



# Forecasting Health and Wellbeing for Shift Workers Using Job-Role Based Deep Neural Network

Han Yu<sup>1</sup>(✉), Asami Itoh<sup>2</sup>, Ryota Sakamoto<sup>2</sup>, Motomu Shimaoka<sup>2</sup>,  
and Akane Sano<sup>1</sup>

<sup>1</sup> Rice University, Houston, TX 77005, USA

{Han.Yu, akane.sano}@rice.edu

<sup>2</sup> Mie University, Mie 514-8507, Japan

{amasui, sakamoto}@clin.medic.mie-u.ac.jp, shimaoka@doc.medic.mie-u.ac.jp

**Abstract.** Shift workers who are essential contributors to our society, face high risks of poor health and wellbeing. To help with their problems, we collected and analyzed physiological and behavioral wearable sensor data from shift working nurses and doctors, as well as their behavioral questionnaire data and their self-reported daily health and wellbeing labels, including alertness, happiness, energy, health, and stress. We found the similarities and differences between the responses of nurses and doctors. According to the differences in self-reported health and wellbeing labels between nurses and doctors, and the correlations among their labels, we proposed a job-role based multitask and multilabel deep learning model, where we modeled physiological and behavioral data for nurses and doctors simultaneously to predict participants' next day's multidimensional self-reported health and wellbeing status. Our model showed significantly better performances than baseline models and previous state-of-the-art models in the evaluations of binary/3-class classification and regression prediction tasks. We also found features related to heart rate, sleep, and work shift contributed to shift workers' health and wellbeing.

**Keywords:** Shift workers · Health · Wellbeing · Wearables · Mobile sensor · Deep learning

## 1 Introduction

Around 20% of the workforce in the world involves in shift work [48]. Their irregular shift work brings a high risk of poor health and wellbeing. For example, shift work disrupts workers' circadian rhythms and causes problems such as sleep

---

Supported by National Science Foundation # 1840167 and Japan Agency for Medical Research and Development.

disorder and insomnia [10]. In addition to the sleep issues, decreased alertness levels were found in healthy shift workers [9], which could lead to occupational errors and accidents. Previous studies also showed the potential associations between shift work and pathological disorders such as fatigue, gastrointestinal malfunction [19], and an increased risk of colorectal cancer in night shift nurses [38]. Moreover, more adverse mental health outcomes, emotional exhaustion, and burnout were observed in shift workers compared to daytime workers [5, 15, 40, 43, 46]. In health care domain, physician burnout is estimated to cost 4.6 billion USD per year [13].

To support shift worker’s health and wellbeing, monitoring and predicting their day-to-day health and wellbeing trajectories and providing aids to help them prepare for challenging situations might be useful. Besides, mobile devices, such as smartphones and wearable sensors, have become parts of people’s daily life, and have been used to detect and predict self-reported health and wellbeing with the help of machine learning models [2, 16, 21, 41, 42, 49]. These previous works targeted health and wellbeing detection or prediction as binary classification [2, 41], 3-class classification [25, 49], and regression tasks [1, 16, 49]. Some of these works developed personalized models by taking participants’ demographic information into account [41] or fine-tuning general models to specific users [50]. Correlations among self-reported multi-dimensional labels - including subjective mood, health, and stress- were also used in building multilabel neural network models [41]. In addition, there are some prior works in monitoring shift workers using wearable sensors. Feng *et al.* extracted a behavioral consistency feature from shift worker wearable data and estimated anxiety levels with an accuracy of 57.8% in binary classification. Mulhall *et al.* used sensors integrated in the vehicles to monitor shift workers’ eye blinking as a marker of alertness [27] while driving. Actigraphy has been also used widely for studying sleep for shift work nurses [11, 17].

Although these previous works have achieved promising results, there is no work to thoroughly monitor and analyze different job types of shift workers’ multidimensional wellbeing and forecast them using machine learning. Furthermore, the models developed previously considered the heterogeneity among participants and correlation among wellbeing labels separately; however, since these two characteristics ubiquitously co-exist, modeling them simultaneously for different job types of shift workers might improve prediction model performance.

In this work, we collected physiological and behavioral data from hospital shift workers, then we developed machine learning models to predict their next day’s wellbeing in binary/3-class classifications and regression tasks. We also verified the rationale of leveraging job role information and multi wellbeing labels simultaneously in the models by analyzing the data. Then, we proposed a multitask multilabel deep learning model that leveraged job role information and correlations among self-reported health and wellbeing labels.

Our contributions can be summarized as: (i) we collected physiological and behavioral data from hospital shift workers, including nurses and doctors, (ii) we analyzed their physiological and behavioral patterns and found similarities

and differences, (iii) we developed a multitask multilabel deep learning model to predict participants' near future wellbeing using wearable sensor, surveys, their job role information, and correlations among wellbeing labels. The details of our proposed model structure, implementation and hyper-parameter information are shared on: <https://github.com/comp-well-org/multitask-multilabel-wellbeing-prediction>.

## 2 Related Work

There are numerous studies on shift workers' health and wellbeing. Heath *et al.* collected survey data from shift work nurses, and applied statistical analysis in exploring the association among their work shift types, sleep, mood, and diet [14]. They showed that shift work was significantly negatively related to shift workers' diet, sleep efficiency, and stress levels. Similarly, Books *et al.* analyzed questionnaire data from shift-working nurses and showed an increased risk of sleep deprivation, family stressors, and mood changes due to the night work shift [3].

