Multivariate Masked Autoencoders for Obstructive Sleep Apnea Diagnosis

Thanh-Cong Do[1], *Hyung-Jeong Yang[1], Hyeong-Chae Yang[21]
[1]Dept. of Artificial Intelligence Convergence, Chonnam National University
[2] Otolaryngology Department, Chonnam National University

Abstract

Obstructive sleep apnea (OSA) is a common sleep disorder, causing disrupted sleep and reduced oxygen levels in the blood. Full-night polysomnography (PSG) and home sleep apnea tests (HSAT) may offer acceptable diagnosis results but have shown some limitations. In recent years, deep learning and data analysis methods are progressively employed on electronic health records, and various methods have been developed for OSA event detection. Self-supervised learning has shown some advantages over supervised training methods, by learning more generalized feature representation of data. Unlike text or image processing, the high information density in multivariate time-series data makes it more challenging to utilize self-supervised learning. In this research, based on the characteristics of sleep dataset, we propose a self-supervised approach with Masked Autoencoder, which masks portions of the input and attempting to reconstruct them. This enables the model to learn more generalized features from unlabeled data. We evaluate our proposed framework with three independent sleep datasets, which have shown significant improvement compared to supervised learning models.

1. Introduction

Obstructive sleep apnea (OSA) is the most common type of sleep related breathing disorder [1]. Several research indicates that OSA affects 425 million adults aged 30–69 years. [2]. It is characterized by repeated episodes of partial or complete blockage of the upper airway during sleep. Recent research shows that OSA increases the risk of stroke, cardiovascular and mortality [3]. The most frequent symptom of OSA is excessive daytime sleepiness, which can result in accidents and low work efficiency.

In hospitals globally, full-night polysomnography (PSG) is regarded as the gold technique for detecting a clinical suspicion of OSA. It is a multi-channel monitoring method that examines the electrophysiological and cardio-respiratory patterns during patients' sleep [4]. However, it is believed that long PSG recordings are time-consuming with low cost-effectiveness and low availability [1]. Recently, home sleep apnea tests (HSAT) have been considered as a cost-efficient alternative to in-lab PSG. Although this method has become a standard technique for several countries, recent research revealed an inaccurate diagnosis rate of 39% in HSAT compared to PSG [5]. These factors have motivated the development of a deep learning (DL) based comprehensive framework for OSA detection. Several studies have focused on OSA screening tests with binary classification task (OSA/non-OSA) [6], [7] but failed to forecast the apnea-hypopnea index (AHI), which is calculated and used to diagnose OSA patients [1].

Self-supervised learning offers the advantage of utilizing vast amounts of unlabeled data, reducing the reliance on costly and time-consuming manual labeling [8]. By learning to predict missing or masked data, models can capture richer representations that generalize well to various downstream tasks. However, applying pre-trained masked models to multivariate time-series (MTS) data presents challenges due to its complex temporal dependencies and varying feature dimensions.

In this work, we introduce a masked autoencoder model for multivariate time-series forecasting (MMAE). Our main contributions are as follows:

- We propose MTSMAE, a novel masked autoencoder model specifically designed for multivariate time-series forecasting, which efficiently handles high-dimensional data and captures long-term dependencies.
- A new patch embedding method based on the idea of Vision Transformer (ViT) is introduced, significantly reducing data redundancy and allowing the model to process longer sequences with less memory consumption.
- We design an overall framework for OSA diagnosis with extensive experiments on multiple sleep datasets, which consistently outperforms state-of-the-art models.

