Appendix: Dynamical Label Augmentation and Calibration for Noisy Electronic Health Records

Yuhao Li[1], Ling Luo[1], and Uwe Aickelin[1]

School of Computing and Information Systems, University of Melbourne
{yuhao.li4, ling.luo, uwe.aickelin}@unimelb.edu.au

A. Data Information

Table 1: Summary Statistics for the Datasets

Dataset	#Classes	#Instances	#Dimensions	Length	Type
eICU	2	89123	29	24	EHR
MIMIC	2	7701	9	6	EHR
ArrowHead	3	211	1	251	IMAGE
CBF	3	930	1	128	SIMULATED
FaceFour	4	112	1	350	IMAGE
MelbournePedestrian	10	3650	1	24	TRAFFIC
OSULeaf	6	442	1	427	IMAGE
Plane	7	210	1	144	SENSOR
Symbols	6	1020	1	398	IMAGE
Trace	4	200	1	275	SENSOR
Epilepsy	4	275	3	207	HAR
NATOPS	6	360	24	51	HAR
EthanolConcentration	4	524	3	1751	OTHER
FaceDetection	2	9414	144	62	IMAGE
FingerMovements	2	416	28	50	EEG

B. Analysis of results on EHR datasets

As shown in Table 2, our model achieves the best performance across almost all noise levels, including Symmetric 30%, Symmetric 40%, Asymmetric 30%, and IDN. However, for the few noise ratios where ACTLL does not achieve the best result, MixUp-BMM, SREA, and even SIGUA predict only a single class for the eICU dataset, resulting in an inflated weighted F1-score of 0.8595 (0) for each entry. This behavior significantly overestimates the actual performance, indicating that these models are heavily influenced by label noise and therefore cannot produce reliable results under EHR-related datasets for robust learning.

C. Hyperparameter Analysis

The hyperparameter analysis is conducted under the setting of 50% symmetric noise (Symm. 50) for all datasets, varying t_{corr} values from $50, 100, 150, 200, 250$.

Table 2: Comparison of weighted F1-score with corresponding standard deviation for baseline methods on the average of ***2 EHR datasets***. Best results are shown in **bold**

Methods	Symmetric Noise (%)					Asymmetric Noise (%)			IDN Noise (%)	
	10	20	30	40	50	10	20	30	30	40
Vanilla	0.805(0.006)	0.777(0.008)	0.734(0.014)	0.687(0.031)	0.563(0.022)	0.812(0.006)	0.781(0.006)	0.737(0.019)	0.729(0.012)	0.665(0.015)
SIGUA	0.822(0.005)	0.817(0.007)	0.813(0.006)	**0.798(0.013)**	0.464(0.231)	0.825(0.006)	0.813(0.003)	**0.811(0.007)**	0.809(0.006)	0.784(0.009)
Co-teaching	0.815(0.008)	0.790(0.006)	0.751(0.010)	0.686(0.014)	0.582(0.026)	0.817(0.004)	0.794(0.005)	0.753(0.012)	0.671(0.010)	0.677(0.017)
Mixup-BMM	0.768(0.010)	0.725(0.114)	**0.818(0.004)**	0.796(0.012)	0.364(0.142)	0.770(0.018)	0.788(0.025)	0.788(0.032)	0.819(0.007)	0.757(0.039)
Dividemix	0.779(0.014)	0.781(0.015)	0.697(0.188)	0.597(0.229)	0.479(0.227)	0.778(0.008)	0.772(0.029)	0.740(0.150)	0.799(0.016)	0.657(0.205)
SREA	0.812(0.020)	0.816(0.015)	0.811(0.022)	0.797(0.017)	0.367(0.169)	0.814(0.015)	0.809(0.005)	0.808(0.018)	**0.821(0.005)**	**0.799(0.021)**
CTW	0.819(0.013)	0.800(0.018)	0.773(0.024)	0.729(0.061)	0.569(0.049)	0.817(0.014)	0.807(0.013)	0.777(0.018)	0.781(0.019)	0.725(0.046)
ACTLL	**0.833(0.013)**	**0.818(0.012)**	0.817(0.012)	0.737(0.034)	**0.597(0.066)**	**0.832(0.010)**	**0.821(0.013)**	0.791(0.018)	0.795(0.019)	0.757(0.026)

We extract statistics from the EHR datasets to perform specific data analysis and identify whether there is a significant variation compared to other benchmark datasets.

