Attention Networks for Multi-Task Signal Analysis

David Ahmedt-Aristizabal[1,2], Mohammad Ali Armin[1], Simon Denman[2], Clinton Fookes[2], Lars Petersson [1]

***Abstract*— Recent advances in deep learning have enabled the development of automated frameworks for analysing medical images and signals. For analysis of physiological recordings, models based on temporal convolutional networks and recurrent neural networks have demonstrated encouraging results and an ability to capture complex patterns and dependencies in the data. However, representations that capture the entirety of the raw signal are suboptimal as not all portions of the signal are equally important. As such, attention mechanisms are proposed to divert focus to regions of interest, reducing computational cost and enhancing accuracy. Here, we evaluate attention-based frameworks for the classification of physiological signals in different clinical domains. We evaluated our methodology on three classification scenarios: neurogenerative disorders, neurological status and seizure type. We demonstrate that attention networks can outperform traditional deep learning models for sequence modelling by identifying the most relevant attributes of an input signal for decision making. This work highlights the benefits of attention-based models for analysing raw data in the field of biomedical research.**

I. INTRODUCTION

Modeling physiological observations plays an invaluable role in assessing disease detection and treatment [1]. The abundance of digital clinical data and the need to identifying patterns that are unambiguous has increased research interest in developing applications to learn from multi-modal data [2]. Classical approaches for time-series analysis have been centered around extracting hand-engineered features from time and frequency domains. However, there are challenges related to expert knowledge, irregular sampling and generalisation [2]. In recent years, machine learning models such as convolutional neural networks (CNNs) and recurrent neural networks (RNN) have become popular for multi-modal time series analysis, achieving high detection accuracy for different case studies [3]. One disadvantage of these sequence models is that the structure operates over the entire sample signal, which is inefficient when processing long sequences [4]. The inherent sequential nature makes also parallelization challenging. Additionally, these methods show minimal resilience in the presence of high levels of noise from sensor recordings [5]. Therefore, it is desirable to develop algorithms capable of processing raw signals that automatically learns where to focus the attention. The family of methods that emphasize important task-relevant features of a given signal are termed attention mechanisms [6].

Attention mechanisms are established in neuroscience, but have only recently become effective in sequence-to-sequence modeling tasks [4] such as natural language processing (NLP) [7]. Soft-attention mechanisms (global attention) with a memory based approach (RNNs) can divide the signal into partitions by emphasising important portions. These networks focus on the most relevant parts of the input to make a decision, and suppress uninformative features in the observed data [8]. Attention mechanisms are an integral network component, often placed between encoders and decoders. They have shown promising results for analysis of physiological signals such as electrocardiograms (ECG) [5], phonocardiograms (PCG) [9], and polysomnography (PSG) [10]. On the other hand, recent research in self-attention mechanisms [11] indicates that models that rely entirely on attention computations without using recurrent architectures can achieve similar performance. This structure has been proposed in the area of biomedical text mining [12]. Nevertheless, its application to time series prediction has not been investigated sufficiently [4].

In this paper, we explore the feasibility of adapting attention-based frameworks for analysis of multi-task data and compare the results with baseline methods such as CNNs and RNNs. We demonstrate the potential of these architectures for the classification of neurogenerative disorders, neurological status and seizure type.

Our main contributions are summarized as follows:

1) We compare and introduce multiple deep learning models including CNNs, LSTMs, soft- and self-attention mechanisms for the purpose of classifying raw physiological signals.
2) We show the effectiveness of attention mechanisms for mapping discriminative features across sequential data.

II. METHODOLOGY

In this paper, we conduct a systematic evaluation of convolutional and recurrent architectures commonly used for sequence modeling. Then, we introduce attention-based frameworks to analyse the time series signals. We aim to determine if the success of attention networks for classifying physiological signals is confined to specific application domains. In this study, we design experiments that use raw signals and eschew handcrafted features and preprocessing phases (*i.e.* obtaining image-based representations). We also compare the performance of adapted baseline methods using our dataset configuration, experimental plan and evaluation metric. A block diagram of the proposed methods is displayed in Fig. 1.

A. Sequence modeling with CNNs and RNNs

1) Temporal Convolutional Networks (TCNs): Recent research shows that variations of convolutional neural networks can achieve impressive results for sequential data.

