On Time Series Cross-Validation for Mental Workload Classification from EEG Signals

Kunjira Kingphai* **and Yashar Moshfeghi***

NeuraSearch Laboratory, Department of Computer and Information Sciences, University of Strathclyde, Glasgow, Scotland

Correspondence*:
kunjira.kingphai@strath.ac.uk

yashar.moshfeghi@strath.ac.uk

ABSTRACT

Mental workload (MWL) is one of the key concepts in ergonomics and human factors for understanding human performance, which can be affected by workload being too high or too low (Holm et al., 2009). The automated warning or assisting system could help people prevent a decline in MWL performance if it could be noticed in time (Holm et al., 2009). In this direction, Electroencephalogram (EEG) has been used as a tool for evaluating MWL (van Erp et al., 2015). To date, machine learning techniques, especially powerful models like deep learning, have been employed to capture variance characteristics in the EEG signals to classify MWL status accurately (Chakladar et al., 2020). This paper investigates the cross-validation procedure employed in evaluating such models.

Cross-validation (CV) is one of the main techniques to train and test deep learning models (Schaffer, 1993). In K-fold CV (Stone, 1974), an entire data set is divided into K equally-sized subsets or folds. Then, a model is trained on $K-1$ folds (called the training set), and the remaining fold is used for model testing (called the testing set). To perform CV, the data should meet the independent and identically distributed (i.i.d.) assumption, i.e. all observations have been drawn from the same probability distribution and are statistically independent of each other (Hoadley, 1971). However, an EEG signal is time-series data generated by a process over time. Therefore, using traditional cross-validation for time-series model evaluation could be violating the key i.i.d. assumption (Bergmeir and Benítez, 2012), particularly if the aim is to predict a future event, such as the subject's MWL. Moreover, the random shuffling before splitting data into K-fold, a common practice in the machine learning area, could lead us to an unreliable classification model performance (Cerqueira et al., 2020).

This is because, in time series data, the available past and present values is used to forecast future values of X. To be more precise, the estimated value $\hat{X}_{t+\tau}$ of X at time $t+\tau$ can be obtained from a function **F** where function **F** computed from given value of X up to time t, (plus additional time-independent variables in multivariate time series analysis) $\hat{X}_{t+\tau} = \mathbf{F}(X_t, X_{t-1}, ...)$ where τ is the lag for prediction. Then, the function of continuous time series will be mapped onto binary-valued of N classes for a classification approach: $\mathbf{F}_c(X_t, X_{t-1}, ...) \rightarrow \hat{c}_i \in C$ where C is the set of class labels (Dorffner, 1996). Hence, future information could be leak into a training dataset when the cross-validation breaks the temporal order of time series data by randomly

splitting. Consequently, it would cause an overfitting problem (Bergmeir and Benítez, 2012) in model training. These problems extend to evaluating deep learning models on EEG Signals.

This paper proposes to use a modification of cross-validation techniques that takes into account the time series characteristics of EEG signals. Several modifications of time series cross-validation techniques have been proposed in the past (Bergmeir and Benítez, 2012; Bergmeir et al., 2014; Cerqueira et al., 2020). Among these, a blocked form of cross-validation with a rolling window has been shown to be an effective approach since the model will be rebuilt in every window, discarding the old values that tend to disturb the model generation Bergmeir and Benítez (2012). However, to the best of our knowledge, that technique has not been investigated in EEG analysis. Figure 1 shows data being split into training, validation and testing set using the rolling window approach. The previous window comprising n% of time t is used for training (Yellow box), the following window of j% of time t is used for validation (Brown box) and the following window of k% of time t is used for testing (Blue box). However, the research question to be investigated is how to implement this method for EEG data and, in particular, “what should be the size of the training, validation and testing window?”.

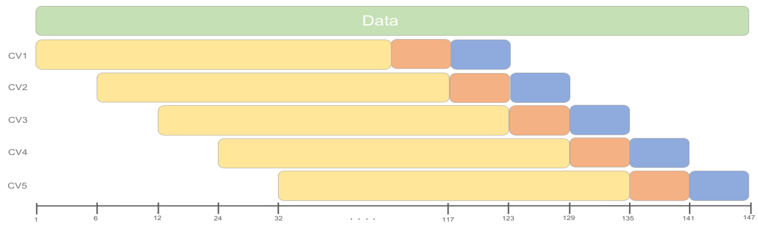

Figure 1. A 5-fold time-series CV with the rolling window size of 6.