2. Related Works

Numerous studies have implemented DL-based techniques

*Corresponding Author

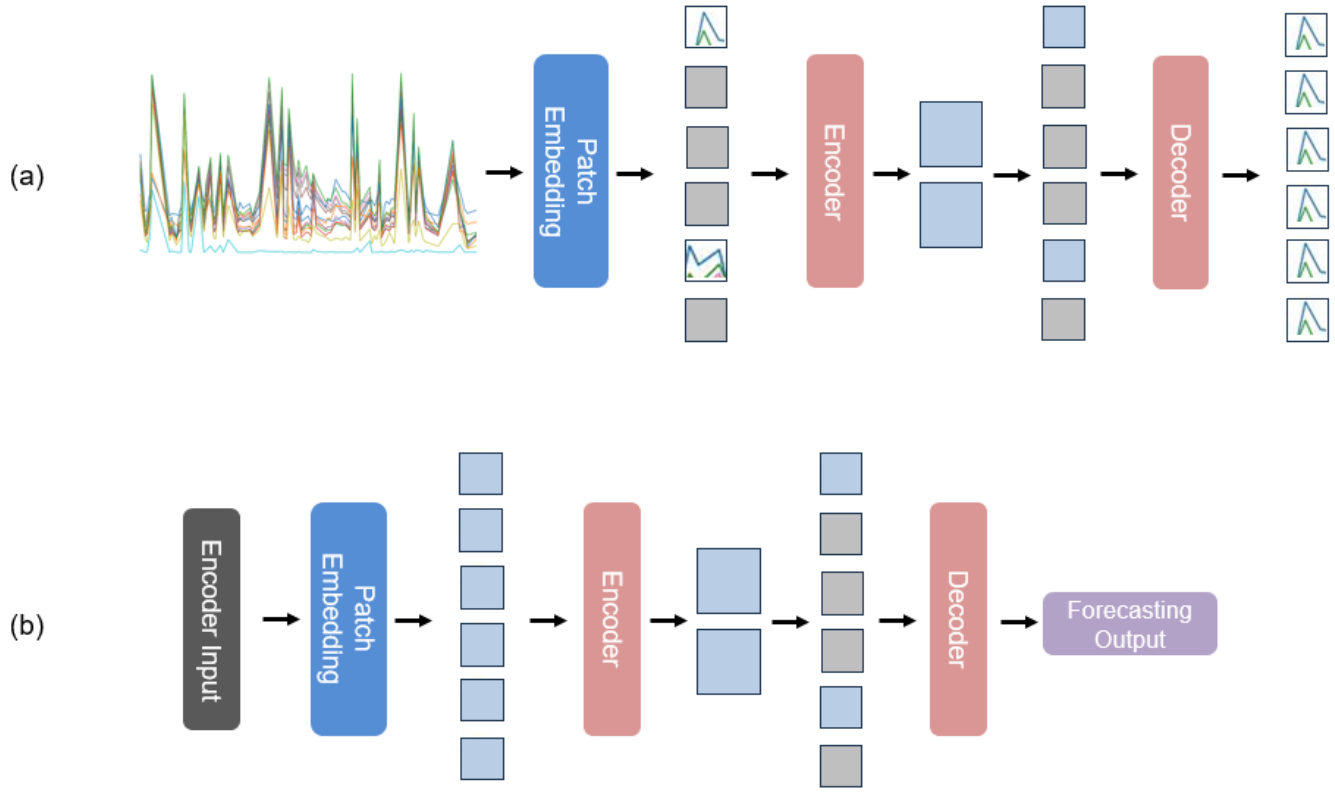

Figure 1: Overall architecture of the proposed approach. (a): Pre-training phase, (b): Fine-tunning phase

for the detection of OSA. Research in [9] employed the Discriminative Hidden Markov Model (HMM) to identify OSA from ECG signals. Furthermore, [10] utilized a continuous wavelet transform (CWT) to create a two-dimensional scalogram image from each minute-long segment of ECG data. They then analyzed this data using CNN and AlexNet models. Another study in [11] conducts the experiments using a single-lead ECG signal with different classifiers such as Artificial Neural Networks (ANN), SVM, and HMM. The main limitation of this work is the absence of classification and illness detection. Several studies apply RNN-based models to capture the sequential information in MTS. [12] using a 2D-CNN model combined with Long Short-Term Memory (LSTM) to recover the spatial and temporal properties. The study in [1] utilizes the combination of 1D-CNN and LSTM, along with some techniques such as deep time-distributed layers and scalograms. These existing RNN-based methods may struggle with the problem of redundancy, as they process sequences step-by-step, leading to an overload of redundant information and difficulty in capturing long-term dependencies effectively [13].

3. Methodology

In MTS, given an input sequence $X_t = \{x_1^t, x_2^t, \dots, x_{L_x}^t\}$, where each x_i^t belong to a feature space R^{d_x}, the goal is to forecast the corresponding future sequence $Y_t = \{y_1^t, y_2^t, \dots, y_{L_y}^t\}$ with each y_i^t in R^{d_y}. Here, L_x and L_y represent the lengths of the input and output sequences, respectively, while d_x and d_y denote the dimensionality of the input and output features. The overall framework architecture is shown in Fig. 1. Our proposed MMAE follows the standard autoencoding framework, where the training procedure is divided into two key stages: pre-training and fine-tuning. During pre-training, the model randomly masks portions of the input time-series data, and the encoder learns to map only the visible patches to a latent representation, while the decoder reconstructs the masked data. In the fine-tuning stage, the encoder processes the entire input sequence, and the decoder is adapted to predict future data based on the learned latent representations from pre-training.

