Fusion of Learned Representations for Multimodal Sensor Data Classification

Lee B. Hinkle[0000−0001−7346−6344], Gentry Atkinson[0000−0001−7049−3734], and Vangelis Metsis[0000−0002−7371−8887]

Department of Computer Science, Texas State University
San Marcos, TX 78666, USA
{leebhinkle, gma23, vmetsis}@txstate.edu

Abstract. Time-Series data collected using body-worn sensors can be used to recognize activities of interest in various medical applications such as sleep studies. Recent advances in other domains, such as image recognition and natural language processing have shown that unlabeled data can still be useful when self-supervised techniques such as contrastive learning are used to generate meaningful feature space representations. Labeling data for Human Activity Recognition (HAR) and sleep disorder diagnosis (polysomnography) is difficult and requires trained professionals. In this work, we apply learned feature representation techniques to multimodal time-series data. By using signal-specific representations, based on self-supervised and supervised learning, the channels can be evaluated to determine if they are likely to contribute to correct classification. The learned representation embeddings are then used to process each channel into a new feature space that serves as input into a neural network. This results in a better understanding of the importance of each signal modality as well as the potential applicability of newer self-supervised techniques to time-series data.

Keywords: self-supervised learning · learned representations · physiological sensor data · polysomnography · human activity recognition · multimodal.

1 Introduction

Multimodal, time-series data is generated by various types of sensors that record movement, electrical potentials, temperature, sound, and other information. It is widely used, and various methods for collecting and fusing the data, including signal processing, feature engineering, and feature extraction, have been utilized [13] [21]. The focus of this work is on the exploration of self-supervised techniques such as nearest-neighbor contrastive learning (NNCLR) [5], which learn representations from unlabeled image data to multimodal time-series data. Specifically, the data from sensors collecting human attributes such as movement, heart rate, and respiration in the area of Human Activity Recognition (HAR) and sleep disorder diagnosis is used as input to deep neural networks with supervised and self-supervised learning configurations.

Diagnosing sleep apnea using polysomnography (PSG) data is an important but expensive task. These data must currently be collected in a controlled environment and interpreted by trained professionals. Machine learning (ML) could potentially make apnea diagnoses cheaper and more widely available. One difficulty with applying ML to the classification of apneic events in sleep data is the wide variety of sensors involved. ML models may struggle to correctly represent and identify the wide range of signals which include: electroencephalography (EEG), electromyography (EMG), electrooculography (EOG), thoracic and abdominal belts, electrocardiography (ECG), and temperature sensors.

In this work, we propose an ML architecture for detecting apneic events in PSG data. This architecture can leverage unlabeled data to drive representation learning of individual signal channels. The importance of these channels is determined during training, meaning that the model's architecture is not tied to any set of sensors but can be easily adapted to fit different data sets and even domains. Signal channels that are more useful to the prediction outcome automatically receive a higher weight at the fusion layer. To demonstrate the flexibility of the model, we use a small labeled HAR dataset and a larger unlabeled dataset collected with the same device. This shows that our system is highly adaptable to its inputs and can be trained on a small set of labeled data by using a larger, unlabeled dataset to boot-strap the feature learners.

The contributions of this work are as follows. An Empatica E4 Wristband (UE4W) dataset containing over 250 hours of unlabeled data for evaluating semi-supervised learning data pipelines. A methodology and code[1] for efficiently processing the physiological signal data in the publicly available PSG-Audio dataset [11]. A Tensorflow adaptation of NNCLR for use with time series.A data pipeline that allows for the seamless fusion of different signal modalities. An evaluation of the predictive performance of three architectures: 1) Concatenation of all channels at the input level. 2) Late fusion of convolutional layers trained on each individual channel. 3) Pretraining on unlabeled data of each signal type, and late fusion and fine-tuning on smaller labeled datasets.

2 Background and Related Work

Collecting physiological data from sleeping subjects is important for diagnosing irregularities in breathing and heart rate. The need to collect this data in a controlled environment and the requirement that it be interpreted by trained experts makes the process expensive and time-consuming. ML can improve this process by automatically interpreting collected data. Apnea detection by convolutional neural networks (CNNs) using a face-mounted nasal pressure device has achieved 96.2% accuracy [3]. Approaches using body-worn belt-type sensors have achieved 84% accuracy [19]. Other studies have shown that ML techniques for apnea detection generalize across humans, rats, and pigeons [1].

