End-to-end Sleep Staging with Raw Single Channel EEG using Deep Residual ConvNets

Ahmed Imtiaz Humayun¹, Asif Shahriyar Sushmit¹, Taufiq Hasan¹, and Mohammed Imamul Hassan Bhuiyan¹,²

Abstract—Humans approximately spend a third of their life sleeping, which makes monitoring sleep an integral part of well-being. In this paper, a 34-layer deep residual ConvNet architecture for end-to-end sleep staging is proposed. The network takes raw single channel electroencephalogram (Fpz-Cz) signal as input and yields hypnogram annotations for each 30s segments as output. Experiments are carried out for two different scoring standards (5 and 6 stage classification) on the expanded PhysioNet Sleep-EDF dataset, which contains multisource data from hospital and household polysomnography setups. The performance of the proposed network is compared with that of the state-of-the-art algorithms in patient independent validation tasks. The experimental results demonstrate the superiority of the proposed network compared to the best existing method, providing a relative improvement in epoch-wise average accuracy of 6.8% and 6.3% on the household data and multi-source data, respectively. Codes are made publicly available on https://github.com/mHealthBuet/ASSCGithub.

I. INTRODUCTION

Sleep disorders are becoming a global health problem and more prominent in clinical practice as the global population and life expectancy increases. 50-70 million adults have a sleep disorder in the United States according to the American Sleep Association. Sleep problems were marked as “An Emerging Global Epidemic” following a study by the World Health Organization (WHO) conducted with 40,000 adults from developing countries across Africa and Asia [1]. Sleep plays an essential role in our health and well-being. It affects the regulation of glucose metabolism and loss of sleep could develop a resistance towards insulin, which may result in Type 2 diabetes [2]. In older adults, sleep disorders could be contributing to neuro-degenerative disorders [3].

Polysomnogram (PSG) is widely used as a diagnostic tool to determine sleep disorders. Polysomnography incorporates the recording and inspection of multiple physiological signals during sleep, like electroencephalogram (EEG), electro-oculogram (EOG) and electrocardiogram (ECG). The physiological variables and body movements help experts determine what stages of sleep the subjects have reached and their duration, depending on standard classification guidelines [4].

Many studies in the past few decades have focused on developing automatic sleep stage classification (ASSC) frameworks. Multi-channel systems incorporating EEG, EOG and EMG [5], [6], single-channel systems with EEG [7] and multimodal systems with smartphone and wearable sensors [8] have been explored. While state-of-the-art systems mostly employ feature extraction based methods along with machine learning algorithms [9], recent papers have adopted the use of Deep Neural Networks (DNN) in an end-to-end fashion [10]. Digital filtering [11] and Wavelet Transform [12], [13], along with its variants [14], have been common pre-processing steps utilized in previous studies. A number of these methods extract time [11], frequency [14] or time-frequency features from each decomposed band of the signal [15]. These features are hand-engineered depending on the study population and therefore may not generalize well. Moreover, there is no consensus on which features are the most discriminatory for the sleep stage classification task [9]. Deep learning methods are known to automatically learn generalizing features, which have been leveraged in a number of different tasks in the past [16]. Recent deep learning approaches for sleep staging have explored end-to-end systems using 1-dimensional Convolutional Neural Networks (1D CNN), Bi-directional Long Short Term Memory Units [10], [17], and also 2 DCNNs with prior spectral decomposition to represent EEG as 2D images [7]. In [18], the EEG signal was first pre-processed with a notch and bandpass filter and followed by two separate CNN and Multi-layer Perceptron (MLP) architectures using time domain and frequency domain signals. Deep complex valued CNNs with C3-A2 EEG signals were employed in [19]. A 12-layer deep 1DCNN architecture for feature extraction was explored in [20], while in [10] a 4-layer deep architecture was deployed.

In this paper, we introduce a 34-layer deep 1DCNN residual architecture for end-to-end classification. Residual architectures (ResNet) were proposed to resolve the vanishing gradient problem arising from the training of deeper CNN models [16]. As the number of layers used in our model is significantly higher compared to previous studies on sleep staging, we utilize the ResNet architecture for our end-to-end system and compare its performance with state-of-the-art algorithms.

