HTTMS: A Hierarchical Transformer for Sleep Quality Prediction using Temporal Multimodal Sensor Data

Uihyeon Maeng
Sangmyung University
maengu1217@gmail.com

Abstract*—Recent advances in mobile and wearable sensing have enabled continuous monitoring of human behavior and physiology, providing new opportunities for sleep quality prediction. However, many existing approaches rely on handcrafted or aggregated features, which overlook fine-grained temporal dependencies and multimodal interactions. In this paper, we propose HTTMS (Hierarchical Temporal Transformer for Multimodal Sensor Data), an end-to-end deep learning framework designed to model daily sensor streams through a hierarchical local-global architecture. HTTMS first segments a day into behaviorally meaningful intervals and applies Local Transformers to capture intra-segment patterns, followed by a Global Transformer to integrate dependencies across segments. To further improve personalization, we incorporate user ID embeddings that encode habitual patterns and user-specific traits, enabling the model to adapt predictions to individual contexts. To the best of our knowledge, this is the first work to jointly apply hierarchical Transformers and user embeddings for multi-label sleep quality prediction in an end-to-end architecture. We evaluate HTTMS on the ETRI Lifelog Dataset 2024, comparing it against 1D CNN, Bi-LSTM, and vanilla Transformer baselines. HTTMS achieves 60.74 percent accuracy and a 0.5954 Macro F1-score, outperforming baselines by over +7 percent in accuracy and +9 points in Macro F1-score, with consistent gains across all sleep-related indicators. These findings demonstrate the effectiveness of hierarchical Transformers and user ID embeddings for multimodal sensor streams and highlight the promise of end-to-end Transformers in advancing both methodological research and real-world applications of digital sleep health monitoring.

Keywords—Sleep quality prediction, Multimodal sensor data, Hierarchical Transformer, End-to-end learning, Time-series modeling, Personalization

I. INTRODUCTION

Recent advances in smartphone and smartwatch sensing technology have enabled continuous monitoring of human behavior and physiology across an entire day [1], [2], [3]. Sleep quality, one of the most critical indicators of physical and mental well-being, is significantly affected by various daytime and nighttime factors, including physical activity, environment, and biological rhythm [4], [5], [6]. However, most existing methods rely on aggregated daily features or statistical summaries, thereby losing fine-grained temporal and contextual information embedded in sensor data [7].

To address this limitation, we propose **HTTMS (Hierarchical Temporal Transformer for Multimodal Sensor Data)**, a deep learning framework that leverages hierarchical temporal structure to model multi-sensor time-series data for sleep quality prediction. Rather than treating the entire day as a single sequence, HTTMS partitions it into fixed-length temporal segments (e.g., 2-hour, 4-hour, or 8-hour windows), allowing the model to learn behaviorally meaningful patterns within localized contexts while preserving long-range temporal dependencies. Each segment is represented as a 2D sensor-time matrix, in which sensor values are aligned along the temporal axis.

The core architecture of HTTMS consists of a two-stage Transformer design: Local Transformer Encoders followed by a Global Transformer Encoder. In the first stage, each temporal segment is independently processed by a Local Transformer, which captures intra-segment dynamics through multi-head self-attention and positional encoding. In the second stage, the locally encoded representations are sequentially passed to a Global Transformer Encoder, which learns inter-segment dependencies across the entire day. This global layer enables the model to capture how patterns evolve throughout the day and how cumulative effects influence sleep-related outcomes. Ultimately, this hierarchical structure of HTTMS allows it to effectively model both short-term micro patterns and long-term behavioral rhythms. To further enhance personalization, HTTMS incorporates user ID embeddings that encode habitual routines and individual traits, enabling the model to adapt predictions to user-specific contexts.

We formulate the task as a multi-label classification problem, aiming to predict six key sleep and pre-sleep health indicators (Q1-Q3 and S1-S3) using multimodal lifelog sensor data. In this study, we selectively utilize 12 sensor items collected from smartphones and smartwatches. Each sensor stream is resampled to a uniform temporal resolution, and missing values are interpolated when necessary. The detailed description of the sensor modalities and target sleep-related indicators can be found in ***Section III. B (Datasets)***.