In addition to the challenges of collecting a large number of signals while a subject is trying to sleep in a clinical site, manually detecting and labeling the

[1] `https://github.com/imics-lab/fusion-of-learned-representations`

episodes of apnea (cessation of breathing in excess of 10 seconds) and hypopnea (a reduction in airflow measured directly or reflected in lowered blood oxygen saturation) is difficult [14]. The PSG-Audio dataset [11, 12] used in this work contains data recorded on 212 individuals in a hospital setting for sleep apnea syndrome (SAS) diagnosis. Five categories of abnormal events were annotated by a medical team. At this time of this writing, two studies have applied machine learning techniques and published results for the PSG-Audio dataset. In [4] Mel-spectrograms of the audio signals were input into a pre-trained VGG19 (19-layer deep CNN+Max Pooling) model followed by a long-short-term memory (LSTM) fused model. The leave-one-out subject accuracy was 66.29%. In [10], the three EEG signals are converted into spectrograms which are input into a model with three 1D-CNN layers followed by a Gated Recurrent Unit (GRU) with a goal to determine the start and stop time of each episode as the duration of the apnea is significant for diagnosis.

A common shortcoming of ML approaches to PSG data processing is that they rely on the presence of a large body of labeled data and the burden of professionally labeling these data makes them expensive and difficult to publish. Unsupervised and self-supervised approaches that do not rely on labeled data have only recently been applied to diagnosing sleep apnea. Clustering has been used to demonstrate an association between PSG data and the risk of future cardiovascular events [24]. Representation learning of PSG data using self-supervised learning has also been investigated [23]. This approach is very promising, but the remaining difficulty is the heterogeneous nature of the channels. The PSG signals are physically dissimilar and may not have much to contribute to the disorder being predicted.

Self-supervised learning removes the need for labeled data by training models to recognize labels that are generated automatically by the algorithm. A sub-field of self-supervised learning, contrastive learning, generates these labels by augmenting anchor samples of data and training an encoder to recognize augmented copies of the same instance [17]. This work has applied a deep representation learner called NNCLR [5], which was originally developed as a tool for image classification. NNCLR maintains knowledge of nearest neighbors in the learned feature space to improve the process of augmenting data. Other works have had success in adapting NNCLR to time series data [20], but this technique has not been tested in the field of PSG. SimCLR and the accompanying analysis [2] show that for image processing, contrastive learning results for image recognition are improved by: proper data augmentations, learnable non-linear transformations, larger batch sizes, and longer training. The SeqCLR framework [16] applies contrastive learning to time-series EEG signals for Emotion Recognition, EEG classification, and sleep-scoring tasks. LIMU-BERT [22] builds upon the concepts of the natural language model BERT to process inertial measurement unit (motion) data for HAR. Our data pipeline and learned representation model utilize these promising techniques for multimodal HAR and PSG data.

3 Methodology

3.1 Data Processing

To demonstrate our data processing pipeline two multimodal datasets were used. The TWristAR dataset [7] consists of multimodal channels recorded with an Empatica E4 Wristband [6]. The channels are acceleration (movement), blood volume pulse (BVP), electrodermal activity (EDA), and peripheral skin temperature. Per Empatica "the BVP signal is obtained from the PPG sensor by a proprietary algorithm that combines the light signals observed during both green and red exposure." All of the signals were re-sampled to 32Hz. The TWristAR dataset is small with only three subjects performing six common HAR activities in a scripted manner for ease of labeling and to produce a balanced dataset. The data were split into sliding windows of three-second duration (96 samples) with a step size of one second. For early evaluation of channel contributions, subject three was reserved for testing only, while subjects one and two are used for training and validation. Due to the presence of both subjects in the validation set and the overlapping windows, the validation result represents subject-dependent accuracies and the expected test accuracy is lower as described in our previous work [8]. An associated dataset, the unlabeled E4 wristband (UE4W) dataset [9] provides significantly more unlabeled data for self-supervised training. Fig. 1 shows an example of signals collected by the four sensors in the E4 wristband.

