

# Sleep Quality Prediction from Wearables using Convolution Neural Networks and Ensemble Learning

OZAN KILIÇ\*, BERRENUR SAYLAM\*, and ÖZLEM DURMAZ İNCEL, Computer Engineering Department, Boğaziçi University, Turkey

Sleep is among the most important factors affecting one's daily performance, well-being, and life quality. Nevertheless, it became possible to measure it in daily life in an unobtrusive manner with wearable devices. Rather than camera recordings and extraction of the state from the images, wrist-worn devices can measure directly via accelerometer, heart rate, and heart rate variability sensors. Some measured features can be as follows: time to bed, time out of bed, bedtime duration, minutes to fall asleep, and minutes after wake-up. There are several studies in the literature regarding sleep quality and stage prediction. However, they use only wearable data to predict or focus on the sleep stage. In this study, we use the NetHealth dataset, which is collected from 698 college students' via wearables, as well as surveys. Recently, there has been an advancement in deep learning algorithms, and they generally perform better than conventional machine learning techniques. Among them, Convolutional Neural Networks (CNN) have high performances. Thus, in this study, we apply different CNN architectures that have already performed well in the human activity recognition domain and compare their results. We also apply Random Forest (RF) since it performs best among the conventional methods. In future studies, we will compare them with other deep learning algorithms.

CCS Concepts: • Computer Methodologies; • Deep Learning; • Artificial Intelligence; • Convolutional Neural Networks; • Ensemble Learning; • Wearables; • Student's Daily Life Activities;

Additional Key Words and Phrases: student's life dataset, convolutional neural networks, random forest, wearables, pervasive health

## ACM Reference Format:

Ozan Kılıç, Berrenur Saylam, and Özlem Durmaz İncel. 2023. Sleep Quality Prediction from Wearables using Convolution Neural Networks and Ensemble Learning. In . ACM, New York, NY, USA, 8 pages. <https://doi.org/XXXXXXX.XXXXXXX>

## 1 INTRODUCTION

Understanding the underlying factors of sleep quality may significantly improve people's living standards. There could be many factors that need to be discovered that affect sleep quality metrics. These factors can be exposed with the help of big data and accurate deep-learning models. This study aims to find meaningful relations between sleep measurements with comprehensive multi-modal data, including personal surveys and everyday wearable device data.

All the mentioned data are derived from the NetHealth open-source dataset <sup>1</sup>[11], which was collected from 698 students in total during eight semesters. This comprehensive study includes data on communication patterns from mobile phones, sleep, and physical activity routines, students' family backgrounds, living conditions, personality, etc., from surveys. This study uses some parts of the basic survey and all the wearable device data. Therefore, to extract the required data, preprocessing is applied to construct a sub-dataset for the analysis.

\*Both authors contributed equally to this research.

<sup>1</sup><http://sites.nd.edu/nethealth/>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

© 2023 Association for Computing Machinery.

Manuscript submitted to ACM

In the original NetHealth dataset, several sub-components, such as communication, wearable, survey, courses and grades, and calendar are included. Wearable measurements are performed for activity and sleep tracking using a Fitbit device. We used wearable and survey datasets for this study to construct a sub-dataset to apply deep learning and ensemble learning. However, this raw data form is incompatible with training deep learning algorithms. The preprocessing step will be further discussed in the paper.

We applied several CNN architectures which perform well in the human activity recognition domain [19]. As our measurements in the dataset are collected from wearables on a daily basis, we decided to focus on the architectures utilized for daily activity recognition tasks. Details of these architectures and their differences are further discussed in the paper. We also applied Random Forest (RF) as an ensemble learning method to compare the performance with deep learning architectures. RF is also used to extract important features for sleep quality prediction. Even though deep learning algorithms, especially CNNs, perform better compared to conventional machine learning techniques, our results show that in this context, RF performs better. We believe that it is highly related to dataset characteristics since in [2], authors also applied different deep learning techniques such as Multilayer Perceptron (MLP), Long-Short Term Memory Network (LSTM), Gated Recurrent Unit (GRU), Elastic Net (EN), and compared them with conventional machine learning technique; RF. They found RF to be performing better than other methods.

