Details on Experimental Results
A Learning System for Automatic Berg Balance Scale Score Estimation

This document provides supplementary material for the experimental results described in [1], by extensively reporting the details on the predictive performance achieved on the Berg Balance Scale (BBS) score estimation tasks. Specifically, here we describe the results achieved in all the experimental settings and by all the learning models that have been considered, namely: Leaky Integrator Echo State Network with root state mapping (LI-ESN-R) and with mean state mapping (LI-ESN-M), Multi Layer Perceptron (MLP), Time Delay Neural Network (TDNN), Simple Recurrent Network (SRN) and k nearest neighbor (k-NN).

For information on the BBS score estimation tasks, on the methodologies adopted for data collection and pre-processing, and for details on the experimental settings and on the double cross-fold validation schema used for model selection, the reader is referred to [1].

The organization of this document is as follows. We first report on the performance achieved on the 3 BBS tasks corresponding to the exercises #6, #7 and #10 of the BBS test, in the following referred to as BBS-6, BBS-7 and BBS-10, respectively. Then, we show the impact on the predictive performance singly due to the use of a weight sharing (WS) approach and to the augmentation of the input with users' clinical data, i.e. age (BBS-10-A task) and weight (BBS-10-W task). Finally, we consider the synergistic effect of the joint use of WS and clinical data, and we report the results achieved by the selected model, i.e. LI-ESN-R, in the selected setting, i.e. BBS-10-W task with WS. A summary of these results is reported in [1].

Selection of the BBS Exercise

As a further support to the clinical considerations reported in [1] on the selection of the BBS exercise to be performed by the users, we compared the predictive performance achieved by the learning models on the 3 tasks BBS-6, BBS-7 and BBS-10. The values of Mean Absolute Error (MAE) and R obtained on the BBS-6, BBS-7 and BBS-10 tasks are respectively reported in Tables 1, 2 and 3.

Model	*TR MAE*	*VL MAE*	*TS MAE*	*TS R*
LI-ESN-R	4.43 ± 0.13	5.04 ± 0.19	5.53 ± 0.43	0.53
LI-ESN-M	3.52 ± 0.08	5.09 ± 0.15	5.67 ± 0.27	0.40
MLP	1.86 ± 0.21	4.09 ± 0.68	5.52 ± 0.59	0.32
TDNN	2.78 ± 0.23	4.76 ± 0.72	6.19 ± 0.87	0.23
SRN	3.61 ± 0.54	4.51 ± 0.68	5.36 ± 0.65	0.37
k-NN	1.56 ± 0.00	3.51 ± 0.00	5.30 ± 0.00	0.46
Overall averages	2.96	4.50	5.59	0.39

Table 1: Training (TR), validation (VL), test (TS) MAEs, achieved by LI-ESN-R, LI-ESN-M, MLP, TDNN, SRN and k-NN on the BBS-6 task. R values on the test set are reported as well. The last row reports the average over the results achieved by all the models.

Model	*TR MAE*	*VL MAE*	*TS MAE*	*TS R*
LI-ESN-R	3.74 ± 0.11	4.74 ± 0.17	5.05 ± 0.32	0.51
LI-ESN-M	3.40 ± 0.11	4.98 ± 0.16	5.71 ± 0.41	0.40
MLP	3.22 ± 0.13	4.37 ± 0.30	5.26 ± 0.33	0.45
TDNN	4.30 ± 0.39	5.50 ± 0.96	5.94 ± 0.87	0.31
SRN	3.85 ± 0.44	4.80 ± 0.49	5.25 ± 0.71	0.57
k-NN	0.70 ± 0.00	3.47 ± 0.00	5.36 ± 0.00	0.42
Overall averages	3.20	4.64	5.43	0.44

Table 2: Training (TR), validation (VL), test (TS) MAEs, achieved by LI-ESN-R, LI-ESN-M, MLP, TDNN, SRN and k-NN on the BBS-7 task. R values on the test set are reported as well. The last row reports the average over the results achieved by all the models.