In addition, with the rapid development of mobile devices and mobile applications, objective data from wearables and smartphones have been used for studying shift workers. For example, Pereira *et al.* collected wearable accelerometer data from hospital shift workers and detected 4 levels of their physical activity intensity with an 83% accuracy score [31]. Feng *et al.* used wearable devices to collect physiological data from shift work nurses for ten weeks and applied a clustering method for extracting behavioral consistency, which intuitively captures unique behavioral patterns between different groups of nurses [7]. They further found that behavioral consistency can help predict self-reported work behaviors and anxiety levels. In another work, Feng *et al.* analyzed physiological and indoor location data from nurses with Fitbit wrist-wearable devices and Bluetooth hubs [6]. They extracted mutual information features and demonstrated the dependency between an individual's movement patterns and physiological responses.

Machine learning models have been designed for detecting or predicting health and wellbeing using mobile and sensor data. For example, Bogomolov *et al.* developed daily stress detection algorithms based on five-month-long weather, mobile phone data (e.g., calls, SMS, and screen usage), and personality survey data from 117 participants [2]. They obtained stress detection accuracy up to 72% in binary classification tasks. In Moodscope paper, mood (1: negative to 5: positive) was detected with the best mean squared error of 0.229 using the data from the mobile phone and a personalized linear regression model [21]. Similarly, Asselbergs *et al.* detected the current mood using mobile phone data with a mean squared error of 0.15 out of  $-2$  to  $2$  mood scale [1]. For further improving the model performance, Taylor *et al.* developed a multitask machine learning model to predict high/low self-reported stress, mood, and health and separately used (i) the demographic information such as gender and personalities of participants and (ii) correlations among labels [41]. This work also inspired

us to use the combination of job role information and label correlations. In this work, we study the differences in daily self-reported health and wellbeing, physiology, and behavior between nurses and doctors, and focus on estimating shift workers' health and wellbeing using the data from mobile sensors and surveys and job-role based deep learning models.

### 3 Methods

#### 3.1 Data Collection

Two hundred and forty-one days of multi-modal data were collected from 14 shift workers, including 10 nurses (one male) and 4 doctors (all males) in a hospital in Japan. The average age of all participants was 31.4 years old, with a standard deviation (SD) of 4.2. For each study day, participants wore a Fitbit wristwatch (Fitbit Charge 3) for monitoring their physiological and behavioral activities such as heart rate, sleep, and step counts. The data sampled every 1 min was downloaded from the Fitbit server for data analysis and modeling. In addition, participants filled out daily morning and evening questionnaires to record their behavioral activities, including sleep, work schedule, and caffeinated drinks, alcohol & drug intake.

Self-reported health and wellbeing labels - including alertness, happiness, energy, health, and stress - were also collected in the morning questionnaire using 0 to 100 scales, with 0 to the most negative and 100 being the most positive (sleepy-alert, sad-happy, sluggish-energetic, sick-healthy, stressed-calm).

#### 3.2 Features

We calculated the following features from the Fitbit data and daily questionnaires:

**Heart Rate.** Heart rate and heart rate variability are related to work stress [44] and mood [39]. Based on heart rate collected from Fitbit sensor every 1 min, we computed features including daily mean, standard deviation (SD), and entropy of heart rates. We computed sample entropy of heart rate, which represented the self-similarity of a sequence and has been used in physiological time-series data analysis [34]. To calculate the sample entropy, we first need to set an embedding dimension  $m$ . Using the given  $m$ , our sequence  $X$  with length  $N$  can be divided into  $N - m + 1$  sliding windows  $\{X_1, \dots, X_{N-m+1}\}$ , where  $X_i = \{x_i, x_{i+1}, \dots, x_{i+m-1}\}$ . The equation of sample entropy is:

$$SampEn = -\log \frac{U^m}{U^{m+1}} \quad (1)$$

where

$$U^m = -\frac{1}{N-m} \sum_{i=1}^{N-m} U_i^m \quad (2)$$

$$U_i^m = \frac{[\# \text{ of } j \mid d(X_i, X_j) < r]}{N - m - 1} \quad (3)$$

In our case, the distance  $d$  is:

$$d(X_i, X_j) = \max |X_i - X_j| = \max_{k=1, \dots, m} |x_{i+k-1} - x_{j+k-1}| \quad (4)$$

Generally,  $m = 2$  and  $r = [0.2 * (\text{SD of } X)]$ .

**Sleep.** From Fitbit sensors, we obtained sleep duration and sleep efficiency. Then, we calculated the mean and SD values of sleep duration and sleep efficiency across the previous 7, 5, and 3 days. Moreover, using sleep data in one-minute resolution, we calculated sleep regularity with sliding windows across 7 days of participants' data. Sleep regularity is a value of 0–1 based on the likelihood of sleep/wake state being the same time-points 24 h apart, and is associated with health, wellbeing, and academic performance in college students [8, 32, 37]. From daily surveys, we obtained a daily feature of the time taken to fall asleep in minutes. Participants also reported how they woke up in the morning: waking up naturally, being awakened by the alarm, or other than alarm. Naps have been shown a positive impact on shift workers' performance, alertness [33], and wellbeing [20]. From participants' questionnaires, we summarized the times and total duration of naps across a day.