This study makes the following key contributions:

- **Hierarchical temporal modeling.** We design a two-stage Transformer framework that combines Local and Global encoders to simultaneously capture short-term micro-patterns and long-term behavioral rhythms.
- **User-aware personalization.** We introduce user ID embeddings that encode habitual routines and individual traits, enabling personalized adaptation.
- **End-to-end multi-label prediction.** We propose a unified framework that directly processes multimodal smartphone and smartwatch streams to jointly predict six sleep-related indicators.

We believe that these contributions lay the groundwork for more personalized, scalable, and actionable digital sleep health systems in real-world applications.

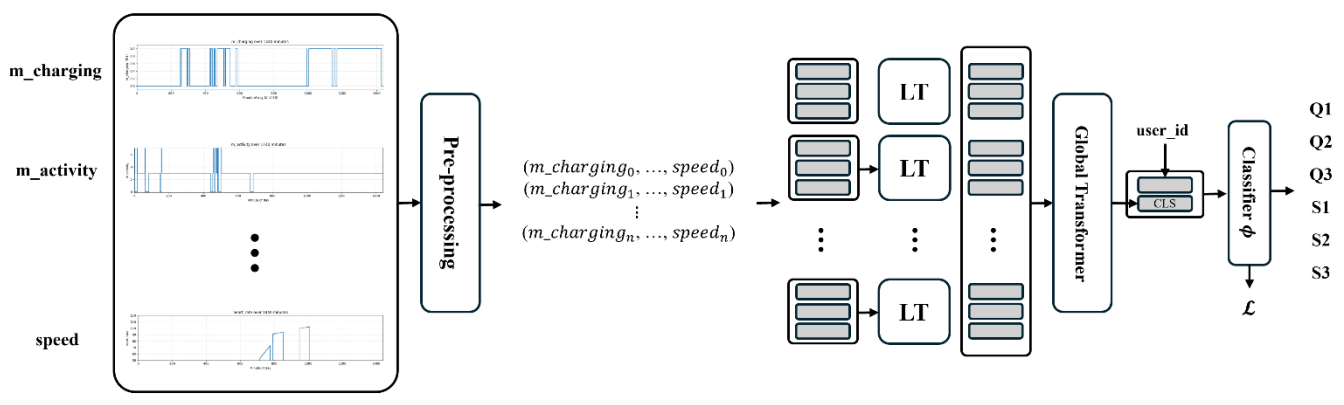

Fig. 1. Overall pipeline of HTTMS. The framework takes multi-sensor time-series data (e.g., smartphone and smartwatch signals) as input, which are first pre-processed into a sequence of feature vectors. Each temporal segment is encoded by Local Transformers (LT) to capture intra-segment temporal dependencies. The segment-level representations are then aggregated by a Global Transformer to model inter-segment relationships, while incorporating user embeddings for personalization. Finally, the [CLS] token representation is fed into a classifier to jointly predict subjective (Q1–Q3) and objective (S1–S3) sleep quality indicators, optimized by the loss function $\mathcal{L}$.

II. Related Work

A. Hierarchical Transformers for Time-Series Data

Initial methods for time-series modeling often relied on RNNs and LSTMs [8], which were capable of capturing temporal dependencies through recurrent mechanisms. While these models performed well on short sequences, they suffered from long-term dependency issues due to vanishing gradients. Transformers [9] addressed this limitation by introducing global self-attention mechanisms that allow direct connections between distant time points, leading to significant improvements in long-range forecasting. However, standard Transformers lack the inductive biases necessary to efficiently model localized temporal patterns, which are critical in sensor-based physiological time-series data [10].

To resolve this, hierarchical Transformer architectures [11] have emerged. These models segment the input time series into smaller windows or patches, allowing local attention to focus on short-term dynamics. Local encoders capture intra-segment information using self-attention mechanisms with positional encoding, while global encoders integrate representations across segments to model long-range dependencies. This local-global design preserves fine-grained detail while enabling contextual understanding at a broader temporal scale.

