Time-Series Forecasting of Physical and Mental Fatigue Based on Wearable Data

JUN NI (JENNY) DU, University of Toronto, Canada

1 ABSTRACT

This paper addresses the prediction of physical and mental fatigue levels through wearable data using a Long Short-Term Memory with Attention (LSTM-AT) algorithm. Building upon a previous study (Luo et al. [2020]), we aim to enhance predictive models by considering historical data over multiple days. Key research questions include exploring the impact of LSTM-AT models on predictive performance, examining how different fatigue label formats influence model performance, and assessing the consequences of patient stratification on model outcomes. The evaluation involves k-fold cross-validation, revealing promising preliminary results. The LSTM-AT model demonstrates competitive or superior performance in predicting binary fatigue labels (physical fatigue: 73.56% ± 0.02; mental fatigue: 64.01% ± 0.08). Ordinal predictions show correlation with the targets, indicating a degree of alignment. However, it's worth noting that the current LSTM-AT model is not sufficiently accurate in predicting numerical fatigue scores. Additionally, subject-based cross-validation (patient stratification) shows a significant drop in performance, especially for mental fatigue predictions. Our findings suggest the potential of time-series forecasting models in improving fatigue prediction using wearable data. However, future work is needed to further refine the model for predicting fatigue scores on novel data.

2 INTRODUCTION

Fatigue is a broad, multifactorial concept that is generally defined by the feelings of reduced physical and mental energy levels. It can strongly impact a patient's health-related quality of life (HRQOL), and has been a primary focus in various treatments [1]. The cause of fatigue can be categorized into pathological and non-pathological types. Pathological fatigue can be attributed to medical conditions like anemia, heart disease, cancer, and other chronic illnesses [6], while non-pathological fatigue can occur in healthy individuals under a variety of circumstances and environmental factors, making it harder to track. Fatigue can also be categorized as physical or mental. Physical fatigue typically occurs in response to extensive muscle activity, such as exercise, and often recovers quickly with rest. In contrast, mental fatigue tends to accumulate over time due to prolonged exposure to stress, anxiety, or excessive stimulation [10].

Existing approaches for studying fatigue often rely solely on patient-reported outcomes and single performance tests performed in a laboratory setting, such as tracking eye movements in a car-driving simulation to assess driver fatigue[14] or utilizing clinical technologies like mechanomyography, electromyography, and ultrasounds to detect muscle fatigue at rest and after specific exercise activities [2]. These approaches exhibit limitations in capturing the complete spectrum of factors that contribute to fatigue, including physiological, psychological, and contextual elements. Furthermore, they face challenges such as data sparsity (caused by the limitation of collecting data only in laboratory settings), limited reproducibility, and sub-optimal compliance (in attending lab sessions). An emerging approach to improving fatigue assessment is the use of wearable sensors to collect continuous and objective data over extended periods in real-life environments. These sensor data, complemented by patient-reported outcomes, are driving a new trend in mobile health research [8].

In a pilot study conducted by Luo et al. [2020], the authors explored the relationship between self-reported non-pathological physical and mental fatigue and multisensor wearable data in healthy subject as they went about their daily routines. They collected sensor data from 28 healthy adult participants, using an armband-sized wearable device that was continuously worn on each participant's non-dominant arm for one or more week-long periods, along with a daily questionnaire on physical and mental fatigue levels. The authors found that, for both physical and mental fatigue, combining multiple sensor parameters resulted in a stronger correlation with the fatigue labels compared to using individual sensor data. Moreover, clustering analysis has revealed distinct top predictive features for physical fatigue in comparison to mental fatigue. Physical fatigue is primarily linked to physical activity, encompassing variables related to energy expenditure. Conversely, mental fatigue is predominantly associated with vital signs, including respiration rate and heart rate variability. This observation suggests disparities in the underlying mechanisms of these two fatigue types.

Luo et al.'s study [2020] was among the first to analyze multimodal data sources related to physical activity, vital signs, and other physiological parameters in the context of their relationship to self-reported non-pathological physical and mental fatigue in real-world settings. This study offers a glimpse into the potential of a machine learning-driven framework connecting multisensor wearable data with patient-reported outcomes, potentially enhancing the understanding of HRQOL factors in clinical trials and daily medical practice. However, the flip-side of its novelty lies in the limitations of the model's performance, which also serves as the underlying motivation for this project. The best models in the pilot study achieved only 70% weighted accuracy for physical fatigue and 65% weighted accuracy for mental fatigue. The authors also simplified each of the fatigue measures into binary values (fatigue/no fatigue), despite the original questionnaire response containing fatigue scores across five levels (never; sometimes; regularly; often; always) and a numerical score ranging from 1 to 10. Furthermore, each participant's daily sensor data and questionnaire response are treated as independent samples, disregarding the time-series nature of the data which could reveal any potential long-term patterns. These long-term patterns could prove especially critical in the prediction of mental fatigue, since accumulated stress or over-stimulation may be reflected in the chronic changes that are not observable within the data from just one day. Therefore, we propose a project to enhance the models presented by Luo et al. [2020]. Specifically, we aim to develop a time-series forecasting algorithm to predict physical and mental fatigue levels using continuous data from wearable sensors collected over multiple days. In addition, we wish to explore alternative approaches for improving the model's accuracy, while eliminating some of the simplifications that were applied to the original models.

