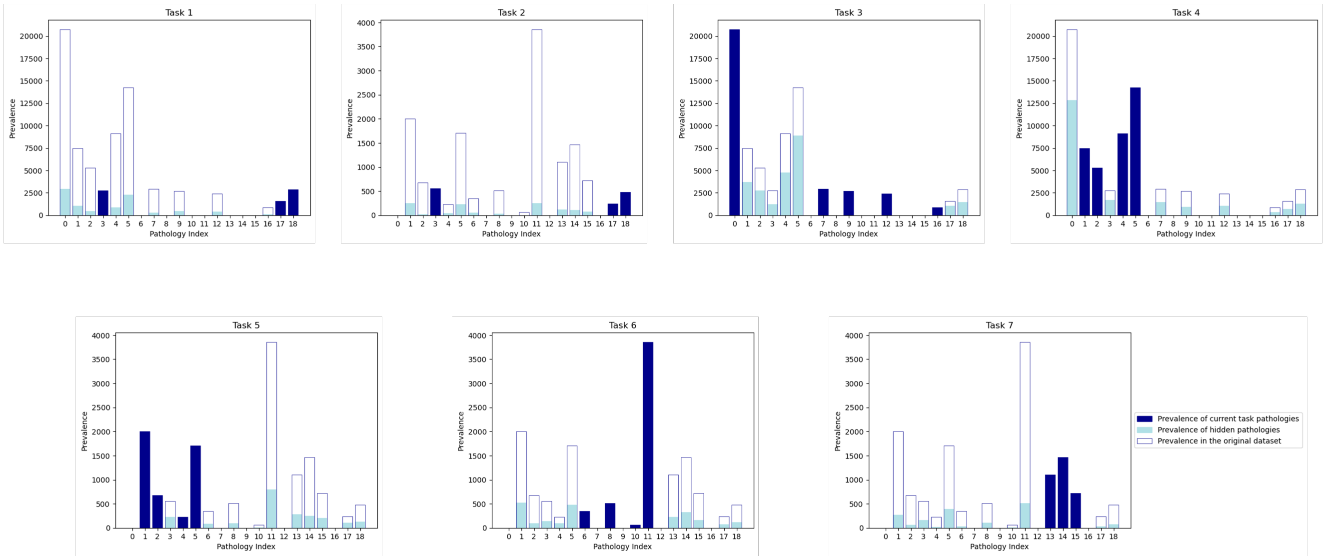
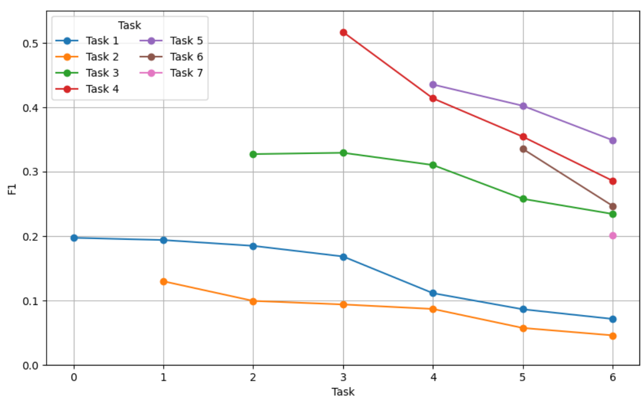
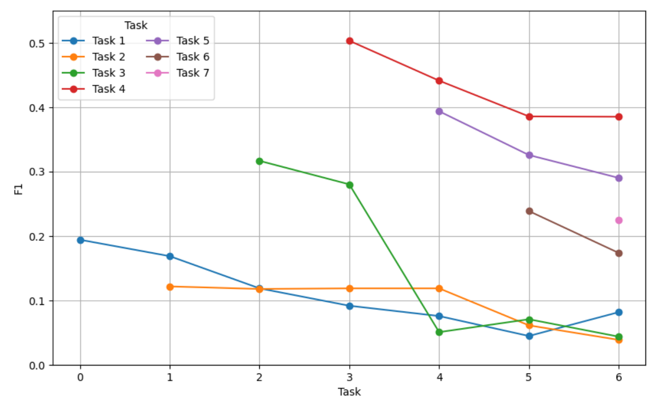


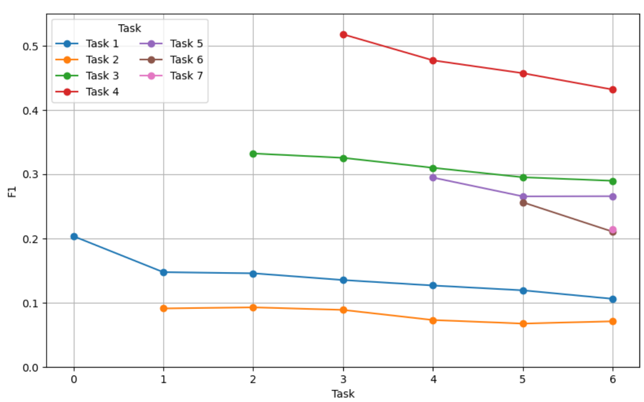
Figure A.1. Visual representation of the frequency of each pathology in each task



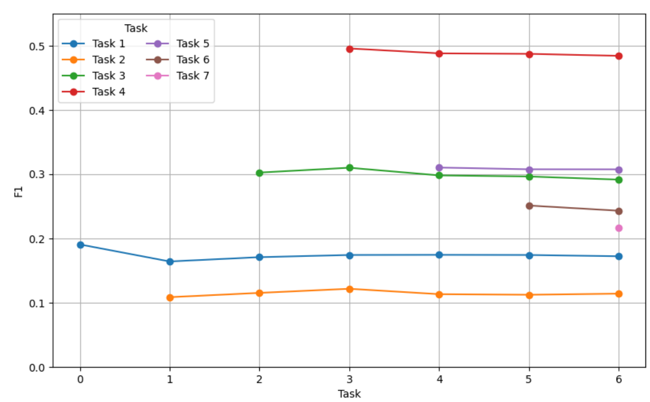
(a) Learning without Forgetting (LwF)



(b) Learning without Forgetting (LwF) Replay



(c) Pseudo-Label



(d) RCLP

Figure A.2. Comparison between the average F1 score on each task between RCLP and prior methods.

modifies only the memory buffer. Consequently, the interference without the forward step is considerably more pronounced.

These results underscore the importance of each of the

Table 2. Experimental results of the ablation study. The results of the four modifications of our method (RCLP) are reported together with the results of three baselines methods.

	Metrics		
Strategy	**Avg. F1 ↑**	**Forgetting F1 ↓**	**Relative gap ↓**
Replay	0.15	59%	60%
Pseudo-Label	0.24	21%	37%
RCLP w/o Backward step	0.25	18%	34%
RCLP w/o Forward step	0.17	48%	55%
RCLP w/o Masked Loss	0.26	9.3%	32%
RCLP w/o Distillation Loss	0.26	11%	32%
RCLP (ours)	**0.27**	**2.4%**	**29%**

four components in mitigating forgetting, with the label propagation technique emerging as the most critical factor.