In recent literature, some studies adopt pyramid-based designs [12], [13] to reduce sequence length by hierarchically aggregating temporal features, enabling efficient global modeling while preserving local detail. Another study [14] utilizes patch-wise embedding, dividing long sequences into fixed-length chunks and applying local and global attention in a two-stage fashion. Additionally, frequency-aware models [15] apply spectral decomposition (e.g., Fourier or wavelet transforms) to separate temporal components before processing them through specialized attention layers. These approaches have shown strong performance in tasks such as electricity demand forecasting and climate prediction, where hierarchical temporal structures can enhance sequence understanding.

In summary, hierarchical Transformers for time-series modeling overcome the limitations of both traditional recurrent models and flat self-attention by explicitly structuring local and global representations. This enables more effective processing of complex, multi-scale patterns found in real-world sequential data [16]. However, most existing models still assume flat or uniform temporal structures and do not incorporate semantic segmentation aligned with human behavioral patterns—such as dividing a day into meaningful activity-based windows (e.g., 2-hour or 4-hour). Furthermore, they typically process multimodal signals in a flattened or aggregated form, lacking explicit mechanisms to jointly model heterogeneous sensor modalities. These limitations reduce their applicability in complex domains such as personalized health monitoring, where temporal structure and multimodal sensor integration are critical.

B. Methods for Sleep Quality Prediction

Early data-driven approaches [17], [18] for predicting sleep quality focused on extracting handcrafted features from wearable sensors such as accelerometers, heart rate monitors, or temperature sensors. These features—such as total movement, sleep duration, sleep consistency, or activity variability—were then used to train classical machine learning models like SVMs, decision trees, or ensemble methods. Although useful for binary classification of good vs. poor sleep quality, these approaches were often limited by shallow representations and a lack of generalization.

The introduction of deep learning [19], [20] allowed more direct use of raw time-series signals from wearables. CNNs were employed to automatically extract meaningful features from 1D sequences like actigraphy or HRV data, often outperforming traditional models in classification accuracy. RNNs and LSTMs were also explored to capture temporal dependencies, though their performance varied depending on the input signal and sequence length.

More recently, Transformer-based models [21], [22], [23] have been introduced to model long-range dependencies in multimodal sensor data. These architectures can capture daily patterns and temporal interactions across multiple sensors simultaneously. In some studies, these models have been used to predict sleep efficiency or subjective sleep quality using pre-sleep behavioral and physiological data. The global

receptive field of Transformers helps detect trends and periodicity, while fine-tuning on segment-level inputs enables attention to local changes.

Furthermore, smartphone- and smartwatch-based sensing has emerged as a practical and scalable alternative to traditional physiological monitoring. These systems leverage data passively collected during everyday life—such as screen status, ambient light levels, and basic motion—to estimate sleep quality without requiring complex or high-resolution physiological signals. While such data may lack clinical granularity, it captures behavioral cues that are strongly correlated with sleep patterns. This lightweight sensing approach enables large-scale, non-intrusive sleep monitoring, making it especially suitable for real-world applications where user compliance and device availability are critical.

III. METHODS

A. Overview

Our proposed architecture, HTTMS, is designed to predict sleep quality indicators by modeling multimodal time-series data collected from wearable and smartphone sensors throughout the day. The model reflects a biologically and behaviorally informed structure by segmenting a day into meaningful temporal intervals and processing each segment with both local and global temporal attention mechanisms. The overall pipeline can be described in the following stages: (1) data preprocessing and embedding of continuous and categorical sensor inputs, (2) segmentation of the input sequence into fixed-length windows, and application of a local encoder to model intra-segment patterns, (3) global Transformer encoding over the full sequence, (4) attention-based temporal pooling, and user-aware classification. This modular architecture is illustrated in ***Fig. 1***.

