Scope and Arbitration in Machine Learning Clinical EEG Classification

Y. Zhu[1], L. Canham[2] and D. Western[1]

1. University of the West of England, UK; 2. North Bristol NHS Trust, UK
yixuan2.zhu@live.uwe.ac.uk, luke.canham@nbt.nhs.uk, david.western@uwe.ac.uk

***Abstract*— A key task in clinical EEG interpretation is to classify a recording or session as normal or abnormal. In machine learning approaches to this task, recordings are typically divided into shorter windows for practical reasons, and these windows inherit the label of their parent recording. We hypothesised that window labels derived in this manner can be misleading – for example, windows without evident abnormalities can be labelled 'abnormal' – disrupting the learning process and degrading performance. We explored two separable approaches to mitigate this problem: increasing the window length and introducing a second-stage model to arbitrate between the window-specific predictions within a recording. Evaluating these methods on the Temple University Hospital Abnormal EEG Corpus, we significantly improved state-of-the-art average accuracy from 89.8 percent to 93.3 percent. This result defies previous estimates of the upper limit for performance on this dataset and represents a major step towards clinical translation of machine learning approaches to this problem. Our study includes electroencephalography (EEG) datasets collected from https://isip.piconepress.com/projects/tuh_eeg/. Our code is shared on https://github.com/zhuyixuan1997/EEGScopeAndArbitration.**

Keywords*— *electroencephalogram, abnormality detection, deep learning.

I. Introduction

I-A. Background

Electroencephalography (EEG) recordings are used for the diagnosis and monitoring of a wide range of neurological conditions. Classification of EEG recordings as normal or abnormal is an essential task in their clinical interpretation. Substantial research has been conducted on the application of machine learning to this task [1–9].

Recent work in this field largely makes use of the Temple University Hospital Abnormal EEG Corpus (TUAB) [10] for training and evaluation. TUAB is a labelled subset of the Temple University Hospital EEG Corpus (TUEG) [11].

Since the presentation of the Deep4 convolutional neural network in 2017 [1] there have been only modest improvements in the accuracy of machine learning approaches to this problem, as measured on TUAB: from 85.4 percent (Deep4) up to 89.8 percent [6] – see Table 1 for further detail. Gemein et al. [5] proposed that there may be an upper limit of around 90 percent accuracy in this task, based on known values of inter-rater agreement between human experts in conventional clinical practice.

A notable but little-discussed difference between conventional clinical practice and virtually all deep learning approaches is that in clinical practice, the label of normal/abnormal is applied to a full EEG session (i.e. a single clinical visit). In clinical practice, experts judge whether the patient exhibits abnormal brain activity based on all the recordings in the session, effectively resulting in a single label for that session. In most recent machine learning approaches, a typical full recording cannot be directly input into the model due to computational constraints – a large input vector length necessitates a large number of parameters in the model. Instead, the recording is divided into smaller windows, with the added advantage of increasing the total number of examples available for training. For training purposes, each window inherits the label of its recording, while evaluation is typically performed on a per-recording basis by aggregating per-window outputs from the classifier. We refer to this downstream aggregation as 'arbitration'.

Western et al. [13] noted that this inheritance of window/recording labels from broader session labels was potentially confounding to the machine learning process. For example, a session may be labelled as 'abnormal' based on several temporally isolated abnormal graphoelements. Many windows in this session may be completely free of abnormal activity, yet they will carry 'abnormal' labels in the training process. These labels are arguably false, depending on whether they are considered to apply to the signal within the window or to the wider session from which it is taken.

In this study, we introduce and evaluate two methodologies to tackle the aforementioned issue: extending

Table 1. Summary of state-of-the-art performance metrics for different models applied to abnormal EEG classification

Model	Accuracy	Sensitivity	Specificity
1D-CNN (T5-O1 channel)[12]	79.3 %	71.4 %	86.0 %
1D-CNN (F4-C4 channel)[12]	74.4 %	55.6 %	90.7 %
Deep4 [1]	85.4 %	75.1 %	94.1 %
TCN [5]	86.2 %		
ChronoNet [9]	86.6 %		
Alexnet[2]	87.3 %	78.6 %	94.7 %
VGG-16 [2]	86.6 %	77.8 %	94.0 %
Fusion Alexnet[8]	89.1 %	80.2 %	96.7 %
[6]	89.8 %	81.3 %	**96.9 %**
Proposed	**93.3 %**	**92.0 %**	92.9 %

the window length, thereby mitigating the impact of misleading window labels, and optimizing the arbitration step through machine learning. This research carries three distinct advantages:

- It substantially mitigates the issue of misleading labels.
- It improves the state-of-the-art performance on the TUAB dataset by 3.5% accuracy as compared to [6]. Additionally, this is achieved without altering the decision threshold, effectively addressing the issue of low sensitivity prevalent in previous models.
- These methodologies bear minimal cost and are largely compatible with most existing models and algorithms, providing a complementary role in the broader context of the field.

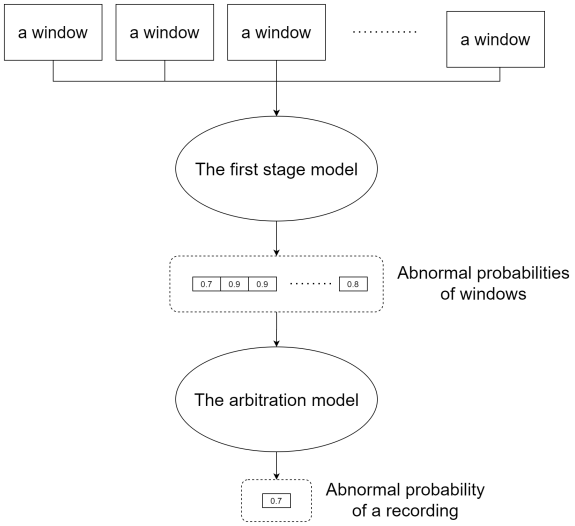


Figure 1. Generic diagram of a typical deep learning approach to clinical EEG classification, as used e.g. by [1]

II. Method

II-A. Data

TUEG [11] is a rich archive of over 30,000 clinical EEG recordings collected at Temple University Hospital (TUH) from 2002 – present, using the standard 10-20 system of electrode placement. TUAB [10] is a subset of 2993 recordings from TUEG that have been labelled as normal/abnormal and divided into training and evaluation sets. The training set contains 1371 normal sessions and 1346 abnormal sessions. The test set contains 150 normal sessions and 126 abnormal sessions. Also, only one file from each session was included in this corpus. All our current results are trained and tested on TUAB. Due to computational cost, all experiments related to window length are carried out on TUAB, including one-stage models, two-stage models and window length. TUAB has already been marked, so we use the original label and its original test-training split. To compare the results with other studies, we followed the pre-processing method in the Deep4 article, which TCN [5] and Fusion Alexnet [8] also used.

II-B. First-Stage Model

A recent study by [5] demonstrated that the Temporal Convolutional Network (TCN) and Deep4 architectures offer near-state-of-the-art performance on TUAB, so we experimented with both of these as the first-stage model. Both are composed of blocks with convolutional and pooling layers. However, TCN replaces common convolution with dilated convolution and introduces a residual structure in the temporal block [14]. To achieve baseline performance consistent with past studies, we use the hyperparameters from [1] for Deep4 and from [5] for TCN.

To explore whether our proposals are applicable to first-stage architectures other than convolutional networks, we also implemented a Vision Transformer (ViT) [15]. For simplicity, the majority of our experiments focussed on Deep4 only. We use Deep4 with the 60s, 180s, 300s, 400s, and 600s windows to perform reproducibility experiments on TUAB. For each Window length, five experiments were performed to avoid the influence of chance factors.

II-C. Second-Stage Models for Arbitration

As shown in Figure 1, the purpose of the arbitration stage is to combine the per-window class probabilities into a single classification of the EEG session. Previous work does not discuss arbitration, although some form of arbitration is inevitable where models are for trained on windows and evaluated on a per-recording basis (e.g. [1, 5]). When looking through the code of Deep4 [1], we found that they used the ‘Mean’ method to integrate the results of windows. In some studies based on time-frequency images, they use a method, such as the Fourier transform[8], to freely choose the size of the image, thus eliminating the need for windowing.