[1] CSIRO, DATA61, Canberra, Australia. Corresponding author: david.ahmedtaristizabal@data61.csiro.au

[2] Image and Video Research Laboratory, SAIVT, Queensland University of Technology, Brisbane, Australia.

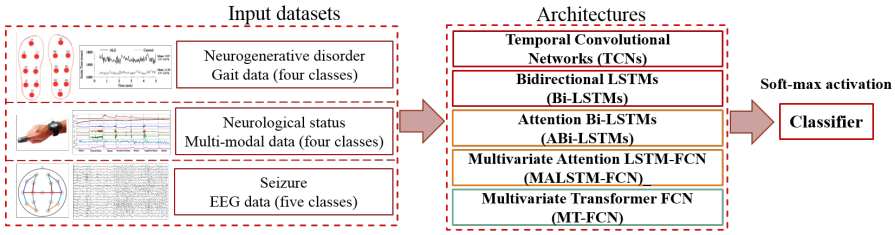

Fig. 1. Overview of the models for classifying raw physiological signals: sequence modeling with CNNs and RNNs, and attention-based frameworks.

The present architectures, temporal convolutional networks (TCN) [13], use dilated causal convolution layers where an output at time t is convolved only with elements from time t or earlier in the previous layer, *i.e.* inputs have no influence on output steps that proceed them in time. In a dilated convolution layer, a filter is sequentially applied to inputs by skipping input values with a certain step (dilatation rate). This allows the network to consider temporal order and capture long-term dependencies without an explosion in model complexity [14]. In our experiments, the baseline model is based on the wavenet implementation [15].

2) Bidirectional LSTMs (Bi-LSTMs): Recurrent neural networks such as Long Short Term Memory (LSTM) [16] have proven to be stable in modeling dependencies in sequential data by employing an external memory cell state. Bidirectional LSTMs (Bi-LSTMs) [17] extend traditional LSTMs and can improve model sequence classification performance by training two LSTMs on the input sequence. This can provide additional context to the network and result in faster and more complete learning [18]. The hidden state dimension of the encoder LSTMs in our baseline implementation is determined empirically and is set to 60 units for each dataset. The Bi-LSTM layer is followed by a classification layer.

B. Attention-based frameworks

1) Bidirectional LSTMs with attention (ABi-LSTM): Soft-attention mechanisms are end-to-end approaches that can be learned by gradient-based methods [7], [8]. They have been proposed to further improve an encoder-decoder performance for domain-specific applications by directing emphasis to different parts of the encoder output in each step of decoding. Here, we implement a Bi-LSTM to capture temporal information from sequences that consider the previous and future input information simultaneously. The attention network [8] allows the model to learn the most relevant parts of the input sequence during training. The Bi-LSTM encodes a feature vector from the raw signal into a hidden representation h_t. We leverage attention mechanisms to capture the attributes of a signal that influence the decision, and then form a dense vector by considering the weights of different input vectors. This can be formulated as follows:

$$\begin{aligned} u_t &= \tanh(Wh_t + b), \\ \alpha_t &= \frac{\exp(u_t^T u_w)}{\sum_{j=1}^{n} \exp(u_t^T u_w)}, \\ s_t &= \sum_t \alpha_t h_t, \end{aligned} \tag{1}$$

where h_t is the concatenation output of the Bi-LSTM model and W and b are the weight and bias of a single multilayer perceptron (MLP), respectively. We measure the importance of each element in h_t by estimating the similarity between u_t and h_t, which is randomly initialized. Then, we infer a normalized importance weight α_t from a softmax function. Finally, these scores are multiplied by the hidden states to calculate the weighted combination s_t. The attention layer is followed by dense and classification layers. Dropout is used between these layers to prevent potential overfitting.

2) Multivariate LSTM-FCN with attention (MALSTM-FCN): Fully convolutional networks (FCN), comprised of a TCN, are typically used as feature extractors. Augmented FCNs with attention LSTMs have dramatically improved the performance on univariate time series classification tasks [19]. Here, we adopt a multivariate attention LSTM-FCN to enhance the analysis of raw physiological signals [20]. The FCN consists of three stacked temporal convolutional blocks with a global average pooling after the final convolution block. The first two convolutional blocks conclude with an attention mechanism known as squeeze-and-excitation block [21]. This block allows the network to perform feature recalibration, by which it can learn to use global information to selectively emphasise informative features and suppress less useful ones. The LSTM block comprises an attention LSTM layer [7]. The output of the FCN and the LSTM block is concatenated and passed to a softmax classification layer.