II. DATASET

A number of sleep health studies have been compiled into datasets that are used in quantitative experiments, i.e. the Sleep Heart Health Study (SHHS) [20], Montreal Archive of Sleep Studies (MASS) [10] and the Physionet Sleep-EDF Database [21]. We use the Physionet Sleep-EDF Expanded Database in our experiments for its data volume and commonality in existing literature [9]. The database contains a total number of

¹mHealth Research Group, Department of Biomedical Engineering, Bangladesh University of Engineering and Technology (BUET), Dhaka - 1205, Bangladesh. Email: taufiq@bme.buet.ac.bd
²Department of Electrical and Electronics Engineering, Bangladesh University of Engineering and Technology, Dhaka - 1205, Bangladesh Email: imamul@eee.buet.ac.bd
61 polysomnography recordings from 42 subjects, sampled at 100 Hz. 30s data epochs from each recording were scored by sleep experts into one of the eight classes: AWA (Awake), S1-S4 (intermediate stages), REM (Rapid Eye Movement), MVT (Movement Time) and ‘Unscored’ [4]. The MVT and unscored data epochs have been removed prior to our experiments since they don’t correspond to any stages of sleep. We have also fused the S3 and S4 stages for 5-stage classification according to AASM [22].

The PhysioNet Sleep-EDF Expanded Database is assembled from two different studies. The first 39 recordings are from the Sleep Cassette (SC) subset, which was collected from in-home sessions to study the effect of aging on health [23]. This portion of the dataset includes sleep data from 20 healthy subjects without any sleep-related medication. Each patient contributed two different recordings from two nights, except for patient number 13. The latter 22 data samples are from Sleep Telemetry (ST) recording subset. It was collected during a hospital study on the effects of Temazepam on sleep. The data was acquired from 22 subjects for two different nights. Our experiments incorporate recordings from the nights without medication. The distribution of samples from various classes are shown in Table 1.

### III. Methodology

#### A. Experimental Setup

We design our experiments to address the robustness of deep learning models towards patient independent classification tasks. We have devised two test cases; the first is a multi-source task where a patient independent classification of sleep was performed on data from both SC and ST, from here on referred to as Random Split task (RS-task). There are 42 recordings (25 from 13 SC subjects & 17 from 17 ST subjects) in the training set and 19 recordings (14 from 7 SC subjects & 5 from 5 ST subjects) in the test set of the RS-task. The second case is when we train and test with the Sleep Cassette subset of the dataset with a 70% – 30% random patient independent split, from here on referred to as Sleep Cassette task (SC-task). This task has 27 recordings from 14 SC subjects in the training set and 12 recordings from 6 SC subjects in the test set. We perform each of these experiments with the FPz-Cz EEG channel data. The specific recording IDs and codes for each test case is available on Github 1.

#### B. Proposed Method

Residual architectures have previously been employed for both image classification tasks [24] and biomedical signal processing tasks [25]. They were introduced to make optimization and convergence of very deep neural networks plausible. Training deeper neural networks poses challenges related to saturation and degradation of training accuracy with the increase of layers. Skip-connections in the residual blocks allow deeper CNN layers to approximate an identity operation instead of performing redundant transformations, by only passing the input tensor to the next block. We, therefore, employ skip-connections in our model for sequence-to-sequence sleep staging of each 30s EEG data epochs. Our proposed architecture has 34 1D CNN layers with the number of kernels in each CNN layer increasing with depth. Our model takes 30s unprocessed EEG time-domain signals as input and outputs hypnogram predictions for each instance. Fig. 1 portrays the high-level network architecture. The residual blocks with skip-connections follow a pre-activation design.