Hence we employ ‘Mean’ as a baseline arbitration model. As alternatives, we explore several implementations of multi-layer perceptrons as the arbitration model. These are distinguished from each other by the pre-processing of the input data (per-window scores) and the specific architecture used. The input pre-processing methods considered are as follows:

Raw: As shown in Figure 2, this approach inputs all the results in each recording directly(‘Raw’). Since the number of windows in each recording is different, padding 0 at the end is required for less than 20 windows data. The value of 20 is chosen here to reflect

the approach of [1, 5], which used 1-minute windows and a maximum of 20 minutes per recording.

Histogram: Being intended as a flexible approach to handling variations in recording length, a histogram (2) was calculated from the per-window scores. The range of potential per-window abnormality scores (0-1) was divided into ten equal-width bins. As shown in Figure 2, the input to our model is represented as a histogram vector, where each element of the vector corresponds to the height of a histogram bin. This histogram captures the distribution of anomaly scores, which range from 0 to 1, across all windows in a recording. For instance, if we have a vector of length 10, we segment the 0-1 range into ten equal intervals, such as 0-0.1, 0.1-0.2, and so on. Each entry in the vector then reflects the count of windows that fall within its corresponding interval. As an example, for a recording with window anomaly scores of 0.8, 0.7, 0.3, and 0.3, the resulting histogram vector would be [0, 0, 0, 2, 0, 0, 0, 1, 1, 0].

Hybrid: Additionally, we considered a hybrid of the 'Raw' and 'Histogram' methods. As shown in Figure 3, in this approach, the 'Raw' and 'Histogram' input forms are concatenated.

The architecture of the arbitration models we propose is a fully-connected layer followed by a softmax layer. We experimented with multi-layer perceptrons of different depths (from 1 layer to 4 layers), the hidden layers of different lengths (from 5 to 20), convolutional layers instead of fully connected layers, activation functions (RELU, ELU, GELU), but these parameters were found not to significantly influence performance.

For each arbitration model architecture and hyperparameter setting, we conduct five experiments on the results of each first-stage model experiment. So, when we consider the two-stage model as a whole, we run $5 \times 5 = 25$ experiments for each architecture and hyperparameter setting.

II-D. Evaluation Method

We focus on accuracy as the most commonly used metric in the preceding reference studies we compare against. Our analysis of sensitivity and specificity provides more detailed insight. We have considered alternatives such as F1 score and ROC. F1 is valuable for unbalanced dataset but provides no additional insight for a well-balanced dataset such as we have used. We did not incorporate the AUC and ROC primarily due to space constraints.

III. Results

III-A. Performance of Our Two-Stage Model

As shown in Figure 4 and 5, all our proposed machine learning arbitration methods outperform the baseline methods (No arbitration' and Mean'), regardless of window length. The highest average accuracy (25 experiments) for the whole two-stage model achieved by any approach was 93.3%, while the highest average accuracy for a single instance of the first-stage model (5 experiments) was 96.2%, both achieved by the 'Hybrid' approach with a window length of 600 s.

When employing a batch size of 64 for the second-stage model (without forcibly increasing the batch size to enhance computation speed), the time cost for a single training session is approximately 2.3 seconds. This provides an insight into the efficiency of our method in terms of computational demands.

In addition, from Figure 5, we can find that the arbitration stage and increased window length both greatly improve the sensitivity of the model with relatively little compromise in specificity.

III-B. Effect of Window Length

As shown in Figure 5, both the performance of the one-stage model ('No arbitration') and the two-stage approaches gets better as the window length increases,

