Deep Learning for Time-Series Analysis of Sleep State Detection in Wrist-Worn Devices

Armando Bringas-Corpus
Escuela de Ingeniería y Ciencias
Tecnológico de Monterrey
Querétaro, México
Email: a01200230@tec.mx

***Abstract*—This investigation proposes a project that utilizes deep learning techniques to address the complex task of detecting sleep states using accelerometer data from wrist-worn devices. This research include a review and critical evaluation of various deep learning architectures, comparing them with other machine learning strategies to determine the most effective approach. With an initial focus on examining state-of-the-art methods, central to this proposal is the development of a deep learning solution for analyzing and predicting time-series data from these wrist-worn devices to accurately identify sleep states. The overarching aim is to implement a reliable predictive model that significantly enhances the precision and dependability of sleep state classification using data from wearable technology.**

I. Introduction

This paper presents a project proposal rooted in the practical application of a Kaggle competition, aimed at addressing a challenge in pediatric health and neuroscience: the identification of sleep states in children through wrist-worn accelerometer data. The competition serves as a catalyst for innovation, prompting the development of a sophisticated machine learning model capable of discerning the subtle onset of various sleep states and periods of wakefulness with precision.

The significance of this project lies in its potential to deepen our understanding of sleep and to provide further insights into its importance. For instance, understanding how environmental factors influence sleep, mood, and behavior can aid in formulating personalized strategies tailored to the unique needs of each child [1]. Furthermore, the outcomes of this project could enable researchers to undertake more comprehensive, large-scale sleep studies across a variety of populations and contexts, which could yield even more valuable information about sleep [1].

II. Literature Review

For this type of problem, a Machine Learning approach based on Random Forest has proven effective in detecting sleep-wake states, non-wear versus wear, and sleep stage classification [2]. An initial attempt was made using a Residual Neural Network (RNN), specifically a ResNet initialized with a Glorot uniform initializer and employing LeakyReLU activations [2]. However, when testing this approach with the Amsterdam dataset, which consists of data collected from 114 individuals recruited by the VU University Medical Center in Amsterdam, The Netherlands [3], it was observed that the Random Forest approach outperformed the ResNet heuristic. Notably, the ResNet had difficulties with wake state prediction [2]. However, it is important to remark that most the used data was collected with the GENActiv accelerometer brand, it should be consider future studies to assess model transferability across other accelerometer brands [2].

We explored another approach that analyzes data from the VU University Medical Center in Amsterdam, The Netherlands [3]. Unlike machine learning methods, this study employed Latent Class Analysis (LCA), a statistical method rather than a machine learning or deep learning technique. LCA, commonly used in social, psychological, and behavioral sciences, identifies latent classes within populations sharing common characteristics [4]. In their study, LCA was utilized to differentiate subtypes of sleep misperception among individuals using data from actigraphy and sleep diaries from the people in their studies [3]. Sleep misperception, in this context, is the discrepancy between subjectively perceived and objectively measured sleep, crucial for understanding disorders like Insomnia Disorder [3]. This study, focusing on a statistical categorical approach to classify sleep misperception, falls outside our scope.

Also, it is important to note that the analyzed information for sleep state detection can come from different types of devices, from example from an optical plethysmography and accelerometer signals [5] [6], depending on the research. Specifically, there are cases in which the data is extracted from accelerometers among other types of sensors. Therefore, the methods utilized in each case serve as a baseline for their respective projects, this means that couldn't be possible a suitable that generalize over the data coming from different sources.

Understanding the clinical implications of devices employed for sleep state detection is crucial. Accelerometry is frequently utilized as a cost-effective method for assessing sleep states. However, it appears that traditional machine learning algorithms may have limitations in accuracy, particularly in pa-

tients with insomnia [2]. There exists both an opportunity and a growing interest in refining these algorithms by enhancing the accuracy of accelerometry could transform it into a more clinically valuable tool, enabling the measurement of sleep and wakefulness over prolonged periods [2].