---

1Codes and supplementary material available at https://github.com/mHealthBuet/ASSC
sub-sampling the incoming data by layers. Each pooling layer performs a max-pooling operation, cross-validation (LOO-CV) scheme where data from only one splitting can be performed at random [7] or in a leave-one-out splitting where patients are split between training and testing independent [11]–[13]. The other way is to perform patient to augment the training data for improved generalization. Class imbalance. During training, rolling shifts were performed the dominance of classes in the training population to tackle training epochs. Cross-entropy loss is weighted according to the train and test sets; therefore the splits are data epochs from the same patient are allowed to be in both divided into two types. The first is example splitting, where the method on the SC-task of trained using Adam optimizer with a maximum learning rate $0.001$ and we divide the learning rate by $10$ after every $10$ training epochs. Cross-entropy loss is weighted according to the dominance of classes in the training population to tackle class imbalance. During training, rolling shifts were performed to augment the training data for improved generalization.

IV. RESULTS

We compare the accuracy metric obtained by our proposed method on the SC-task with that reported on recent papers in Table II. The evaluation methods used in literature can be divided into two types. The first is example splitting, where data epochs from the same patient are allowed to be in both the train and test sets; therefore the splits are not patient independent [11]–[13]. The other way is to perform patient splitting where patients are split between training and testing to ensure a patient independent test set [7], [10], [17]. Patient splitting can be performed at random [7] or in a leave-one-out cross-validation (LOO-CV) scheme where data from only one patient is used for evaluation [15]. The number of epochs used in the experiments also require consideration. The performance achieved by our proposed method in the SC-task is comparable with the accuracy acquired by state-of-the-art methods on the SC subset of the Physionet Sleep-EDF database, surpassing most except for [15] in patient independent tasks. For a fair comparison between patient independent and not independent splitting, we implement the method proposed by Aboalayon et al. [11], which obtained the best accuracy reported for this problem. In our implementation, we used a Balanced Bagging ensemble of 71 Decision Trees to compensate for class imbalance. Table III contains comparison of its acquired metrics with our proposed model. Our model exhibits a relative improvement of 6.8% in epoch-wise accuracy on the SC-task and 6.25% on the RS-task. We also evaluate how well our model is performing for individual patients; our model exhibits 6.76% relative increase in patient-wise accuracy on the SC-task and 6.45% increase on the RS-task.

V. DISCUSSION

A. Number of Epochs and Split

The Sleep-EDF database, though widely employed in quantitative experiments, lacks a standard test bench for fair comparison between different ASSC frameworks. Evaluation metrics are susceptible to changes in class distributions, number of epochs used in the evaluation task and also the epoch selection scheme [20]. We can see a 8.8% reduction in the accuracy of [11] in the 6-stage SC-task compared to the paper reported metric, which can be due to an increased number of data epochs and patient independent splitting during evaluation. Our SC-task uses data from multiple patients (Fig 2) during test, which is more challenging compared to LOO-CV. This

| Approach                  | No. of Epochs | Split               | Method               | 5-Stage Acc | 6-Stage Acc |
|---------------------------|---------------|---------------------|----------------------|-------------|-------------|
| K. Aboalayon 2016 [11]    | 23806 Fpz-Cz 30s | 80%-20% Not independent | IIR-MMD-DT         | 95.5%       | 93.1%       |
| N. Liu 2017 [12]          | 60647 Fpz-Cz 30s | 90%-10% Not independent | DWT-CNN-MLP        | 93%         | 91.1%       |
| M. Sharma 2018 [13]      | 85900 Single EEG 30s | 10 fold CV Not independent | Wavelet-SVM    | 91.7%       | 91.5%       |
| A. Vilamala 2017 [7]      | Not Specified | 15-4 Independent    | Spectral 2D-VGG16   | 86.0%       | -           |
| A. Supratak 2017 [10]     | 41950 Fpz-Cz 30s | 20 patients LOO-CV  | CNN-BLSTM          | 82.0%       | -           |
| H. Phan 2018 [17]         | 46236 Fpz-Cz 30s | 20 patients LOO-CV  | BLSTM-SVM          | 82.5%       | -           |
| D. Jiang 2019 [15]        | 54728 Fpz-Cz 30s | 20 patients LOO-CV  | Multi.Decomp.-RF-HMM | 93.0%       | -           |
| **Proposed Method**       | 127008 Fpz-Cz 30s | 70%-30% Independent | Residual CNN       | 91.4%       | 90.1%       |

| Approach                  | Exp. Task | Sl Sens. | Avg Sens. | Avg Spec. | Epoch-Wise Acc | Patient-Wise Acc |
|---------------------------|-----------|----------|-----------|-----------|----------------|-----------------|
| IIR-MMD-DT [11]           | RS-task   | 15.6     | 58.2      | 95.6      | 83.8           | 80.3            |
| IIR-MMD-DT [11]           | SC-task   | 26.4     | 61.9      | 94.8      | 84.3           | 84.3            |
| **Proposed Method**       | RS-task   | 43.8     | 66.5      | 97.9      | 88.8           | 85.5            |
| **Proposed Method**       | SC-task   | 43.3     | 67.9      | 97.0      | 90.1           | 90.0            |

