Heterogeneous Multi-Modal Sensor Fusion with Hybrid Attention for Exercise Recognition

Anjana Wijekoon
School of Computing and Digital Media
Robert Gordon University
Aberdeen, UK
a.wijekoon@rgu.ac.uk

Nirmalie Wiratunga
School of Computing and Digital Media
Robert Gordon University
Aberdeen, UK
n.wiratunga@rgu.ac.uk

Kay Cooper
School of Health Sciences
Robert Gordon University
Aberdeen, UK
k.cooper@rgu.ac.uk

***Abstract*—Exercise adherence is a key component of digital behaviour change interventions for the self-management of musculoskeletal pain. Automated monitoring of exercise adherence requires sensors that can capture patients performing exercises and Machine Learning (ML) algorithms that can recognise exercises. In contrast to ambulatory activities that are recognisable with a wrist accelerometer data; exercises require multiple sensor modalities because of the complexity of movements and the settings involved. Exercise Recognition (ExR) pose many challenges to ML researchers due to the heterogeneity of the sensor modalities (e.g. image/video streams, wearables, pressure mats). We recently published MEx, a benchmark dataset for ExR, to promote the study of new and transferable HAR methods to improve ExR and benchmarked the state-of-the-art ML algorithms on 4 modalities. The results highlighted the need for fusion methods that unite the individual strengths of modalities. In this paper, we explore fusion methods with a focus on attention and propose a novel multi-modal hybrid attention fusion architecture mHAF for ExR. We achieve the best performance of 96.24% (F1-measure) with a modality combination of a pressure mat, a depth camera and an accelerometer on the thigh. mHAF significantly outperforms multiple baselines and the contribution of architecture components are verified with an ablation study. The benefits of attention fusion are clearly demonstrated by visualising attention weights; showing how mHAF learns feature importance and modality combinations suited for different exercise classes. We highlight the importance of improving deployability and minimising obtrusiveness by exploring the best performing 2 and 3 modality combinations.**

***Index Terms*—Attention, Heterogeneous Multi-Modal Fusion, Exercise Recognition**

I. Introduction

Currently, adherence monitoring in digital behaviour change interventions for the self-management of musculoskeletal conditions rely on self-reported exercises which often leads to incorrect or inconsistent reporting. As a result, reasoning algorithms that rely on self-reported exercises for recommending interventions become ineffective causing the users to lose trust. Automated Exercise Recognition (ExR) and performance assessment with sensors have been a research challenge dedicated to addressing this issue.

ExR and Exercise Performance Assessment are the main functional requirements that are needed to automate self-reporting. Activity recognition through sensor data or Human Activity Recognition (HAR) is a well-established area in Artificial Intelligence and Machine Learning (ML) research [1], [2]. ExR is viewed as a sub-domain, which involves the interpretation of sensor data like in HAR but must also consider multi-modal sensors. Nevertheless, the advancements in HAR are yet to be realised for ExR. Looking at recent literature we recognise ExR is mostly attempted with early ML algorithms that use manual feature extraction methods [3]–[5].

The proprietary nature of applications with bespoke sensor setups, data and algorithms are seen as the main barrier to furthering research and development in ExR. We recently published the Multi-modal Sensor Exercise Dataset MEx [1] addressing these challenges. Our previous work [6] presented MEx as a dataset for bench-marking ExR algorithms and HAR algorithms in general, evaluating each modality for the ExR task. Not surprisingly single modality based models were found to be inadequate for ExR, given the complexity of movements in exercises. Accordingly in this paper we explore Fusion [7] and Attention [8] algorithms to reason with heterogeneous multi-modal sensors for ExR.

We are keen on advancing research in ExR by introducing novel algorithms while highlighting the need for developing ExR systems that are unobtrusive and easy to deploy in the real-world. In addition, the limitations of obtaining training data for ExR calls for algorithms that use minimal sensor configurations and only require a comparatively smaller amount of training data (i.e. less train-able parameters). To this end, we propose a novel hybrid attention fusion architecture mHAF for ExR.

We summarise the key contributions made in this paper below:

- We propose a novel hybrid attention fusion architecture mHAF for heterogeneous multi-modal sensor fusion in the ExR task, benched-marked with the MEx Dataset;
- We verify the contribution of architecture components with an ablation study and visualise the capacity of the architecture to learn feature importance and modality combinations that are optimal for different exercise classes; and,

This work is part funded by SELFBACK which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 689043.

[1] [6], Publicly available at https://archive.ics.uci.edu/ml/datasets/MEx

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

- We identify the best 2, 3 and 4 modality combinations to cater to user preferences and unobtrusive deployment in a digital behaviour change intervention.