Model	*TR MAE*	*VL MAE*	*TS MAE*	*TS R*
LI-ESN-R	3.56 ± 0.12	4.21 ± 0.14	4.80 ± 0.40	0.68
LI-ESN-M	3.65 ± 0.06	4.26 ± 0.11	4.44 ± 0.25	0.67
MLP	2.21 ± 0.21	3.96 ± 0.28	4.96 ± 0.49	0.57
TDNN	2.79 ± 0.16	3.72 ± 0.34	4.69 ± 0.70	0.54
SRN	3.85 ± 0.34	4.02 ± 0.42	4.86 ± 0.56	0.57
k-NN	1.93 ± 0.00	4.61 ± 0.00	7.03 ± 0.00	0.16
Overall averages	2.99	4.13	5.13	0.53

Table 3: Training (TR), validation (VL), test (TS) MAEs, achieved by LI-ESN-R, LI-ESN-M, MLP, TDNN, SRN and k-NN on the BBS-10 task. R values on the test set are reported as well. The last row reports the average over the results achieved by all the models.

From the results in Tables 1, 2 and 3, we can observe that all the learning models, except for k-NN, achieve a better validation performance on the BBS-10 task. Noteworthy, for LI-ESN-R, LI-ESN-M, MLP, TDNN and SRN, the BBS-10 task also corresponded to the smallest generalization MAE and to the larger R value on the test set. Moreover, we can see that for all the BBS tasks, the k-NN model results in a particularly disadvantageous ratio between training and validation MAE, denoting an overfitting behavior that is confirmed by looking at the obtained test results. Thereby, we consider that k-NN is not a good model to make judgments about the selection of the balance exercise to be performed. As a further point, it is also possible to note that the overall average on validation MAE achieved by all models is clearly better in correspondence of the BBS-10 task, representing a trend confirmed also when looking at the generalization performance (with smallest test MAE and highest test R value).

Overall, results in Tables 1, 2 and 3 provide an experimental evidence that BBS exercise #10 generally enables a more accurate estimation of the total BBS score, allowing us to focus only on the BBS-10 task in the successive stages of performance assessment.

Inter-subject Correlation

Table 4 shows the inter-subject Pearson correlation between signals gathered during exercise execution by the different subjects (averaged over the exercise repetitions in the datasets). The mean correlation values achieved for exercises #6, #7 and #10 are respectively 0.42, 0.44 and 0.40, clearly showing that exercise #10 is featured by a lower inter-subject correlation.

Task	*inter-subject correlation*
BBS ex. #6	0.42
BBS ex. #7	0.44
BBS ex. #10	0.40

Table 4: Inter-subject correlation computed on signals gathered during the execution of exercises #6, #7 and #10 by the different subjects.

Weight Sharing Approach on Input Connections

The performance achieved by LI-ESN-R, LI-ESN-M, MLP, TDNN and SRN on the BBS-10 task by adopting the WS technique is reported in Table 5. Details on the implementation of the WS are described in [1].

Model	*TR MAE*	*VL MAE*	*TS MAE*	*TS R*
LI-ESN-R	3.43 ± 0.04	4.09 ± 0.08	4.03 ± 0.13	0.71
LI-ESN-M	3.20 ± 0.06	4.27 ± 0.24	4.69 ± 0.24	0.63
MLP	2.73 ± 0.16	4.10 ± 0.42	4.80 ± 0.56	0.61
TDNN	3.13 ± 0.16	3.91 ± 0.51	4.57 ± 0.49	0.61
SRN	3.60 ± 0.30	3.75 ± 0.31	4.31 ± 0.54	0.61
Overall averages	3.22	4.02	4.48	0.63

Table 5: Training (TR), validation (VL), test (TS) MAEs, achieved by LI-ESN-R, LI-ESN-M, MLP, TDNN, SRN with WS on the BBS-10 task. R values on the test set are reported as well. The last row reports the average over the results achieved by all the models.

Results show the general positive impact due to the WS approach (whenever it can be applied), as can be observed e.g. by comparing the overall validation performance with and without the adoption of WS (see Tables 3 and 5). It is also possible to note that in the case of the selected model, i.e. LI-ESN-R, the performance shows a particularly low variability (as measured by the standard deviation) and a significant improvement on the validation result. On the other hand, in cases in which validation results in the WS settings are (slightly) lower than in the basic settings (especially for MLP and TDNN), the high performance variability indicate that such difference has a reduced significance. Overall, the positive effect of the WS approach is reflected also on the test results, as it can be seen from Table 5.