**Steps.** Total daily number of steps and minute by minute number of steps were recorded in the Fitbit dataset. To measure the variability of participants' physical activities, we computed the mean and SD to indicate step variability across the previous 7, 5, and 3 days. Excluding the sleep time, we counted the minutes of: (i) duration of segments without steps (stationary segments) and (ii) duration of segments with continuous steps (active segments) in 1-min bins. We used the following information entropy equation to calculate the entropy of the two types of physical activity based stationary and active segments:

$$En = - \sum_i p_i \log p_i \quad (5)$$

where  $p_i$  represents the probability that the  $i_{th}$  item was observed.

**Work.** Work schedules and work hours are directly related to symptoms such as sleep disorders and chronic fatigue [4]. Also, excessive work hours are harmful to workers' health and wellbeing [12]. We engineered work related features such as daily work shifts, total work duration per day, and overwork duration in minutes according to participants' answers in the questionnaires. There were three different work shifts, and each shift was for eight hours (1: 8:30-16:30, 2: 16:30-0:30, 3:0:30-8:30). Total work duration was actual work time, and the overwork duration was the difference between the actual work hours and the scheduled hours.

**Caffeine, Alcohol and Drug Use.** Considering caffeine, alcohol, and drug intake affects workers’ alertness [29,35], we computed features related to the intake of caffeinated drinks, drug, and alcohol based on the participants’ reports: the number of caffeinated drinks per day, and a binary feature for indicating whether the participant had drug or alcohol each day.

### 3.3 Statistical Analysis of Physiological and Behavioral Features Between Nurses and Doctors

We applied statistical tests to analyze the differences of physiological and behavioral features between two groups, nurses and doctors. Seventy-seven days of data were in the group of doctors, and 164 days of data were in the nurses’ group. Between 2 groups, we compared the numeric features such as daily average heart rate, steps, and overwork time using Mann-Whitney U test (non-normally distributed features) [23] and Welch’s t-test (normally distributed features) [45], whereas the categorical features such as awakening types and working shifts were compared with chi-square test [30].

### 3.4 Job-Role Based Multitask Multilabel Neural Network

Neural networks have been widely used in various areas, including face detection [36], mood, health, and stress prediction [41]. These previous outstanding works showed that the design of neural network structure needs the consideration of unique characteristics of data sets used in different applications. As discussed briefly in Sect. 1, in this work, we considered two important aspects: (1) different distributions in health and wellbeing labels based on our participants’ demographic information and (2) correlations among health and wellbeing labels. We observed differences in the distributions of self-reported health and wellbeing labels from two job roles, nurses and doctors. Also, there are correlations among the five labels. The details of the data statistics will be discussed in Sect. 5.1.

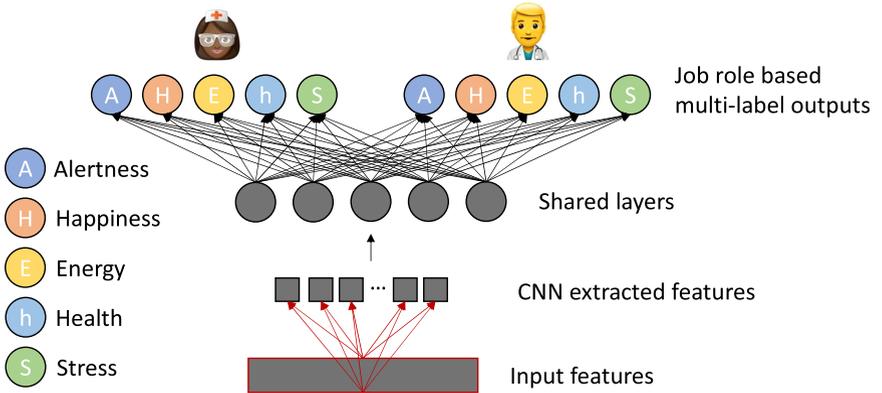
To learn different representations corresponding to participant job roles, we applied a multitask learning method, which divided tasks according to participants’ job roles. Furthermore, as another form of multitask learning, we used multilabel learning for considering different health and wellbeing labels as tasks. In this way, the model would also fit the correlation among labels. In this work, we designed a job-role based multitask and multilabel neural network model that leveraged user demographic information and correlations among labels at the same time. Figure 1 shows a simplified version of our model. When training the model, there might be redundant features in our input data that would not help health and wellbeing prediction. In contrast, some non-linear combinations of features might improve our model performance. Thus, we applied a one-dimension convolutional neural network (CNN) layer to extract auto-features from our inputs. As shown in Fig. 1, we designed convolutional kernels to learn higher-level features across every day feature vectors: 32 row-wise convolutional kernels embedded 32 channels of new features. Then, the CNN extracted features were fed into the multitask neural network. The shared layers in the network

learn the representation from all participant data, and the divided branches of the network structure learn the representation independently from participants in different job roles, nurses and doctors. When doctors' data are fed into the model for training, the weights of loss and optimizer of the nurse branch will be set to 0, and vice versa. Furthermore, each branch of the network outputs all five labels (alertness, happiness, energy, health, and stress) from the shared network layers. Therefore, the outputs of our model simultaneously provide the prediction of all five labels for nurses and doctors. The batch loss function of our model can be represented as:

$$L = \sum_{nurse} L_{ml} + \sum_{doctor} L_{ml} \quad (6)$$

$$L_{ml} = \sum_{l=\{alert,happy,energy,health,stress\}} loss(x, y_l) \quad (7)$$

Where  $x$  and  $y$  represent the input data and the expected output target, respectively.  $loss$  is mean squared error loss in regression tasks and cross-entropy loss in the classification tasks.