As described earlier, the second dataset used, PSG-Audio [11, 12] contains data recorded in a hospital setting for SAS diagnosis. Several elements of this dataset that led us to use it in this work: it is fully open-access, multimodal, and the labeling was completed and reviewed by trained personnel. It should be noted that the namesake signals, two channels of high-definition audio, were not used in this work. The 12 channels that were used are shown in Fig. 2b. Nine of the channels (3 x EEG, 2 x EMG, 2 x EOG, 2 leg sensors, and ECG) were downsampled from 200 Hz to 100 Hz to match the remaining three sensors used (flow thermistor plus thoracic and abdominal respiratory belts). Each subject's data, contained within a standard format EDF file were segmented into non-overlapping sliding windows with 500 samples each representing five seconds of time. The event data for labeling were derived from the provided "clean" rml file for patients 995-1494 (not all numbers are present). The respiratory events for "Obstructive Apnea", "Central Apnea", "Mixed Apnea", "Hypopnea", and three other categories were treated as "abnormal". All subject data is labeled, so for the purposes of this work data from the first 50 subjects were treated as labeled and data from the second 50 subjects were treated as unlabeled and only used to train the self-supervised models. For supervised learning, the train and validation sets were better balanced by discarding a portion of the normal samples. Neither the test set nor the unlabeled data could be rebalanced without incurring data leakage.

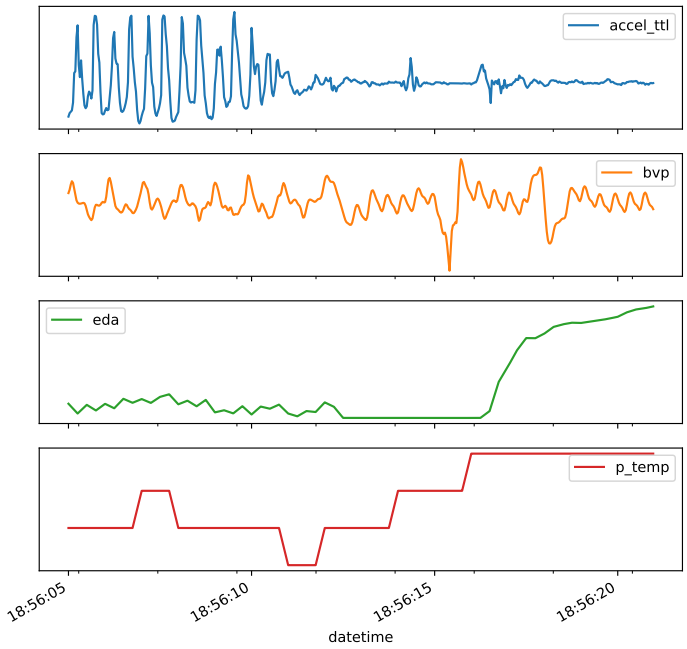
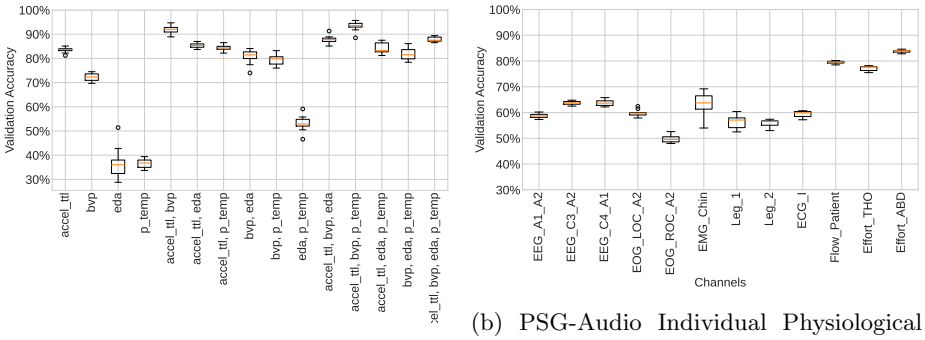


Fig. 1: A 15-second sample of the Empatica E4 sensor data. For this work, the 64Hz BVP data is downsampled and the 4Hz EDA and peripheral temperature signals are upsampled so that all signals match the 32Hz sampling rate of the acceleration data.