Use of Clinical Data

A further significant experimental assessment consisted in the evaluation of the influence on the predictive performance due to the adoption of additional input information including users' clinical data such as age and weight. Information concerning the way in which such information was used to augment the input and the corresponding experimental settings can be found in [1]. Tables 6 and 7 respectively report the performance achieved by the learning models on the BBS-10-A task (input augmented with users' age) and on the BBS-10-W task (input augmented with users' weight).

Model	*TR MAE*	*VL MAE*	*TS MAE*	*TS R*
LI-ESN-R	3.67 ± 0.12	4.23 ± 0.13	4.52 ± 0.27	0.65
LI-ESN-M	3.86 ± 0.09	4.34 ± 0.27	4.76 ± 0.34	0.64
MLP	3.04 ± 0.60	5.08 ± 0.45	6.07 ± 0.72	0.31
TDNN	2.18 ± 0.23	3.93 ± 0.37	5.42 ± 0.94	0.38
SRN	3.15 ± 0.31	3.91 ± 0.44	4.67 ± 0.52	0.58
k-NN	2.51 ± 0.00	4.51 ± 0.00	7.20 ± 0.00	0.23
Overall averages	3.07	4.33	5.44	0.47

Table 6: Training (TR), validation (VL), test (TS) MAEs, achieved by LI-ESN-R, LI-ESN-M, MLP, TDNN, SRN and k-NN on the BBS-10-A task. R values on the test set are reported as well. The last row reports the average over the results achieved by all the models.

Model	*TR MAE*	*VL MAE*	*TS MAE*	*TS R*
LI-ESN-R	3.50 ± 0.08	4.08 ± 0.09	4.62 ± 0.30	0.69
LI-ESN-M	4.06 ± 0.14	4.24 ± 0.17	4.43 ± 0.34	0.66
MLP	1.59 ± 0.23	4.36 ± 0.56	5.53 ± 1.00	0.56
TDNN	1.69 ± 0.25	3.62 ± 0.52	5.38 ± 1.22	0.59
SRN	2.89 ± 0.51	3.84 ± 0.47	4.70 ± 0.89	0.66
k-NN	2.89 ± 0.00	4.48 ± 0.00	6.05 ± 0.00	0.23
Overall averages	2.77	4.10	5.12	0.57

Table 7: Training (TR), validation (VL), test (TS) MAEs, achieved by LI-ESN-R, LI-ESN-M, MLP, TDNN, SRN and k-NN on the BBS-10-W task. R values on the test set are reported as well. The last row reports the average over the results achieved by all the models.

Results in Tables 6 and 7 show that for all the considered learning models the use of the users' weight information as additional input for the task (corresponding to the BBS-10-W task case) leads to better validation results (i.e. smaller validation MAE) than the use of users' age (corresponding to the BBS-10-W task case). Furthermore, a comparison between Tables 3 and 7 shows the positive impact of augmenting the input with users' weight data with respect to the standard setting (i.e. corresponding to the BBS-10 task), in terms of a general improvement of the validation performance (except for MLPs, in which case, however, the results are within the same range of variability). This improvement is particularly evident in the case of the selected model, i.e. LI-ESN-R, resulting in a smaller standard deviation on the validation set and a better trade-off between training and validation, confirmed also by the performance gain on the test set.

Joint Use of Weight Sharing and Clinical Data

The combined effect of the WS approach and the use of users' weight information to augment the input of the BBS score estimation task is investigated by means of experiments on the BBS-10-W task with WS. Results achieved under this setting by LI-ESN-R, LI-ESN-M, MLP, TDNN and SRN are reported in Table 8.