To investigate this research question, we adopted the long short-term memory (LSTM) model, which is one of the most well-known and widely used deep learning models to predict MWL level (Nagabushanam et al., 2019; Jeong et al., 2019; Chakladar et al., 2020). This model has been successfully used to capture sequential information in EEG data (Zhang et al., 2017). To train our model, we used a publicly available mental workload dataset named STEW (Lim et al., 2018). We performed the analysis in two classification tasks; Task 1: resting-state vs working state and Task 2: low vs moderate vs high MWL. In this study, we applied a 5-fold time-series CV with a rolling window approach. There are two parameters to be optimised for the time-series CV, namely the size of the training and testing window. We varied the training window size from 20% to 90% of the data with an incremental step of 10%. The rolling window strategy is implemented by shifting forward the training and test data at each fold by a constant window size of m seconds (s). The test data of the previous fold will be used as a validation set for tuning the deep learning model's parameters. To investigate the effect of the test (and validation) window size, we varied the rolling window size from 3 s to 24 s with an incremental step of 3 s.

Our results show that in Task 1, the LSTM model achieved the highest accuracy of 87.44% when it was trained using 80% of data with a window size of 6 s for validation and testing. However, in Task 2, the highest model accuracy of 88.79% was obtained when the model was trained by using 90% of data with a window size of 3 seconds for validation and testing. This shows there is not an optimal general training and test window size, and they need to be tuned when time-series CV is being applied to train deep learning models, particularly for EEG signals to predict MWL.

Keywords: EEG, Deep Learning, Mental Workload, Time Series, Cross-validation

REFERENCES

Bergmeir, C. and Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences* 191, 192 – 213. doi:https://doi.org/10.1016/j.ins.2011.12.028. Data Mining for Software Trustworthiness

Bergmeir, C., Costantini, M., and Benítez, J. M. (2014). On the usefulness of cross-validation for directional forecast evaluation. *Computational Statistics & Data Analysis* 76, 132–143

Cerqueira, V., Torgo, L., and Mozetič, I. (2020). Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning* 109, 1997–2028

Chakladar, D. D., Dey, S., Roy, P. P., and Dogra, D. P. (2020). Eeg-based mental workload estimation using deep blstm-lstm network and evolutionary algorithm. *Biomedical Signal Processing and Control* 60, 101989

Dorffner, G. (1996). Neural networks for time series processing. In *Neural network world* (Citeseer)

Hoadley, B. (1971). Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case. *The Annals of mathematical statistics* , 1977–1991

Holm, A., Lukander, K., Korpela, J., Sallinen, M., and Müller, K. M. (2009). Estimating brain load from the eeg. *TheScientificWorldJOURNAL* 9, 639–651

Jeong, J.-H., Yu, B.-W., Lee, D.-H., and Lee, S.-W. (2019). Classification of drowsiness levels based on a deep spatio-temporal convolutional bidirectional lstm network using electroencephalography signals. *Brain sciences* 9, 348

Lim, W., Sourina, O., and Wang, L. (2018). Stew: simultaneous task eeg workload data set. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 26, 2106–2114

Nagabushanam, P., George, S. T., and Radha, S. (2019). Eeg signal classification using lstm and improved neural network algorithms. *Soft Computing* , 1–23

Schaffer, C. (1993). Selecting a classification method by cross-validation. *Machine Learning* 13, 135–143

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)* 36, 111–133

van Erp, J. B., Brouwer, A.-M., and Zander, T. O. (2015). Using neurophysiological signals that reflect cognitive or affective state. *Frontiers in neuroscience* 9, 193

Zhang, Y., Liu, B., Ji, X., and Huang, D. (2017). Classification of eeg signals based on autoregressive model and wavelet packet decomposition. *Neural Processing Letters* 45, 365–378