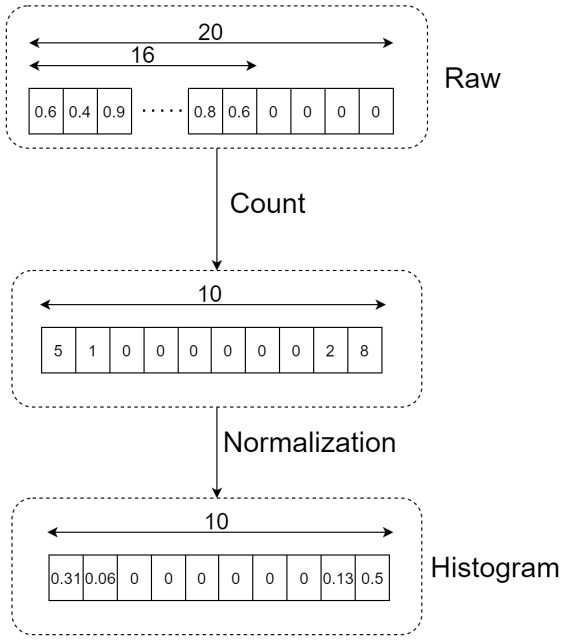


Figure 2. 'Raw' and 'Histogram' pre-processing for the arbitration model. Each small square in 'Raw' is the output of the first-stage model (probability of 'abnormal') for one window. In this example there are 16 windows in the recording. In the general case, since we use the data between 1 and 21 minutes in a recording at most, a recording contains at most 20 windows with a length of 1 minute. When there are fewer than 20, we pad zero at the end. Then we count the 'Raw' into a histogram of ten equal bins across the range 0-10.

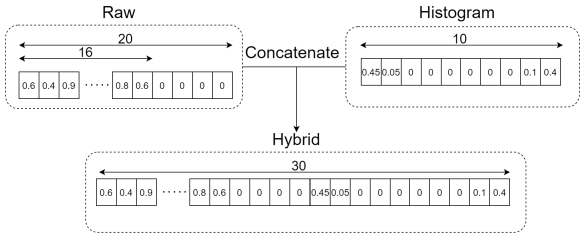
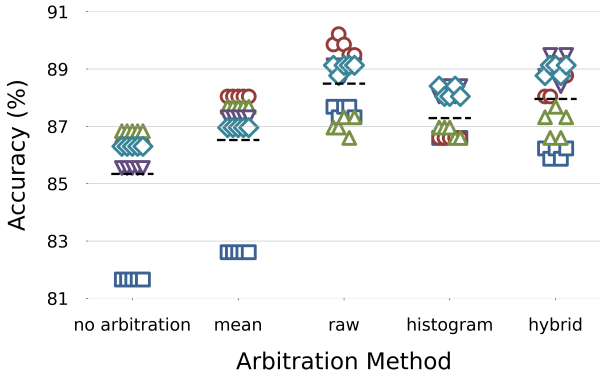
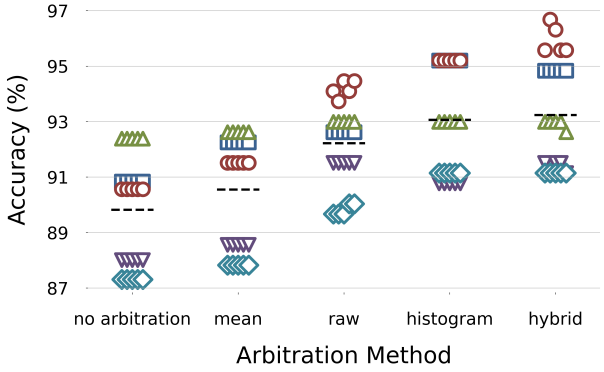


Figure 3. 'Hybrid' pre-processing for the arbitration model.

