

# Semi-Supervised Learning for Sparsely-Labeled Sequential Data: Application to Healthcare Video Processing Supplementary Materials

Florian Dubost\*<sup>1</sup>, Erin Hong\*<sup>1</sup>, Siyi Tang<sup>1</sup>  
Nandita Bhaskhar<sup>1</sup>, Christopher Lee-Messer<sup>1</sup>, and Daniel Rubin<sup>1</sup>

<sup>1</sup>Stanford University

{floriandubost1, erin.hong17}@gmail.com, {cleemess, rubin}@stanford.edu

\*equal contribution

Table 1. **Statistics of the full hospital video dataset.**

Label	Nbr of windows	Average length (sec)	Median length (sec)	Standard Deviation (sec)	Total length (sec)	Number of patients
Suctioning	45	14.36	10	14.46	646	20
Chewing	15	89.07	45	91.23	1336	12
Rocking	21	54.67	24	71.08	1148	10
Cares	44	88.66	46	170.59	3901	23
Patting	33	38.64	21	40.49	1275	9
All	158	52.57	25	104.47	8306	59

Table 2. **Statistics of the training split of the hospital video dataset.**

Label	Nbr of windows	Average length (sec)	Median length (sec)	Standard Deviation (sec)	Total length (sec)	Number of patients
suctioning	12	20.92	16	14.92	251	5
chewing	5	147.4	130	98.04	737	4
rocking	6	92.33	56	97.53	554	2
cares	18	122.56	41	255.59	2206	8
patting	10	42.2	18	56	422	3

Table 3. **Statistics of the validation split of the hospital video dataset.**

Label	Nbr of windows	Average length (sec)	Median length (sec)	Standard Deviation (sec)	Total length (sec)	Number of patients
suctioning	18	13.5	11	10.96	243	7
chewing	6	44.33	20	42.13	266	5
rocking	8	45.75	24	53.7	366	4
cares	15	75.33	65	50.24	1130	10
patting	17	31.41	17	28.8	534	4

Table 4. **Statistics of the testing split of the hospital video dataset.**

Label	Nbr of windows	Average length (sec)	Median length (sec)	Standard Deviation (sec)	Total length (sec)	Number of patients
suctioning	15	10.13	7	15.88	152	8
chewing	4	83.25	44	96.14	333	3
rocking	7	32.57	17	43.74	228	4
cares	11	51.36	34	50.54	565	5
patting	6	53.17	52	32.53	319	2

Table 5. **Results on the HMDB51 dataset.** Brushing hair vs. draw sword. We observe an early positive contribution of N with a late and progressively increasing negative contribution (same as CIFAR) with  $\alpha_1 = 3, \beta_1 = 1, \alpha_2 = 0.5,$  and  $\beta_2 = 9.$

Risk Level (N)	1	2	3	4	5
mAP, mean (std)	71 (6)	61 (9)	85 (5)	77 (11)	72 (16)
mF1, mean (std)	61 (27)	60 (23)	83 (14)	58 (34)	39 (43)

Task:	Risk Level (N)								
car vs.	1	2	3	4	5	6	7	8	9
airplane	70.6 (11.4)	66.3 (4.6)	79.7 (4.2)	82.8 (4.7)	80.9 (5.8)	87.6 (3.3)	72.9 (3.2)	85.3 (5.5)	<b>88.0 (9.2)</b>
bird	83.2 (4.4)	75.1 (15.6)	<b>90.6 (3.3)</b>	83.7 (4.6)	86.3 (4.4)	71.7 (7.9)	89.9 (3.0)	88.8 (3.3)	85.4 (4.9)
cat	84.5 (5.7)	84.7 (3.6)	84.7 (3.7)	87.9 (3.5)	89.7 (2.3)	<b>89.9 (1.7)</b>	82.5 (5.2)	86.9 (1.8)	81.2 (7.0)
deer	87.3 (3.1)	89.7 (2.5)	85.2 (4.3)	87.8 (2.4)	84.9 (4.3)	81.0 (4.4)	86.4 (3.9)	85.1 (3.1)	85.4 (4.6)
dog	87.5 (3.4)	88.7 (1.5)	89.8 (1.6)	93.7 (1.2)	87.5 (3.8)	90.4 (2.8)	<b>95.0 (1.0)</b>	92.5 (1.1)	88.3 (2.9)
frog	81.3 (5.6)	88.0 (2.9)	<b>92.2 (2.6)</b>	92.8 (2.1)	88.3 (1.6)	85.3 (4.1)	88.6 (1.4)	83.7 (4.8)	76.9 (5.4)
horse	83.0 (3.4)	88.8 (5.2)	87.4 (2.8)	91.3 (2.5)	90.8 (3.0)	84.1 (5.1)	78.2 (4.1)	77.7 (9.4)	<b>91.6 (2.3)</b>
ship	70.5 (4.6)	59.0 (7.0)	68.9 (4.6)	81.1 (2.7)	77.2 (3.7)	<b>84.6 (6.4)</b>	76.1 (4.6)	72.8 (1.8)	75.9 (6.1)
truck	66.4 (6.1)	69.7 (5.9)	71.5 (4.6)	57.6 (16.9)	<b>74.5 (6.0)</b>	72.8 (6.2)	73.6 (7.4)	58.9 (7.0)	55.7 (8.5)