A. Patch Embedding

Applying masked modeling to multivariate time-series (MTS) data is more challenging compared to text and images due to the complex temporal dependencies and high-dimensional nature of the data. In this work, we apply the idea of ViT, which processes images by dividing them into smaller patches. Similarly, for multivariate time-series data, instead of working with individual time steps, MTSMAE segments the time-series into patches along the time axis. These patches represent continuous segments of the input sequence, capturing both temporal and feature dependencies. The input embedding for our MMAE includes two parts as in (1): the scalar projection (SP) with 1-D convolutional filters to

transforms raw time-series into a structured format that can be processed by the encoder; the positional encoding (PE) as in [14] to capture the position information of sequential data:

$$X = SP + PE \tag{1}$$

where:

$$SP = Conv1d(x_i^t) \tag{2}$$

$$PE_{(i,2j)} = \sin\left(i/10000^{2j/d_{model}}\right) \tag{3}$$

$$PE_{(i,2j+1)} = \cos\left(i/10000^{2j/d_{model}}\right) \tag{4}$$

where $i \in \{1, \dots, L_x\}$, $j \in \{1, \dots, d_{model}/2\}$, d_{model} is the feature dimension. After the embedding, we patch the time-series as follow:

$$X_p = Conv1d(X) \in \mathbb{R}^{L_x/P \times d_{model}} \tag{5}$$

$$X_{pt} = Conv1d\left(X_p\right) \in \mathbb{R}^{L_x/P^2 \times d_{model}} \tag{6}$$

where P is the kernel size of $Conv1d$, X_{pt} is the final output of the patch embedding module. Based on (5) and (6), the time-series data is divided into smaller continuous segments along the time axis. Each patch contains multiple consecutive time points, allowing the model to capture temporal patterns within a segment rather than processing individual points. This patching method reduces data complexity and enables the model to handle longer sequences more efficiently.

B. The Pre-training Phase

In this work, we use the architectures of Transformer encoder and decoder for our MMAE. After the patch embedding, the model begins by masking a portion of the time-series data, selecting random patches to be hidden from the model. In MTS, each time point typically contains a large amount of information that is similar to the data from adjacent time points, leading to heavy spatial redundancy. In our proposed MMAE, we select the masking ratio of 85%, which is higher than common models for text and image. The high masking rate is well-suited to MTS due to the redundancy and smoothness of the data over time. Only 15% of the patches are served as input for the Transformer encoder, which consists of two sub-blocks: a multi-head self-attention layer (MSA); and a fully connected network (MLP) with layer normalization as in [14].

In the pre-training phase, the input to the decoder consists of both the latent representations generated by the encoder for the visible patches and the tokens for the masked patches. The output of the decoder is the reconstructed data for the masked patches, which attempts to recover the original values of the hidden parts based on the learned representations from the visible patches. This process helps the model learn to develop a deeper understanding of the MTS structure.

B. The Fine-tunning Phase

In the fine-tuning phase, the encoder processes the entire input sequence, based on the representations learned during pre-training to capture comprehensive temporal and feature patterns from the unmasked data. This enables the model to fine-tune its understanding of the time-series, adjusting to the specific forecasting task. The decoder is now responsible for predicting future time points rather than reconstructing missing data. It takes the encoded representations from the full sequence and generates the expected future values, aligning its output to the specific forecasting horizon. This phase allows the model to improve its predictive capabilities by leveraging the detailed information learned during pre-training and refining it for the target task.

4. Experiment

A. Dataset

In this research, we utilize three independent public datasets to train and evaluate the performance of MMAE for the task of AHI estimation. AHI is defined as the average number of all apneas and hypopneas per hour of sleep, following the rule of American Academy of Sleep Medicine (AASM) [15]. Three sleep datasets include: **Sleep Heart Health Study** (SHHS), a prospective cohort study aimed at investigating the relationships between sleep-disordered breathing and cardiovascular diseases; **Multi-Ethnic Study of Atherosclerosis** (MESA) a comprehensive medical database aimed at understanding the development and progression of cardiovascular disease; **Osteoporotic Fractures in Men Study** (MrOS), a comprehensive, multi-center study designed for the osteoporotic fractures in men. A total of 10,915 PSG recordings have been collected for our experiment. We follow previous research [4] for data preprocessing with the exclusion criteria including recordings with technical faults, patients with total sleep time (TST) less than 4 hours and patients under 18 years old.

B. Experiment Setting

Two baseline machine learning (ML) models are used to benchmark our proposed MMAE. The first model includes the oxygen desaturation index (ODI) with a threshold at 3% and the second model uses digital oximetry biomarkers (OBM) as input. We also compare our performance with OxiNet, a DL-based model for OSA diagnosis based on single channel oximetry [4]. For the task of AHI estimation, the models are trained with Mean Squared Error (MSE) loss. We use Adam optimizer, an initial learning rate of 0.001, weight decay of 10-4. L2 regularization is applied to prevent over-fitting. The experiments are performed with GPU RTX3090 with the total training time for 5-fold cross-validation is 6 hours.