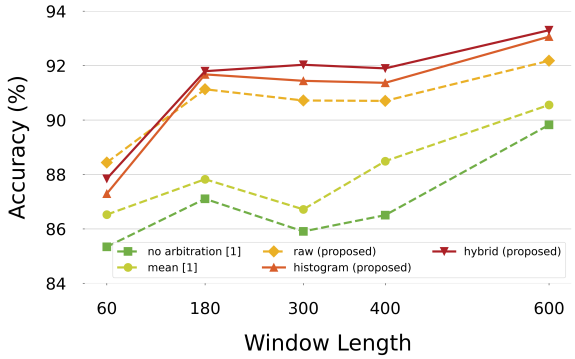


(a) 60 s window length

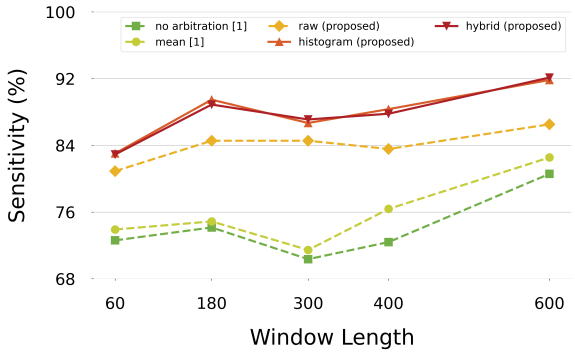


(b) 600 s window length

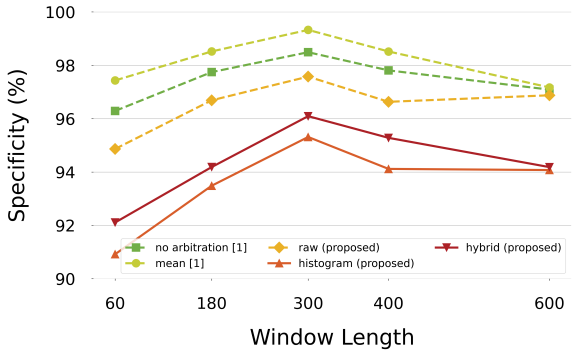
Figure 4. Performance of different arbitration models using window lengths of (a) 60 s and (b) 600 s. Points with the same marker shape come from the same instance of the first-stage model. The dashed lines represent the mean for each arbitration method.



(a) accuracy



(b) sensitivity



(c) specificity

Figure 5. Effect of window length on (a) accuracy, (b) sensitivity, and (c) specificity. Note that the accuracy of the 'no_arbitration' approach is calculated across all windows ($4340 \leq N \leq 57\,482$, depending on window length), whereas the accuracy of the arbitration models is calculated across all recordings ($N = 2993$).

although the performance is not strictly monotonically increasing; all models have the worst performance at 60 s and the best performance at 600 s. For the effect on sensitivity and specificity, we can see similar effects to the two-stage model.

III-C. The Search for the Arbitration Model Architecture

As shown in Table 2, we examined the effect of hidden layer depth and length on the performance of the arbitration model. The results show that they have no significant impact on the model performance, although when the model depth is greater than or equal to three, the model is hard to train. (When the model parameters are initialised in a high loss position, the model will be difficult to train, that is, maintaining a high loss, although when it is initially in a relatively low loss position, the model can reach the same performance as the shallower architectures.). We also experimented with varying the activation function (RELU [16], ELU[17], GELU[18]), but the results show that they did not affect the model performance substantially. Therefore,

we finally chose to use a simple fully-connected layer and a SoftMax layer to form our model to pursue the optimisation of computational efficiency.

IV. Discussion

IV-A. Window Length (Scope)

As noted in Section I-A, it can be argued that EEG window labels inherited from the full recording/session label are often misleading. The smaller the window, the less it represents the wider recording. In particular, a single transient abnormal event may be sufficient for a recording to be labelled as 'abnormal' even when the majority of the windows contain no discernible abnormality. Hence we hypothesised that increasing the window length would improve training (of the first stage model) by making the window labels more accurate.

Figure 5 shows that the performance of the one-stage model ('no_arbitration') roughly increases with increasing window length, which confirms that hypothesis. For a window length of 600 s, the accuracy increases by about 5 percentage points compared with the 60 s window. The performance of the two-stage models show similar trends.

IV-B. The Second-Stage Model (Arbitration)

As seen in Figures 4 and 5, all our proposed variants of the arbitration model improve accuracy substantially compared with the one-stage model ('No arbitration') and baseline arbitration model ('Mean'). 'Mean' offers minor improvement over 'No arbitration': less than two percentage points. As a simple non-parametric algorithm, 'Mean' can mitigate occasional anomalous outputs but cannot learn more complex or finely tuned decisions. All neural network arbitration models outperformed the baseline methods, confirming our hypothesis.

As illustrated in Figure 5, it is evident that with the employment of a substantial temporal window (specifically, one that is greater than or equal to 180 seconds in duration), the Hybrid model consistently outperforms both the Histogram and Raw models in terms of Accuracy. The intrinsic advantage of the hybrid model over the histogram model essentially mirrors the advantage of the raw approach over the histogram approach. During the process of converting raw data into a histogram, all positional information is lost. A hybrid approach is shown to be superior in preserving such crucial details.