2.1 Research Question

Building upon the pilot study, the core research questions of this project are as follows:

(1) Can the predictive performance of fatigue prediction models be improved by employing a Long Short-Term Memory (LSTM) model that incorporates historical data from each participant, such as all sensor data gathered over the preceding five days?
(2) How does the format of the fatigue labels—specifically, whether to binarize them or maintain the original ordinal format—affect the performance of the model?

3 RELATED WORK

3.1 Multi-Modal Fatigue Prediction Using Wearable Data

The issue of solely relying on self-reporting for fatigue assessment has been an ongoing concern, and numerous novel approaches have been explored to measure fatigue in an objective manner. Given the prevalence of digital health technologies and wearable devices, continuous measurements of physiological data, such as heart rate, respiration rate, sleep, and physical activities in an individual's daily life, can be collected. On top of Luo et al.'s study, in another study done by Rao et al. [2023], the authors focused on using digital health technologies, specifically wearable devices (FitBit), to study the relationship between physical activity and self-reported fatigue in individuals with chronic inflammatory rheumatic diseases compared to healthy individuals. The study collected observational data from 296 participants in the United States, including continuous multimodal digital data from Fitbit devices (heart rate, physical activity, sleep features) and app-based daily and weekly questions on various health-related quality of life factors, including fatigue. The Fitbit data was aggregated and averaged on a daily basis. Common machine learning methods, including XGBoost, logistic regression, support vector machine, and random forest (RF), were employed to predict fatigue levels. The findings of the study include the identification of different digital phenotypes based on wearable parameters, as well as the identification of a set of key predictive features to quantify the impact of physical activity on fatigue. [12] Bai et al. [2021] also conducted a similar study, utilizing mutli-modal physiological data (ECG, accelerometer, skin temperature, respiratory rate) to develop fatigue assessment models utilizing random forest-based mixed-effects models. [12] These studies are highly similar to Luo et al. [2020], signifying the potential of using multi-modal data collected using wearables for fatigue prediction.

3.2 Fatigue Prediction From Time-Series Data

In another study done by Bai et al. [2020], the authors introduced an approach for predicting fatigue by employing recurrent neural networks (RNNs) on time-series sensor data. They collected continuous ECG and Actigraphy data from 9 participants using two medical-grade devices (Actigraph GT9X Link and Vital Patch), along with questionnaire response where the users rate their general fatigue levels, from 1 to 10, four times a day over 7 days. The authors proposed two types of models for fatigue prediction: interpretable models and deep learning models. The interpretable models used linear regression based on high-level features and a feature selection method, whereas the deep learning models utilized LSTM networks with different attention mechanisms, including self-attention (SA) and consistency self-attention (CSA). The models underwent evaluation through a 5-fold cross-validation process, employing metrics such as mean absolute error (MAE) and root mean square error. The findings indicated that the LSTM-CSA model demonstrated the best performance when integrating ECG and Actigraphy data. [4] This study demonstrates the potential of time series data over daily aggregations in capturing the dynamic patterns and nuances of fatigue prediction. Nevertheless, the authors acknowledged specific limitations in their study, particularly the relatively small dataset and the noted sensitivity to specific subjects in both the training and testing sets. As data segments increase in size, such as using a 5-day window instead of a 1-day window, it is inevitable that the sample size will decrease. Therefore, it is crucial to consider the trade-off between data segment size, sample size reduction, and the generalizability of the findings.

4 SYSTEM DESIGN

4.1 Data Pre-processing

The dataset for this project has been published by the authors [9] and is accessible at the following link: https://zenodo.org/record/4266157. It includes data from 28 healthy adult participants aged 26 to 55 years (average age: 42 years; 11/16/1 female/male/unknown gender), with a total of 973 recording days. A detailed breakdown of the participant demographic, including age, gender, and mean fatigue scores can be found in Supplementary Table 1 from Luo et al.'s study. [9] The recorded data consists of continuous sensor data on physical activity, vital signs, and other physiological parameters at 1-minute intervals, as well as responses to the daily questionnaires on fatigue. The sensor data are collected using an armband-sized clinical-grade multisensor wearable, Everion (Biovotion AG, Switzerland), worn continuously on each participant's non-dominant arm. The Everion device combines a 3-axis accelerometer, barometer, galvanic skin response electrode, and temperature and photo sensors. The daily fatigue questionnaire is delivered using a mobile app, SymTrack (Gastric GmbH, Switzerland). For this study, two parameters (Activity Class and Galvanic Skin Response) and one questionnaire question ("Did you do sport today?") were excluded from the dataset due to insufficient coverage. These parameters did not emerge as top predictive features for either physical or mental fatigue [9]. The fatigue labels were transformed from word responses to ordinal encodings. A comprehensive list of sensor parameters and questionnaire questions utilized in this study is detailed in **Table 1**.