3) Multivariate Transformer-FCN (MT-FCN): The transformer [11] follows the encoder-decoder paradigm and has demonstrated an ability to capture temporal dependencies. It is solely based on self-attention (intra-attention) and computes representations without using sequence-aligned RNNs. In self-attention, the queries, keys, and values (Q, K, and V) are all created using encodings of the sequence [4]. Here, we modify the MALSTM-FCN [20] model by changing the attention LSTM block to a transformer model. This block combines a 1D convolutional layer to compute an input embedding, multi-head attention, and a fully connected feed-forward network. By packing Q, K, and V, together in a matrix, the output of the self-attention layer is computed by,

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{n}}V), \tag{2}$$

where n is the input sequence length. The output is a weighted sum of the values, where the weight assigned to each value is determined by the dot-product of the query with all the keys [11]. The multi-head mechanism runs through

the scaled dot-product attention multiple times in parallel. Therefore, for each Head_i (8 in this work), the attention is

$$\text{Head}_i = \text{Attention}(Qw_i^Q, KW_i^K, VW_i^v). \quad (3)$$

These attention functions are concatenated and projected resulting in the multi-head attention output,

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{Head}_i, ..., \text{Head}_h)W^O, \quad (4)$$

This module allows an attention mechanism to concentrate different parts of the input. Further technical information on the multi-head attention module can be found in [4], [11].

III. EVALUATION

A. Datasets and experimental setup

To evaluate the performance of each proposed method, we selected publicly available datasets with particular characteristics: multi-modal and multi-channel data.

1) Neurogenerative disorder: We distinguish neurogenerative disorders by analysing the gait cycle. Gait dynamics from 15 patients with Parkinson's disease, 20 with Huntington's disease, 13 with amyotrophic lateral sclerosis and 16 healthy control subjects were recorded [22], [23]. The raw data was obtained while the participant was walking at their usual pace along a 77 meter long hallway for 5 minutes, with accelerometers sampling at 300Hz placed in the subject's left and right shoes. Raw data per patient is split into sequences of one second. Thus each sample has a dimension of $[2 \times 300]$. Then, we combine all samples across patients with the same disorder.

2) Neurological status classification: The database contains non-EEG physiological signals for the assessment of induced stress [23], [24]. We aim to distinguish responses of 20 healthy participants by analysing physiological signals collected from two wrist-worn biosensors while performing the following tasks in order: relaxation (5min), physical stress (5min), relaxation (5min), cognitive stress (3min), relaxation (5min), emotional stress (5min) and relaxation (5min). A Nonin 3150, sampling at 1Hz, recorded the arterial oxygen level (SpO2) and heart rate (HR), and an Affectiva sensor sampling at 8Hz sensed the electrodermal activity (EDA), temperature and acceleration. As such, the dataset contains 7 channels of multi-modal data. We split the recording of each participant according to the label of each neurological status and concatenate all samples of the same class across all participants. We consider only the first session of relaxation to avoid class imbalance. All data is re-sampled to 1Hz to ensure uniform input dimension and each variable is independently normalized using a min-max scaling. A sliding window of 20 seconds is used to create samples, with each sample being of dimension $[7 \times 20]$.

3) Seizure type classification: We use the most recent release of the TUH EEG seizure corpus (v1.5.0) [25]. The seizure classes are simple partial seizure (n=52), complex partial seizure (n=361), absence seizure (n=99), tonic seizure (n=68) and tonic-clonic seizure (n=60). The general classes, focal non-specific (FN) and generalized non-specific (GN) seizures, were not considered in this analysis as in these recordings the seizure origin was not precisely identified. The experimental dataset is created by combining the official training and test set with non-overlapping patients, resulting in a total of 640 seizures. One class is defined as the combination of all seizure recordings for the same seizure type. Data was re-sampled to 200 Hz from the following 19 common EEG channels: FP1, F7, T3, T5, O1, F3, C3, P3, FP2, F8, T4, T6, O2, F4, C4, P4, FZ, CZ, PZ. EEG sequences are created from the raw EEG signal by a) sliding a one second window over the signals to obtain clips, b) concatenating all clips into an EEG sequence, and c) making all samples the same sequence length (200) by padding with zeros or truncating. Thus, each EEG sequence has a dimension of $[200 \times 19 \times 200]$.