The rest of the paper is organized as follows: Section 2 explains state of the art on sleep recognition studies from the point of the wearable domain. Section 3 explains dataset details and the preprocessing steps for further analyses. In Section 4, we present model results with different proposed architectures. Finally, Section 5 discusses our findings with other future study ideas.

## 2 RELATED WORKS

Several related studies exist about the sleep quality prediction of people that use wearable devices and questionnaires as measurement metrics. We can emphasize the studies [1, 3, 7, 9, 11–13] as the most related ones to our study. Although the personal survey is not included in all the mentioned works, wearable devices are the key for all of them.

In [1], authors use three different sleep quality metrics: Daily Sleep Quality, Weekly Sleep Quality, and Sleep Consistency. CNN and MLP methods are applied. CNN is found to be the outperforming method compared to MLP on all these three quality metrics. There were three classes for daily and weekly sleep quality metrics indicating good to poor quality; and four classes for sleep consistency metric which is obtained by the weighted sum of the standard deviations of sleep, rest, bedtime, out-of-bedtime, and sleep inconsistency between weekdays and weekends. CNN obtains the best results on weekly sleep quality classification with 97.30% accuracy.

In [3], the main aim was correctly predicting the future sleep duration, which will help identify future sleep quality. As a model, they utilized a generalized linear mixed model (GLMM) to consider individual variability on sleep duration, which provides better estimates, hence better sleep quality predictions, in addition to a generalized linear model (GLM) for considering fixed effects.

In [7], sleep and activity metrics are used to estimate university students' well-being. The same data set is used in this study, NetHealth data. Daily sleep and activity data is used to train a classification algorithm. The study aims to predict the "Subjective Well-being" score of people from average daily steps, average heart rate, heartbeat standard deviation, average sleep duration, and sleep duration deviation gathered from wearable devices. They used Naive Bayes, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Ensemble classifiers as methods to train.

In [9], data from 39 students using Fitbit devices is collected during 106 days. They focused on predicting two classes: good quality and poor quality of sleep based on seven features, such as calories, steps, distance, sedentary, lightly active,

fairly active, very active. They applied Random Forest as a conventional machine learning method and LSTM, GRU, and CNN as deep learning methods. Their obtained results indicate that CNN could improve sleep quality prediction from activities. But the best performing model is reported as GRU.

In [11], authors discuss the compliance of Fitbit usage among college students and the reasons for non-compliance. They observed patterns for general health, sleep, and exercise. Mainly, a contrastive pattern between weekdays and weekends, regular school days, and deadline-intensive patterns are observed among the ones with a compliance rate higher than 60%.

Study [12] has a wide range of target values while using personality traits, wearable device data, and mobile phone data. They collected multi-modal data from 66 people for a month. This multi-modal data includes perceived stress, sleep, personality, physiological, behavioral, and social interaction data to train a model to divine academic performance, sleep, stress, and mental health scores. They used SVM-L and SVM-RBF models to train and compared the GPA (Academic Performance), PSQI (Pittsburg Sleep Quality Index), PSS (Perceived Stress Scale), and MCS (Mental Health Composite Score) results for different training datasets. As sleep was not a target in the scope of this study, they examined only its relation with other terms. It is found that poor sleep quality and high-stress levels are highly correlated. In addition, the poor sleep quality group has been found to have more screen time around 3-6 am and less around 9 am-12 pm.