B. Datasets

In this study, we utilized the ETRI Lifelog Dataset 2024 [24], a multimodal dataset designed to capture various aspects of participants' daily behaviors, environments, and physiological states. The dataset was collected using a combination of smartphones, smartwatches, sleep sensors, and self-reporting mobile applications over the course of 2024. It includes a total of 700 days of lifelog records from 10 participants, which are publicly available for non-commercial and academic research purposes only.

The dataset comprises 12 sensor items that reflect real-world user contexts. Sensor keys prefixed with *'m-'* denote data collected from mobile devices (smartphones), whereas those with *'w-'* indicate data from wearable devices (smartwatches). The 12 sensor items are as follows:

1) ***mACStatus:*** Indicates whether the smartphone is currently charging.

2) ***mActivity:*** Activity labels as estimated by the Google Activity Recognition API.

3) ***mAmbience:*** Detected ambient sound types (provided by Google AudioSet) and corresponding confidence levels.

4) ***mBle:*** Information on surrounding Bluetooth devices.

5) ***mGps:*** smartphone-based GPS coordinates collected multiple times per minute.

6) ***mLight:*** Ambient light levels as measured by the smartphone.

7) ***mScreenStatus:*** Indicates the on/off state of the smartphone screen.

8) ***mUsageStats:*** App usage patterns, including duration and application name.

9) ***mWifi:*** List of surrounding Wi-Fi devices (not used in this study due to its high variability and limited relevance).

10) ***wHr:*** Heart rate measurements from the smartwatch.

11) ***wLight:*** Ambient light levels as measured by the smartwatch.

12) ***wPedo:*** Step-related activity metrics recorded via the smartwatch's pedometer.

Additionally, the dataset includes six sleep-related health metrics, derived from a combination of sensor-based sleep monitoring and self-reported survey responses:

1) ***Q1*****:** Self-perceived overall sleep quality upon waking (0: Below individual average, 1: Above individual average)..

2) ***Q2*****:** Self-assessed physical fatigue before sleep (0: High level of fatigue, 1: Low level of fatigue).

3) ***Q3*****:** Self-assessed stress levels before sleep (0: High level of stress, 1: Low level of stress).

4) ***S1*****:** Degree of adherence to total sleep time (TST) guidelines (0: Not recommended, 1: May be appropriate, 2: Recommended).

5) ***S2*****:** Degree of adherence to sleep efficiency (SE) guidelines (0: Inappropriate, 1: Recommended).

6) ***S3*****:** Degree of adherence to sleep onset latency (SOL/SL) guidelines (0: Inappropriate, 1: Recommended).

Each of these target metrics is discretely labeled into two or three ordinal levels (0, 1, 2), depending on the item. In this study, these six sleep-related metrics serve as the prediction targets (labels) for our models.

The specific classification threshold of target metrics and data collection procedure closely follows methodologies established in prior ETRI datasets, ensuring continuity and comparability with earlier research on multimodal lifelogging and behavioral health analysis.

C. Input Representation and Preprocessing

The input data consists of multimodal sensor signals collected over a single day. Each day spans from 6:00 AM to 6:00 AM of the following day and is aligned to a fixed temporal resolution (1-minute granularity), resulting in a sequence of 1,440 time steps. We extract 15 features from 12 sensor items, including both smartphone- and smartwatch-based modalities. The signals are divided into 10 continuous features, 3 binary features, and 2 categorical features. Continuous and binary signals are normalized and input directly as tensors of shape $B \times 13 \times T$, where $T = 1440$. The categorical signals are embedded using trainable embedding layers ($Embedding(9, 4)$ and $Embedding(254, 8)$, respectively) and concatenated with continuous features to form a fused representation of dimension $B \times (13 + 12) \times T$.

D. Segment-wise Encoding with Local Transformer

To enhance the model's sensitivity to region-specific temporal changes, we divide each daily sequence into fixed-length segments of 2, 4, or 6 hours. This segmentation scheme enables the model to independently process temporally localized behavioral patterns that might differ across phases of the day. We segment the time-series based on uniformly

sized intervals, which reflect behaviorally meaningful cycles without requiring explicit annotations. This strategy improves the model's ability to detect short-term deviations associated with sleep quality outcomes.

