After getting a univariate time series $S = \langle S_j \rangle_{j=1}^{n}$, we learn an attention weight for a specific time interval j by:

$$\alpha_j = \boldsymbol{W}_j^{\top} \boldsymbol{S}, \tag{7}$$

where $\boldsymbol{W}_j \in \mathbb{R}^n$ is an n-dimensional vector of parameters. To obtain the attention weights $\boldsymbol{\alpha} = \{\alpha_j\}_{j=1}^{n}$ for n time intervals, we need n^2 parameters. Together with the number of embedding parameters, the total number of parameters needed is $(n^2 + \kappa + \beta + 1)$.

After that, a softmax operator is applied on $\boldsymbol{\alpha}$ and we get the time-aware attention mechanism specially for limited training samples. For each time interval in $TC = \langle \boldsymbol{tc}_j \rangle_{j=1}^{n}$, we combine the feature representation $\boldsymbol{tc}_j$ with the corresponding attention weight α_j, and get the time-aware representation as $\boldsymbol{TC_\alpha} = \langle [\alpha_j, \boldsymbol{tc}_j] \rangle_{j=1}^{n}$. Note that this time-aware attention is learned using backpropagation, thus it is a data-driven approach to evaluating how important a specific time interval is in contributing to the prediction of AD.

3.3 Temporal Dependency Modeling

Recurrent Neural Networks (RNN), such as Long Short-Term Memory recurrent neural network (LSTM) and Gated Recurrent Unit (GRU) [9], are famous for their excellent performance in modeling the long-term temporal dependencies in time series data. In our solution, we use GRU to capture the temporal dependencies, since GRU can achieve the same level of performance but requires fewer parameters compared with LSTM. The consideration to take the temporal dependencies into modeling is that subjects have some long-term patterns and these patterns cannot be represented by short-term representations such as $\boldsymbol{TC_\alpha}$. For example, [10] finds that some severe AD subjects have some disturbances in sleep-wake cycles, and these kinds of patterns cannot be captured by $\boldsymbol{TC_\alpha}$.

GRU can capture the long-term temporal dependencies by keeping the recurrent hidden states inside. The hidden state h_j is updated upon the previous hidden state h_{j-1} with the input $\boldsymbol{TC_\alpha}$ and the personal particulars of a subject, denoted as $\boldsymbol{d}$, as defined in Section 2.2:

$$h_j = \phi(h_{j-1}, \boldsymbol{TC_\alpha}, \boldsymbol{d}), \tag{8}$$

where ϕ is a nonlinear function such as composition of a logistic sigmoid with an affine transformation.

4 EXPERIMENTS

In this section, we validate the effectiveness of TATC for predicting AD and MCI with actigraphy data.

4.1 Experimental Setup

4.1.1 Baselines and Metrics. We use three methods on time series classification as baselines: (1) Dynamic Time Warping (DTW) [4], implemented as a sum of squared DTW distances in each dimension; (2) BOSS [23]. As BOSS works on univariate time series only, we train a base classifier on each univariate time series and build an ensemble of four base classifiers; and (3) SMTS [3] which ranked second in the gesture recognition competition organized by 2nd ECML/PKDD Workshop on Advanced Analytics and Learning on Temporal Data[4]. For fair comparison with these baselines, personal particulars of subjects are excluded from TATC and only actigraphy time series is used.

[4]https://aaltd16.irisa.fr/challenge/

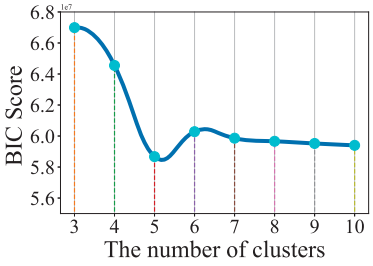


Figure 3: BIC score corresponding to different number of clusters

Table 4: Number of parameters in the TATC architecture

Component	Input	Output	#Parameters
CNN representation	24×16	12×2	64
max-pooling	12×2	6×2	0
TICC representation	24×5	24×5	0
max-pooling	24×5	6×5	0
TC representation	$6 \times (2 + 5)$	6×7	0
attention	6×7	6×1	44
time-aware representation	$6 \times (1 + 7)$	6×8	0
GRU	6×8	6×2	66
fully connected	6×2	2	26
logistic regression	2	1	3
			total: 203

Two binary classifiers are constructed: one on AD versus NC, and the other on MCI versus NC. To handle the imbalanced class distribution, we perform oversampling by SMOTE [6] on the AD and MCI samples in the training set.