![Fig. 2. Epoch-wise accuracy for different test set recordings in each task: (a) Random-split task (RS-task) (b) Sleep Cassette task (SC-task).](image)
explains the accuracy difference between Jiang et al. [15] and our proposed method (Table II).

B. Difference between SC & ST subsets

From Fig. 2-a we observe that our model exhibits a drop in accuracy for recordings which are from the ST subset compared to the SC subset. No such trend is visible when the model is trained and evaluated on the SC subset (Fig 2-b). This can be attributed to the presence of data from two different sources in the RS-task which was also discussed in [20]. To explore whether the SC and ST subsets in the Sleep-EDF database exhibit heterogeneous properties or not, we perform one-way ANOVA on two time-domain features (MMD & EnergySis [11]) and two spectral features (Spectral Rolloff & Spectral Spread) extracted from 5 frequency bands (δ, θ, α, β, γ) of the Fpz-Cz channel. Except for the γ-band EnergySis feature (Fig. 3-a), all of the features exhibit a significant difference in the distribution with a \(p \leq 0.001\). Fig 3-b shows significant difference between the density functions of γ-band Spectral Rolloff between SC and ST. This strengthens our argument regarding heterogeneity between the subsets & explains the accuracy difference between RS-task & SC-task.

VI. CONCLUSION

In this paper, we have proposed a residual architecture based ASSC framework. Our evaluation task comprised of an increased number of data epochs and subjects from multiple sources, compared to previously reported works. Experimental results showed that our model is robust towards noise perturbations which demonstrates relative improvement in epoch-wise average accuracy of 6.8% on the SC-task and 6.3% on the RS-task compared to Aboalayon et al. [11].

ACKNOWLEDGMENT

We would like to thank the Department of EEE & BME, BUET and Brain Station 23 (Dhaka, Bangladesh) for supporting this research. The TITAN Xp GPU used for this work was donated by the NVIDIA Corporation.

REFERENCES

[1] S. Stranges, W. Tigbe, F. X. Gómez-Olivé, M. Thorogood, and N.-B. Kandala, “Sleep problems: an emerging global epidemic? Findings from the in-depth WHO-SAGE study among more than 40,000 older adults from 8 countries across Africa and Asia,” Sleep, vol. 35, no. 8, pp. 1173–1181, 2012.

[2] K. Spiegel, K. Knutson, R. Leproult, E. Tasali, and E. V. Cauter, “Sleep loss: a novel risk factor for insulin resistance and Type 2 diabetes,” J. of App. Physiol., vol. 99, no. 5, pp. 2008–2019, 2005.

[3] S. L. Naismith, S. J. Lewis, and N. L. Rogers, “Sleep–wake changes and cognition in neurodegenerative disease,” in Prog. Brain Res., 2011, vol. 190, pp. 21–52.

[4] J. A. Hobson, “A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects,” Electroencephalogr. Clin. Neurophysiol., vol. 26, no. 6, p. 644, 1969.

[5] L. Cen, Z. L. Yu, Y. Tang, W. Shi, T. Kluge, and W. Ser, “Deep learning method for sleep stage classification,” in Int. Conf. Neural Information Processing, 2017, pp. 796–802.

[6] J. Vrulkala, J. Hasan, A. Vári, S.-L. Himanen, and K. Müller, “Automatic sleep stage classification using two-channel electro-oculography,” J. of neuroscience methods, vol. 166, no. 1, pp. 109–115, 2007.