(a) TWristAR Channel Powerset

(b) PSG-Audio Individual Physiological Signals

Fig. 2: Boxplots showing the accuracy of individual channels. In the case of TWristAR with four channels, the powerset of all combinations concatenated was evaluated. For PSG-Audio each of the 12 channels was evaluated individually.

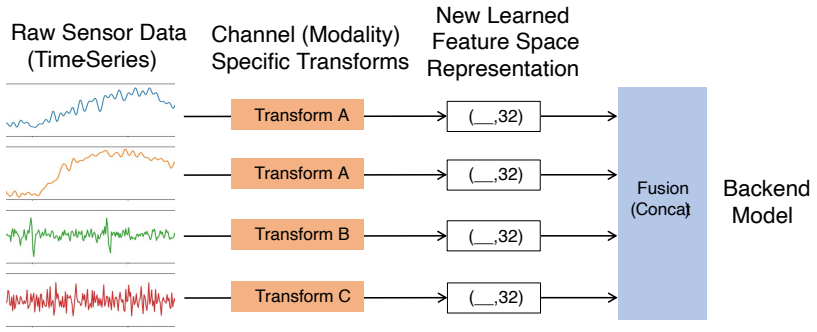


Fig. 3: Each time-series signal is transformed into a new n-dimensional feature space by a learned representation model. The models can be trained in a supervised or in a self-supervised manner on labeled or unlabeled data, respectively.

3.2 Fusion of Multimodal Sensor Data

ML models can learn from several channels of data simultaneously, but when those channels are dissimilar from one another or non-uniform in their contribution to the output of the model, it may not be beneficial to apply the same process of representation learning to every channel. To test this hypothesis, a baseline classifier was trained that applied separate convolutional layers to subsets of the available channels from our investigated datasets. The outputs of these convolutional layers were concatenated and then passed a global average pooling layer, and finally classified by dense layers.

The baseline classifier was also used to conduct a per-channel analysis. Since the network architecture other than shape does not vary with the number of input channels it is possible to run this classifier with one or a subset of multiple channels. For the four channels in the TWristAR dataset, the full powerset (all possible channel combinations) were run. The results of 10 runs in Figure 2 show that the accuracy with the acceleration and blood volume pulse are much higher than the electro-dermal activity and peripheral temperature signals. The subject-dependent accuracy was highest with all signals except EDA. Figure 2 also shows channel data for PGS-Audio. Due to the number of possible combinations the powerset was not run. The highest accuracies for respiration classification were the airflow and respiration band signals which serve as the basis for the subsequent model evaluations. The EMG_Chin signal showed much greater variability across the 10 runs for unknown reasons.

The models trained as our baseline for comparison were trained using standard back-propagation of the loss relative to the assigned labels. While they will still benefit from learning channel-wise representations of the input data, they cannot capitalize on the large body of unlabeled data that we had available for the two datasets. For the final comparison between the concatenated, multi-headed, and self-supervised experiments sklearn [18] GroupKFold was used

to perform hold-one-subject-out cross-validation. For the PSG-Audio dataset, a similar methodology was used but the 192 subjects were split group-wise for 5-fold cross-validation.