In [13], the aim is to classify sleep quality with a deep learning model trained with wearable device data. They used the physical activity data collected during awake time. The target values of the prediction model are poor or good sleep quality. The wearable measurements are done using an Actigraphy device which is accepted as a gold-standard device for clinical studies of sleep and physical activity patterns. Data is collected from 92 adolescents over one whole week. Their methods are traditional logistic regression as a conventional ML method and more advanced deep learning methods: multilayer perceptron (MLP), convolutional neural network (CNN), simple Elman-type recurrent neural network (RNN), long short-term memory (LSTM-RNN), and a time-batched version of LSTM-RNN (TB-LSTM). At the end of the study, they revealed that the CNN has the highest accuracy with 0.9449 while the traditional logistic regression has a score of 0.6463 accuracies.

Our study differs from the literature with its extensive feature sets and its nature to predict sleep efficiency, contrary to already applied classification techniques. We contribute to the literature by applying CNN architectures which already performed well in the human activity recognition domain. We also compare the performance with an ensemble model, i.e., RF. It is found that RF performs better compared to CNN architectures in the scope of our experiments.

### 3 METHODOLOGY

#### 3.1 Dataset and Features

The data used in this study is gathered from the Nethealth dataset, which is collected at Notre Dame University. This data covers eight semesters between Fall 2015 and Spring 2019. There are data on approximately 400 students from 2015 to 2017 and 300 from 2015 to 2019. Data collection comprises social networks, physical activity, sleep data, basic survey data, communication data, courses and grades data, and academic calendar data. Basic survey data include family background & demographic variables, student background, self-reported course, grade, major information, activities, clubs, musical tastes, personality - Big five, and social-psychological scales tests. Besides these, physical and psychological health-related questions, PSQI sleep quality test, political attributes and views, and computer, phone, and Fitbit device usage of students are included.

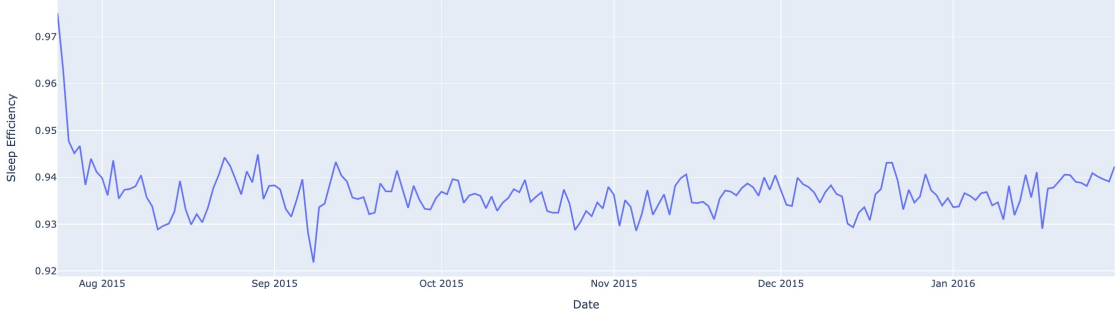


Fig. 1. Sleep Efficiency for first and second semesters

NetHealth dataset is a comprehensive, multi-year longitudinal dataset. So, some survey questions were only asked in some wave. The missing information restricts this study from using every wave in deep learning training. In this study, we focus on Fitbit’s sleep quality scores as our target value.

In the scope of this study, we chose all wearable data modalities, i.e., activity and sleep. However, wearable sleep data is chosen as the target value. Thus, it is not in the feature space. In addition, we added questionnaire responses about bad habits, personality (BigFive), exercise, health, mental health, personal information, origin, and sex. The used sub-datasets and corresponding features are listed in study [14] in Table 1. Due to space limit we were not able to add the adjusted version in this paper. However, the parameter space remains the same but removed the courses and grades features. There, the underscored numbers correspond to the semester’s indications. The feature space size before pre-processing is 89 without participantID, date, and the target value, which is sleep efficiency.

### 3.2 Preprocessing

As a target value in the scope of this study, we use the Fitbit sleep efficiency score that is collected from the wearable - Fitbit. The device uses its sensors to detect if someone is awake or asleep. Besides this information, total time in bed, awake time in bed, and sleeping time are also detected. Thanks to these values, the sleep efficiency score can be calculated, i.e.,  $\text{minsasleep} / (\text{minsasleep} + \text{minsawake})$ . The score is the ratio of asleep time in bed to the total time in bed and ranges between 0 and 1.