**Fig. 1.** A simplified version of our job-role based multitask multilabel neural network. Convolutional neural network kernels are applied for extracting high-level features. Our health and wellbeing prediction is designed for nurses and doctors using a portion of the network trained only using data from either nurses or doctors. Shared layers learn representation from all participants. The final output layers provide the prediction of all five labels simultaneously.

## 4 Experiments

Our tasks are formulated in two ways for evaluation: regression and classification tasks. The regression task is to predict the next day's health and wellbeing scores,

each in the range of 0–100, whereas the classification task is to predict next day’s high/low (binary classification, defined as 100-51, 50-0) or high/mid/low health and wellbeing levels, and high/ mid/low (3-class classification, defined as 100-67, 66-34, or 33-0). Our models use the wearable and survey data up to and including the current day for predicting nurses’ and doctors’ next day health and wellbeing labels.

We compared our job-role based multitask multilabel model (MTML-NN) with following approaches to evaluate the benefits of using demographic information and the correlation among labels: (1) random forest (RF), (2) RBF kernel based support vector machine (SVM), (3) multitask neural network (MT-NN) that used clusters of participants and achieved state-of-the-art performance in a previous study [41], (4) multitask neural network with labels as tasks (ML-NN). In addition to applying ML-NN to all participants (ML-NN (all)), we also calculated the prediction results for nurses (ML-NN(N)) and doctors (ML-NN(D)) separately.

For training and testing our models, we randomly split the dataset into training and testing data in a ratio of 80% to 20%. We applied 10-fold cross-validation and grid search to finalize the hyperparameters for all models mentioned above in the training set. Then, we tested models in the testing set. To make the evaluation process more robust, we repeated the random data split strategy (training/testing : 80%/20%) 10 times to evaluate the model performance. As the evaluating metrics, we use mean absolute errors for the regression models and f1-scores for classification tasks. Furthermore, we adopt focal loss[22] as the objective function in the classification tasks to mitigate the unbalanced sample size in both binary and 3-class tasks. The Adam optimizer [18] was used in training the neural networks, with a learning rate of 0.005 and 0.9, 0.999 for  $\beta_1$  and  $\beta_2$ .

#### 4.1 Model Weights Analysis

In addition to the prediction performance, interpretability is also an essential part of machine learning models. Ideally, we would like to provide our prediction results along with reasonable explanations to our participants or health/medical stakeholders. First, from the weights in the RF model, we analyzed the importance of input features. Then, in our deep learning MTML-NN model, we analyzed the importance of the features by examining the parameters in the first CNN layer before the non-linear activation function. Since the CNN kernel we designed is in one-dimension with a size of the number of features, and parameters in the CNN kernel would correspond to the input features. We calculated the average value of each feature on all channels to check the importance of the features. Also, we computed the correlations between the output of the CNN layer and the input features. Features that have higher correlations with the CNN outputs would also be considered important features.

## 5 Results and Discussion

### 5.1 Data Statistics

As shown in Table 1, the average score of alertness label was the lowest among all five labels; while the stress label (0: pressure-1: calm) showed the highest average score. Compared with other labels, the SD of happiness score was lower. Moreover, the distribution of health and wellbeing labels for nurses and doctors were different. For example, doctors generally had higher subjective alertness and energy than nurses in the morning. In addition, we computed correlations among the five health and wellbeing labels. Figure 2 shows the correlation coefficients matrix of all labels, and there are different degrees of correlation among the labels. The Pearson test [26] showed that all five labels were significantly correlated. The linear fitting coefficient of determination ( $r^2$ ) values [28] between the alert label and other labels ranged from 0.19 to 0.28, while the  $r^2$  values among the happy, energy, health and stress labels were all higher than 0.55, with the highest value being 0.70 (happy and stress).

We also compared feature distributions between nurses and doctors (Table 2). We found that the mean heart rate of doctors was significantly higher than that of nurses; whereas the variability of heart rate, defined as SD and sample entropy, was higher in nurses than doctors. In terms of sleep, we found that doctors showed higher sleep efficiency and lower sleep irregularity than nurses. Further, We found statistical differences between nurses and doctors in movement features, including mean and SD of daily steps across the previous 7 days, and the entropy for stationary/active segments. We did not observe any statistical differences between nurses and doctors in working shifts and total work hours among shift work features. However, we found that overwork was more common among doctors.

**Table 1.** Mean (SD) of daily wellbeing & P-values from Welch’s t-test

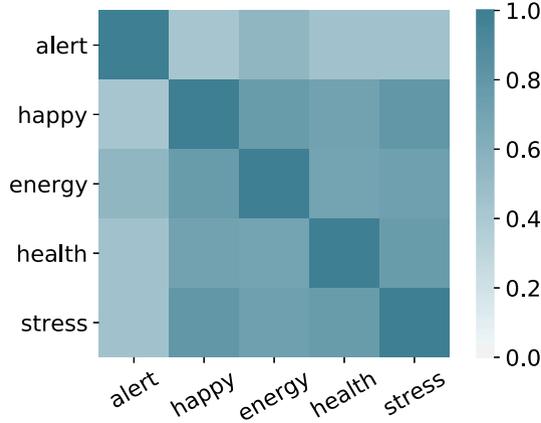
	Nurse	Doctor	p-value
Alertness	38.5 (22.9)	52.8 (23.5)	<0.05
Happiness	57.2 (20.9)	59.2 (17.1)	0.38
Energy	54.0 (22.9)	60.5 (22.4)	<0.05
Health	63.3 (22.1)	63.9 (22.4)	0.83
Stress	63.5 (23.4)	65.6 (17.5)	0.39