Each temporal segment is encoded using a dedicated Local Segment Encoder that combines CNNs and Transformer layers. The objective is to extract short-term behavioral dynamics that are often masked when treating the entire day as a single sequence. Local segment $x_{seg} \in \mathbb{R}^{B\times L\times D}$ is first passed through a 1D convolution projection module. 1D convolution layers are applied along the time axis for each sensor channel. The per-sensor outputs are then concatenated and passed through a 1D projection layer to unify dimensions across sensors. The projected features are transposed and enhanced with positional encoding.

Next, the segment is encoded using a Local Transformer. This stack captures temporal dependencies within the segment. The result retains full temporal resolution, allowing the model to preserve fine-grained local variations.

E. Global Temporal Encoder

After local encoding of each temporal segment, all outputs are concatenated to form a unified sequence representing the entire day, preserving minute-level resolution: $\mathbb{R}^{B\times 1440\times d_{model}}$. Unlike approaches that reduce each segment to a single token before global aggregation, our design retains the full temporal granularity to preserve subtle transitions and variations that may be predictive of sleep quality. To encode long-range dependencies and behavioral trends across the full day, we apply sinusoidal positional encoding to the concatenated sequence, maintaining the temporal identity of each time step. The resulting sequence is passed through a Global Transformer Encoder. This module enables information flow across distant time points, allowing the model to capture high-level routines (e.g., evening wind-down behavior following afternoon activity).

To convert this sequence into a fixed-length vector suitable for downstream prediction, we employ an attention-based temporal pooling mechanism. This pooling operation dynamically assigns learned attention weights to each time step based on its relevance to the task, producing a context-sensitive summary vector $h_{pooled} \in \mathbb{R}^{d_{model}}$. This vector encapsulates both local and global behavior patterns indicative of sleep quality.

F. Additional User Embedding and Classification

We further enhance the model's personalization capacity by incorporating a user embedding vector $e_{user} \in \mathbb{R}^{256}$, which captures individual-specific traits and habitual patterns. The user embedding is concatenated with the pooled temporal representation $h_{pooled} \in \mathbb{R}^{d_{model}}$ to form a personalized feature vector:

$$z = [h_{pooled} \oplus e_{user}] \in \mathbb{R}^{d_{model}+256} \quad (1)$$

This vector is passed to six parallel classification heads, each responsible for predicting one of the sleep quality indicators: Q1, Q2, Q3, S1, S2, S3. Each head is a multi-layer perceptron (MLP) composed of fully connected layers. The final output layer produces logits of shape $\mathbb{R}^{C_k}$, where C_k is the number of classes for task k.

TABLE I. ACCURACY AND MARCO F1-SCORE COMPARISON

Model	Accuracy (%)	Macro F1-score
1D-CNN	54.07	0.3318
Bi-LSTM	52.59	0.5062
Transformer	49.07	0.4633
HTTMS (Ours)	**60.74**	**0.5954**

TABLE II. ACCURACY COMPARISON ON EACH LABEL

Model	Accuracy (%)					
	Q1	Q2	Q3	S1	S2	S3
1D-CNN	50.00	51.11	50.00	47.78	58.89	**66.67**
Bi-LSTM	57.78	**55.56**	54.44	37.78	55.56	54.44
Transformer	**62.22**	50.00	53.33	28.89	53.33	46.67
HTTMS (Ours)	60.00	**55.56**	**66.67**	**51.11**	**67.78**	63.33

We use the categorical cross-entropy loss for each task, regardless of the number of classes:

$$\mathcal{L} = \frac{1}{K}\sum_{k=1}^{K} CE(y_k, \hat{y}_k) \quad (2)$$

Where $\hat{y}_k \in \mathbb{R}^{C_k}$ is the logit vector predicted for task k, and $y_k \in \{0, 1, \ldots, C_k - 1\}$ is the corresponding ground-truth label. This formulation enables joint multi-task optimization across binary and multi-class tasks simultaneously.