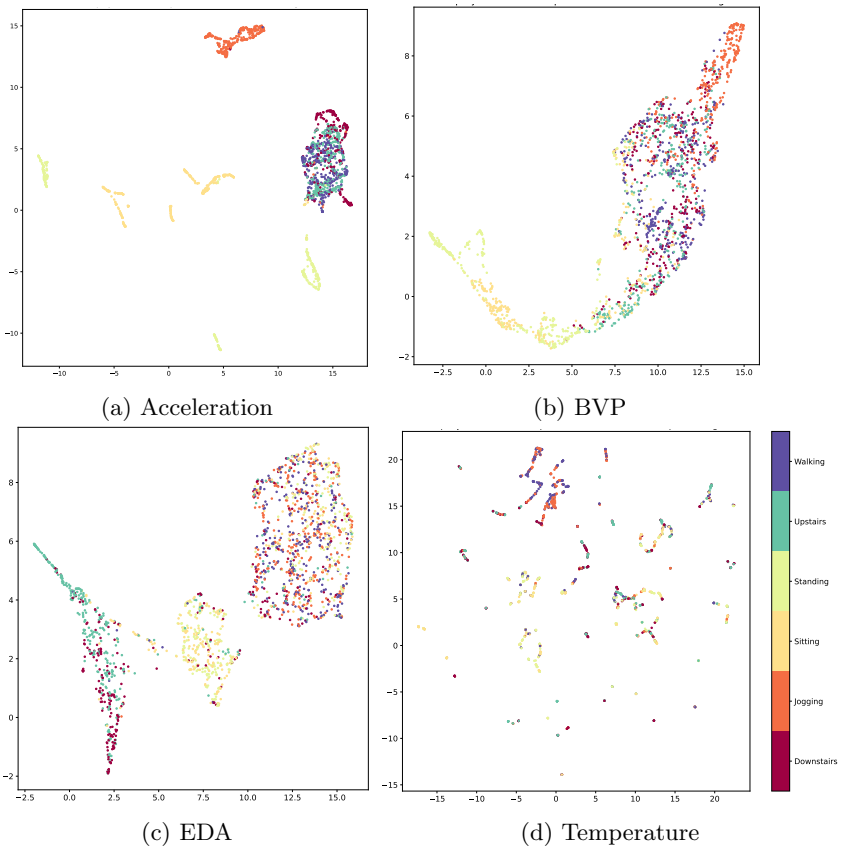
3.3 Self-Supervised Learned Representations

NNCLR [5] is a training method that uses labels automatically generated at train time to fit feature extraction layers to training data with or without assigned labels. As NNCLR is a training technique, not a model architecture, it can be applied to a variety of neural network architectures. The model architecture used to learn the feature representations is usually called an encoder. For this project, NNCLR was used as the training platform for an encoder composed of two layers of 1D convolutional networks. This CNN-based encoder architecture was chosen because it is simple, effective, and highly flexible for time-series data.

Each feature encoder was trained on one channel of the unlabeled and labeled training data. A validation group was held out, and the loss on that set was used for early stopping. The trained encoders were then used as pre-trained models to generate feature vectors for one channel of the labeled training data, with a fixed output vector dimensionality of 32 elements. Pairs of channels that were univariate, such as the left and right eye in the EOG data, were encoded using a single encoder and represented by a single vector. The output of all encoders is fused together and fed to dense layers for supervised learning. The weights of the encoder can be frozen or allowed to continue to be trained during the supervised training of the fused model. In our case, we did not freeze the encoder. Due to data availability limitations, especially for the PSG dataset, the encoders were trained on roughly twice as much data as the labeled dataset, in other domains considerably more unlabeled data are typically used.

Figure 4 shows the UMAP [15] 2D projection of the 32-dimensional self-supervised learned representation feature vectors for the TWristAR dataset. The encoder model used to produce the feature vectors was trained on the unlabeled E4 wristband (UE4W) dataset [9]. The labels were used to color code each data point for the figure, but not during the training of the encoder. The features learned from the acceleration channel create a separation of the different classes in the feature space. The data points of “standing”, “sitting”, and “jogging” are far apart from the points of the other classes, which suggests they would be easy to classify. The points of “walking”, “upstairs”, and “downstairs” are closer together, which would make them harder to separate. The BVP channel appears to be the second best, although there is some overlap between different classes they do not completely overlap. The EDA and Temperature channels seem to be the worst with points with all classes scattered around.

Figure 5 shows similar plots for the different channels of the PSG dataset. The projected features were processed in a similar manner. The PSG-Audio dataset contains a larger number of instances, and thus the graphs are much denser. The distribution of data points coming from the two classes (normal and abnormal breathing) largely overlap for most channels. The channels for which the point distributions overlap the least are the airflow and the thoracic plus abdominal



(a) Acceleration (b) BVP

(c) EDA (d) Temperature

Fig. 4: UMAP projections of self-supervised learned representations of sequences from the different channels of the TWristAR dataset.

effort. This is expected, as these channels are directly associated with breathing patterns. These three channels also produced the highest classification accuracy when used individually to detect abnormal breathing, as shown in Figure 2b.