Despite evidence suggesting that classical methods, such as Support Vector Machines, effectively detect sleep states and may outperform Deep Learning methods in specific scenarios with data from actigraph devices that recorded signals using a microelectromechanical system (MEMS) accelerometer [3], similar to the GENActiv accelerometer [2], were analyzed., we explored alternative approaches. Our investigation revealed the deployment of Deep Learning and other Machine Learning models for detecting sleep states from various source signals. These include electrocardiography (ECG), which measures the heart's electrical activity in combination of respiratory and movement signals [7], or electroencephalogram (EEG), employing electrodes on the scalp to monitor brain electrical signals [8]. Other sources involve a combination of optical plethysmography and accelerometer signals [5] [6], and more complex inputs like multi-channel polysomnogram (PSG) [9]. PSG encompasses diverse physiological continuous-time signals from multiple sensors, including EEG, electrooculography (EOG) for eye movement, electromyography (EMG) for muscle contractions, and monitors for respiration and body oxygen levels [9].

Although the different source signals, overally, they shared in common the complexity, high-dimensionality, noise and temporo-spatial dependency structure of these data types that make them suitable for analysis using deep learning models with their respective challenges [10]. A potential challenge could yield with working with raw signals introduced with the network without doing feature extraction [7]. From the work of [11] they compare different methods and algorithms based on data from heart rate and wrist actigraphy, but they did low-level and mid-level feature extraction before introducing to the network, the best results was obtained from a Long Short-Term Memory LSTM network for 2-class classification with an accuracy of 83.56% [11].

A Long Short-Term Memory network (LSTM) is a specialized type of Recurrent Neural Network (RNN). It's important to note that RNNs are a natural choice for time-series forecasting and prediction [12]. LSTM enhances the RNN architecture by modifying the recurrence conditions of how the hidden states are propagated [12]. Additionally, it addresses the challenges of vanishing and exploding gradients, which are critical issues in neural network training [11], [12].

In their research, Sekkal et al. [8] utilized a LSTM network for analyzing the Physionet Sleep-EDF Database, containing 153 polysomnograms (PSGs) from two-night recordings. Feature extraction was performed to refine the data within the PSG signals. The study's relevance stems from its comparison of classical machine learning techniques with deep learning approaches, notably focusing on LSTM due to its proficiency in identifying temporal relationships between sleep stages [8]. In their bi-directional LSTM classifier implementation, an accuracy of 87.8% was achieved on the PSG (Fpz-Cz + Pz-Oz + EOG) classifier. Contrasting this, Fraiwan & Alkhodari [13] reported a peak accuracy of 97.1% using a bi-directional LSTM with a single channel. However, their training set showed significant class imbalance, with the 'awake' stage constituting 68% of the epochs. This imbalance likely skewed the classifier's performance, favoring the more prevalent sleep stage.

The bi-directional LSTM classifier model, referenced in the studies by Sekkal et al. [8] and Fraiwan & Alkhodari [13], represents an enhancement of the standard LSTM. This variant eliminates the one-step truncation inherent in the original LSTM design. Bidirectional training possesses an architectural advantage over unidirectional training if used to classify phonemes [14], it incorporates a comprehensive error gradient calculation, enabling training through standard backpropagation through time (BPTT) [15].

Is important to remark that Fiorillo et al. [16] present a systematic and comprehensive review of deep learning algorithms applied to sleep scoring. The paper thoroughly examines the latest applications and compares them to other methodologies, it concluded that deep learning methods offer numerous advantages. Notably, these algorithms can be applied directly to raw data in comparison with other approaches that involved feature extraction [7] [11], therefore requiring minimal artifact removal. Furthermore, they are capable of unveiling hidden information that traditional feature-based approaches might overlook.

Finally, in the work of [7] they implemented a LSTM architecture with the raw signals from EEG without any prior feature extraction. The preprocessing involved reducing noise and smoothing the signal using a median filter and, in some cases, applying signal flipping or baseline subtraction. LSTM were designed to learn features directly from these raw, preprocessed signals without any feature extraction step, they achieved an accuracy of 80% in 5-class classification.

To address our particular challenge, we propose initially implementing an LSTM architecture without feature extraction. This approach is selected based on revisited literature that emphasizes its effectiveness in capturing temporal dependencies in sequential data, a key characteristic of our problem domain. Subsequent stages of our research will involve evaluating the performance of this model and discussing potential enhancements. These may include the integration of feature extraction techniques or the exploration of other Deep Learning architectures, as needed.

III. Future Directions

The ultimate goal is to apply a developed Deep Learning architecture to the dataset provided by [1]. Our primary motivation lies in the use of deep learning methods for time series analysis, particularly with the "Detect Sleep States" dataset furnished by the Child Mind Institute [1].