We perform 5-fold cross validation and report the average results. In the medical domain, *sensitivity*, *specificity*, and Area Under the receiver operating characteristics Curve (*AUC*) are most commonly used metrics for evaluating the classification performance. In our problem, sensitivity measures the recall of positives (i.e., AD or MCI), and specificity measures the recall of negatives (i.e., NC).

4.1.2 Implementation details. We use minibatch based Adam [15] to minimize the binary cross-entropy loss. He-normal [13] is used as the initializer for CNN. The drop-out strategy is used in the full connected layer with a rate of 0.3.

We use Bayesian Information Criterion (BIC) to decide the optimal number of clusters in TICC, which is set to 5 as shown in Figure 3. According to BIC, we also discover the optimal penalty β to be 400. The number of temporal features α in CNN is set to 2. The number of time intervals n is set to 24. Detailed information of TATC's parameter size is listed in Table 4.

4.2 Results

The experimental results for predicting AD and MCI are listed in Table 5 and Table 6 respectively. For predicting AD, TATC achieves the best performance among 4 approaches, with a good balance

Table 5: Quantitative comparison of different classifiers to predict AD

Approach	Sensitivity	Specificity	AUC
DTW	90.3%	47.5%	68.9%
BOSS	38.7%	91.3%	76.1%
SMTS	45.2%	92.5%	84.5%
TATC	80.6%	86.3%	86.2%

Table 6: Quantitative comparison of different classifiers to predict MCI

Approach	Sensitivity	Specificity	AUC
DTW	70.0%	48.3%	59.1%
BOSS	5.7%	95.5%	58.8%
SMTS	5.0%	91.0%	58.5%
TATC	42.3%	81.3%	61.7%

between sensitivity and specificity. It shows great promise of being put into practice for early detection of AD. SMTS comes second, with a high specificity but a low sensitivity of 45.2%. DTW is biased towards AD, leading to a high sensitivity, but a low specificity and AUC.

For predicting MCI, TATC still achieves the best performance. We note that the classification performance is not as good as that for predicting AD. In the literature of MCI research, MCI is further categorized into stable MCI (sMCI) and progressive MCI (pMCI) [25]. Subjects who convert to AD within 36 months are classified as pMCI, and those who do not convert to AD are classified as sMCI. sMCI subjects are physically as active as NC subjects. We also observe in Figure 1 that the circadian activity of MCI and NC subjects is very close, making it hard to differentiate these two groups based on their physical activity. To improve the performance of detecting MCI, we plan to explore the possibility of incorporating other measurements besides actigraphy data.

We further evaluate the effectiveness of different components of TATC in predicting AD. In this group of experiments, personal features are included. Specifically we apply chi-squared test on the features listed in Table 1 for the purpose of feature selection, and select MASCH, MHDIAB, and GRIPAM as discriminative personal features. The classification performance is listed in Table 7. In *simple classifier*, only personal features are used, which achieves 65.6% in AUC. *cnn* shares the same architecture with TATC, but does not use TICC representation or time-aware attention. *cnn* achieves 79.7% in AUC. *ticc+cnn* uses both TICC and CNN representation, but lacks the time-aware attention mechanism. It achieves 83.6% in AUC. TATC achieves the best performance by including all components, demonstrating the effectiveness of the composite feature representation with TICC and CNN, the time-aware attention mechanism, as well as the personal features.

4.3 Hidden State Interpretation

In this subsection, we interpret the hidden states learned by TATC. As the optimal hidden state number is 5 according to BIC, 5 Markov

Table 7: Quantitative comparison of different components of TATC to predict AD

Approach	Sensitivity	Specificity	AUC
simple classifier	22.5%	92.5%	65.6%
cnn	64.5%	86.2%	79.7%
ticc+cnn	71.0%	81.3%	83.6%
TATC	73.5%	94.3%	88.2%

In *simple classifier*, only personal features are used. In *cnn*, there is no TICC or time-aware attention. In *ticc+cnn*, there is no time-aware attention.