### 5.2 Wellbeing Prediction

The classification and regression performance using different models is shown in Table 3. Our proposed job-role based MTML-NN performed the best for four labels in binary classification and all wellbeing labels in 3-class classification and regression (ANOVA, Tukey,  $p < 0.05$ ). Our results showed the benefits of our proposed simultaneous job role and correlated label modeling, especially in

**Table 2.** List of main features and the statistics of nurses and doctors. The statistics of numeric features are shown in mean (SD) values, and the differences are tested with Mann-Whitney U-test (non-normally distributed features) and Welch’s t-test (normally distributed features, indicated with  $\star$ ); whereas we use percentages to reveal the statistics of categorical features and apply the chi-square test to check their statistical differences.

Source	Daily Features	Nurses (N = 10)	Doctors (N = 4)	P-value	
Fitbit	Heart Rate (HR) - Mean $\star$	78.5 (7.1)	70.6 (6.8)	<0.05	
	Heart Rate(HR) - SD	13.5 (3.5)	12.3 (3.2)	< 0.05	
	Heart Rate(HR) - Entropy $\star$	0.64 (0.23)	0.69 (0.20)	<0.05	
	Sleep Duration (mins)	374.3 (134.0)	363.1 (106.3)	0.36	
	Sleep Efficiency (0–100)	93.1 (4.9)	95.5 (2.9)	<0.05	
	Sleep Regularity $\star$	0.31 (0.25)	0.26 (0.19)	<0.05	
	Sleep Duration - Mean across previous 7 days	370.5 (74.4)	359.0 (46.1)	0.21	
	Sleep Efficiency - Mean across previous 7 days	94.8 (3.3)	93.2 (3.6)	<0.05	
	Sleep Duration - SD across previous 7 days	106.8 (39.1)	88.4 (37.8)	<0.05	
	Sleep Efficiency - SD across previous 7 days	2.19 (1.31)	2.20 (1.02)	0.21	
	Steps $\star$	8931.2 (4030.3)	8139.9 (3350.8)	0.21	
	Steps - Mean across previous 7 days $\star$	8684.8 (2372.8)	9063.4 (2236.9)	<0.05	
	Steps - SD across previous 7 days $\star$	3099.5 (1017.8)	2582.5 (849.7)	<0.05	
	Entropy (stationary segments)	2.17 (0.53)	2.61 (0.24)	<0.05	
	Entropy (active segments)	1.67 (0.45)	1.65 (0.19)	<0.05	
	Survey	Number of Naps	0.55 (0.68)	0.37 (0.66)	<0.05
		Duration of Naps (mins)	31.1 (68.4)	19.1 (49.4)	0.12
		# of Cups of Caffeinated Drinks	0.47 (0.73)	0.45 (0.62)	0.45
		Wake-up Type	-	-	<0.05
		- Natural	35.5%	35.6%	
- Alarm		60.9%	46.5%		
- Other than alarm		3.6%	17.8%		
Time to Fall Asleep (mins)		-	-	<0.05	
- 0-5		34.0%	40.6%		
- 6–15		31.4%	41.6%		
- 16-30		17.3%	11.9%		
- 31-45		6.6%	3.0%		
- 45-60	3.6%	3.0%			
- 60+	7.1%	0%			
Work Shifts	-	-	0.12		
- Shift 1	53.8%	64.3%			
- Shift 2	30.4%	19.8%			
- Shift 3	15.7%	15.8%			
Work Time (hours)	8.0 (0.0)	8.4 (1.7)	0.08		
Overwork Time (mins)	11.0 (41.8)	202.6 (320.1)	<0.05		



**Fig. 2.** Correlation coefficients matrix of wellbeing labels

3-class classification and regression. However, according to the performance in 3-class classification, we found poor classification performance for some classes. For example, in the 3-class alertness classification, the high-alertness class precision and recall values were only 0.16 and 0.27 in respectively; and our low-energy class prediction was also relatively low with a precision of 0.33 and a recall of 0.24. These errors might come from the data imbalance problem. In the 3-class classification tasks, the high alertness labels accounted for only 20% of all labels, and the low energy labels accounted for 15% of all labels.

Furthermore, we also found the benefits of using job role information or multiple labels separately. For example, in the alertness prediction, job-role based MT-NN showed significant improvement from NN for both binary and 3-class classification. Besides the overall f1-score, we observed some improvements revealed in each class. For example, in the 3-class alertness classification tasks, MT-NN model provided significantly higher recall and precision scores in low and middle alertness classification compared to NN model (Welch's t-test,  $p < 0.05$ ). We did not observe any significant improvement in the regression tasks. However, the average prediction MAE of MT-NN was lower than that of NN. Significant improvements were observed in ML-NN compared to NN in almost all tasks. For example, in the regression tasks, the ML-NN (all) performed statistically significantly better than NN in predicting alertness, happiness, energy, and stress labels.