TABLE I

EVALUATION RESULTS ON THE EXPERIMENTAL DATASETS

Dataset	Methods	F1-score	Adapted baseline Methods	F1-score
Neurogenerative disorder (Four classes)	TCN	0.832	LSTM (based on [26])	0.825
	BiLSTM	0.886	LSTM-DNN (based on [27])	0.846
	ABiLSTM	0.902		
	MALSTM-FCN	0.924		
	MT-FCN	0.914		
Neurological status (Four classes)	TCN	0.934	CNN+FC (based on [28])	0.882
	BiLSTM	0.928	MLP (based on [29]	0.742
	ABiLSTM	0.940		
	MALSTM-FCN	0.988		
	MT-FCN	0.890		
Seizure type (Five classes)	TCN	0.942	CNN (based on [30]	0.875
	BiLSTM	0.948	CNN-LSTM (based on [31]	0.923
	ABiLSTM	0.954		
	MALSTM-FCN	0.964		
	MT-FCN	0.961		

B. Evaluation metric and implementation

All models were evaluated on each dataset with 80% of data samples allocated for training, 10% for validation, and 10% for testing, with a 10-fold cross-validation (CV). We tested multiple window length to split each raw signal and adopted the best one based on the validation set. We evaluate the performance using the F1 score which takes both false positives and false negatives into account [9], [10]. Categorical cross-entropy loss and the Adam optimizer [32] (learning rate=0.002, beta1= 0.9, beta2= 0.999) are used to train the models. Models are trained for 100 epochs with a mini-batch size of 32. Training terminates early if validation loss does not improve for over 15 epochs. All models were implemented in Keras [33] with a Tensorflow backend.

C. Experimental results and discussion

An evaluation of all proposed models and baseline methods for each dataset is shown in Table I. The results showed improved performance of all attention-based frameworks in comparison to convolutional and recurrent networks. The MALSTM-FCN model, which uses two blocks with different soft-attention mechanisms [7], [21], showed the best F1-score on all datasets, and the proposed methods do not rely on the accuracy of frequency domain representation of raw signals. We note that the transformer model (MT-FCN) achieved similar results to the soft-attention mechanism, but its performance is limited when encountering a small input sequence (*e.g.*stress classification). The TCN architecture can reach a similar accuracy to the BiLSTM for analysing multi-modal signals. However, with a time series that has a large number of steps, using causal convolutions to learn from the

entire history makes the model computationally complex. On the other hand, a critical disadvantage of encoder-decoder Bi-LSTM is that they focus more on the recent history. For this reason, incorporating attention mechanisms aims to resolve issues caused by noisy multivariate time series data, which is common in physiological data. The main disadvantage of the attention mechanism is the increase in model parameters, which increases training time, especially if the input sequences are long. Although self-attention mechanisms are popular in the NLP domain, their adaptation to time series classification has remained limited. This is likely due to the difficulty of defining how the query, key and values are formed in different domains. Overall, one key aspect of attention mechanisms is that they simultaneously give more weight to related parts of each input sequence and consider the whole recording to extract consecutive dependencies. This can be useful when distinguishing contact relationships in the evaluation of seizure type propagation.

An interesting direction for future research is that graph structures may have more capacity to encode complicated pair-wise relationships between signals. As such attention-based architectures with graph-structured data [34] merit investigation for use with complex physiological recordings such as intracranial EEG and *f*MRI.

IV. CONCLUSIONS

We introduce and compare convolutions and recurrent structures with attention-based frameworks for modelling spatiotemporal dependencies in raw physiological signals. Our analysis showed that attention-based models outperform RNN and CNN-based models when applied to multi-modal data such as gait, SpO2, HR, EDA, temperature, acceleration and EEG recordings for the purpose of neurogenerative disorder, neurological status and seizure type classification.

REFERENCES

[1] O. Faust and M. G. Bairy, "Nonlinear analysis of physiological signals: a review," *Journal of Mechanics in Medicine and Biology*, vol. 12, no. 04, p. 1240015, 2012.

[2] O. Faust, Y. Hagiwara, T. J. Hong, O. S. Lih, and U. R. Acharya, "Deep learning for healthcare applications based on physiological signals: A review," *Computer methods and programs in biomedicine*, vol. 161, pp. 1–13, 2018.