Table 8: PageRank and mean values for five hidden states

State	Interpretation	Measure	X_{acc}	Y_{acc}	Z_{acc}	Lux
*1	**good sleep**	PageRank	0	0	0	0
		mean	0	0	0	0
*2	**sedentary activity**	PageRank	0.37	0.37	0.26	0
		mean	704	746	917	0
*3	**light activity**	PageRank	0.22	0.34	0.20	0.24
		mean	932	972	1240	23
*4	**moderate activity**	PageRank	0.31	0.30	0.25	0.14
		mean	1293	1281	1600	118
*5	**exercising**	PageRank	0	0	0	0
		mean	2563	2238	2270	628

random field (MRF) are generated, each corresponding to a hidden state. An MRF is a weighted undirected graph which consists of 4 vertices, and each vertex represents a variable from $I = [X_{acc}, Y_{acc}, Z_{acc}, Lux]$. If there is a partial correlation between two variables within a hidden state, there is an edge connecting them in the corresponding MRF. A large edge weight indicates that the two variables are heavily dependent on each other. Notice for each hidden state, its MRF will not change throughout time, thus the MRF can be treated as a unique signature for each hidden state. PageRank [21] is a commonly used graph analytic measure to quantify the relative importance of each vertex inside a graph. We apply PageRank to each of the 5 MRF to measure how influential a variable is inside a hidden state. If a variable has a very high PageRank value in an MRF, it means it has a strong capacity to influence other variables. In addition to PageRank, mean value of each variable reflects the intensity of body movement and ambient light. We list the calculated PageRank and mean values in Table 8.

We infer an interpretation for each hidden state as follows. For *1, as the PageRank and mean of *Lux* are both zero, we can infer that ambient light in this hidden state does not change along with body movement, meaning the circumstances do not change. Together with the fact that the mean of all variables is zero, we interpret *1 as *good sleep*. For *2, with the same observation on *Lux*, we infer the circumstances do not change. We also observe that *2 has the second lowest movement as measured by the mean values, and X_{acc}, Y_{acc} and Z_{acc} have a clear dependence of each other as reflected by PageRank. Thus we infer *2 as *sedentary activity* such as disturbed sleep. As for *5, all mean values reach the maximum indicating intensive body movement and high ambient light, whereas

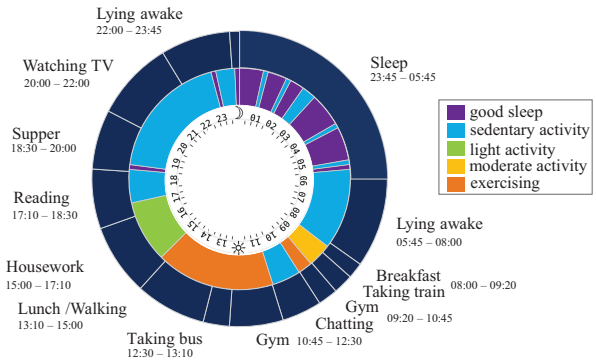
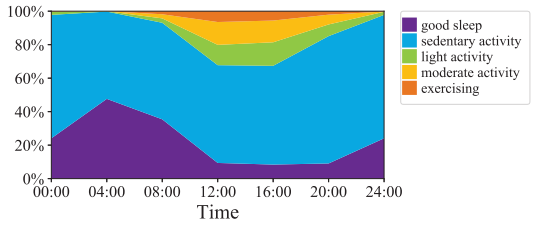
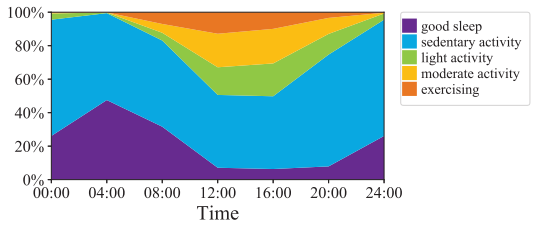


Figure 4: Self-report daily activities along with inferred hidden states. Inner circle represents hidden states inferred by TATC. Outer circle represents self-report activities.



(a) Distribution of hidden states for AD subjects



(b) Distribution of hidden states for NC subjects

Figure 5: Circadian activity comparison between AD and NC subjects

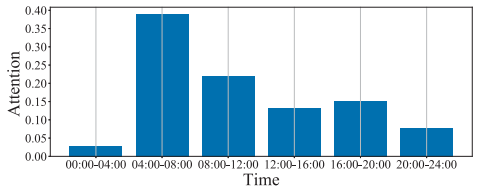


Figure 6: Average attention weight learned from NC and AD subjects

all PageRank values are 0 indicating no clear dependence between the variables. Thus we interpret it as *exercising*. For hidden states *3 and *4, their patterns are quite similar. The difference between these two is that acceleration and *Lux* are smaller for *3, which is interpreted as *light activity*, whereas *4 is interpreted as *moderate activity*. These can be regarded as a reasonable interpretation of the 5 hidden states learned by TATC.