Model	*TR MAE*	*VL MAE*	*TS MAE*	*TS R*
LI-ESN-R	3.11 ± 0.05	3.85 ± 0.08	3.80 ± 0.17	0.76
LI-ESN-M	3.05 ± 0.06	3.99 ± 0.11	3.95 ± 0.21	0.73
MLP	3.36 ± 0.16	4.62 ± 0.48	5.76 ± 0.66	0.56
TDNN	2.99 ± 0.26	3.89 ± 0.69	5.19 ± 0.91	0.57
SRN	3.72 ± 0.38	3.69 ± 0.45	4.34 ± 0.69	0.68
Overall averages	3.25	4.00	4.60	0.66

Table 8: Training (TR), validation (VL), test (TS) MAEs, achieved by LI-ESN-R, LI-ESN-M, MLP, TDNN and SRN with WS on the BBS-10-W task. R values on the test set are reported as well. The last row reports the average over the results achieved by all the models.

Results point out that the joint use of WS and users' weight in input turns out to represent the selected experimental setting, leading to overall better validation results and a preferable ratio between training and validation errors. More in detail, from the perspective of the best results of individual learning models, we can note that, with respect to BBS-10, BBS-10 with WS and BBS-10-W without WS, BBS-10-W with WS is the setting in which the smallest validation MAE are achieved in most cases[1].

Moreover, results in Table 8 enable a performance comparison among the different learning models under the final (complete) experimental setup adopted for the assessment of users' balance abilities, as described in [1]. Results show that the MAE achieved by LI-ESN-R, LI-ESN-M, TDNN and SRN on the validation set are in the same range of variability, while MLP obtained worse results (close to those of k-NN on the BBS-10-W task). Furthermore, we can observe that in this setting LI-ESN-R presents the smallest standard deviation on the validation set (even smaller than in the basic setting, reported in Table 3) and a good ratio between training and validation MAE, thereby resulting in the

[1] Exceptions are represented by MLP and TDNN, which reached the MAE values of 3.96 ± 0.28 and 3.62 ± 0.52 on the validation set, respectively in the BBS-10 (see Table 3) and BBS-10-W without WS (see Table 7) task settings. The results of MLP and TDNN in these cases, however, are obtained at the cost of a worse ratio between training and validation errors.

selected model. The goodness of such choice is also confirmed by the generalization results on the test set reported in Table 8, showing that LI-ESN-R finally results in the smallest MAE and highest R value on test (unseen) data.

Results of the Selected Model

The experimental analysis discussed so far in this document highlighted that the selected learning model and experimental settings for the BBS score estimation task correspond to the use of LI-ESN-R with WS, using data gathered during the execution of BBS exercise #10 and complemented by the subject personal weight information (i.e. BBS-10-W task with WS). The predictive performance results achieved in the final setting are summarized in Table 9. Note that in this table we also report the standard deviations computed on the test set under different perspectives. In particular, the uncertainty of the BBS score estimation is represented by the standard deviation on the external folds of the cross-validation, denoted by STDf. Standard deviations with respect the reservoir guesses, the different sequences (i.e. the exercise repetitions), and the different users are respectively denoted by STDg, STDs and STDu.

VL MAE	*TS MAE*	*STDg*	*STDs*	*STDf*	*STDu*	*TS R*
3.85	3.80	0.17	2.92	1.64	2.01	0.76

Table 9: Performance results achieved in the final setting for BBS score estimation, i.e. with LI-ESN-R using WS on the BBS-10-W task. The table reports validation (VL) and test (TS) MAE, along with the standard deviation computed on the test set with respect to: the reservoir guesses (STDg), the different sequences (STDs), the external folds of the double cross-validation scheme (STDf), the different users (STDu). The R value on the test set is reported as well.

As it can be seen, the selected model achieved a validation MAE (mean of errors over the folds of the cross-validation) of 3.85 BBS score points (corresponding to the 6.88% of the total BBS score range), a test MAE of 3.80 BBS score points (corresponding to the 6.79% of total BBS score range). Such a result is indeed extremely good, considering that the generalization error is even below the threshold of 4 BBS score points, that is the score range of each individual BBS exercise.

References

[1] D. Bacciu, S. Chessa, C. Gallicchio, A. Micheli, L. Pedrelli, E. Ferro, L. Fortunati, D. La Rosa, F. Palumbo, F. Vozzi, O. Parodi, A Learning System for Automatic Berg Balance Scale Score Estimation, Engineering Applications of Artificial Intelligence 66 (2017) 60 – 74. doi:10.1016/j.engappai.2017.08.018.