Sensor	Parameter Name	Description
Accelerometer	ActivityCounts	Intensity of motion (counts per minute)
	Steps	Number of steps
	EnergyExpenditure	Amount of energy a person uses to complete all regular bodily functions (calories/min)
Photoplethysmography	HR	Heart rate (beats per minute)
	HRV	Heart rate variability; indicates the beat to beat variations (milliseconds)
	RESP	Respiration rate (breaths per minute)
	BloodPerfusion	Percentage change in blood volume in local tissue resulting from a heartbeat
	BloodPulseWave	Amplitude of pulse wave generated during the contraction of the heart
Temperature	SkinTemperature	Skin temperature ($^{\circ}C$)
Barometer	Barometer	Atmospheric pressure and altitude (mbar)

Measure	Questionnaire Question	Possible Answers
Physical fatigue score (PhF)	"Physically, today how often did you feel exhausted?"	0 (never); 1 (sometimes); 2 (regularly); 3 (often); 4 (always)
Mental fatigue score (MF)	"Mentally, today how often did you feel exhausted?"	(same as above)
Visual analogue scale score (VAS)	"Describe fatigue on a scale of 1–10, where 1 means you don't feel tired at all and 10 means the worst tiredness you can imagine."	a number from 1 to 10
Indicator of relative perception (RelP)	"Are you feeling better, worse, or the same as yesterday?"	1 (better); -1 (worse); 0 (same)

Table 1. List of parameters measured by the wearable device (Everion) and the daily questionnaire (SymTrack).

For each set of fatigue labels, the wearable data from the preceding five days is aggregated up to the timestamp of the questionnaire response. This choice of a five-day segment aligns with the consideration that several subjects have limited sensor data, ranging from 5 to 7 days. To address issues related to missing data, stemming from subjects not wearing the device or loss of skin contact, segments with over 20% missing data are systematically excluded to ensure optimal performance in downstream analyses. The choice of the 20% threshold was determined experimentally, aiming to establish an optimal balance between ensuring an adequate data set size and minimizing the presence of missing data (missing data threshold:dataset size: 10%:164. 20%:261, 30%:286). Following this data filtering step, the dataset is refined to contain a total of 25 subjects and 261 segments of data. Notably, Subject 6, 22, and 28 are excluded from the analysis as they were not able to yield any data after filtering due to insufficient information. However, these subjects may still serve as test samples later on to evaluate the model's performance in situations with insufficient data. The sensor data is aggregated to the means of hour intervals to enhance computational efficiency. With 10 wearable sensor features and 120 time steps, each resulting sample is structured as (120, 10). In **Figure 1**, the distribution of each fatigue label after filtering is illustrated. It is worth noting that the fatigue labels, specifically PhF, MF, and VAS, are unevenly distributed, with the high fatigue labels having the lowest prevalence. For the scope of this study, we will retain the data in its current state. However, additional approaches, such as over-sampling or under-sampling, can be considered to potentially address the issue of class imbalance. In addition to the ordinal format labels, a binarized rendition of each fatigue label is imputed for the comparative analysis of model performance between the original and binarized formats. The conversion scale is adopted from Luo et al.'s study, which is also the source of the dataset:

$$PhF_{binary}, MF_{binary} = \begin{cases} 0 & \text{for } PhF, MF = 0 \\ 1 & \text{for } PhF, MF \in \{1; \ldots; 3\} \end{cases} \qquad VAS_{binary} = \begin{cases} 0 & \text{for } VAS \in \{1; \ldots; 4\} \\ 1 & \text{for } VAS \in \{5; \ldots; 10\} \end{cases}$$

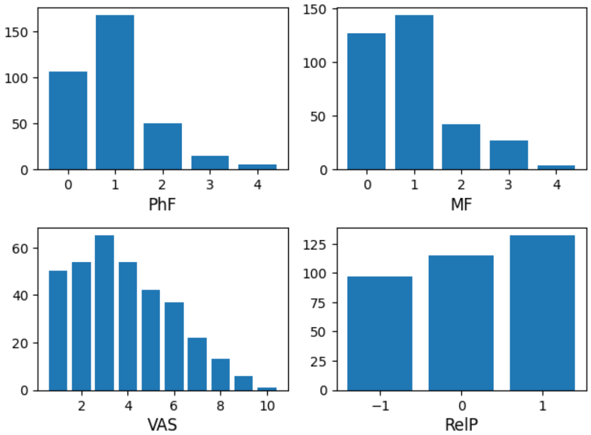

Fig. 1. Distribution of each fatigue label after filtering.

For the remaining missing data in the filtered datasets, they are left as-is. In Luo et al.'s study, the authors employed BRITS (Bidirectional Recurrent Imputation for Time Series) [5] to impute missing data.[9] The primary rationale for imputing missing data stems from the requirement of typical machine learning models, such as CNN or logistic regression models, for input data to maintain a consistent shape without missing values. In these models, the presence of missing values can disrupt the training process and overall functionality. Replacing missing values with zeros is not a viable option since valid data points, such as step counts, can legitimately be zero. On the contrary, LSTM and other RNN models processes input sequences in a step-by-step manner, enabling them to handle input data of varying lengths. LSTMs can also handle gaps in the input sequence, to an extent, by processing the available parts of the sequence. Therefore, the chosen approach here is to leave the missing data as they are.