IV. EXPERIMENTS

A. Experimental Setup

We evaluate our proposed model HTTMS on a sleep quality prediction task using multi-sensor daily life data, ETRI Lifelog Dataset 2024 [24], introduced in *Section III. B*.

We aim to assess the model's performance from three perspectives: (1) comparison with conventional sequence modeling baselines, (2) analysis of the contribution of core architectural components through ablation, and (3) sensitivity to the choice of input features and temporal segmentation.

To quantify model performance, we employ the following evaluation metrics: Accuracy. Captures the proportion of correct predictions among total predictions for each task. Marco-averaged F1 scores. Measures class-balanced performance across all labels, regardless of class imbalance. Per-label F1 scores. Evaluates the classification performance for each of the six sleep quality indicators (Q1-Q3, S1-S3).

In this study, we do not use data augmentation or ensemble methods, focusing instead on comparing architecture-level contributions under consistent training conditions.

B. Baseline Model Comparison

To validate the effectiveness of our proposed HTTMS model, we compare it against several widely used time-series baselines, namely 1D CNN [25], Bi-LSTM [26], and vanilla Transformer [9]. **Table I** reports overall accuracy and Macro F1-scores across models. Among the baselines, the Bi-LSTM achieves the best Macro F1-score (0.5062) and 1D CNN

yields the highest accuracy (54.07%), while the vanilla Transformer performs relatively poorly with 49.07% accuracy and 0.4633 F1-score. Our proposed HTTMS model substantially outperforms all baselines, reaching an accuracy of 60.74% and a Macro F1-score of 0.5954. The improvement margins over the strongest baselines exceed 6% in accuracy and nearly 9 points in F1-score, demonstrating the effectiveness of incorporating hierarchical architecture representations.

We further analyze per-label performance in **Table II.** HTTMS achieves the best results across most categories, with particularly large gains on S1 (51.11%), compared to Bi-LSTM (37.78%) and vanilla Transformer (28.89%). It also provides consistent improvements on S2 and S3, outperforming all baselines. On relatively easier labels such as Q3, HTTMS achieves 66.67% accuracy, substantially higher than both 1D CNN and Bi-LSTM. These results indicate that the hierarchical transformer in HTTMS captures fine-grained temporal cues that flat architectures fail to model.

In summary, both overall and per-label comparisons confirm that HTTMS consistently surpasses strong baselines. The hierarchical transformer architecture not only enhances global predictive performance but also ensures more balanced accuracy across different sleep-related labels, underscoring its utility in modeling complex physiological time-series data.

TABLE III. ABLATION STUDY OF SENSOR SELECTION

Sensor	feature	8	15
Smartphone	m_charging		✓
	m_activity	✓	✓
	m_ambience	✓	✓
	m_ble		✓
	m_gps	✓	✓
	m_light	✓	✓
	m_screen_use	✓	✓
	m_usage_stats	✓	✓
Smartwatch	heart_rate	✓	✓
	w_light	✓	✓
	step		✓
	step_frequency		✓
	distance		✓
	speed		✓
	burned_calories		✓
Accuracy (%)		**60.74**	**60.00**

TABLE IV. ABLATION STUDY OF USER ID EMBEDDING

User ID Embedding	Accuracy (%)	Marco F1-score
with	**60.74**	**0.5954**
without	53.37	0.5051

C. Ablation Study

Sensor Selection. To examine the contribution of each sensor modality, we conducted an ablation study, selectively including different subsets of smartphone and smartwatch signals. As shown in Table III, using a reduced set of 8 sensor features achieved an overall accuracy of 60.74%, which was slightly higher than the 60.00% obtained when all 15 features were included. This suggests that while additional features such as step-related and calorie-related measures may provide complementary information, they can also introduce redundancy or noise that does not necessarily translate into improved accuracy. Notably, heart rate, ambient light, and activity sensors consistently contributed to maintaining performance, indicating their robustness as core modalities. These results highlight that a carefully chosen subset of multimodal inputs can sometimes outperform the full set by reducing irrelevant variation, though incorporating diverse signals remains advantageous in scenarios where robustness across conditions or long-term generalization is prioritized.