A. Data

The dataset consists of approximately 500 multi-day recordings from wrist-mounted accelerometers. The accelerometer

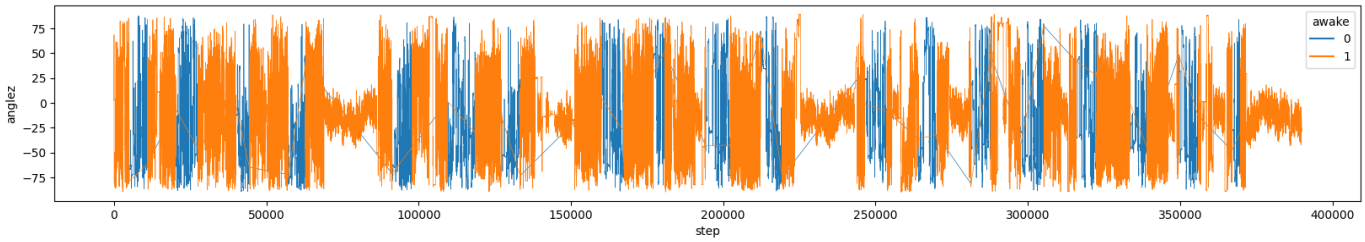

Fig. 1. Accelerometer training data. Plot of `step` againts `anglez` showing `event` state

data in the dataset was processed using R with the GGIR package [17]. The recordings are labeled with two event types: 'onset', indicating the start of sleep, and 'wakeup', marking its end. The primary objective is to identify these two events within the accelerometer data series, this primarily represents a binary classification task. As shown in Figue 1 The data represents series of continuous recording of the accelerometer data for a single subject spanning many days, where we can find:

- `series_id`: Unique identifier for each accelerometer series.
- `step`: An integer timestep for each observation within a series. It is uniqute within a series
- `event`: The type of event, whether onset or wakeup.
- `anglez`: As calculated and described by the GGIR package, z-angle is a metric derived from individual accelerometer components that is commonly used in sleep detection, and refers to the angle of the arm relative to the vertical axis of the body [17].
- `enmo`: As calculated and described by the GGIR package, ENMO is the Euclidean Norm Minus One of all accelerometer signals, with negative values rounded to zero. While no standard measure of acceleration exists in this space, this is one of the several commonly computed features [17].

B. Model

We are planning to implement an architecture that incorporates an LSTM cell into our sequence of input data. In this cell architecture, the intermediate variables $\overline{i}, \overline{f}, \overline{o}$, corresponding to the *input*, *forget*, and *output* gates, play crucial roles in updating the cell and hidden states. The determination of the hidden state vector $\overline{h}_t^{(k)}$ and the cell state vector $\overline{c}_t^{(k)}$ is a multi-step process that starts by computing the intermediate variables, followed by the computation of the hidden states from these intermediates [12]. The updates are as follows:

[Setting up intermediates]

$$\begin{matrix}\text{Input Gate:}\\ \text{Forget Gate:}\\ \text{Output Gate:}\\ \text{New C.-State:}\end{matrix}\begin{bmatrix}\overline{i}\\ \overline{f}\\ \overline{o}\\ \overline{c}\end{bmatrix} = \begin{bmatrix}\text{sigm}\\ \text{sigm}\\ \text{sigm}\\ \tanh\end{bmatrix} W^{(k)} \begin{bmatrix}h_t^{(k-1)}\\ h_{t-1}^{(k)}\end{bmatrix}$$

[Selectively forget and add to long-term memory]

$$\overline{c}_t^{(k)} = \overline{f} \odot \overline{c}_{t-1}^{(k)} + \overline{i} \odot \overline{c}$$

[Selectively leak long-term memory to hidden state]

$$\overline{h}_t^{(k)} = o_t \odot \tanh(\overline{c}_t^{(k)})$$

Our neural network architecture employs an LSTM to incorporate the classifications of previous inputs into the current sample's classification. This approach is suitable to looking back to inform the present, a method particularly suited for cyclical patterns like sleep stages [7], which recur throughout the night. Therefore, LSTMs, with their recurrent nature, could be an ideal choice for this type of cyclic temporal problems.