Before model evaluation, we concatenated all separated responses according to timestamps and participants. In addition, we converted the survey questions into numeric models. Besides, there were missing values in some survey responses for different semesters. In this study, we are concentrating on the first and second semester’s data due to not being requested from the participants since the decided features have responses. Furthermore, students who still need to answer the questionnaires are subtracted. In the end, we have 200615 rows and 92 columns. In Figure 1, the change of our target variable, i.e., sleep efficiency, over the first and second semesters can be observed. It consists of all participants’ overall patterns in two semesters. Its range is between 0.93 and 0.95. There were two peaks; one at the beginning of the study in August 2015 and the other around September 2015. However, we can not comment on the possible reasons since we do not know about the events that occur these days.

### 3.3 Model Details and Performance Metrics

In this study, we apply deep learning techniques as they perform better in literature [19]. More specifically, we are concentrating on Convolutional Neural Networks (CNN) architectures that perform well in the human activity recognition domain [5, 6, 16–18]. As our measurements in the dataset are collected daily from wearables, we decided to focus on the architectures utilized for daily activity recognition tasks. The selected architectures are listed in Table 1. C, P, and FC stand for convolution layer, max-pooling layer, and fully connected layer, respectively. We applied these selected architectures to both scenarios. The models differ from each other primarily on their complexities, and also, we chose different kernel sizes already performed well, which are 3, 5, 12, 20. We used ReLU activation for convolution and linear activation for fully connected layers. Filter and pool size remained the same for all architectures, 32 and 2, respectively.

Since our problem is prediction, mean squared error (MSE) and mean absolute error (MAE) can be utilized as performance metrics. MAE is calculated by considering the absolute average distance. MSE is calculated by averaging the squared difference between the estimated and actual values. In our context, we used MAE due to its interpretability.

## 4 EVALUATION

### 4.1 Comparison of CNN models and Random Forest

In the dataset, we have continuous measurements for sleep efficiency. Thus, we modelled them as a prediction problem. Again, we employed the same architectures, and obtained results are presented in Table 2. Architecture 2 is the best performer based on MAE values. The change of MAE for the best one (C-P-C-P-FC) is given in Figure 2. In addition, we implemented RF with 10 estimators; its performance was found to be better than the CNN models. We think that this may be related to the dataset characteristics since in [2], RF outperformed different deep learning models on the same dataset but on a different task.

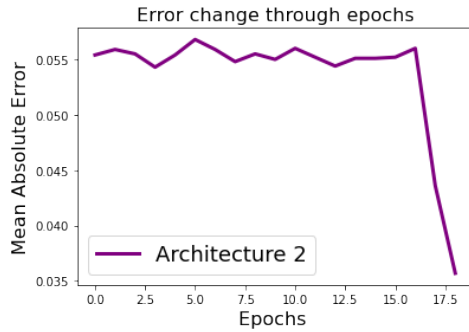


Fig. 2. Evaluation of the best architecture

Architecture	Mean Absolute Error
C-P-FC-FC	0.0876
C-P-C-P-FC	<b>0.0357</b>
C-P-C-P-FC	0.0429
C-C-P-C-C-P-FC	0.0844
C-P-C-P-FC-FC	0.0679
C-P-C-P-C-P	0.9361
Random Forest	<b>0.0282</b>

Table 2. Sleep Quality Prediction Results

### 4.2 Important Factors

To understand the most affecting factors to sleep efficiency, we employed RF feature importance. 20 top affecting ones are presented in Figure 3. They are mostly related to Fitbit activity measurements. We got SRQE\_introj (introjective self-regulation for exercise), CESDOverall (CES depression score), Neuroticism (personality trait) from survey responses. Also, we found mom age affects sleep efficiency somehow.