Table 6. **Recall for the CIFAR-10 dataset.** Means (standard deviations) are computed for 10 repetitions of the experiments with different random initialization of the weights. Best statistically significant results (p-value < 0.05) are highlighted in bold.

Task:	Risk Level (N)								
car vs.	1	2	3	4	5	6	7	8	9
airplane	72.3 (5.5)	72.9 (4.1)	82.0 (2.0)	81.2 (2.2)	78.8 (2.7)	<b>82.5 (0.9)</b>	77.4 (1.7)	80.0 (3.0)	76.4 (3.8)
bird	77.5 (2.0)	67.9 (9.2)	77.9 (0.6)	75.5 (1.3)	77.5 (1.2)	70.9 (2.7)	<b>79.5 (1.7)</b>	76.7 (0.8)	77.0 (1.4)
cat	68.1 (1.6)	67.4 (1.1)	72.4 (1.7)	74.7 (1.2)	73.7 (1.1)	<b>75.5 (1.2)</b>	66.9 (1.8)	69.0 (1.3)	68.3 (3.0)
deer	72.5 (2.6)	74.9 (2.1)	71.4 (0.7)	<b>75.7 (1.6)</b>	72.8 (1.1)	70.0 (1.4)	71.2 (1.4)	73.2 (0.7)	73.2 (0.7)
dog	71.6 (0.9)	70.0 (1.0)	<b>73.9 (1.0)</b>	72.8 (0.7)	70.7 (1.5)	72.5 (1.7)	71.5 (1.4)	73.8 (0.8)	72.5 (1.9)
frog	61.8 (2.6)	65.4 (0.9)	70.7 (1.9)	68.1 (1.6)	66.6 (1.3)	<b>71.2 (2.2)</b>	66.7 (0.4)	63.7 (0.7)	59.7 (2.2)
horse	67.2 (1.1)	66.0 (2.2)	67.2 (1.2)	69.5 (2.2)	67.2 (1.2)	71.7 (2.9)	<b>74.1 (1.7)</b>	72.2 (3.0)	69.4 (1.3)
ship	72.1 (2.3)	67.4 (3.9)	74.9 (2.4)	78.1 (1.0)	77.4 (1.7)	<b>80.1 (2.2)</b>	76.0 (2.0)	73.5 (1.1)	75.7 (2.0)
truck	62.1 (2.6)	<b>68.7 (3.8)</b>	64.2 (2.0)	55.7 (6.4)	61.1 (2.8)	68.5 (2.5)	57.3 (2.9)	61.4 (3.1)	63.5 (4.5)

Table 7. **F1-scores for the CIFAR-10 dataset.** Means (standard deviations) are computed for 10 repetitions of the experiments with different random initialization of the weights. Best statistically significant results (p-value < 0.05) are highlighted in bold.