Table 2. The effect of hidden layer depth and length on the performance (Accuracy) of the arbitration model

Hidden layer depth	Hidden layer length			
	5	10	15	20
0	0.9330	0.9330	0.9330	0.9330
1	0.9333	0.9342	0.9334	0.9339
2	0.9176	0.9334	0.9328	0.9325
3	0.6686	0.6732	0.6102	0.7045

Comparing panels in Figure 5 indicates that the improvements in accuracy (both for increased window length and the proposed arbitration models) are underpinned by improved sensitivity with relatively little compromise, if any, in specificity. This supports our supposition that, when many small windows inherit an abnormal label from their parent recording, many of the resulting labels are misleading; the window many contain no evident abnormalities, leading to increased 'false negative' results.

As discussed in the Appendix, we experimented with applying the arbitration models to alternative first case architectures. Similar benefits were observed when using a Vision Transformer (ViT), but not when using a Temporal Convolutional Network (TCN).

Figure 4 and 5 show that machine learning arbitration models outperform the baseline methods across all window lengths in our experiments, although the effect is less pronounced at a window length of 60 s.

Although evidence of arbitration stages can be found in the codebase of previous studies such as that of [1], the concept and the selection of the model are not discussed, suggesting they were not considered to be important. We have proved that using a machine learning arbitration method can substantially exceed the baseline performance of the same model.

In some other approaches, such as that of [8], arbitration of classifier outputs is not applicable because a single model fuses features across all windows to achieve a single classifier output for the recording. This approach has sound justification, but the increased complexity of the first stage model poses a challenge for optimisation. The results we present using relatively simple architectures demonstrate substantially greater accuracy. This may simply reflect the ease of achieving relatively thorough optimisation for our approach. Alternatively, the arbitration approach may present some distinct advantage in terms of robustness to transient non-clinical anomalies, which might dominate the decision in an architecture with upstream fusion of features across windows. Confirmation of an explanation for the performance differences between these methods would require a more extensive case-by-case comparison.

IV-C. Label Quality and Performance Ceilings of Machine-Learning-Based Models on EEG binary Classification Problems

Gemein et al. [5] suggested that EEG pathology decoding accuracies observed [1, 9, 19] at approximately 86% were approaching the theoretical optimum imposed by label noise. This suggestion was based on

the observation that inter-rater agreement in the binary classification of EEGs into pathological and non-pathological has been reported as 86–88% [20, 21], although these scores were based on EEG ratings of only two neurologists. In a more recent, broader study, Beuchat et al. [22] found interrater agreement to be even lower, $82-86\%$. However, our study demonstrated that the performance of machine learning-based models in EEG binary classification could be much greater than 86%. Although different raters may give different labels to the same EEG signal, a machine learning model can learn to replicate the judgement of one rater (or team of raters, as used in the curation of TUAB [10]). Now that machine learning approaches can, in some sense, match human expert performance in this task, future work should include the curation of datasets that combine a diverse range of human expert judgements and/or data on clinical outcomes to optimise label accuracy.

IV-D. Future Work

In this study, we explored a limited range of arbitration model architectures to demonstrate the importance of arbitration in windowed EEG classifiers. In immediate future work, we will explore a wider range of arbitration models, such as random forests. Arguably, the inputs to the arbitration model can be thought of as tabular data. Random forests are frequently found to outperform neural networks on tabular data.

It is likely that our pre-processing of the first-stage outputs can also be optimised further. We will explore the use of overlapping windows to increase the resolution of information available to the arbitration model. Further enhancement may be achieved by optimising the binning of the 'Histogram' pre-processing. Rather than using a simple linear spacing of windows, it may be more effective to use narrower bins in ranges with a higher density of samples and wider bins (coarser resolution) elsewhere.

We will also extend the application of arbitration to cases in which the classification task spans multiple recordings from a single clinical visit, using the wider TUEG dataset in combination with automated labelling based on the text reports [13].