### 5.3 Weight Analysis

From the RF model, for both the binary and 3-class classification happiness prediction tasks, we found features including mean heart rate and heart rate sample entropy across the day, sleep duration, sleep regularity, and the SD of sleep efficiency across the previous seven days, were the most important. In the alertness prediction tasks, work shifts, stationary segment entropy, mean step,

**Table 3.** Prediction performance (f1-score for classification; mean absolute error (MAE) for regression) of different algorithms. Bold entries represent statistically significantly better results over the other models.

Tasks	Algorithms	Alertness	Happiness	Energy	Health	Stress
Binary	RF	50% $\pm$ 7%	78% $\pm$ 4%	65% $\pm$ 4%	<b>84% <math>\pm</math> 3%</b>	82% $\pm$ 3%
	SVM	52% $\pm$ 4%	69% $\pm$ 4%	62% $\pm$ 6%	80% $\pm$ 4%	77% $\pm$ 5%
	NN	55% $\pm$ 4%	71% $\pm$ 5%	65% $\pm$ 3%	82% $\pm$ 3%	80% $\pm$ 3%
	MT-NN	60% $\pm$ 4%	76% $\pm$ 3%	69% $\pm$ 5%	<b>83% <math>\pm</math> 4%</b>	<b>83% <math>\pm</math> 4%</b>
	ML-NN (all)	55% $\pm$ 7%	74% $\pm$ 7%	68% $\pm$ 4%	80% $\pm$ 4%	<b>83% <math>\pm</math> 3%</b>
	ML-NN (N)	55% $\pm$ 9%	69% $\pm$ 5%	64% $\pm$ 6%	79% $\pm$ 7%	79% $\pm$ 7%
	ML-NN (D)	59% $\pm$ 8%	75% $\pm$ 5%	67% $\pm$ 7%	<b>85% <math>\pm</math> 5%</b>	<b>85% <math>\pm</math> 7%</b>
	MTML-NN	<b>64% <math>\pm</math> 7%</b>	<b>79% <math>\pm</math> 3%</b>	<b>71% <math>\pm</math> 4%</b>	81% $\pm$ 3%	<b>84% <math>\pm</math> 3%</b>
3-class	RF	53% $\pm$ 5%	39% $\pm$ 5%	46% $\pm$ 6%	49% $\pm$ 7%	44% $\pm$ 5%
	SVM	47% $\pm$ 7%	40% $\pm$ 5%	43% $\pm$ 5%	49% $\pm$ 7%	46% $\pm$ 4%
	NN	51% $\pm$ 5%	45% $\pm$ 6%	45% $\pm$ 5%	53% $\pm$ 6%	51% $\pm$ 4%
	MT-NN	57% $\pm$ 6%	46% $\pm$ 5%	48% $\pm$ 6%	53% $\pm$ 5%	50% $\pm$ 5%
	ML-NN (all)	52% $\pm$ 8%	<b>53% <math>\pm</math> 7%</b>	48% $\pm$ 7%	55% $\pm$ 3%	54% $\pm$ 4%
	ML-NN (N)	45% $\pm$ 7%	<b>54% <math>\pm</math> 5%</b>	49% $\pm$ 5%	56% $\pm$ 2%	53% $\pm$ 5%
	ML-NN (D)	54% $\pm$ 5%	<b>51% <math>\pm</math> 7%</b>	45% $\pm$ 6%	54% $\pm$ 3%	52% $\pm$ 7%
	MTML-NN	<b>59% <math>\pm</math> 5%</b>	<b>52% <math>\pm</math> 4%</b>	<b>51% <math>\pm</math> 4%</b>	<b>58% <math>\pm</math> 7%</b>	<b>57% <math>\pm</math> 5%</b>
Regression	SVR	20.6 $\pm$ 2.9	19.7 $\pm$ 2.5	21.3 $\pm$ 2.4	18.9 $\pm$ 1.8	21.7 $\pm$ 2.1
	NN	19.9 $\pm$ 1.8	19.0 $\pm$ 1.9	20.3 $\pm$ 2.2	19.5 $\pm$ 1.9	20.4 $\pm$ 1.9
	MT-NN	19.4 $\pm$ 2.1	18.8 $\pm$ 2.3	20.7 $\pm$ 1.6	19.3 $\pm$ 2.0	20.3 $\pm$ 1.7
	ML-NN (all)	18.6 $\pm$ 1.3	16.9 $\pm$ 3.1	18.6 $\pm$ 2.0	18.7 $\pm$ 2.6	19.4 $\pm$ 2.6
	ML-NN (N)	18.0 $\pm$ 1.1	15.9 $\pm$ 1.9	17.3 $\pm$ 1.7	15.6 $\pm$ 1.7	17.7 $\pm$ 1.3
	ML-NN (D)	20.4 $\pm$ 2.1	16.0 $\pm$ 2.0	19.3 $\pm$ 2.0	17.4 $\pm$ 2.0	19.4 $\pm$ 3.1
	MTML-NN	<b>17.4 <math>\pm</math> 1.4</b>	<b>15.1 <math>\pm</math> 1.6</b>	<b>17.7 <math>\pm</math> 1.2</b>	<b>15.4 <math>\pm</math> 1.5</b>	<b>15.6 <math>\pm</math> 1.9</b>

and mean sleep duration across the previous 7, 5 days played important roles. The analysis of the parameters in the CNN layer in the MTML-NN model and the correlations between the CNN output and input features indicated that features including heart rate sample entropy, sleep regularity, sleep efficiency, work shifts, steps, and active segment entropy - contributed to health and wellbeing prediction. For example, from the correlation analysis, we found that the sleep efficiency, sleep regularity, and daytime work shift were positively related to the wellbeing (Pearson test, p-value  $<$  0.05/(# of features)); whereas the step and the entropy of active segments were negatively related to the wellbeing (Pearson test, p-value  $<$  0.05/(# of features)). Our findings were consistent with some prior results. For example, according to the previous works, sleep influences physical and psychological health [47], and stress [24]; sleep regularity is associated with mood [37]. Previous studies also indicated the association between work shifts and stress levels [46].