Table 1. Details of Model Parameters

Architecture	Layer	Activation	Kernel Size	Pool Size	Filters	Input Size	Output Size
C-P-FC-FC	1D Conv.	ReLU	20	-	32	(93, 1)	(74, 32)
	Pooling	-	-	2	-	-	(37, 32)
	Dense	Linear	-	-	-	(93, 1)	(37, 16)
	Dense	Linear	-	-	-	-	(37, 16)
	Flatten	-	-	-	-	-	592
C-P-C-P-FC	1D Conv.	ReLU	3	-	32	(93, 1)	(91, 32)
	Pooling	-	-	2	-	-	(45, 32)
	1D Conv.	ReLU	3	-	32	-	(43, 32)
	Pooling	-	-	2	-	-	(21, 32)
	Dense	Linear	-	-	-	(50, 1)	(21, 16)
	Flatten	-	-	-	-	-	336
C-P-C-P-FC	1D Conv.	ReLU	5	-	32	(93, 1)	(89, 32)
	Pooling	-	-	2	-	-	(44, 32)
	1D Conv.	ReLU	5	-	32	-	(40, 32)
	Pooling	-	-	2	-	-	(20, 32)
	Dense	Linear	-	-	-	(50, 1)	(20, 16)
	Flatten	-	-	-	-	-	320
C-C-P-C-C-P-FC	1D Conv.	ReLU	5	-	32	(93, 1)	(89, 32)
	1D Conv.	ReLU	5	-	32	-	(85, 32)
	Pooling	-	-	2	-	-	(42, 32)
	1D Conv.	ReLU	5	-	32	-	(38, 32)
	1D Conv.	ReLU	5	-	32	-	(34, 32)
	Pooling	-	-	2	-	-	(17, 32)
	Dense	Linear	-	-	-	(93, 1)	(17, 16)
	Flatten	-	-	-	-	-	272
C-P-C-P-FC-FC	1D Conv.	ReLU	5	-	32	(93, 1)	(89, 32)
	Pooling	-	-	2	-	-	(44, 32)
	1D Conv.	ReLU	5	-	32	-	(40, 32)
	Pooling	-	-	2	-	-	(20, 32)
	Dense	Linear	-	-	-	(93, 1)	(20, 16)
	Dense	Linear	-	-	-	-	(20, 16)
	Flatten	-	-	-	-	-	320
C-P-C-P-C-P	1D Conv.	ReLU	12	-	32	(93, 1)	(82, 32)
	Pooling	-	-	2	-	-	(41, 32)
	1D Conv.	ReLU	12	-	32	-	(30, 32)
	Pooling	-	-	2	-	-	(15, 32)
	1D Conv.	ReLU	12	-	32	-	(4, 32)
	Pooling	-	-	2	-	-	(2, 32)
	Flatten	-	-	-	-	-	64

We re-ran the indicated architectures using only the selected most important factors. However, we got higher MAE values compared to utilizing all factors. We can conclude that in this context even though the final performance contribution is low for some features, when we exclude them, the overall prediction performance decreases.