As shown in Figure 1, the model performance across all datasets improves as the correction time increases, peaking at $T_{\text{corr}} = 200$, followed by a sharp decline at $T_{\text{corr}} = 250$. The results for the EHR datasets demonstrate a similar pattern but with a larger standard deviation. This phenomenon justifies our choice of $T_{\text{corr}} = 200$ for the model, as early correction does not ensure sufficient learning of clean patterns, while late correction hinders the model's ability to learn from the corrected labels. For the choice of k and T for λ_{corr}, we plotted

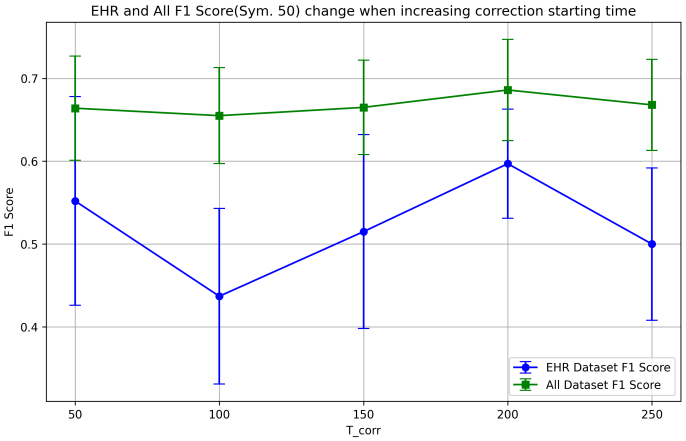

Fig. 1: Performance change when increasing T_{corr}

the corresponding function from T_{corr} to N, with $k = 0.1, 0.5$ and $T = 1, 10, 100$, as shown in Figure 2. To ensure a gradual increase of λ_{unce} with a relatively small initial value, the default setting best matches the requirement.

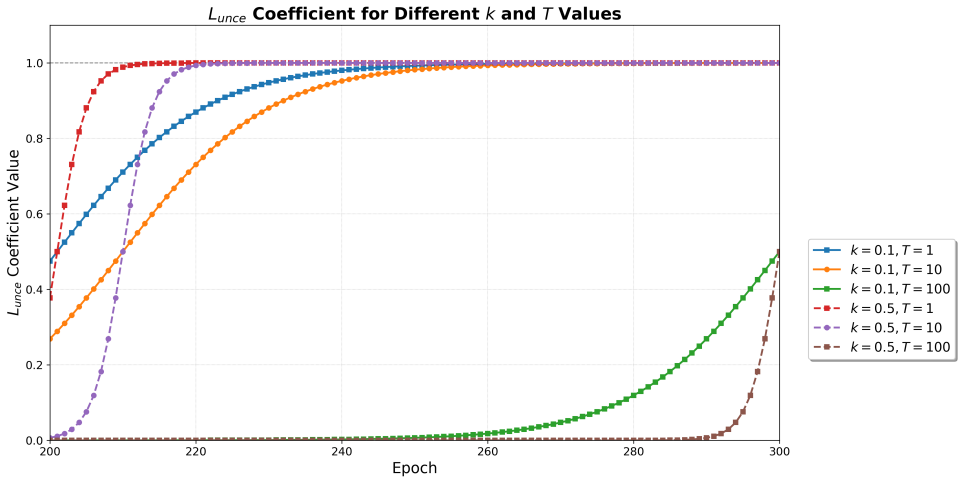

Fig. 2: Value of λ_{corr} from the time when reaching correction time until the end of the training, with $\epsilon = 1$, n is a current epoch, $T_{corr} = 200$

D. Statistic Test(Mann-Whitney U) Results

Table 3 summarizes the Mann-Whitney U Test results for various models, comparing their performance against ACTLL. Models such as Dividemix, Vanilla, MixUp-BMM, SIGUA, and Co-teaching show significant differences, as indicated by their low p-values (≤ 0.011). Dividemix has the highest U-statistic (100.0, $p = 0.000$), demonstrating a strong divergence from the baseline.

In contrast, SREA and CTW are not significantly different, with p-values of 0.076 and 0.678, respectively. These results suggest that ACTLL outperforms most other models, as it was used as the reference for comparison. Although the Mann-Whitney U Test results for SREA and CTW do not reject the null hypothesis, indicating no statistically significant difference compared to the reference (p-values = 0.076 and 0.678, respectively), they still show comparable performance to ACTLL. For further statistical analysis, Figure 3 illustrates that

Table 3: Mann-Whitney U Test Results for Different Models

Model	**U-Statistic**	**P-Value**
SREA	74.0	0.076
CTW	56.0	0.678
MixUp-BMM	87.0	0.006
Co-teaching	84.0	0.011
Dividemix	100.0	0.000
SIGUA	84.0	0.011
Vanilla	88.0	0.005

ACTLL has the lowest mean rank, highlighting its superior performance compared to all other models. Although the Mann-Whitney U Test results for SREA and CTW did not reject the null hypothesis, the critical difference (CD) diagram shows that ACTLL is still statistically better, as it is positioned farthest to the left with the lowest rank value of 1.20. The distance between ACTLL and the other models, including SREA and CTW, further demonstrates its advantage. This emphasizes that ACTLL is the top-performing model in terms of rank, even when significance tests suggest no statistical difference for specific model pairs.