As shown in Figure 2 We are proposing a starting neural network architecture with the following blocks where consists of a LSTM layer of 64 units, ideal for processing sequences by capturing dependencies from prior inputs. This is followed by a Dense layer, the size of which matches the number of classification categories in your problem, in this case for our binary classification problem $n_classes = 2$. The final component is a softmax activation function, applied to convert the output into a probability distribution across the predicted classes.

$$\xrightarrow{\text{input}} \boxed{\text{LSTM}(64)} \rightarrow \boxed{\text{Dense}(n_{\text{classes}})} \xrightarrow{\text{softmax}}$$

Fig. 2. Neural Network for accelerometer data classification

Is important to check if the impact to work directly with raw signals, the effects of the noise and the implications of not doing feature extraction do not affect performance in the training of the model. Subsequently, with the results of the first trials we might plan work more in data pre-processing if needed or even transition to more complex architectures. Other point of consideration is that one of the common problems is that the time-series sequences can be extremely long and therefore can be certain limitations with its performance [12].

C. Evaluation

To evaluate model classification, predictions that align with the ground truth and exceed the threshold are labeled as True Positives (TP). Predictions that do not match are labeled as False Positives (FP), while ground truths without a corresponding prediction are labeled as False Negatives (FN). Where,

$$\text{Score}(x) = \begin{cases} \text{TP} & \text{if } x \text{ matched and } x > \text{thresh.} \\ \text{FP} & \text{if } x \text{ unmatched pred.} \\ \text{FN} & \text{if } x \text{ unmatched truth} \\ \text{TN} & \text{otherwise} \end{cases}$$

It is necessary to compute the average precision [1]. By collecting the events within each `series_id` from the data the Average Precision score will be computed for each event $\times$ tolerance group [1]. The Average Precision is a measure that combines recall and precision for ranked retrieval results, for one information need, the average precision is the mean of the precision scores after each relevant document is retrieved [18]. Average Precision is defined as:

$$\text{AP} = \sum_n (R_n - R_{n-1}) P_n,$$

where P_n and R_n are the precision and recall at the nth threshold. With random predictions, the AP is the fraction of positive samples, the value from this function ranges from 0 and 1 and higher is better computed from the prediction scores, [19] [20]. Precision and Recall are computed as follows:

$$\text{precision} = \frac{tp}{tp + fp}, \quad \text{recall} = \frac{tp}{tp + fn}$$

IV. Conclusion

To address this challenge, we want to continue over implementing deep learning techniques, despite evidence from [2] indicating that traditional methods such as Random Forest outperform ResNets. Our motivation stems from the insights gained from [21], where the authors present the Time-Series Neural Network (TNN). This method, which incorporates a Kernel Filter alongside a Time Attention Mechanism [21], has demonstrated high accuracy in forecasting, will be interesting to evaluate their perfomance on classification. In our specific context, we are in the process of evaluating whether Long Short-Term Memory Networks (LSTM), Graph Neural Networks (GNN), or potentially an attention-based mechanism would be most suitable for our application.

Acknowledgment

This project is being proposed during the Machine Learning Fall 2023 course with the Tsinghua University as part of the hybrid classroom experience in collaboration with Alexis Guerrero (alexis.guerrero@ug.uchile.cl) from the Universidad de Chile.

References

[1] N. Esper, M. Demkin, R. Hoolbrok, Y. Kotani, L. Hunt, A. Leroux, V. van Hees, V. Zipunnikov, K. Merikangas, M. Milham, A. Franco, and G. Kiar, "Child mind institute - detect sleep states," 2023. [Online]. Available: https://kaggle.com/competitions/child-mind-institute-detect-sleep-states

[2] K. Sundararajan, S. Georgievska, B. H. W. te Lindert, P. R. Gehrman, J. Ramautar, D. R. Mazzotti, S. Sabia, M. N. Weedon, E. J. W. van Someren, L. Ridder, J. Wang, and V. T. van Hees, "Sleep classification from wrist-worn accelerometer data using random forests," *Scientific Reports*, vol. 11, no. 1, p. 24, 2021. [Online]. Available: https://doi.org/10.1038/s41598-020-79217-x

[3] B. H. W. te Lindert, T. F. Blanken, W. P. van der Meijden, K. Dekker, R. Wassing, Y. D. van der Werf, J. R. Ramautar, and E. J. W. Van Someren, "Actigraphic multi-night home-recorded sleep estimates reveal three types of sleep misperception in insomnia disorder and good sleepers," *Journal of Sleep Research*, vol. 29, no. 1, p. e12937, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/jsr.12937

[4] H. Qing, "Latent class analysis by regularized spectral clustering," *arXiv preprint arXiv:2310.18727*, 2023.