In addition to efforts to improve the arbitration stage, we will continue to explore alternative first-stage architectures. Figure 4a, 6a, and 6b indicate the degree of improvement achieved by arbitration varies significantly between different first-stage architectures. It is possible that the best first-stage architecture for use with arbitration is not the same as the best single-stage architecture (for per-window classification, i.e. 'No arbitration'). Furthermore, as we move on from TUAB to the larger TUEG dataset, we may find that data-hungry architectures such as transformers may outperform those that have achieved previous state-of-the-art results on TUAB.

The arbitration principle is likely to be transferable to other time-series applications where a holistic classification is to be applied to a windowed signal. For example, in ECG arrhythmia detection, end-to-end training of architectures with densely connected output layers is common [23], but we are not aware of other cases where this final classification layer is trained separately. Our results suggest that this approach is an effective way to increase the input scope of the system with minimal added computational expense. For cardiac electrophysiology, enabling the application of machine learning classifiers to holistic analysis of long-term Holter recordings could be important for the detection of subtle abnormalities that cannot be discerned from shorter signals.

V. Conclusions

Our proposed approach, combining increased window length and a machine learning arbitration stage, substantially improved upon previous state-of-the-art performance in clinical EEG classification. The results support our premise that the inheritance of window labels from recording labels compromised the sensitivity of previous state-of-the-art solutions. Given the importance of sensitivity for promising applications such as routine screening or accelerating the workflow of human EEG interpreters, this improvement presented here is an important step towards the broader translation of machine learning EEG classifiers into clinical practice. The principles may also be transferable to other time-series classification problems.

Acknowledgements

This work was supported by a PhD studentship funded by Southmead Hospital Charity and the University of the West of England.

References

[1] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human brain mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[2] S. U. Amin, M. S. Hossain, G. Muhammad, M. Alhussein, and M. A. Rahman, "Cognitive smart healthcare for pathology detection and monitoring," *IEEE Access*, vol. 7, pp. 10 745–10 753, 2019.

[3] H. Banville, O. Chehab, A. Hyvärinen, D.-A. Engemann, and A. Gramfort, "Uncovering the structure of clinical eeg signals with self-supervised learning," *Journal of Neural Engineering*, vol. 18, no. 4, p. 046020, 2021.

[4] H. Banville, S. U. Wood, C. Aimone, D.-A. Engemann, and A. Gramfort, "Robust learning from corrupted eeg with dynamic spatial filtering," *NeuroImage*, vol. 251, p. 118994, 2022.

[5] L. A. Gemein, R. T. Schirrmeister, P. Chrabąszcz, D. Wilson, J. Boedecker, A. Schulze-Bonhage, F. Hutter, and T. Ball,

"Machine-learning-based diagnostics of eeg pathology," *NeuroImage*, vol. 220, p. 117021, 2020.

[6] G. Muhammad, M. S. Hossain, and N. Kumar, "Eeg-based pathology detection for home health monitoring," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 2, pp. 603–610, 2020.

[7] N. Wagh and Y. Varatharajah, "Eeg-gcnn: Augmenting electroencephalogram-based neurological disease diagnosis using a domain-guided graph convolutional neural network," *Machine Learning for Health*. PMLR, 2020, pp. 367–378.

[8] M. Alhussein, G. Muhammad, and M. S. Hossain, "Eeg pathology detection based on deep learning," *IEEE Access*, vol. 7, pp. 27 781–27 788, 2019.

[9] S. Roy, I. Kiral-Kornek, and S. Harrer, "Chrononet: a deep recurrent neural network for abnormal eeg identification," *Artificial Intelligence in Medicine: 17th Conference on Artificial Intelligence in Medicine, AIME 2019, Poznan, Poland, June 26–29, 2019, Proceedings 17*. Springer, 2019, pp. 47–56.

[10] S. López, I. Obeid, and J. Picone, "Automated interpretation of abnormal adult electroencephalograms," Ph.D. dissertation, 2017.

[11] I. Obeid and J. Picone, "The temple university hospital eeg data corpus," *Frontiers in neuroscience*, vol. 10, p. 196, 2016.