3.4 Model comparisons

To demonstrate the strength of channel-specific representation learning using self-supervised learning, a classifier was trained on the fused feature vectors output by the NNCLR feature learners. The three encoders used for three different channels share the same network architecture, but their weights have been separately pre-trained on unlabeled data of that channel. Figure 6 shows the architecture of the model that utilizes the self-supervised learning encoders.

Table 1 shows the final comparison of the three architectures evaluated for the TWristAR and PSG-Audio datasets. The train/test split was made using GroupKfold, with each subject/patient representing a different group, with

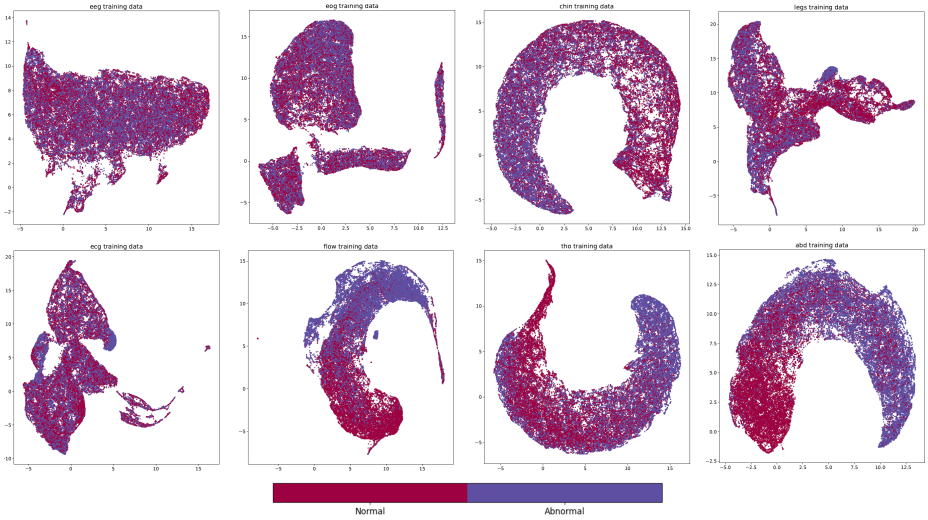


Fig. 5: UMAP projections of learned representations of sequences from different channels of the PSG dataset. From left to right, top to bottom the images show EEG, EOG, Chin EMG, Legs EMG, ECG, Air Flow, Thoracic Effort, and Abdominal Effort.

Table 1: Classification percent accuracies of the three different architectures.

	Simple Concat	Multi-head Supervised	Multi-head Pretrained
TWristAR	79.1	80.9	80.6
PSG-Audio	88.9	88.3	86.3

three and five-fold cross-validation, respectively. The results show that the Multi-head Supervised model slightly outperforms the simple concatenation of the inputs. This is likely due to the ability to have different hyperparameters for each channel which is not possible in the concatenated model. The results with PSG-Audio show that the simple concatenation slightly outperformed the multi-headed model. Further tuning may be required. For both datasets, the pre-training did not improve the performance compared to fully supervised training. Based on the usage of self-supervised models in image recognition, it is possible that a much larger set of unlabeled data than is provided by these two datasets is required along with potentially more tuning, including alternate augmentations. Note that the CNN model architecture was initially tuned for the single-head model and was subsequently copied and used as an encoder for each of the different heads of the pre-trained model.

Figure 7 the left is the confusion Matrix for the Concatenated Model, which uses accelerometer and BVP channels as flat inputs into a single CNN. On the right is the confusion Matrix for the Multihead Model, which uses two separately tuned CNNs for the accelerometer and BVP channels. Input is the TWristAR dataset evaluated with hold-one-subject out. The TWristAR dataset is balanced,