4.2 Model Design

To predict the fatigue response, we proposed a Long Short-Term Memory with Attention (LSTM-AT) model. The architecture of the LSTM-AT model comprises an LSTM encoder[7] that processes input sequences to effectively capture contextual information. Subsequently, an attention layer[13] is employed to assign weights to various segments of the input sequence, enhancing the model's ability to focus on pertinent features during the decoding phase. The LSTM decoder leverages these weighted context representations to generate an output sequence. The final prediction is obtained by passing the output sequence through a fully connected layer with an output dimension of 1, representing the fatigue label prediction. For ordinal fatigue labels, the predictions are used as-is, whereas for the binarized fatigue labels, the output is fed through a sigmoid activation function and binarized based on a threshold of 0.5. This integrated approach harnesses the strengths of both LSTM and attention mechanisms to improve the accuracy and interpretability of fatigue predictions and has proven to have better performance than LSTM alone. [4]

In this study, we used the predefined ‘`LSTM()`’ function available in the PyTorch library [11]. To measure model performance on ordinal labels, we utilized the mean squared error (MSE) via the `MSELoss()` function available in the PyTorch library [11], which quantifies the squared difference between predicted and actual responses. For binarized labels, the loss is calculated using the `BCEWithLogitsLoss()` function, also from the PyTorch library [11]. This is equivalent to applying a Sigmoid layer to the output of the fully connected layer, which converts the output to a number between 0 and 1. Then, the binary cross-entropy (BCE) loss between the prediction and the target is calculated. **Figure** 2 contains a visual representation of the model architecture. This evaluation framework accommodates both ordinal and binarized label scenarios, providing a versatile and robust solution for assessing the model's performance on tasks with varying degrees of label granularity. We trained models individually for each of the four fatigue labels, as well as separately for ordinal and binary labels. Each model underwent 30 epochs, featuring a hidden size of 64. We utilized the Adam optimizer (`optim.Adam` [11]) with a default learning rate set at 0.001.

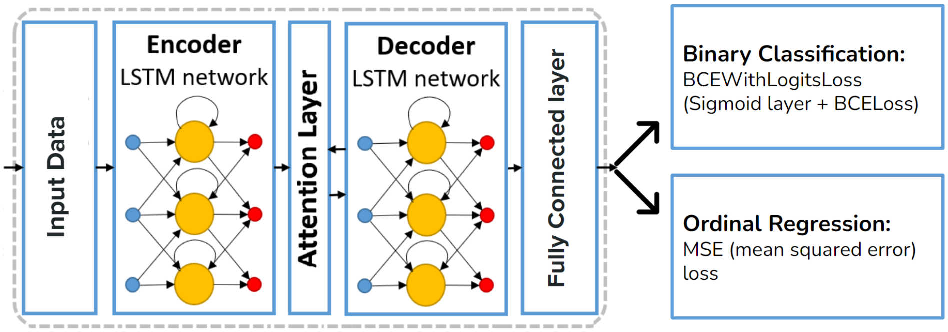

Fig. 2. Visualization of the LSTM-AT model. Figure modified from Aqueel et al.[3]

5 EXPERIMENT DESIGN

5.1 K-Fold Cross-Validation

To enhance the robustness of our results, we incorporated k-fold cross-validation (CV) with k=5 to alleviate the influence of dataset variability and produce more dependable performance metrics. This method involves dividing the dataset into five distinct subsets. The evaluation process is iteratively performed five times, with each iteration employing a different subset as the validation set and the remaining data for training. The resulting performance metrics are subsequently averaged across these iterations to provide a more comprehensive understanding of the variation in the estimates of the model's predictive capabilities.

We employed two types of k-fold CV: (1) data-based and (2) subject-based. In data-based CV, the 261 segments were divided into 5 subsets—a common approach to cross validation but susceptible to potential issues related to data leakage. Data leakage occurs when information about the test data unintentionally exists in the training dataset, leading to instances where multiple data segments from the same participant end up in both the training and testing datasets, resulting in model over-fitting. Specifically, in scenarios like wearable sensor data, individuals exhibit diverse baselines and trends in their physiological signals, which means that data leakage can lead to the model learning individual-specific characteristics rather than capturing generalizable patterns. To address this concern, we also conducted subject-based CV, also known as patient stratification, dividing the 25 subjects into 5 subsets. Given that not all participants contributed the same amount of data in this dataset, we implemented an iterative approach where each subject was randomly assigned to a subset, and the process was repeated until each partition contained approximately 20% of the data, ensuring that every subject belonged strictly to one subset. Despite our awareness of potential issues related to data leakage, we still chose to include data-based cross-validation (CV) for the evaluation of our LSTM-AT model. The main motivation behind this decision is to enable a fair comparison of LSTM-AT performance against traditional models developed by Luo et al.[9] The traditional models were also assessed using the data-based k-fold CV approach, and we aim to maintain consistency in the evaluation methodology to facilitate a more direct and meaningful comparison.