[12] Ö. Yıldırım, U. B. Baloglu, and U. R. Acharya, "A deep convolutional neural network model for automated identification of abnormal eeg signals," *Neural Computing and Applications*, vol. 32, pp. 15 857–15 868, 2020.

[13] D. Western, T. Weber, R. Kandasamy, F. May, S. Taylor, Y. Zhu, and L. Canham, "Automatic report-based labelling of clinical eegs for classifier training," *2021 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*. IEEE, 2021, pp. 1–6.

[14] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018. (available at: *https://arxiv.org/abs/1803.01271*).

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[16] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 315–323.

[17] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.

[18] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[19] K. Van Leeuwen, H. Sun, M. Tabaeizadeh, A. Struck, M. Van Putten, and M. Westover, "Detecting abnormal electroencephalograms using deep convolutional networks," *Clinical neurophysiology*, vol. 130, no. 1, pp. 77–84, 2019.

[20] E. E. Houfek and R. J. Ellingson, "On the reliability of clinical eeg interpretation," *The Journal of nervous and mental disease*, vol. 128, no. 5, pp. 425–437, 1959.

[21] S. W. Rose, J. K. Penry, B. G. White, and S. Sato, "Reliability and validity of visual eeg assessment in third grade children," *Clinical Electroencephalography*, vol. 4, no. 4, pp. 197–205, 1973.

[22] I. Beuchat, S. Alloussi, P. S. Reif, N. Sterlepper, F. Rosenow, and A. Strzelczyk, "Prospective evaluation of interrater agreement between eeg technologists and neurophysiologists," *Scientific Reports*, vol. 11, no. 1, p. 13406, 2021.

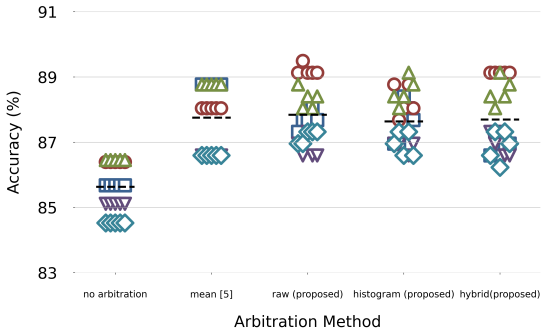
[23] Z. Ebrahimi, M. Loni, M. Daneshtalab, and A. Gharehbaghi, "A review on deep learning methods for ECG arrhythmia classification," *Expert Systems with Applications: X*, vol. 7, p. 100033, Sep. 2020. (available at: *https://www.sciencedirect.com/science/article/pii/S2590188520300123*).

APPENDIX

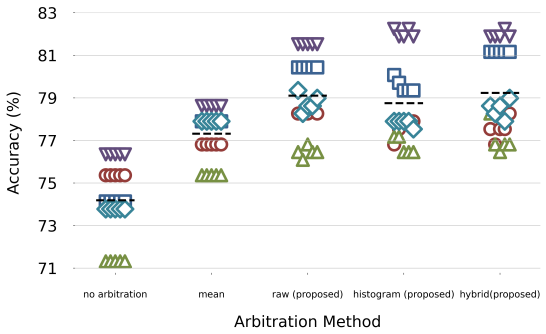
Results on Alternative First-Stage Architectures

As shown in Figure 6, we also explored the effect of the arbitration models on two alternative first-stage architectures: a temporal convolutional networks (TCN) [5, 14] and vision transformer (ViT) [15]. Full details of the implementation and hyperparameter tuning are beyond the scope of this paper, but the implementations are available in our code repository. We present the results briefly here to demonstrate the extent to which our method is transferrable to other first-stage models.

For TCN, the performance of the proposed arbitration models is not substantially different from the baseline ('Mean'). For ViT, our proposed arbitration models can provide about two percentage points of performance improvement. Based on the present evidence, the proposed methods appear to offer a safe improvement in the sense that no cases were observed in which accuracy was substantially worsened. We will test the effect of the arbitration models on a wider selection of first-stage models as well as longer window lengths in future work.



(a) TCN



(b) ViT

Figure 6. Performance of different arbitration methods using (a) TCN and (b) ViT as the first-stage architecture with a window length of 60 s.