Rest of the paper is organised as follows. In Section II we present the recent research in ExR. In Section III we formalise the problem of ExR with heterogeneous multi-modal sensors and in Section IV we present our novel hybrid attention fusion architecture mHAF. Section V presents the evaluation methodology and the results. Our conclusions appear in Section VI with plans for future work.

II. Related Work

Research in Exercise Recognition (ExR) spans a number of application areas such as callisthenics, weight exercises, yoga and sports. Inertial Measurement Units (IMUs) are the most widely used data stream in literature [9]–[12], but some have explored sensors such as Pressure mats [4], [13], Channel State Information [3] and Electrocardiograms [5]. ExR is often viewed as the classification of many discrete labels given sensor data streams. Often these recognition algorithms use a manual feature extraction pipeline followed by a classification algorithm such as k-NN [3], [13], Random Forest [9], Decision Trees [4] or HMM [5].

While Deep Learning methods (CNN and LSTM) are the state-of-the-art in Human Activity Recognition (HAR) [14], [15], literature suggests that they are rarely considered with ExR [6], [10]. For instance, authors of [10] use a recurrent architecture to recognise shoulder rehabilitation exercises with wrist-worn IMU data streams and achieve 88.9% accuracy; their dataset is not publicly available and their methods cannot be transferred to other exercise domains due to lack of sensors that capture movements from other body parts except the wrist. In contrast, the public available dataset MEx [6] recently reported single sensor performance on physiotherapy exercise recognition using four sensors; two accelerometers placed on the wrist and the thigh, a pressure mat and a depth camera achieving F1-measure 63.35%, 90.15%, 74.08% and 87.20% respectively. These results highlight that a single sensor is inadequate to recognise a wide range of exercises with high precision. Accordingly, we recognise the need for multi-modal learning strategies to achieve higher performance in recognition tasks. This calls for fusion architectures and methods such as attention [8], to combine heterogeneous multi-modal sensor data.

Fusion algorithms are explored at different feature representation levels, mainly Early, Mid and Late levels. Early-fusion is the concatenation of raw features and learning a shared feature representation for all modalities. [15]. With mid-fusion and late-fusion, each modality learns a feature representation individually and are later concatenated; mid-fusion additionally learns a shared feature representation. While homogeneous sensor modalities find shared feature representations learned in early and mid-fusion advantages, we argue that it can be detrimental for the fusion of heterogeneous sensor modalities.

Fig. 1. 7 exercises in the MEx dataset

III. Problem Definition

We formalise the problem of heterogeneous multi-modal sensor fusion for Exercise Recognition (ExR) with the MEx dataset here. The sensor-rich dataset Mex [6] contains 7 exercises that have been selected by a physiotherapist for the self-management of chronic musculoskeletal pain (Figure 1). MEx contains data recorded with 4 sensor modalities, from 30 participants, and each participant performed all 7 exercises, each one for the duration of 60 seconds (maximum). It is noteworthy that this dataset is smaller in comparison to other deep learning datasets in computer vision, text or HAR. Table I summaries the data type of each sensor modality.

TABLE I
Sensor Modalities

Modality	Frequency	Raw features(m)
Depth Camara(DC)	$15Hz$	240×320
Thigh Accelerometer(ACT)	$100Hz$	(x, y, z)
Wrist Accelerometer(ACW)	$100Hz$	(x, y, z)
Pressure Mat(PM)	$15Hz$	16×32

We assume, there exists a modality combination S of k modalities and a deep neural network θ that is optimised to recognise exercises in a multi-modal setting. Let M represent the time-series data from a sensor modality where each time-stamp is associated with the set of raw features m.

$$S = \{M_1, M_2, ..., M_k\} \tag{1}$$

$$M_i = [m_1, m_2, ..., m_t, ...] \tag{2}$$

The classifier θ predicts the exercise class when the input is the set of modalities S within the window of n timestamps.

$$y = \theta(S_{[t,t+n]}) \tag{3}$$

Selecting the θ architecture involve three main design aspects: firstly, how to represent individual sensor modalities; secondly, when to aggregate multiple modalities and create shared representations and; lastly, how to attend to features of such representations using an attention mechanism. Furthermore, as with any Deep Learning architecture, the amount of train-able parameters in θ is constrained by the amount of training data available to avoid under-fitting. In the next