5.2 Model Performance Evaluation

We trained separate models to predict fatigue labels in both the binarized version (PhF_{binary}, MF_{binary}, VAS_{binary}) and their original (ordinal) format (PhF, MF, VAS). For the binarized fatigue labels, we computed performance metrics such as accuracy, precision, recall, and F1 scores for each model. The latter three metrics were weighted based on class size to address the class imbalance issue. These scores were then compared to those achieved by the best model (cCNN + RF) developed by Luo et al. [9]. This comparative analysis aims to assess the impact of incorporating time-series data, as opposed to aggregating to daily averages, on fatigue prediction. Additionally, for the LSTM-AT models, we evaluated the difference between data-based and subject-based k-fold CV to observe the effect of patient stratification on model performance. Regarding ordinal fatigue labels, we assessed model performance using mean squared error and compared differences in performance between data-based and subject-based cross-validation. Furthermore, we plotted the predictions against their corresponding ground truths, calculating the Pearson correlation coefficient and p-value for each fatigue label.

6 EVALUATION

6.1 Binary Fatigue Label Prediction Performances

Table 2 presents the performance metrics for predicting binary fatigue labels (PhF_{binary}, MF_{binary}, VAS_{binary}) using both cCNN+RF and LSTM-AT models (cCNN+RF results are replicated from Luo et al. [9]). The reported results correspond to both data-based and subject-based k-fold CV approaches.

When comparing the performance of LSTM-AT models to cCNN+RF models, both using data-based k-fold cross-validation, LSTM-AT models outperform cCNN+RF models for physical fatigue (**73.56% ± 0.02** vs. 71.85% ± 1.44), exhibiting higher precision, recall, and F1 score. In the case of mental fatigue, LSTM-AT performs slightly less favorably than cCNN+RF models in terms of mean accuracy but shows smaller variation (**66.20% ± 4.58** vs. 64.01% ± 0.08). The weighted precision, recall, and F1 score between the two models are comparable. Unfortunately, no performance metrics were available for the cCNN+RF model in the case of the visual analogue scale score, but the LSTM-AT model achieved an accuracy of 66.06%.

When comparing the performance of LSTM-AT models between data-based and subject-based k-fold cross-validation, all models for physical and mental health exhibit significant performance drops after applying subject-based k-fold CV. The drop is particularly pronounced for mental health, with accuracy decreasing from 64.01% ± 0.08 to 47.83% ± 0.05. Conversely, subject-based k-fold CV surprisingly resulted in a slightly higher accuracy for the visual analogue scale score, although the increase is not very substantial.

In summary, the LSTM-AT model generally demonstrated competitive or superior performance compared to the cCNN+RF model in predicting binary fatigue labels. However, this performance dropped significantly after applying patient stratification, with a notable impact on mental health predictions.

6.2 Ordinal Fatigue Label Prediction Performances

Table 3 presents the mean squared error results for ordinal fatigue label predictions using the LSTM-AT model, highlighting distinct performances across four fatigue labels. In the context of data-based cross-validation, physical fatigue (PhF) demonstrates the lowest MSE at 0.58 ± 0.11, followed closely by mental health (MF) with

Label	Approach	K-fold CV	Accuracy, %	Weighted Precision	Weighted Recall	Weighted F1 score
PhF_{binary}	cCNN+RF	data-based	71.85 ± 1.44	0.70 ± 0.03	0.73 ± 0.03	0.71 ± 0.04
	LSTM-AT	data-based	**73.56 ± 0.02**	**0.76 ± 0.05**	**0.74 ± 0.02**	**0.74 ± 0.03**
	LSTM-AT	subject-based	69.95 ± 0.17	0.72 ± 0.17	0.70 ± 0.17	0.69 ± 0.18
MF_{binary}	cCNN+RF	data-based	**66.20 ± 4.58**	0.65 ± 0.05	**0.66 ± 0.05**	**0.65 ± 0.05**
	LSTM-AT	data-based	64.01 ± 0.08	**0.67 ± 0.07**	0.64 ± 0.08	0.65 ± 0.07
	LSTM-AT	subject-based	47.83 ± 0.05	0.56 ± 0.22	0.48 ± 0.06	0.48 ± 0.12
VAS_{binary}	LSTM-AT	data-based	65.54 ± 0.09	**0.51 ± 0.17**	**0.66 ± 0.09**	**0.54 ± 0.13**
	LSTM-AT	subject-based	**66.06 ± 0.18**	0.47 ± 0.23	0.66 ± 0.18	0.54 ± 0.22

Table 2. Binary fatigue label prediction performances. cCNN+RF results are copied from Luo et al. [9]

an MSE of 0.67 ± 0.11. It is important to note that the visual analogue scale score (VAS) and the indicator of relative perception ($RelP$) have different ranges of possible values, making direct comparisons to other fatigue labels inappropriate. When comparing between data-based and subject-based cross-validation, subject-based CV generally exhibits a larger mean MSE and greater variations compared to its data-based CV counterpart. The exception to this is $RelP$; however, considering that the values only range from -1 to 1, an MSE of 0.65 is still relatively high.

Figure 3 and **Figure** 4 plot the model predictions against their ground truths for each of the four fatigue labels under data-based k-fold CV and subject-based k-fold CV, respectively. Under data-based k-fold CV (**Figure** 3), both physical and mental fatigue predictions exhibit significant correlations with their ground truths (PhF: r=0.55, p<0.0001; MF: r=0.47, p=0.0004). However, this correlation is significantly weakened in subject-based k-fold CV (**Figure** 4). While predictions for physical fatigue still show a significant correlation with the ground truth (r=0.40, p=0.0051), the correlation for mental fatigue is no longer significant (r=0.24, p=0.1064). Additionally, the ordinal fatigue label prediction model performed poorly for the visual analogue scale score and indicator of relative perception under both types of k-fold CV, indicating the model's insufficiency in predicting these measures.