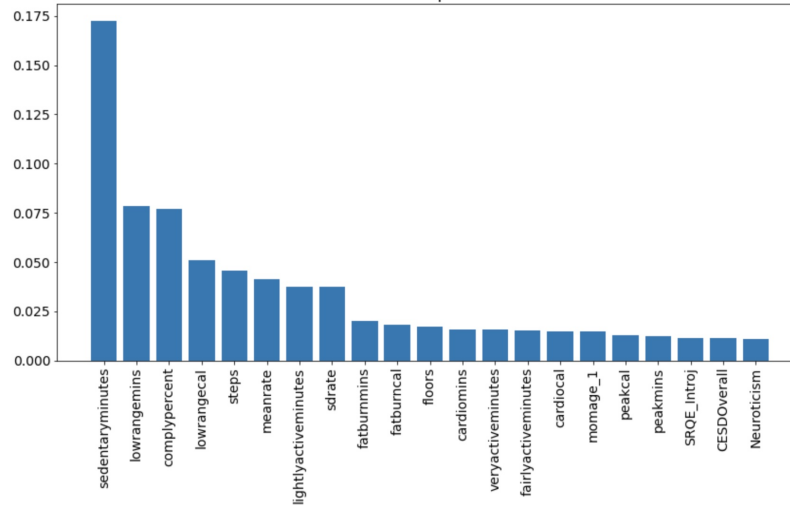


Fig. 3. Most Important Factors for Sleep Quality Prediction

## 5 DISCUSSION AND CONCLUSION

In this study, we predict sleep efficiency values collected from Fitbit devices. We applied the most promising CNN architectures in the human activity domain. We also implemented conventional RF to be able to compare deep learning results. We also extracted the most affecting factors using the feature importance technique.

Even though CNN predictions are promising and the best architecture is found as C-P-C-P-FC, RF performs the best in our context but the difference is small in terms of MAE. This is probably due to the nature of the dataset since in [2], authors applied different deep learning techniques, and it was found that RF performs the best in all cases.

We also experimented with the reduced size of features in the same scenario. However, it is found that deep learning without feature extraction performs better compared to manually extracted features. Nevertheless, this feature ranking provides us with a meaning for the interpretation. Among the best 20 factors, sleep efficiency is affected by the measurements from wearables related to the activity. There are also some survey-related factors, such as exercise, depression, and personality-related responses. These are expected since, in literature, it is already shown that there is a relation between sleep, stress, physical activity, and mood [4, 8, 10, 15]. We contribute by working all these modalities in one study in a multi-modal manner by applying deep learning methods over wearable and survey data collected in an unrestricted environment.

We want to give ideas for future perspectives. In this study, we worked on a prediction problem, but the target value range was narrow (between 0.93 and 0.95). There were no high fluctuations. This may be the reason for obtaining very accurate predictions and low errors. In addition, we focused on the application of algorithms by not considering personal differences. This aspect may be examined in future studies. Furthermore, we did not consider the time-based differences for a participant. One may examine personal changes over time periods, especially during exam dates and holidays.

## 6 ACKNOWLEDGMENTS

Tübitak Bideb 2211-A academic reward is gratefully acknowledged. This research has been supported by the Boğaziçi University Research Fund, project number: 19301P.