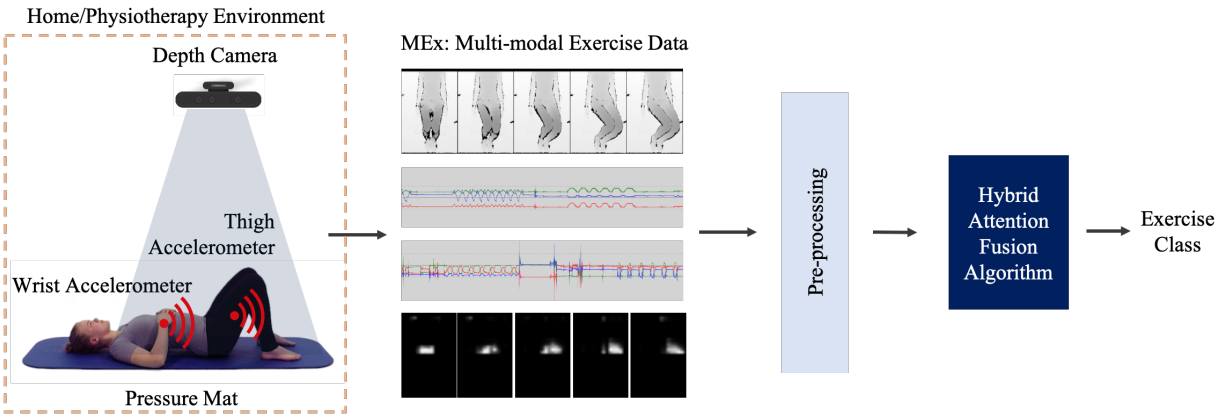

Fig. 2. Exercise recognition with heterogeneous multi-modal sensors

section, we detail our approach to achieving heterogeneous multi-modal sensor fusion for ExR while considering above design aspects and constrains.

IV. Methods

The proposed Exercise Recognition (ExR) system is shown in Figure 2. There are two sensor modalities installed in the environment, a pressure mat (PM) and a depth camera (DC); and two further on-body sensor modalities, accelerometer sensors located on the wrist (ACW) and the thigh (ACT). All sensor modalities continuously emit data streams that are synchronised using timestamps. The objective of the Multi-modal Hybrid Attention Fusion (mHAF) architecture is to predict the exercise class using a modality fusion strategy applied to the pre-processed data streams.

A. Data Pre-processing

We apply the sliding window method with 5 second time window and 2 second stride (3 seconds of overlap) to all modality data streams to create data instances. This allows the system to make a prediction (i.e. exercise class) every 2 seconds in real-time. The following modality specific pre-processing decisions were made in-line with previous studies in the literature.

- **DC:** Reduce the frame rate to 1 frame/second, and reduce the frame size to 12×16.
- **ACW and ACT:** Segment the data stream to 1 second windows(100 timestamps) and apply Discrete Cosine Transformation (DCT) on each axis, Select the most significant 60 DCT coefficients and concatenate all axes to obtain the feature representation of length 180.
- **PM:** Reduce the frame rate to 1 frame/second, and reduce the frame size to 16×16.

Specifically, the hyper-parameters values for window, overlap, frame rate and frame size are adapted from the comparative study published in [6]. The DCT feature transformation is adapted from literature where it has been shown to significantly outperform when using raw features and other feature transformation methods [6], [16].

B. mHAF Architecture

The mHAF architecture in Figure 3 has three main components; Modality specific feature representations, Hybrid Attention Fusion and Classification. Firstly, the individual modalities learn independent feature representations with the most optimal feature representation method identified for each respective modality. Next, the Hybrid Attention Fusion (HAF) module learns a shared feature representation by exploiting two attention approaches, Hard Attention and Soft Attention. Lastly a softmax layer predicts the exercise class label. mHAF architecture is trained end to end using the cross-entropy loss which is minimised towards correctly predicting exercise classes. Importantly, we want to minimise the number of trainable parameters in the architecture to encourage convergence during model training given a relatively small set of training data.

1) Modality Specific Feature Representations: The heterogeneity of sensor modalities calls for feature representations that are modality specific instead of modality agnostic [6]. Accordingly, we refer to the benchmark performances published by the authors of [6] where multiple feature representations and classifiers are compared for ExR with individual modalities of the MEx dataset. The best performing architecture with each modality from that study is listed in Table II. Accordingly, in this paper, we will take forward the 1D-CNN-LSTM model for DC, ACT and ACW modalities and the 2D-CNN model for the PM modality. Architecture details of the two models are presented in Table III.