[7] A. Vilamala, K. H. Madsen, and L. K. Hansen, “Deep convolutional neural networks for interpretable analysis of eeg sleep stage scoring,” arXiv preprint arXiv:1710.00833, 2017.

[8] W. Chen, A. Sano, D. L. Martinez, S. Taylor, A. W. McHill, A. J. Phillips, L. Barger, E. B. Klerman, and R. W. Picard, “Multimodal ambulatory sleep detection,” in Proc. IEEE BHI, 2017, p. 465.

[9] R. Boostani, F. Karimzadeh, and M. Nami, “A comparative review on sleep stage classification methods in patients and healthy individuals,” Computer methods and programs in biomedicine, vol. 140, pp. 77–91, 2017.

[10] A. Supratik, H. Dong, C. Wu, and Y. Guo, “DeepSleepNet: a model for automatic sleep stage scoring based on raw single-channel EEG,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 25, no. 11, pp. 1998–2008, 2017.

[11] J. Aboalayon, M. Faezipour, W. Almuhammadi, and S. Moslehpour, “Sleep stage classification using eeg signal analysis: a comprehensive survey and new investigation,” Entropy, vol. 18, no. 9, p. 272, 2016.

[12] N. Liu, Z. Lu, B. Xu, and Q. Liao, “Learning a convolutional neural network for sleep stage classification,” in Proc. CISP-BMEI, 2017, pp. 1–6.

[13] M. Sharma, D. Goyal, P. Achuth, and U. R. Acharya, “An accurate sleep stages classification system using a new class of optimally time-frequency localized three-band wavelet filter bank,” Computers in biology and medicine, vol. 98, pp. 58–75, 2018.

[14] A. R. Hassan and M. I. H. Bhuian, “A decision support system for automatic sleep staging from eeg signals using tunable q-factor wavelet transform and spectral features,” J. Neurosci. Methods, vol. 271, pp. 107–118, 2016.

[15] D. Jiang, Y.-n. Lu, M. Yu, and W. Yuanyuan, “Robust sleep stage classification with single-channel eeg signals using multimodal decomposition and hmm-based refinement,” Expert Syst. Appl., vol. 121, pp. 188–203, 2019.

[16] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in Proc. IEEE ICCV, 2015, pp. 1026–1034.

[17] H. Phan, F. Andreotti, N. Cooray, O. Y. Chen, and M. De Vos, “Automatic sleep stage classification using single-channel eeg: Learning sequential features with attention-based recurrent neural networks,” in Proc. IEEE EMBC, 2018, pp. 1452–1455.

[18] M. Manzano, A. Guillén, I. Rojas, and L. J. Herrera, “Deep learning using eeg data in time and frequency domains for sleep stage classification,” in Int. Work-Conference on Artificial Neural Networks, 2017, pp. 132–141.

[19] J. Zhang and Y. Wu, “Automatic sleep stage classification of single-channel eeg by using complex-valued convolutional neural network,” Biomed. Tech. (Berl), vol. 63, no. 2, pp. 177–190, 2018.

[20] A. Sors, S. Bonnet, S. Mirek, L. Verscheur, and J.-F. Payen, “A convolutional neural network for sleep stage scoring from raw single-channel eeg,” Biomedical Signal Processing and Control, vol. 42, pp. 107–114, 2018.

[21] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. Ch, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, “Physiobank, physiotoolkit, and physionet,” Circulation, vol. 101, no. 23, pp. e215–e220, 2000.

[22] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, C. Marcus, and B. Vaughn, “The AASM manual for the scoring of sleep and associated events,” Rules, Terminology and Technical Specifications, Darien, Illinois, AASM, 2012.
[23] M. Mourtazaev, B. Kemp, A. Zwinderman, and H. Kamphuisen, “Age and gender affect different characteristics of slow waves in the sleep eeg,” Sleep, vol. 18, no. 7, pp. 557–564, 1995.

[24] K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” in Proc. ECCV, 2016, pp. 630–645.

[25] P. Rajpurkar, A. Y. Hannun, M. Haghpanahi, C. Bourn, and A. Y. Ng, “Cardiologist-level arrhythmia detection with convolutional neural networks,” arXiv preprint arXiv:1707.01836, 2017.