We have designed a detailed self-report questionnaire for subjects to record their daily activities. We validate our interpretation of the hidden states by aligning their reported activities along with the hidden states in 24 hours. One NC case is exemplified in Figure 4. With reference to the self-report activities, we get the real-world interpretation of the hidden states. For example, *lying awake* and *watching TV* are both clustered into hidden state *2 *sedentary activity*, which makes sense because both of them involve low body movements and *Lux* remains stable. *Housework* belongs to hidden state *3 *light activity* since the subject needs to walk around inside the house and *Lux* may also change. *Gym* is related to vigorous body movement and is clustered into hidden state *5 *exercising*. One interesting finding is that the subject is unaware of his/her disturbed sleeping at night, while our method can capture light body movement during sleep and interpret those short periods as *sedentary activity*.

Based on the interpretation of the 5 hidden states, we compare the circadian activity of AD and NC subjects. Along 24 hours, we calculate the distribution of 5 hidden states for AD subjects (Figure 5a) and NC subjects (Figure 5b). We observe that AD subjects spend nearly 85% of their time on either *good sleep* or *sedentary activity*, versus 75% for NC subjects. Another interesting discovery is that the difference between AD and NC subjects is more obvious during 4:00am – 12:00noon. NC subjects spend 20% time on *exercising* while AD subjects spend only 9% in the same period. NC subjects spend 20% time on *moderate activity* during 8:00am – 12:00noon while AD subjects spend 13% only. This also proves that we should pay different attention to different time intervals. In the next subsection we give detailed analysis on the time-aware attention mechanism.

4.4 Interpretation of Attention Mechanism

The motivation to apply time-aware attention in TATC is that different time intervals in a day have different degree of importance to differentiate the three groups of subjects. To validate the effectiveness of the attention mechanism, we plot the average attention weight learned from the classification of AD versus NC subjects in a 24-hour time span in Figure 6. Indeed, we observe that different time intervals have different attention weights. The largest attention weight appears at 4:00 – 8:00am, which is the typical wake-up time for the elderly. This is consistent with our observation in Figure 1 that the biggest gap between AD and NC appears during 4:00 – 8:00am. It also justifies our interpretation of hidden states in Figure 5 where we notice NC group exercises more and has less sedentary activity than AD group. We also find that the time intervals 0:00 – 4:00am and 20:00 – 24:00pm have the lowest attention weights, as this is sleeping time with little body movement. It is consistent with our discovery in the hidden state interpretation that the two groups are similar at these two intervals.

5 LESSONS AND INSIGHTS

This paper presents our experiences of applying deep learning techniques to predict Alzheimer's Disease based on the collected

actigraphy data, and realize our ideas in a solution called TATC. We summarize our lessons and insights gained from this project.

5.1 Data Collection

Data collection is very challenging with respect to the aged cohort. As the average age of our subjects is above 80, engaging them in repeated clinical assessments is not an easy task even though we have standardized the procedures. For data collection from actigraphy devices, the subjects were educated on the proper usage of GT3X in the interview. But many subjects still forgot to put on the device after bathing or swimming, or forgot to wear it for various reasons. In addition, we have designed a detailed self-report questionnaire for subjects to record their daily activities, but the returned pieces are of low quality. In the end we could use only 20 self-report questionnaires for validation. An important lesson is that our data collection procedure should be made simple and bring little disturbance to subjects' daily life, so that more valid and valuable data can be recorded. In future work, we shall examine the inclusion of other devices such as GT9X, which can be worn during bathing and swimming.

5.2 Practical Value to Clinical Diagnosis

Traditional cognitive status diagnosis involves lots of clinical assessments and clinic visits (see Section 2), which heavily rely on the domain knowledge of doctors. These clinical assessments bring much burden to patients and may deteriorate their cognitive status. For those who have been diagnosed MCI, the progression from MCI to AD is unpredictable and there may be years or even dozens of years before AD is developed. In such a long period, frequent clinic visits become infeasible especially for those who refuse or have difficulties in doing so, and patients may lose the best opportunity of timely diagnosis and treatment.