	MSE (Data-based CV)	MSE (Subject-based CV)
Physical Fatigue (PhF)	0.58 ± 0.11	0.65 ± 0.33
Mental Fatigue (MF)	0.67 ± 0.11	0.78 ± 0.35
Visual analogue scale score (VAS)	4.28 ± 0.39	4.82 ± 0.56
Indicator of relative perception ($RelP$)	0.73 ± 0.08	0.65 ± 0.12

Table 3. Mean squared error of ordinal fatigue label predictions using the LSTM-AT model. Note that PhF and MF ranges from 0 to 3, VAS ranges from 1 to 10, and $RelP$ ranges from -1 to 1.

7 DISCUSSION

In this study, treating wearable sensor data as sequential inputs, combined with the utilization of recurrent neural networks—specifically Long Short-Term Memory with Attention (LSTM-AT) models—resulted in noteworthy enhancements in predicting physical fatigue. Additionally, the performance in predicting mental fatigue was comparable, underscoring the effectiveness of this approach. These findings contribute significantly to the expanding body of literature supporting the use of sequential data and advanced neural network architectures

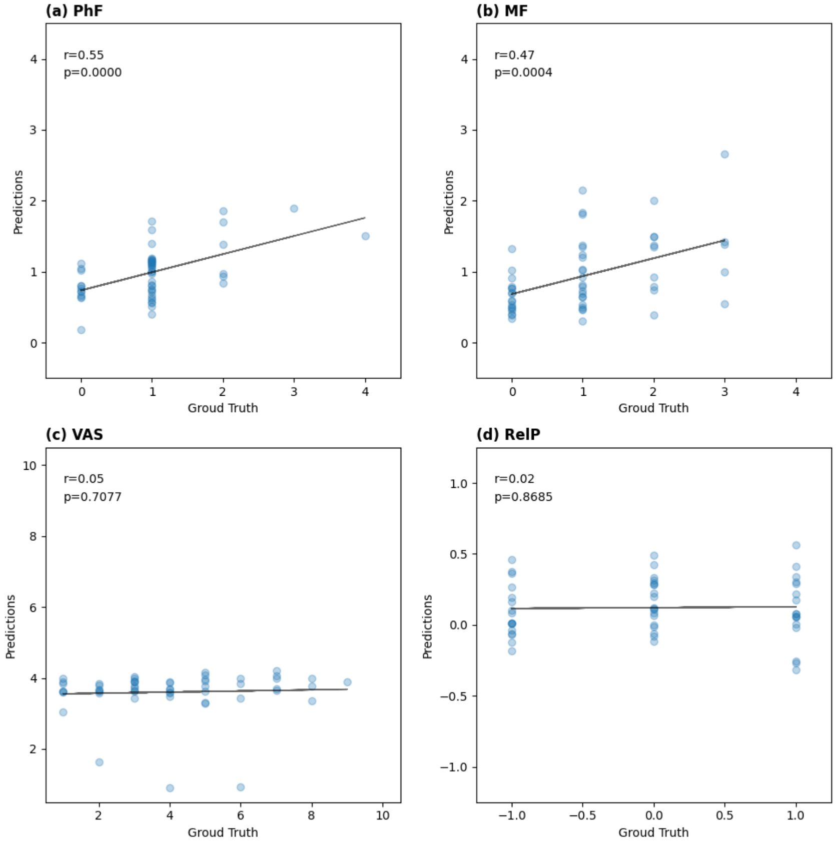

Fig. 3. Scatter plot illustrating the model predictions against ground truths for fatigue labels under data-based k-fold cross-validation. Both physical (PhF) and mental fatigue (MF) predictions show significant correlations with their ground truths (PhF: r=0.55, p<0.0001; MF: r=0.47, p=0.0004).

for fatigue assessment.[4] Consistent with existing literature,[9] our results affirm that physical fatigue is more predictable than mental fatigue based on wearable sensor data. This suggests that physical fatigue is more easily detectable through physiological changes, whereas identifying mental fatigue proves more challenging, possibly due to dynamic mechanisms and the influence of environmental factors.

Ordinal regression outputs exhibited a meaningful correlation with target labels, suggesting the potential for ordinal predictions in fatigue assessment. Nevertheless, the present model falls short in predicting exact ordinal scores. The observed limitation in predicting numerical fatigue scores with the current LSTM-AT model emphasizes the need for further refinement. This insight is pivotal for enhancing the precision and applicability of fatigue severity quantification. By enabling a more detailed score input and output (ordinal versus binarized), we foster a more nuanced representation of fatigue levels, facilitating a finer-grained understanding of the spectrum of fatigue severity.