## 6 Conclusion

In this work, we collected physiological and behavioral wearable sensor data as well as survey data from shift-work nurses and doctors, and compared their physiology and behaviors between two job roles. Then, we proposed a job-role

based multitask and multilabel learning model structure to predict shift workers' health and wellbeing for next day using sensor and questionnaire data. The proposed model outperformed the benchmark models, including RF and SVM as well as the previous state-of-the-art models. The analysis of model weights showed that health rate, work shifts, sleep parameters such as sleep regularity and sleep efficiency contributed to shift workers' health and wellbeing labels. As future work, we will collect more data from shift workers and design a system to improve shift workers' health and wellbeing.

## References

1. Asselbergs, J., Ruwaard, J., Ejdys, M., Schrader, N., Sijbrandij, M., Riper, H.: Mobile phone-based unobtrusive ecological momentary assessment of day-to-day mood: an explorative study. *J. Med. Internet Res.* **18**(3), e72 (2016)
2. Bogomolov, A., Lepri, B., Ferron, M., Pianesi, F., et al.: Daily stress recognition from mobile phone data, weather conditions and individual traits. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 477–486 (2014)
3. Books, C., Coody, L.C., Kauffman, R., Abraham, S.: Night shift work and its health effects on nurses. *Health Care Manager* **36**(4), 347–353 (2017)
4. Bourdouxhe, M., Quéinnec, Y., Guertin, S.: The interaction between work schedule and workload: case study of 12-hour shifts in a Canadian refinery. In: *Shiftwork International Newsletter*, p. 19 (2000)
5. Courtney, J.A., Francis, A.J., Paxton, S.J.: Caring for the carers: fatigue, sleep, and mental health in Australian paramedic shiftworkers. *Aust. J. Organisational Psychol.* **3**, 32–41 (2010)
6. Feng, T., Booth, B.M., Narayanan, S.S.: Modeling behavior as mutual dependency between physiological signals and indoor location in large-scale wearable sensor study. In: *ICASSP 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1016–1020. IEEE (2020)
7. Feng, T., Narayanan, S.S.: Modeling behavioral consistency in large-scale wearable recordings of human bio-behavioral signals. In: *ICASSP 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1011–1015. IEEE (2020)
8. Fischer, D., et al.: Irregular sleep and event schedules are associated with poorer self-reported well-being in us college students. *Sleep* **43**(6), zsz300 (2020)
9. Ganesan, S., et al.: The impact of shift work on sleep, alertness and performance in healthcare workers. *Sci. Rep.* **9**(1), 1–13 (2019)
10. Garbarino, S., et al.: Sleepiness and sleep disorders in shift workers: a study on a group of Italian police officers. *Sleep* **25**(6), 642–647 (2002)
11. Geiger-Brown, J., Rogers, V.E., Trinkoff, A.M., Kane, R.L., Bausell, R.B., Scharf, S.M.: Sleep, sleepiness, fatigue, and performance of 12-hour-shift nurses. *Chronobiol. Int.* **29**(2), 211–219 (2012)
12. Golden, L., Wiens-Tuers, B.: Overtime work and wellbeing at home. *Rev. Soc. Econ.* **66**(1), 25–49 (2008)
13. Han, S., et al.: Estimating the attributable cost of physician burnout in the United States. *Ann. Intern. Med.* **170**(11), 784–790 (2019)
14. Heath, G., Dorrian, J., Coates, A.: Associations between shift type, sleep, mood, and diet in a group of shift working nurses. *Scand. J. Work Environ. Health* **45**(4), 402–412 (2019)