Task:	Risk Level (N)								
car vs.	1	2	3	4	5	6	7	8	9
airplane	73.5 (2.3)	75.5 (3.5)	<b>82.6 (1.4)</b>	80.8 (2.0)	78.3 (2.6)	81.4 (1.0)	78.7 (1.3)	78.7 (3.2)	72.9 (4.1)
bird	75.9 (2.1)	65.4 (7.5)	74.2 (1.6)	72.9 (1.4)	74.9 (1.9)	70.8 (1.5)	76.8 (2.3)	73.1 (0.9)	74.6 (1.3)
cat	60.5 (1.2)	59.1 (1.4)	67.7 (2.4)	<b>70.1 (2.1)</b>	67.9 (1.6)	70.8 (1.8)	59.2 (1.5)	60.8 (2.6)	62.5 (3.3)
deer	66.7 (4.4)	69.8 (3.5)	65.8 (2.4)	<b>71.8 (2.3)</b>	68.4 (1.4)	65.3 (1.0)	65.0 (1.9)	68.8 (1.2)	68.7 (1.0)
dog	65.3 (1.5)	61.9 (2.1)	<b>68.3 (2.0)</b>	64.9 (1.4)	63.7 (2.3)	65.7 (3.4)	62.1 (2.6)	67.2 (1.4)	66.4 (2.9)
frog	49.9 (2.6)	53.5 (1.1)	61.7 (4.0)	56.4 (3.4)	55.7 (2.9)	<b>65.4 (3.9)</b>	55.7 (1.2)	52.3 (2.4)	48.2 (1.6)
horse	59.5 (2.9)	54.4 (2.5)	57.4 (1.4)	59.9 (4.0)	55.7 (2.6)	66.6 (5.6)	<b>72.6 (2.0)</b>	70.3 (2.9)	59.5 (2.8)
ship	72.8 (1.4)	71.8 (1.9)	76.9 (1.6)	77.2 (0.7)	77.6 (1.2)	<b>79.0 (2.2)</b>	76.0 (1.6)	73.7 (1.2)	75.8 (0.8)
truck	59.7 (1.2)	68.1 (5.8)	60.2 (3.3)	56.0 (6.3)	52.7 (3.7)	66.5 (3.4)	45.6 (1.0)	63.2 (2.8)	<b>68.5 (1.8)</b>

Table 8. **AUC for the CIFAR-10 dataset.** Means (standard deviations) are computed for 10 repetitions of the experiments with different random initialization of the weights. Best statistically significant (p-value < 0.05) results are highlighted in bold.

Task: car vs.	Risk Level (N)								
	1	2	3	4	5	6	7	8	9
airplane	75.6 (3.6)	81.2 (4.0)	<b>84.6 (1.4)</b>	79.7 (2.5)	77.2 (4.1)	78.1 (2.1)	82.7 (2.2)	75.5 (3.7)	68.1 (4.9)
bird	<b>72.7 (2.6)</b>	63.0 (6.8)	68.4 (2.5)	69.0 (2.3)	70.5 (3.1)	70.9 (3.3)	71.3 (2.7)	67.7 (1.5)	70.4 (2.2)
cat	57.1 (1.1)	56.0 (1.1)	63.3 (2.4)	65.0 (2.6)	62.5 (1.5)	<b>65.1 (1.7)</b>	56.3 (1.1)	57.2 (2.0)	59.2 (2.8)
deer	62.1 (3.8)	64.4 (3.7)	61.6 (2.8)	<b>66.6 (2.3)</b>	63.9 (1.9)	61.7 (1.1)	60.6 (1.7)	64.2 (1.5)	64.2 (2.0)
dog	60.6 (1.5)	57.8 (1.7)	<b>62.8 (2.0)</b>	59.5 (1.1)	59.4 (1.9)	60.7 (3.2)	57.4 (1.8)	61.5 (1.2)	61.5 (2.5)
frog	49.9 (1.6)	52.1 (0.7)	57.4 (2.9)	53.8 (2.3)	53.6 (1.9)	<b>61.3 (4.1)</b>	53.5 (0.8)	51.5 (1.7)	48.9 (1.1)
horse	56.6 (2.5)	52.7 (1.6)	54.6 (0.9)	56.2 (3.0)	53.4 (1.8)	63.0 (5.7)	<b>70.6 (3.1)</b>	68.3 (4.2)	55.9 (2.0)
ship	73.9 (0.6)	79.5 (2.3)	<b>82.2 (2.4)</b>	75.4 (1.1)	77.8 (1.0)	76.5 (4.1)	76.1 (2.3)	74.2 (2.0)	76.0 (2.4)
truck	58.6 (1.2)	68.5 (7.1)	58.5 (3.5)	60.8 (14.1)	52.0 (2.6)	65.2 (4.4)	47.1 (0.8)	65.2 (4.9)	<b>75.5 (3.9)</b>

Table 9. **Precision for the CIFAR-10 dataset.** Means (standard deviations) are computed for 10 repetitions of the experiments with different random initialization of the weights. Best statistically significant (p-value < 0.05) results are highlighted in bold.