For the evaluation of regression models, we utilize the Intraclass Correlation Coefficient (ICC) as the main evaluation metric. Based on the estimated AHI value, we also discriminate the output into four groups of severity: non-OSA (AHI < 5), mild-OSA (5 ≤ AHI < 15), moderate-OSA (15 ≤

TABLE I. SUMMARY TABLE OF ALL DATASET

Dataset	Number of samples	AHI	Age	%Male
SHHS	5778	10.3 ± 15.6	63.0 ± 17.0	52
MESA	2002	17.0 ± 21.0	68.0 ± 14.0	46
MrOS	3135	14.3 ± 22.0	76.0 ± 5.5	100

AHI < 30) and severe-OSA (AHI ≥ 30). The macro averaged F1 score (F1,M) is reported for the evaluation of the classification task.

C. Experiment Results

The experiment results are shown in Table 2, which indicates the MMAE performance was significantly better than conventional supervised training models. Among the 3 datasets, SHHS achieved the best performance with ICC = 0.92, $F_{1,M}$ = 0.83 on the test set. The performance was decreased but acceptable for the two remaining databases: MROS (ICC = 0.89, $F_{1,M}$ = 0.75) and MESA (ICC = 0.84, $F_{1,M}$ = 0.72).

TABLE II. EXPERIMENT RESULTS

Model	SHHS		MESA		MrOS	
	ICC	$F_{1,M}$	ICC	$F_{1,M}$	ICC	$F_{1,M}$
ODI	0.84	0.69	0.71	0.62	0.80	0.67
OBM	0.86	0.74	0.69	0.66	0.83	0.69
OxiNet	0.90	0.82	0.79	0.68	0.85	0.72
MMAE	**0.92**	**0.83**	**0.84**	**0.72**	**0.89**	**0.75**

5. Conclusion

To conclude, this paper proposed a DL-based architecture to estimate the apnea-hypopnea index and diagnosis of OSA. We introduce MMAE, a novel model designed for multivariate time-series forecasting using masked autoencoders. By leveraging a patch embedding approach and self-supervised pre-training, the model effectively reduces data redundancy and improves long-term dependency modeling. Extensive experiments show that MMAE outperforms traditional supervised models across various datasets. For the future work, we plan to explore the correlation between OSA and the risk of cardiovascular diseases, which is considered as a serious medical condition with a high prevalence.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00256629) grant funded by the Korea government(MSIT)

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT). (RS-2023-00208397)

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2024-RS-2024-00437718) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation)

References

[1] Wang, E., Koprinska, I. , Jeffries, B. Sleep Apnea Prediction Using Deep Learning. IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 11, pp. 5644-5654, Nov. 2023, doi: 10.1109/JBHI.2023.3305980.

[2] Benjafield, A. V. et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. Lancet Respir. Med. 7, 687–698 (2019).

[3] Gottlieb, D. J. & Punjabi, N. M. Diagnosis and management of obstructive sleep apnea: a review. JAMA 323, 1389–1400 (2020).

[4] Levy, J., Álvarez, D., Del Campo, F. et al. Deep learning for obstructive sleep apnea diagnosis based on single channel oximetry. Nat Commun 14, 4881 (2023).

[5] Massie, F., Van Pee, B. & Bergmann, J. Correlations between home sleep apnea tests and polysomnography outcomes do not fully reflect the diagnostic accuracy of these tests. J. Clin. Sleep Med. 18, 871–876 (2022).

[6] Deviaene, M. et al. Automatic screening of sleep apnea patients based on the spo 2 signal. IEEE J. Biomed. Health Inform. 23, 607–617 (2018).

[7] Behar, J. A. et al. Single-channel oximetry monitor versus in-lab polysomnography oximetry analysis: does it make a difference? Physiol. Meas. 41, 044007 (2020).

[8] He, Kaiming, et al. "Masked Autoencoders Are Scalable Vision Learners." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022): 16000-16009.

[9] Alshaer, H.; Hummel, R.; Mendelson, M.; Marshal, T.; Bradley, T.D. Objective Relationship between Sleep Apnea and Frequency of Snoring Assessed by Machine Learning. J. Clin. Sleep Med. 2019, 15, 463–470.

[10] Singh, S.A.; Majumder, S. A novel approach osa detection using single-lead ECG scalogram based on deep neural network. J. Mech. Med. Biol. 2019, 19, 1950026.

[11] Song, C.; Liu, K.; Zhang, X.; Chen, L.; Xian, X. An obstructive sleep apnea detection approach using a discriminative hidden markov model from ECG signals. IEEE Trans. Biomed. Eng. 2015, 63, 1532–1542.

[12] Zarei, A.; Beheshti, H.; Asl, B.M. Detection of sleep apnea using deep neural networks and single-lead ECG signals. Biomed. Signal Process. Control 2022, 71, 103125.

[13] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.

[14] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS), 2017.

[15] Thornton, A. T., Singh, P., Ruehland, W. R. & Rochford, P. D. Aasm criteria for scoring respiratory events: interaction between apnea sensor and hypopnea definition. Sleep 35, 425–432 (2012).