User ID Embedding. We further investigated whether incorporating user-specific representations improves personalization and generalization. As summarized in **Table IV**, without any user embedding, the model achieved only 53.37% accuracy with a 0.5051 Macro F1-score. By contrast, when user ID embeddings were included, performance increased substantially to 60.74% accuracy and 0.5954 Macro F1-score. This gain of over 7 percentage points in accuracy and nearly 9 points in Macro F1 highlights the importance of modeling individual differences in sleep-related behavior. The embeddings allow HTTMS to adapt predictions to habitual routines and personal traits, thereby improving both predictive accuracy and user-specific relevance. These findings underscore the necessity of personalization in digital sleep health systems, where inter-individual variability is significant.

V. CONCLUSION

In this study, we introduced HTTMS, a hierarchical Transformer framework for multimodal sleep quality prediction. Unlike conventional models that rely on handcrafted features or flat sequence encoders, HTTMS explicitly models temporal structures by combining Local Transformers for short-term dynamics with a Global Transformer for long-range dependencies. To enhance personalization, we further incorporated user ID embeddings that encode habitual patterns of individuals, enabling the model to provide more user-specific predictions. Through experiments on the ETRI Lifelog Dataset 2024 , HTTMS achieved the highest accuracy and Macro F1-score among strong end-to-end baselines, including 1D CNN, Bi-LSTM, and vanilla Transformers. Per-label analysis confirmed that the hierarchical design and user-aware adaptation offer substantial improvements on challenging classes such as S1 and S2, demonstrating the ability to capture subtle behavioral cues.

An additional advantage of our framework lies in its end-to-end learning paradigm. By directly processing multimodal sensor streams without manual feature engineering, HTTMS ensures a streamlined workflow for both training and inference. While performance margins may be narrower than task-specific pipelines, the unified hierarchical and personalized design offers scalability and robustness for real-world applications. Overall, HTTMS demonstrates the promise of end-to-end, user-aware hierarchical Transformers

as a foundation for next-generation, behavior-aware digital sleep health systems.

REFERENCES

[1] R. Wang *et al.*, "StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Seattle Washington: ACM, Sept. 2014, pp. 3–14. doi: 10.1145/2632048.2632054.

[2] L. Piwek, D. A. Ellis, S. Andrews, and A. Joinson, "The rise of consumer health wearables: promises and barriers," *PLoS Med.*, vol. 13, no. 2, p. e1001953, 2016.

[3] D. Shome, N. M. Ghahjaverestan, and A. Etemad, "NapTune: Efficient Model Tuning for Mood Classification using Previous Night's Sleep Measures along with Wearable Time-series," in *International Conference on Multimodel Interaction*, San Jose Costa Rica: ACM, Nov. 2024, pp. 204–213. doi: 10.1145/3678957.3685722.

[4] M. A. Kredlow, M. C. Capozzoli, B. A. Hearon, A. W. Calkins, and M. W. Otto, "The effects of physical activity on sleep: a meta-analytic review," *J. Behav. Med.*, vol. 38, no. 3, pp. 427–449, June 2015, doi: 10.1007/s10865-015-9617-6.

[5] M. E. Billings, L. Hale, and D. A. Johnson, "Physical and social environment relationship with sleep health and disorders," *Chest*, vol. 157, no. 5, pp. 1304–1312, 2020.