## REFERENCES

- [1] Anshika Arora, Pinaki Chakraborty, and MPS Bhatia. 2020. Analysis of data from wearable sensors for sleep quality estimation and prediction using deep learning. *Arabian Journal for Science and Engineering* 45, 12 (2020), 10793–10812.
- [2] Brandon M Booth, Hana Vrzakova, Stephen M Mattingly, Gonzalo J Martinez, Louis Faust, and Sidney K D’Mello. 2022. Toward Robust Stress Prediction in the Age of Wearables: Modeling Perceived Stress in a Longitudinal Study with Information Workers. *IEEE Transactions on Affective Computing* 13, 4 (2022), 2201–2217.
- [3] Chih-You Chen, Sudip Vhaduri, and Christian Poellabauer. 2020. Estimating Sleep Duration from Temporal Factors, Daily Activities, and Smartphone Use. In *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*. 545–554. <https://doi.org/10.1109/COMPSAC48688.2020.0-196>
- [4] Shruti Gedam and Sanchita Paul. 2021. A review on mental stress detection using wearable sensors and machine learning techniques. *IEEE Access* 9 (2021), 84045–84066.
- [5] Sojeong Ha and Seungjin Choi. 2016. Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors. In *2016 international joint conference on neural networks (IJCNN)*. IEEE, 381–388.
- [6] Artur Jordao, Leonardo Antônio Borges Torres, and William Robson Schwartz. 2018. Novel approaches to human activity recognition based on accelerometer data. *Signal, Image and Video Processing* 12 (2018), 1387–1394.
- [7] Akif Can Kılıç, Ahmet Karakuş, and Emre Alptekin. 2022. Prediction of University Students’ Subjective Well-Being with Sleep and Physical Activity Data using Classification Algorithms. *Procedia Computer Science* 207 (2022), 2648–2657.
- [8] David Marcusson-Clavertz, Martin J Sliwinski, Orfeu M Buxton, Jinhyuk Kim, David M Almeida, and Joshua M Smyth. 2022. Relationships between daily stress responses in everyday life and nightly sleep. *Journal of behavioral medicine* 45, 4 (2022), 518–532.
- [9] Dinh-Van Phan, Chien-Lung Chan, and Duc-Khanh Nguyen. 2020. Applying Deep Learning for Prediction Sleep Quality from Wearable Data. In *Proceedings of the 4th International Conference on Medical and Health Informatics*. 51–55.
- [10] June J Pilcher and Allen I Huffcutt. 1996. Effects of sleep deprivation on performance: a meta-analysis. *Sleep* 19, 4 (1996), 318–326.
- [11] Rachael Purta, Stephen Mattingly, Lixing Song, Omar Lizardo, David Hachen, Christian Poellabauer, and Aaron Striegel. 2016. Experiences Measuring Sleep and Physical Activity Patterns across a Large College Cohort with Fitbits. In *Proceedings of the 2016 ACM International Symposium on Wearable Computers (Heidelberg, Germany) (ISWC ’16)*. Association for Computing Machinery, New York, NY, USA, 28–35. <https://doi.org/10.1145/2971763.2971767>
- [12] Akane Sano, Andrew J Phillips, Z Yu Amy, Andrew W McHill, Sara Taylor, Natasha Jaques, Charles A Czeisler, Elizabeth B Klerman, and Rosalind W Picard. 2015. Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones. In *2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE, 1–6.
- [13] Aarti Sathyanarayana, Shafiq Joty, Luis Fernandez-Luque, Ferda Ofli, Jaideep Srivastava, Ahmed Elmagarmid, Teresa Arora, Shahradd Taheri, et al. 2016. Sleep quality prediction from wearable data using deep learning. *JMIR mHealth and uHealth* 4, 4 (2016), e6562.
- [14] Berrenur Saylam, Ekrem Yusuf Ekmekci, Eren Altunoğlu, and Ozlem Durmaz Incel. 2022. Academic Performance Relation with Behavioral Trends and Personal Characteristics: Wearable Device Perspective. (2022).
- [15] Andrew Steptoe, Katie O’Donnell, Michael Marmot, and Jane Wardle. 2008. Positive affect, psychological well-being, and good sleep. *Journal of psychosomatic research* 64, 4 (2008), 409–415.
- [16] Triwiyanto Triwiyanto, I Putu Alit Pawana, and Mauridhi Hery Purnomo. 2020. An improved performance of deep learning based on convolution neural network to classify the hand motion by evaluating hyper parameter. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 28, 7 (2020), 1678–1688.
- [17] Yang Xu and Ting Ting Qiu. 2021. Human activity recognition and embedded application based on convolutional neural network. *Journal of Artificial Intelligence and Technology* 1, 1 (2021), 51–60.
- [18] Ming Zeng, Le T Nguyen, Bo Yu, Ole J Mengshoel, Jiang Zhu, Pang Wu, and Joy Zhang. 2014. Convolutional neural networks for human activity recognition using mobile sensors. In *6th international conference on mobile computing, applications and services*. IEEE, 197–205.
- [19] Shibo Zhang, Yaxuan Li, Shen Zhang, Farzad Shahabi, Stephen Xia, Yu Deng, and Nabil Alshurafa. 2022. Deep learning in human activity recognition with wearable sensors: A review on advances. *Sensors* 22, 4 (2022), 1476.

Received 30 January 2023; accepted 10 February 2023