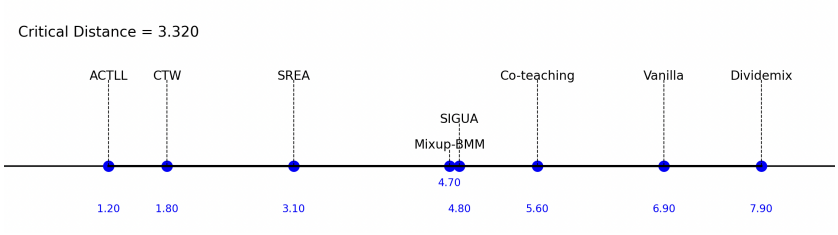

Fig. 3: Critical difference diagram of the comparison with baselines on 15 datasets with the confidence level of 95%

D: Sample Selection Analysis

To further illustrate the distinction between the three sample selection strategies, we have plotted the precision with respect to clean train labels under 60% symmetric label noise. Figure 4 shows the precision changes for the Trace, FaceFour, and MIMIC datasets, respectively.

The first graph illustrates precision changes over epochs for the three methods—Small Loss, GMM, and BMM—on the Trace dataset. The BMM method demonstrates the highest stability, with precision values consistently remaining between 0.50 and 0.60 across epochs, indicating its robustness to noise. In contrast, the GMM method exhibits substantial fluctuations, with frequent peaks and valleys, suggesting lower stability and reduced effectiveness. The Small Loss method shows lower precision than BMM but maintains greater stability compared to GMM, making it a reasonable choice when stability is prioritized.

The second graph shows the precision variation over epochs for the MIMIC dataset. BMM displays a smooth, steady increase in precision, eventually stabilizing around 0.37. GMM achieves slightly higher precision, stabilizing around 0.38, but displays slight instability in the early epochs. The Small Loss method starts at a lower precision but follows a consistent growth trend, gradually adapting to the dataset. Overall, BMM and GMM perform comparably on the MIMIC dataset, with BMM having an edge in stability.

The third graph presents precision changes for the FaceFour dataset. BMM once again proves to be the most stable method, with precision ranging between

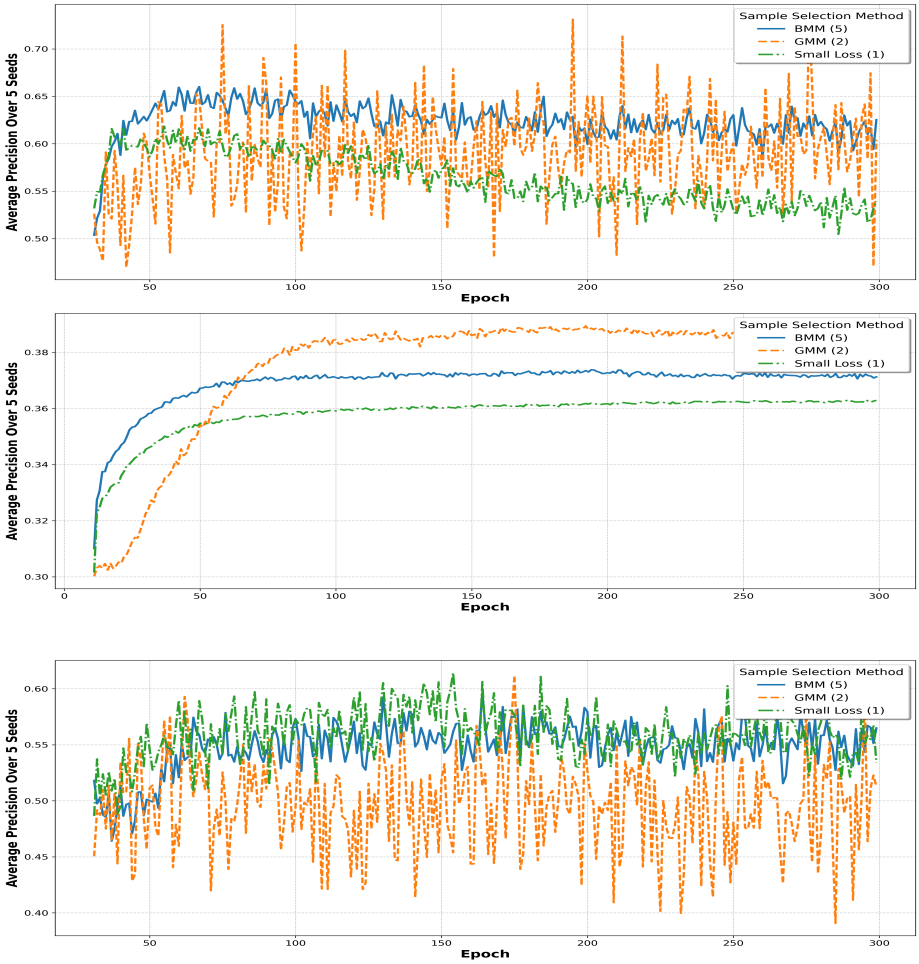

Fig. 4: Precision comparison for three sample selection strategies under the **Trace**(top), **MIMIC**(middle), **FaceFour**(below), dataset. Each graph shows precision changes for different sample selection strategies.

0.60 and 0.70 and minimal fluctuations. GMM shows high variability with erratic changes, making it less reliable. The Small Loss method maintains stable precision around 0.60. Although GMM occasionally reaches higher precision, its lack of consistency across epochs makes BMM the preferable choice across these datasets.