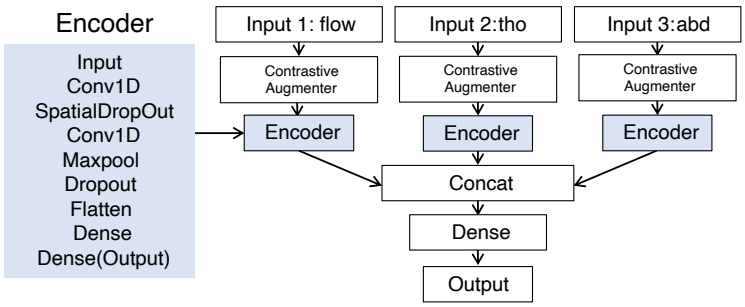
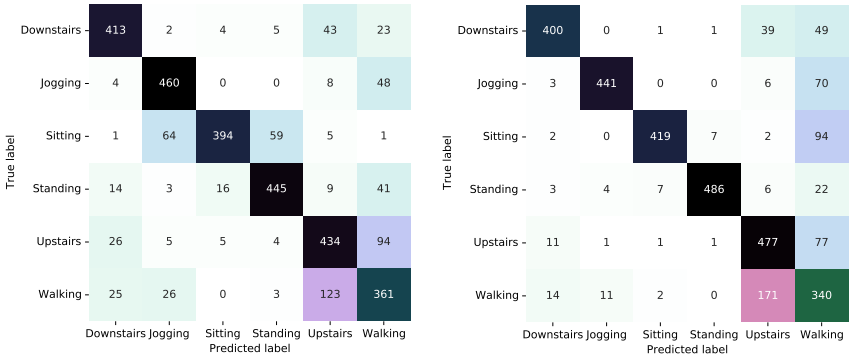


Fig. 6: The self-supervised model architecture



(a) Input Concatenation (b) Multi-Headed

Fig. 7: The confusion matrices for the Concatenated Input model and the Multihead Model.

so the accuracy is reflective of the classifiers' performance. Both architectures struggle to differentiate between some instances of similar activities.

4 Discussion

This work focuses on wearable sensors because of the heterogeneous nature of the signals. A breath flow sensor is much more dissimilar from an ECG signal than the RGB channels of an image. This means that apnea and activity detection can benefit from selecting the information-dense channels and either discarding the others or using ML techniques such as attention to minimize the weights of low-value channels. But this observation is not limited to these specific problem spaces. The methods described in this work could also be applied to other varieties of time-series data with non-uniformly informative channels.

Some degree of domain expertise is still essential when selecting channels of data for training. Ideally, an ML approach to data analysis should rely as little as possible on human expertise. It is challenging to assess which sensors

should be useful for solving a given problem without understanding the physical properties of those sensors. One observation of this work has been that the information density of a signal can be roughly estimated by observing the UMAP projections of the learned representations of the training data. The potential of evaluating the usefulness of a channel without access to labeled data warrants further investigation as being able to reduce the number of channels required by examining the results of a self-supervised model trained only on unlabeled data would be very beneficial. At a minimum, this information could guide more targeted experiments using less of the costly labeled data.

5 Conclusion

In this work, we have proposed a methodology for modeling multimodal time-series data when the different signal channels significantly differ from each other. As larger datasets become available, having the ability to tune learning to the properties of each signal and to replace parts of the network with a model that has been trained elsewhere, is crucial. Furthermore, the ability to leverage large amounts of unlabeled training data can benefit supervised models trained on limited-size labeled datasets. Self-supervised learning methods, which have been popularized in other machine learning domains, can have a significant impact on health-related applications in the near future.

References

1. Allocca, G., Ma, S., Martelli, D., Cerri, M., Del Vecchio, F., Bastianini, S., Zoccoli, G., Amici, R., Morairty, S.R., Aulsebrook, A.E., et al.: Validation of 'somnivore', a machine learning algorithm for automated scoring and analysis of polysomnography data. Frontiers in neuroscience **13**, 207 (2019)
2. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
3. Choi, S.H., Yoon, H., Kim, H.S., Kim, H.B., Kwon, H.B., Oh, S.M., Lee, Y.J., Park, K.S.: Real-time apnea-hypopnea event detection during sleep by convolutional neural networks. Computers in biology and medicine **100**, 123–131 (2018)
4. Ding, L., Peng, J., Song, L., Zhang, X.: Automatically detecting apnea-hypopnea snoring signal based on vgg19+ lstm. Biomedical Signal Processing and Control **80**, 104351 (2023)
5. Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., Zisserman, A.: With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9588–9597 (2021)
6. Empatica: E4 wristband user's manual, rev. 2.0. `https://empatica.app.box.com/v/E4-User-Manual` (2020), accessed: 2022-06-09
7. Hinkle, L.B., Atkinson, G., Metsis, V.: Twristar - wristband activity recognition. online (Jan 2022). https://doi.org/10.5281/zenodo.5911808, `https://doi.org/10.5281/zenodo.5911808`