[3] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (eeg) classification tasks: a review," *Journal of neural engineering*, vol. 16, no. 3, p. 031001, 2019.

[4] H. Song, D. Rajan, J. J. Thiagarajan, and A. Spanias, "Attend and diagnose: Clinical time series analysis using attention models," in *AAAI*, 2018.

[5] S. Mousavi, F. Afghah, A. Razi, and U. R. Acharya, "Ecgnet: Learning where to attend for detection of atrial fibrillation with deep visual attention," in *BHI*, 2019, pp. 1–4.

[6] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015, pp. 2048–2057.

[7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *ICLR*, 2015.

[8] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *NAACL HLT*, 2016, pp. 1480–1489.

[9] T. Fernando, H. Ghaemmaghami, S. Denman, S. Sridharan, N. Hussain, and C. Fookes, "Heart sound segmentation using bidirectional lstms with attention," *IEEE journal of biomedical and health informatics*, 2019.

[10] S. Mousavi, F. Afghah, and U. R. Acharya, "Sleepeegnet: Automated sleep stage scoring with sequence to sequence deep learning approach," *PloS one*, vol. 14, no. 5, 2019.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[12] K. Jiang, T. Chen, R. A. Calix, and G. R. Bernard, "Prediction of personal experience tweets of medication use via contextual word representations," in *EMBC*, 2019, pp. 6093–6096.

[13] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *CVPR*, 2017, pp. 156–165.

[14] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[15] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[16] S. Hochreiter and J. Schmidhuber, "Lstm can solve hard long time lag problems," in *NeurIPS*, 1997, pp. 473–479.

[17] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[18] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[19] F. Karim, S. Majumdar, H. Darabi, and S. Chen, "Lstm fully convolutional networks for time series classification," *IEEE access*, vol. 6, pp. 1662–1669, 2017.

[20] F. Karim, S. Majumdar, H. Darabi, and S. Harford, "Multivariate lstm-fcns for time series classification," *Neural Networks*, vol. 116, pp. 237–245, 2019.

[21] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018, pp. 7132–7141.

[22] J. M. Hausdorff, A. Lertratanakul, M. E. Cudkowicz, A. L. Peterson, D. Kaliton, and A. L. Goldberger, "Dynamic markers of altered gait rhythm in amyotrophic lateral sclerosis," *Journal of applied physiology*, vol. 88, no. 6, pp. 2045–2053, 2000.

[23] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[24] J. Birjandtalab, D. Cogan, M. B. Pouyan, and M. Nourani, "A non-eeg biosignals dataset for assessment and visualization of neurological status," in *SiPS*, 2016, pp. 110–114.

[25] V. Shah, E. Von Weltin, S. Lopez de Diego, J. R. McHugh, L. Veloso, M. Golmohammadi, I. Obeid, and J. Picone, "The temple university hospital seizure detection corpus," *Frontiers in Neuroinformatics*, vol. 12, p. 83, 2018.

[26] A. Zhao, L. Qi, J. Li, J. Dong, and H. Yu, "Lstm for diagnosis of neurodegenerative diseases using gait data," in *ICGIP*, vol. 10615, 2017, p. 106155B.

[27] G. Paragliola and A. Coronato, "Gait anomaly detection of subjects with parkinson's disease using a deep time series-based approach," *IEEE Access*, vol. 6, pp. 73 280–73 292, 2018.

[28] A. Jafari, A. Ganesan, C. S. K. Thalisetty, V. Sivasubramanian, T. Oates, and T. Mohsenin, "Sensornet: A scalable and low-power deep convolutional neural network for multimodal data classification," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 66, no. 1, pp. 274–287, 2018.

[29] A. Arsalan, M. Majid, S. M. Anwar, and U. Bagci, "Classification of perceived human stress using physiological signals," in *EMBC*, 2019, pp. 1247–1250.

[30] N. Sriraam, Y. Temel, S. V. Rao, P. L. Kubben *et al.*, "A convolutional neural network based framework for classification of seizure types," in *EMBC*, 2019, pp. 2547–2550.

[31] M. Golmohammadi, S. Ziyabari, V. Shah, E. Von Weltin, C. Campbell, I. Obeid, and J. Picone, "Gated recurrent networks for seizure detection," in *SPMB*, 2017, pp. 1–5.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[33] F. Chollet *et al.*, "Keras," 2017.

[34] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *ICLR*, 2018.