Our proposed TATC method provides an automatic, low-cost solution for continuously monitoring the change of physical activity of subjects in daily living environment. The actigraphy data is sent to the server on a daily basis, upon which the classification model can be applied on the incoming data for prediction. Once potential risk of AD is identified, doctors will be alerted immediately. Then they can arrange clinic visits for subjects for further diagnosis and treatment. We believe the future deployment of TATC can benefit both doctors and patients in early detection of potential AD risk, particularly for those who have been diagnosed MCI, as monitoring the trajectory of changes is of great importance to them. This is the most important contribution of this study to the medical domain.

6 RELATED WORK

This work is related to time series representation, attention-based neural network and healthcare interpretations.

Time series representation techniques can be generally classified into two groups. The first group is non-transformed techniques. Representative works include SAX [16], Shapelets [26], TSF [11], and DTW [4]. This group of techniques works in the original time domain and represents time series as a common shape (e.g., DTW), or divides time series into several segments and represents them accordingly (e.g., SAX and TSF). In contrast to these non-transformed techniques, the second group transforms time series into another space. Representative works include SVD [5], BOSS [23], SMTS [3], MFCC [18], and CNN [20]. This group, especially CNN-based representation, has become quite popular in recent years for its good performance in classification.

Attention-based neural network has been successfully used especially in machine translation [2] and sentence summarization [22]. In healthcare domain the effectiveness of this mechanism has also been demonstrated by RETAIN [7] and Dipole [19] on Electronic Health Records (EHR). To interpret how attention works, Dipole exemplifies several patient visits and analyses why some visits are more important than others.

Understanding the hidden states behind the observed time series is vital for healthcare applications. In [20], a deep neural network is trained to connect the observed data and the hidden activity. In [12] each hidden state is represented by a Markov random field, and graph analytics such as betweenness centrality is used to achieve reasonable interpretation.

7 CONCLUSION

We design a multivariate time series classification method TATC for predicting AD with actigraphy data. TATC takes a neural deep learning approach with time-aware attention for modeling the effect of circadian rhythm. It achieves promising prediction performance in terms of sensitivity, specificity and AUC. It also generates meaningful interpretation of daily behavior pattern of subjects. TATC shows great potential and practical value in continuous monitoring of physical activity of subjects and in early detection of AD risk. For future work, we plan to explore the possibility of incorporating other measurements for predicting MCI.

ACKNOWLEDGMENTS

The authors would like to thank Xixin Wu and Siyuan Zhang for discussions and suggestions. This work is supported by the Research Grants Council Earmarked Grant CUHK4101/02M, the National Institute of Health R01 Grant AR049439-01A1 and the CUHK Stanley Ho Big Data Decision Analytics Research Centre.

REFERENCES

[1] A. Alberdi, A. Aztiria, and A. Basarab. 2016. On the early diagnosis of Alzheimer's Disease from multimodal signals: A survey. *Artificial Intelligence in Medicine* 71 (2016), 1–29.
[2] D. Bahdanau, K. Cho, and Y. Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).
[3] M. G. Baydogan and G. Runger. 2015. Learning a symbolic representation for multivariate time series classification. *Data Mining and Knowledge Discovery* 29, 2 (2015), 400–422.
[4] D. J. Berndt and J. Clifford. 1994. Using dynamic time warping to find patterns in time series. In *KDD Workshop*. 359–370.
[5] J. A. Cadzow, B. Baseghi, and T. Hsu. 1983. Singular-value decomposition approach to time series modelling. *IEE Proceedings F (Communications, Radar and Signal Processing)* 130, 3 (1983), 202–210.
[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16 (2002), 321–357.
[7] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart. 2016. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. In *NIPS*. 3504–3512.
[8] L. Choi, Z. Liu, C. E. Matthews, and M. S. Buchowski. 2011. Validation of accelerometer wear and nonwear time classification algorithm. *Medicine and Science in Sports and Exercise* 43, 2 (2011), 357–364.
[9] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Deep Learning and Representation Learning Workshop*.