TABLE II
BEST PERFORMING ARCHITECTURES FOR MODALITIES IN MEx

Sensor	F1-measure(%)	Architecture
DC	87.20	2D-CNN
ACT	90.15	1D-CNN-LSTM
ACW	63.35	1D-CNN-LSTM
PM	74.08	1D-CNN-LSTM

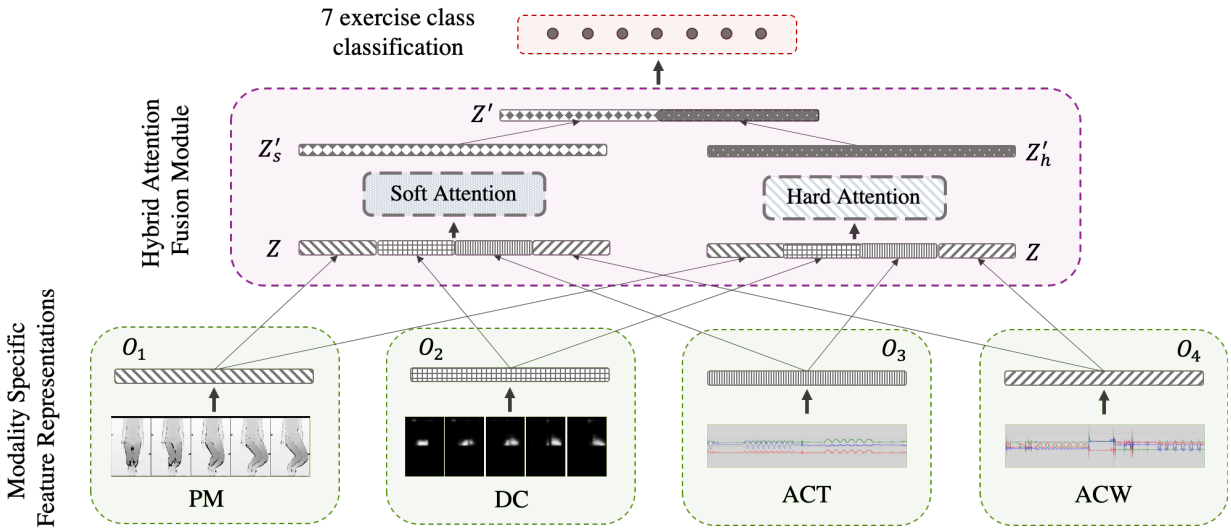

Fig. 3. mHAF architecture for exercise recognition

TABLE III
ARCHITECTURAL DETAILS FOR 2D-CNN AND 1D-CNN-LSTM, td:TIMEDISTRIBUTEDLAYER, $conv(k)n$:CONVOLUTIONALLAYER WITH n KERNELS OF KERNEL SIZE k, $maxpool(k)$:MAXPOOLINGLAYER WITH POOL SIZE k, $dense(k)$:DENSELAYER WITH k UNITS, bn:BATCHNORMALISATION

Model	Architecture
2D-CNN	$conv(3\times3)32 \rightarrow maxpool(2\times2) \rightarrow bn \rightarrow conv(3\times3)64 \rightarrow maxpool(2\times2) \rightarrow bn \rightarrow flatten$ $\rightarrow dense(1200) \rightarrow bn \rightarrow dense(600) \rightarrow bn \rightarrow dense(100) \rightarrow bn$
1D-CNN-LSTM	$td-conv(5)32 \rightarrow maxpool(2) \rightarrow bn \rightarrow td-conv(5)64 \rightarrow maxpool(2) \rightarrow bn \rightarrow td-flatten$ $\rightarrow lstm(1200) \rightarrow bn \rightarrow dense(600) \rightarrow bn \rightarrow dense(100) \rightarrow bn$

2) Fusion: We adapt a late-fusion methodology where we first learn the modality specific feature representations $(\theta_1, \theta_2, ..., \theta_k)$ for individual modalities. Then the output feature vectors o_i (each of size 100 as in Table III) are concatenated to form the fusion feature representation z, where x_i is the pre-processed modality input as discussed in Section IV-A. It is also noteworthy that late-fusion does not introduce any additional parametric components when compared to mid-fusion.

$$\begin{aligned} o_i &= \theta_i(x_i) \\ z &= concat(o_1, o_2, ..., o_k) \end{aligned} \tag{4}$$

3) Hybrid Attention Fusion Module: Attention [8] learns the significance of features for classification during model training. This is particularly beneficial for achieving comparable performance with a shallow parametric model trained on a smaller dataset, that otherwise may only have been achieved using a very deep architecture trained on a larger training dataset for a longer time. In general an attention module learns attention weights w_a in relation to a feature vector z, where each attention weight w_a^i indicates the importance of it's respective feature z^i. w_a is learnt using a parametric model θ_a such that $|w_a| = |z|$ and the weights are normalised using the function $norm$. The output of the attention module is the weighted feature representation of z, z' (Equation 5).