[6] M. Alshareef, "Stress Detection: Leveraging IoMT Data and Machine Learning for Enhanced Well-being," PhD Thesis, Queen Mary University of London, 2025. Accessed: Aug. 25, 2025. [Online]. Available: https://qmro.qmul.ac.uk/xmlui/handle/123456789/110191

[7] S. Deldari, "Learning from multimodal time-series data with minimal supervision," PhD Thesis, RMIT University, 2024. Accessed: Aug. 25, 2025. [Online]. Available: https://research-repository.rmit.edu.au/articles/thesis/Learning_from_multimodal_time-series_data_with_minimal_supervision/27597498

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[9] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, Accessed: Aug. 19, 2025. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[10] S. Li *et al.*, "Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting," *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, Accessed: Aug. 19, 2025. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/6775a0635c302542da2c32aa19d86be0-Abstract.html

[11] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A Time Series is Worth 64 Words: Long-term Forecasting with Transformers," Mar. 05, 2023, *arXiv*: arXiv:2211.14730. doi: 10.48550/arXiv.2211.14730.

[12] S. Liu *et al.*, "Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting," in *Proceedings of the Tenth International Conference on Learning Representations (ICLR 2022)*, 2022. Accessed: Aug. 19, 2025. [Online]. Available: https://repositum.tuwien.at/handle/20.500.12708/135874

[13] H. Zhou *et al.*, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, 2021, pp. 11106–11115. Accessed: Aug. 19, 2025. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17325

[14] Y. Zhang and J. Yan, "Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting," in *The eleventh international conference on learning representations*, 2023. Accessed: Aug. 19, 2025. [Online]. Available: https://openreview.net/forum?id=vSVLM2j9eie

[15] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, "Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting," in *International conference on machine learning*, PMLR, 2022, pp. 27268–27286. Accessed: Aug. 19, 2025. [Online]. Available: https://proceedings.mlr.press/v162/zhou22g

[16] Q. Wen *et al.*, "Transformers in Time Series: A Survey," May 11, 2023, *arXiv*: arXiv:2202.07125. doi: 10.48550/arXiv.2202.07125.

[17] K. Sundararajan *et al.*, "Sleep classification from wrist-worn accelerometer data using random forests," *Sci. Rep.*, vol. 11, no. 1, p. 24, 2021.

[18] A. Logacjov, E. Skarpsno, A. Kongsvold, K. Bach, and P. J. Mork, "A Machine Learning Model for Predicting Sleep and Wakefulness Based on Accelerometry, Skin Temperature and Contextual Information," *Nat. Sci. Sleep*, vol. Volume 16, pp. 699–710, June 2024, doi: 10.2147/NSS.S452799.

[19] A. Sathyanarayana *et al.*, "Sleep quality prediction from wearable data using deep learning," *JMIR MHealth UHealth*, vol. 4, no. 4, p. e6562, 2016.

[20] O. Kilic, B. Saylam, and O. Durmaz Incel, "Sleep Quality Prediction from Wearables using Convolution Neural Networks and Ensemble Learning," in *Proceedings of the 2023 8th International Conference on Machine Learning Technologies*, Stockholm Sweden: ACM, Mar. 2023, pp. 116–120. doi: 10.1145/3589883.3589900.

[21] H. Lee, M. Cho, S. W. Lee, and S. S. Park, "Predicting sleep quality with digital biomarkers and artificial neural networks," *Front. Psychiatry*, vol. 16, p. 1591448, 2025.

[22] Y. Guo, M. Nowakowski, and W. Dai, "FlexSleepTransformer: a transformer-based sleep staging model with flexible input channel configurations," *Sci. Rep.*, vol. 14, no. 1, p. 26312, 2024.

[23] C. Wang *et al.*, "Constructing a Transformer-based Model to Infer Daytime Productivity from Biometric Information During Sleep," *Hum. Syst. Eng. Des.*, vol. 158, p. 175, 2024.

[24] S. W. Oh *et al.*, "Understanding Human Daily Experience Through Continuous Sensing: ETRI Lifelog Dataset 2024," July 18, 2025, *arXiv*: arXiv:2508.03698. doi: 10.48550/arXiv.2508.03698.

[25] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1D convolutional neural networks and applications: A survey," *Mech. Syst. Signal Process.*, vol. 151, p. 107398, 2021.

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