One significant issue is that all models experienced notable performance drops after the implementation of patient stratification (subject-based k-fold CV). This implies that performance scores based on data-based k-fold cross-validation are affected by issues related to data leakage, resulting in a considerable decrease in model

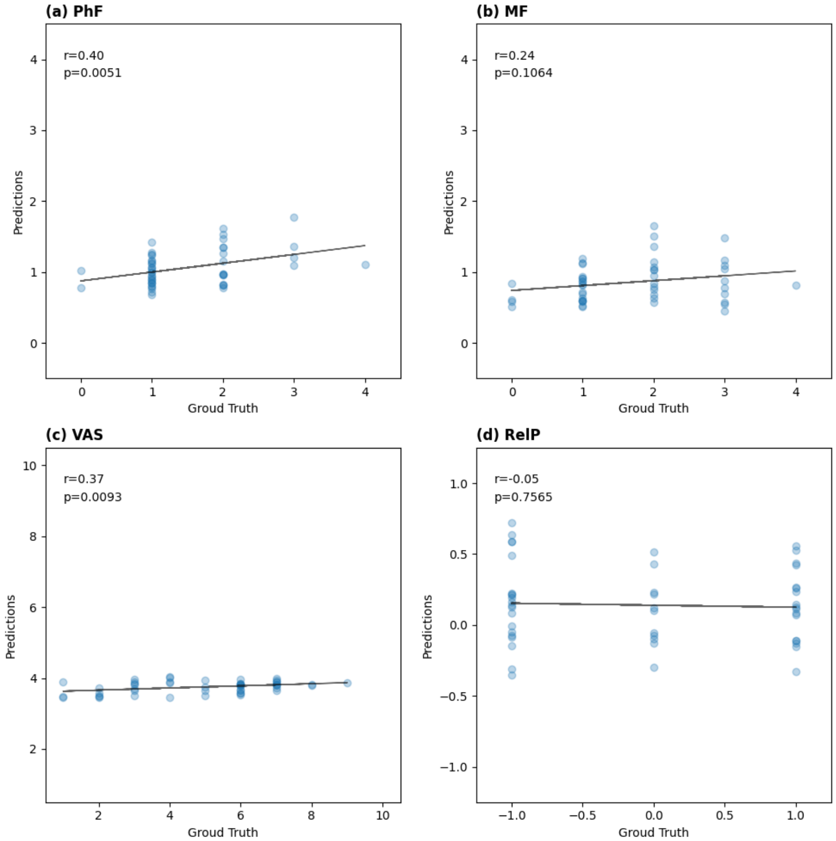

Fig. 4. Scatter plot depicting the model predictions against ground truths for fatigue labels under subject-based k-fold cross-validation. While predictions for physical fatigue maintain a significant correlation with the ground truth (*r*=0.40, *p*=0.0051), the correlation for mental fatigue becomes non-significant (*r*=0.24, *p*=0.1064).

performance when faced with novel, unseen data points. The challenges introduced by participant stratification underscore the influence of individual variability on model performance. Notably, the observed data leakage indicates that a subject's historical data can serve as a valuable predictor for future fatigue prediction. This duality of challenge and opportunity emphasizes the importance of tailoring predictive models to individual characteristics, thereby paving the way for more personalized and accurate fatigue assessments.

7.1 Limitations

Despite the promising results, it is essential to acknowledge several limitations. The relatively small dataset, consisting of 261 samples from a rather homogeneous population (all healthy adults), raises concerns regarding the generalizability of the findings. The over-representation of certain subjects further complicates the issue; for instance, some subjects have over 50 data segments, while most others only have around 7. This necessitates caution in interpreting and extrapolating the results. To address this limitation, a potential solution involves utilizing a larger and more diverse dataset. This may include selectively removing samples from over-represented subjects to achieve a more balanced and representative dataset. This approach aims to enhance the robustness and generalizability of the findings beyond the current sample set.

Another limitation stems from the inherent subjective biases in patient-reported outcomes and the intra- and inter-rater variability in response options. The variability in interpreting terms such as "exhausted" or the frequency of occurrences introduces uncertainty into the data. For instance, in the context of physical and mental fatigue, the response choices transition abruptly from "never" to "sometimes", without intermediary options like "rarely". Different participants may categorize their experiences differently. Moreover, the assumption of a linear ordinal scale for the labels introduces another layer of potential limitation. For example, the gap between "never" and "sometimes" may be greater than the gap between "regularly" and "often", challenging the validity of a strictly linear interpretation.

7.2 Future Work

Moving forward, our study identifies several avenues for future research. Fine-tuning the model with additional data from novel participants emerges as a promising strategy to augment predictive capabilities. Our results demonstrated a significant decline in model performance when the training dataset lacks data segments from the test participant. Conversely, this suggests that incorporating historical data from a new participant into the training set has the potential to enhance the model's predictions. This iterative approach potentially enables the model to adapt and refine its predictions based on individual characteristics, effectively addressing challenges associated with participant stratification. Furthermore, this fine-tuning process has the potential to mitigate issues related to inter- and intra-rater variability, as it learns the user's personal understanding of the responses.

Another general area of future work is to expand the dataset to encompass a more diverse population. The current study's limitations, stemming from a relatively small and homogeneous dataset, highlight the need for a broader range of participants. Including individuals with varying demographic characteristics, lifestyles, and health conditions will contribute to a more comprehensive understanding of fatigue prediction across diverse contexts. Moreover, exploring alternative machine learning approaches beyond the LSTM-AT model could provide valuable insights. Integrating different architectures may capture diverse patterns and relationships within the wearable sensor data, potentially improving predictive performance. Employing a multi-modal approach that combines wearable sensor data with other relevant sources, such as demographic information or environmental factors, could further enhance the model's accuracy. By incorporating these elements, future research can aim for a more holistic and adaptable fatigue assessment system capable of addressing the intricacies associated with diverse populations and varying predictive challenges.