[5] Z. Beattie, Y. Oyang, A. Statan, A. Ghoreyshi, A. Pantelopoulos, A. Russell, and C. Heneghan, "Estimation of sleep stages in a healthy adult population from optical plethysmography and accelerometer signals," *Physiological Measurement*, vol. 38, no. 11, p. 1968, oct 2017. [Online]. Available: https://dx.doi.org/10.1088/1361-6579/aa9047

[6] I. Fedorin, K. Slyusarenko, W. Lee, and N. Sakhnenko, "Sleep stages classification in healthy people based on optical plethysmography and accelerometer signals via wearable devices," in *2019 IEEE 2nd Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, 2019, pp. 1201–1204.

[7] K. Stuburić, M. Gaiduk, and R. Seepold, "A deep learning approach to detect sleep stages," *Procedia Computer Science*, vol. 176, pp. 2764–2772, 2020, knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 24th International Conference KES2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050920321840

[8] R. N. Sekkal, F. Bereksi-Reguig, D. Ruiz-Fernandez, N. Dib, and S. Sekkal, "Automatic sleep stage classification: From classical machine learning methods to deep learning," *Biomedical Signal Processing and Control*, vol. 77, p. 103751, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809422002737

[9] H. Li and Y. Guan, "Deepsleep convolutional neural network allows accurate and fast detection of sleep arousal," *Communications Biology*, vol. 4, no. 1, p. 18, 2021. [Online]. Available: https://doi.org/10.1038/s42003-020-01542-8

[10] A. W. Thomas, H. R. Heekeren, K. Müller, and W. Samek, "Interpretable lstms for whole-brain neuroimaging analyses," *CoRR*, vol. abs/1810.09945, 2018. [Online]. Available: http://arxiv.org/abs/1810.09945

[11] N. Michielli, U. R. Acharya, and F. Molinari, "Cascaded lstm recurrent neural network for automated sleep stage classification using single-channel eeg signals," *Computers in Biology and Medicine*, vol. 106, pp. 71–81, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482519300137

[12] C. C. Aggarwal, *Recurrent Neural Networks*. Cham: Springer International Publishing, 2018, pp. 271–313. [Online]. Available: https://doi.org/10.1007/978-3-319-94463-0_7

[13] L. Fraiwan and M. Alkhodari, "Investigating the use of uni-directional and bi-directional long short-term memory models for automatic sleep stage scoring," *Informatics in Medicine Unlocked*, vol. 20, p. 100370, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914820302161

[14] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602–610, 2005, iJCNN 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608005001206

[15] R. C. Staudemeyer and E. R. Morris, "Understanding LSTM - a tutorial into long short-term memory recurrent neural networks," *CoRR*, vol. abs/1909.09586, 2019. [Online]. Available: http://arxiv.org/abs/1909.09586

[16] L. Fiorillo, A. Puiatti, M. Papandrea, P.-L. Ratti, P. Favaro, C. Roth, P. Bargiotas, C. L. Bassetti, and F. D. Faraci, "Automated sleep scoring: A review of the latest approaches," *Sleep Medicine Reviews*, vol. 48, p. 101204, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1087079218301746

[17] J. H. Migueles, A. V. Rowlands, F. Huber, S. Sabia, and V. T. van Hees, "Ggir: A research community–driven open source r package for generating physical activity and sleep outcomes from multi-day

raw accelerometer data," *Journal for the Measurement of Physical Behaviour*, vol. 2, no. 3, pp. 188–196, 2019.

[18] E. Zhang and Y. Zhang, *Average Precision*. Boston, MA: Springer US, 2009, pp. 192–193. [Online]. Available: https://doi.org/10.1007/978-0-387-39940-9_482

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[20] P. Flach and M. Kull, "Precision-recall-gain curves: Pr analysis done right," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2015/file/33e8075e9970de0cfea955afd4644bb2-Paper.pdf

[21] L. Zhang, R. Wang, Z. Li, J. Li, Y. Ge, S. Wa, S. Huang, and C. Lv, "Time-series neural network: A high-accuracy time-series forecasting method based on kernel filter and time attention," *Information*, vol. 14, no. 9, 2023. [Online]. Available: https://www.mdpi.com/2078-2489/14/9/500