[10] A. N. Coogan, B. Schutová, S. Husung, K. Furczyk, B. T. Baune, P. Kropp, F. Häßler, and J. Thome. 2013. The circadian system in Alzheimer's Disease: Disturbances, mechanisms, and opportunities. *Biological Psychiatry* 74, 5 (2013), 333–339.
[11] H. Deng, G. Runger, E. Tuv, and M. Vladimir. 2013. A time series forest for classification and feature extraction. *Information Sciences* 239 (2013), 142–153.
[12] D. Hallac, S. Vare, S. P. Boyd, and J. Leskovec. 2017. Toeplitz inverse covariance-based clustering of multivariate time series data. In *KDD*. 215–223.
[13] K. He, X. Zhang, S. Ren, and J. Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*. 1026–1034.
[14] L. Huang, Y. Jin, Y. Gao, K.-H. Thung, and D. Shen. 2016. Longitudinal clinical score prediction in Alzheimer's disease with soft-split sparse regression based random forest. *Neurobiology of Aging* 46 (2016), 180–191.
[15] D. Kingma and J. Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014). arXiv:1412.6980
[16] J. Lin, E. Keogh, L. Wei, and S. Lonardi. 2007. Experiencing SAX: a novel symbolic representation of time series. *Data Mining and Knowledge Discovery* 15, 2 (2007), 107–144.
[17] G. Livingston, A. Sommerlad, V. Orgeta, S. G. Costafreda, J. Huntley, D. Ames, C. Ballard, S. Banerjee, A. Burns, J. C. Mansfield, C. Cooper, N. Fox, L. N. Gitlin, R. Howard, H. C. Kales, E. B. Larson, K. Ritchie, K. Rockwood, E. L Sampson, Q. Samus, L. S Schneider, G. Selbæk, L. Teri, and N. Mukadam. 2017. Dementia prevention, intervention, and care. *The Lancet* 390, 10113 (2017), 2673–2734.
[18] B. Logan. 2000. Mel frequency cepstral coefficients for music modeling. In *ISMIR*.
[19] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao. 2017. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *KDD*. 1903–1911.
[20] F. J. Ordóñez and D. Roggen. 2016. Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. *Sensors* 16, 1 (2016), 115.
[21] L. Page, S. Brin, R. Motwani, and T. Winograd. 1999. *The PageRank citation ranking: Bringing order to the web.* Technical Report. Stanford InfoLab.
[22] A. M. Rush, S. Chopra, and J. Weston. 2015. A Neural Attention Model for Abstractive Sentence Summarization. *CoRR* abs/1509.00685 (2015). arXiv:1509.00685
[23] P. Schäfer. 2015. The BOSS is concerned with time series classification in the presence of noise. *Data Mining and Knowledge Discovery* 29, 6 (2015), 1505–1530.
[24] F. Sofi, D. Valecchi, D. Bacci, R. Abbate, G. F. Gensini, A. Casini, and C. Macchi. 2011. Physical activity and risk of cognitive decline: a meta-analysis of prospective studies. *Journal of Internal Medicine* 269, 1 (2011), 107–117.
[25] C. Y. Wee, P. T. Yap, and D. Shen. 2013. Prediction of Alzheimer's disease and mild cognitive impairment using cortical morphological patterns. *Human Brain Mapping* 34, 12 (2013), 3411–3425.
[26] L. Ye and E. Keogh. 2011. Time series shapelets: a novel technique that allows accurate, interpretable and fast classification. *Data Mining and Knowledge Discovery* 22, 1-2 (2011), 149–182.
[27] P. Y. Yeung, L. L. Wong, C. C. Chan, J. L. Leung, and C. Y. Yung. 2014. A validation study of the Hong Kong version of Montreal Cognitive Assessment (HK-MoCA) in Chinese older adults in Hong Kong. *Hong Kong Medical Journal* 20, 6 (2014), 504–510.
[28] J. M. Zeitzer, T. Blackwell, A. R. Hoffman, S. Cummings, S. Ancoli-Israel, and K. Stone. 2018. Daily patterns of accelerometer activity predict changes in sleep, cognition, and mortality in older men. *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences* 73, 5 (2018), 682–687.
[29] J. M. Zeitzer, R. David, L. Friedman, E. Mulin, R. Garcia, J. Wang, J. A. Yesavage, P. H. Robert, and W. Shannon. 2013. Phenotyping apathy in individuals with Alzheimer Disease using functional principal component analysis. *The American Journal of Geriatric Psychiatry* 21, 4 (2013), 391–397.