8 CONCLUSION

In conclusion, this study sheds light on the promising potential of leveraging wearable sensor data, treated as sequential inputs, in conjunction with advanced neural network architectures for fatigue prediction. The LSTM-AT demonstrated superior performance compared to traditional CNN-based models in predicting physical fatigue and performed similarly well in predicting mental fatigue. While ordinal regression outputs show meaningful correlations with target labels, the challenge of predicting numerical fatigue scores highlights the need for continued refinement of models. Significant performance drops were observed after applying patient stratification, pointing to issues related to data leakage in existing data-based k-fold cross-validation approaches. Acknowledging limitations, such as the relatively small and homogeneous dataset, underscores the need for cautious interpretation. Looking ahead, fine-tuning models with individual participant data emerges as a promising avenue that can enhance the accuracy and personalization of fatigue assessments. This study lays the groundwork for future research, emphasizing the importance of collaborative efforts to navigate complexities and drive innovation in the field of fatigue prediction.

REFERENCES

[1] Lauren S. Aaronson, Cynthia S. Teel, Virginia Cassmeyer, Geri B. Neuberger, Leonie Pallikkathayil, Janet Pierce, Allan N. Press, Phoebe D. Williams, and Anita Wingate. 1999. Defining and Measuring Fatigue. *Image: the Journal of Nursing Scholarship* 31, 1 (March 1999), 45–50. https://doi.org/10.1111/j.1547-5069.1999.tb00420.x

[2] Mohamed R. Al-Mulla, Francisco Sepulveda, and Martin Colley. 2011. A Review of Non-Invasive Techniques to Detect and Predict Localised Muscle Fatigue. *Sensors* 11, 4 (March 2011), 3545–3594. https://doi.org/10.3390/s110403545

[3] Anza Aqeel, Ali Hassan, Muhammad Attique Khan, Saad Rehman, Usman Tariq, Seifedine Kadry, Arnab Majumdar, and Orawit Thinnukool. 2022. A Long Short-Term Memory Biomarker-Based Prediction Framework for Alzheimer's Disease. *Sensors* 22, 4 (Feb. 2022), 1475. https://doi.org/10.3390/s22041475

[4] Yang Bai, Yu Guan, and Wan-Fai Ng. 2020. Fatigue assessment using ECG and actigraphy sensors. In *Proceedings of the 2020 International Symposium on Wearable Computers*. ACM. https://doi.org/10.1145/3410531.3414308

[5] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. 2018. BRITS: Bidirectional Recurrent Imputation for Time Series. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2018/file/734e6bfcd358e25ac1db0a4241b95651-Paper.pdf

[6] Josef Finsterer and Sinda Zarrouk Mahjoub. 2013. Fatigue in Healthy and Diseased Individuals. *American Journal of Hospice and Palliative Medicine®* 31, 5 (July 2013), 562–575. https://doi.org/10.1177/1049909113494748

[7] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (Nov. 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[8] Xiao Li, Jessilyn Dunn, Denis Salins, Gao Zhou, Wenyu Zhou, Sophia Miryam Schüssler-Fiorenza Rose, Dalia Perelman, Elizabeth Colbert, Ryan Runge, Shannon Rego, Ria Sonecha, Somalee Datta, Tracey McLaughlin, and Michael P. Snyder. 2017. Digital Health: Tracking Physiomes and Activity Using Wearable Biosensors Reveals Useful Health-Related Information. *PLOS Biology* 15, 1 (Jan. 2017), e2001402. https://doi.org/10.1371/journal.pbio.2001402

[9] Hongyu Luo, Pierre-Alexandre Lee, Ieuan Clay, Martin Jaggi, and Valeria De Luca. 2020. Assessment of Fatigue Using Wearable Sensors: A Pilot Study. *Digital Biomarkers* 4, Suppl. 1 (Nov. 2020), 59–72. https://doi.org/10.1159/000512166

[10] Kei Mizuno, Masaaki Tanaka, Kouzi Yamaguti, Osami Kajimoto, Hirohiko Kuratsune, and Yasuyoshi Watanabe. 2011. Mental fatigue caused by prolonged cognitive load associated with sympathetic hyperactivity. *Behavioral and Brain Functions* 7, 1 (2011), 17. https://doi.org/10.1186/1744-9081-7-17

[11] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[12] Chaitra Rao, Elena Di Lascio, David Demanse, Nell Marshall, Monika Sopala, and Valeria De Luca. 2023. Association of digital measures and self-reported fatigue: a remote observational study in healthy participants and participants with chronic inflammatory rheumatic disease. *Frontiers in Digital Health* 5 (June 2023). https://doi.org/10.3389/fdgth.2023.1099456

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[14] Junli Xu, Jianliang Min, and Jianfeng Hu. 2018. Real-time eye tracking for the assessment of driver fatigue. *Healthcare Technology Letters* 5, 2 (Jan. 2018), 54–58. https://doi.org/10.1049/htl.2017.0020