15. Jamal, M.: Burnout, stress and health of employees on non-standard work schedules: a study of Canadian workers. *Stress Health: J. Int. Soc. Invest. Stress* **20**(3), 113–119 (2004)
16. Jaques, N., Taylor, S., Sano, A., Picard, R., et al.: Predicting tomorrow's mood, health, and stress level using personalized multitask learning and domain adaptation. In: *IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing*, pp. 17–33 (2017)
17. Kato, C., Shimada, J., Hayashi, K.: Sleepiness during shift work in Japanese nurses: a comparison study using JESS, SSS, and actigraphy. *Sleep Biol. Rhythms* **10**(2), 109–117 (2012)
18. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
19. Knutsson, A.: Health disorders of shift workers. *Occup. Med.* **53**(2), 103–108 (2003)
20. Li, H., et al.: Napping on night-shifts among nursing staff: a mixed-methods systematic review. *J. Adv. Nurs.* **75**(2), 291–312 (2019)
21. LiKamWa, R., Liu, Y., Lane, N.D., Zhong, L.: MoodScope: building a mood sensor from smartphone usage patterns. In: *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services*, pp. 389–402 (2013)
22. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988 (2017)
23. Mann, H.B., Whitney, D.R.: On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* 50–60 (1947)
24. Mezick, E.J., et al.: Intra-individual variability in sleep duration and fragmentation: associations with stress. *Psychoneuroendocrinology* **34**(9), 1346–1354 (2009)
25. Muaremi, A., Arnrich, B., Tröster, G.: Towards measuring stress with smartphones and wearable devices during workday and sleep. *BioNanoScience* **3**(2), 172–183 (2013). <https://doi.org/10.1007/s12668-013-0089-2>
26. Mukaka, M.M.: A guide to appropriate use of correlation coefficient in medical research. *Malawi Med. J.* **24**(3), 69–71 (2012)
27. Mulhall, M.D., et al.: A pre-drive ocular assessment predicts alertness and driving impairment: a naturalistic driving study in shift workers. *Accid. Anal. Prev.* **135**, 105386 (2020)
28. Nagelkerke, N.J., et al.: A note on a general definition of the coefficient of determination. *Biometrika* **78**(3), 691–692 (1991)
29. Pasman, W.J., Boessen, R., Donner, Y., Clabbers, N., Boersma, A.: Effect of caffeine on attention and alertness measured in a home-setting, using web-based cognition tests. *JMIR Res. Protoc.* **6**(9), e169 (2017)
30. Pearson, K.: X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **50**(302), 157–175 (1900)
31. Pereira, A., Nunes, F.: Physical activity intensity monitoring of hospital workers using a wearable sensor. In: *12th EAI International Conference on Pervasive Computing Technologies for Healthcare-Demos, Posters, Doctoral Colloquium. European Alliance for Innovation (EAI)* (2018)
32. Phillips, A.J., et al.: Irregular sleep/wake patterns are associated with poorer academic performance and delayed circadian and sleep/wake timing. *Sci. Rep.* **7**(1), 1–13 (2017)

33. Purnell, M., Feyer, A.M., Herbison, G.: The impact of a nap opportunity during the night shift on the performance and alertness of 12-h shift workers. *J. Sleep Res.* **11**(3), 219–227 (2002)
34. Richman, J.S., Moorman, J.R.: Physiological time-series analysis using approximate entropy and sample entropy. *Am. J. Physiol. Heart Circulatory Physiol.* **278**(6), H2039–H2049 (2000)
35. Roehrs, T., Roth, T.: Sleep, sleepiness, and alcohol use. *Alcohol Res. Health: J. Nat. Inst. Alcohol Abuse Alcoholism* **25**(2), 101–109 (2001)
36. Rowley, H.A., Baluja, S., Kanade, T.: Neural network-based face detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(1), 23–38 (1998)
37. Sano, A.: Measuring college students' sleep, stress. *Mental Health and Wellbeing with Wearable Sensors and Mobile Phones*. Ph.D. thesis, MIT (2015)
38. Schernhammer, E.S., et al.: Night-shift work and risk of colorectal cancer in the nurses' health study. *J. Nat. Cancer Inst.* **95**(11), 825–828 (2003)
39. Shapiro, D., Jamner, L.D., Goldstein, I.B., Delfino, R.J.: Striking a chord: moods, blood pressure, and heart rate in everyday life. *Psychophysiology* **38**(2), 197–204 (2001)
40. Srivastava, U.R.: Shift work related to stress, health and mood states: a study of dairy workers. *J. Health Manag.* **12**(2), 173–200 (2010)
41. Taylor, S.A., Jaques, N., Nosakhare, E., Sano, A., Picard, R.: Personalized multitask learning for predicting tomorrow's mood, stress, and health. *IEEE Trans. Affective Comput.* **11**, 200 (2017)
42. Umematsu, T., Sano, A., Taylor, S., Picard, R.W.: Improving students' daily life stress forecasting using LSTM neural networks. In: 2019 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), pp. 1–4. IEEE (2019)
43. Vogel, M., Braungardt, T., Meyer, W., Schneider, W.: The effects of shift work on physical and mental health. *J. Neural Transm.* **119**(10), 1121–1132 (2012)
44. Vrijkotte, T.G., Van Doornen, L.J., De Geus, E.J.: Effects of work stress on ambulatory blood pressure, heart rate, and heart rate variability. *Hypertension* **35**(4), 880–886 (2000)
45. Welch, B.L.: The generalization of student's' problem when several different population variances are involved. *Biometrika* **34**(1/2), 28–35 (1947)
46. Wisetborisut, A., Angkurawaranon, C., Jiraporncharoen, W., Uaphanthasath, R., Wiwatanadate, P.: Shift work and burnout among health care workers. *Occup. Med.* **64**(4), 279–286 (2014)
47. Wong, M.L., Lau, E.Y.Y., Wan, J.H.Y., Cheung, S.F., Hui, C.H., Mok, D.S.Y.: The interplay between sleep and mood in predicting academic functioning, physical health and psychological health: a longitudinal study. *J. Psychosomatic Res.* **74**(4), 271–277 (2013)
48. Wright Jr., K.P., Bogan, R.K., Wyatt, J.K.: Shift work and the assessment and management of shift work disorder (SWD). *Sleep Med. Rev.* **17**(1), 41–54 (2013)
49. Yu, H., Klerman, E.B., Picard, R.W., Sano, A.: Personalized wellbeing prediction using behavioral, physiological and weather data. In: 2019 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), pp. 1–4. IEEE (2019)
50. Yu, H., Sano, A.: Passive sensor data based future mood, health, and stress prediction: User adaptation using deep learning. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 5884–5887. IEEE (2020)