8. Hinkle, L.B., Metsis, V.: Model evaluation approaches for human activity recognition from time-series data. In: International Conference on Artificial Intelligence in Medicine. pp. 209–215. Springer (2021)
9. Hinkle, L.B., Metsis, V.: Unlabeled Empatica E4 Wristband Data (UE4W) Dataset. online (Jul 2022). https://doi.org/10.5281/zenodo.6898244, `https://doi.org/10.5281/zenodo.6898244`
10. Kokkalas, L., Korompili, G., Tatlas, N.A., Mitilineos, S.A., Potirakis, S.M.: Severe obstructive sleep apnea event detection from eeg re-cordings. In: Presented at 2nd International Electronic Conference on Applied Sciences. vol. 15, p. 31 (2021)
11. Korompili, G., Amfilochiou, A., Kokkalas, L., Mitilineos, S.A., Tatlas, N.A., Kouvaras, M., Kastanakis, E., Maniou, C., Potirakis, S.M.: Psg-audio, a scored polysomnography dataset with simultaneous audio recordings for sleep apnea studies. Scientific data **8**(1), 1–13 (2021)
12. Korompili, G., Amfilochiou, A., Kokkalas, L., Mitilineos, S.A., Tatlas, N.A., Kouvaras, M., Kastanakis, E., Maniou, C., Potirakis, S.M.: PSG-Audio (Mar 2022). https://doi.org/10.11922/sciencedb.00345, `\url{https://doi.org/10.11922/sciencedb.00345}`
13. Lahat, D., Adali, T., Jutten, C.: Multimodal data fusion: an overview of methods, challenges, and prospects. Proceedings of the IEEE **103**(9), 1449–1477 (2015)
14. Levy, P., Pépin, J.L., Deschaux-Blanc, C., Paramelle, B., Brambilla, C.: Accuracy of oximetry for detection of respiratory disturbances in sleep apnea syndrome. Chest **109**(2), 395–399 (1996)
15. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)
16. Mohsenvand, M.N., Izadi, M.R., Maes, P.: Contrastive representation learning for electroencephalogram classification. In: Machine Learning for Health. pp. 238–253. PMLR (2020)
17. Mondal, A., Jain, V., Siddiqi, K.: Mini-batch similarity graphs for robust image classification (2012)
18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
19. Piorecky, M., Bartoň, M., Koudelka, V., Buskova, J., Koprivova, J., Brunovsky, M., Piorecka, V.: Apnea detection in polysomnographic recordings using machine learning techniques. Diagnostics **11**(12), 2302 (2021)
20. Qian, H., Tian, T., Miao, C.: What makes good contrastive learning on small-scale wearable-based tasks? arXiv preprint arXiv:2202.05998 (2022)
21. Sleeman IV, W.C., Kapoor, R., Ghosh, P.: Multimodal classification: Current landscape, taxonomy and future directions. arXiv preprint arXiv:2109.09020 (2021)
22. Xu, H., Zhou, P., Tan, R., Li, M., Shen, G.: Limu-bert: Unleashing the potential of unlabeled data for imu sensing applications. In: Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems. pp. 220–233 (2021)
23. Zhao, A., Dong, J., Zhou, H.: Self-supervised learning from multi-sensor data for sleep recognition. IEEE Access **8**, 93907–93921 (2020)
24. Zinchuk, A.V., Jeon, S., Koo, B.B., Yan, X., Bravata, D.M., Qin, L., Selim, B.J., Strohl, K.P., Redeker, N.S., Concato, J., et al.: Polysomnographic phenotypes and their cardiovascular implications in obstructive sleep apnoea. Thorax **73**(5), 472–480 (2018)