$$\begin{aligned} z &= concat(o_1, o_2, ..., o_k) \\ score &= \theta_a(z) \\ w_a &= norm(score) \\ z' &= w_a \times z \end{aligned} \tag{5}$$

The goals for introducing an attention module for ExR is two-fold: firstly, to boost multiple features from modalities that together contribute towards classification accuracy; and secondly, to acutely discriminate between features from modalities that are noisy from others that are important for the classification task.

To achieve the former, we introduce a Soft Attention (SA) module. SA uses the sigmoid function as the $norm$ where w_{as} are less skewed, resulting in a normally weighted feature representation (Equation 6).

$$\begin{aligned} score_s &= \theta_{as}(z) \\ w_{as}^i &= \frac{1}{1+exp(-score_s^i)} \\ z'_s &= w_{as} \times z \end{aligned} \tag{6}$$

To achieve the latter, we introduce a Hard Attention (HA) module. HA uses the softmax function as the $norm$ where the w_{ah} are skewed to attend to only one or few features in z (Equation 7).

$$score_h = \theta_{ah}(z)$$
$$w_{ah}^i = \frac{exp(score_h^i)}{\sum_j exp(score_h^j))} \quad (7)$$
$$z'_h = w_{ah} \times z$$

The HAF module is the concatenation of the two attention modules, HA and SA (Equation 8), where z'_h and z'_s are the outputs from the HA and the SA modules respectively.

$$z' = concat(z'_h, z'_s) \quad (8)$$

We implement the attention at feature granularity level, where each feature is assigned a weight regardless of the modality. This is in contrast to modality level where the attention weights highlight all the features from one modality and discard all features from other modalities. The reasoning behind selecting feature level granularity is two-fold; firstly all features from a single modality are not equally important for ExR; and secondly, more than one sensor modality will contribute features towards improved classification.

A single dense layer parametric model with a non-linear activation is used as the θ_a (Equation 9) in the implementation of both HA and SA modules.

$$score = tanh(w^T z + b) \quad (9)$$

where, w and b are train-able parameters and $w \in \mathbb{R}^{|z|\times|z|}$ and $b \in \mathbb{R}^{1\times|z|}$ such that $|score| = |z|$. We select these hyper-parameters after an exploratory study where we compared models with increasing number of train-able parameters (i.e. hidden layers and nodes). θ_{as} for SA and θ_{ah} for HA are the only parametric component in the mHAF architecture apart from the modality specific feature representations.

4) Classification: The concatenated output of the HAF module z' is further connected to a dense layer with softmax activation for exercise class prediction.

$$y = \underset{e\in E}{\operatorname{argmax}} \left(softmax(w_c^T z' + b_c)\right) \quad (10)$$

where, E is the set of exercise classes, $w_c \in \mathbb{R}^{|z'|\times|E|}$ and $b_c \in \mathbb{R}^{1\times|E|}$.

V. Performance Evaluation

Our evaluation of the Exercise Recognition system is three-fold;

- evaluate mHAF with heterogeneous multi-modal sensors against multiple baselines and variants of mHAF derived through model ablation;
- compare different sensor modality combinations to identify the best performing minimal combinations; and,
- evaluate HAF module's capabilities to learn feature importance and modality combinations for different exercise classes by visualising attention weights.

A. Evaluation Methodology

We evaluate mHAF algorithm using Leave-One-Person-Out (LOPO) methodology. With the MEx dataset, LOPO creates 30 folds; each fold is trained with 29 user data and tested with the data from one user. This methodology emulates a real-life deployment setting where end-user data is not available during training. For a given experiment, on average, the training set has 6032 instances and the test set has 208 instances. The experiments are implemented using TensorFlow and Keras libraries. The models are optimised using following parameters; loss is cross-entropy, optimiser is Adadelta, learning rate is 1.0, batch size is 32 and the number of epochs is 30. We use batch normalisation as the regularisation method of the models.

Macro F1-measure is the selected performance measure for the experiments as it provides a better representation of precision and recall compared to accuracy. F1-measure is first calculated for each label i, and their non-weighted mean is calculated (Equation 11). For a given experiment, we present the mean F1-measure averaged over 30 LOPO folds. Weighted F1-measure is not required in these experiment as the dataset is class balanced. LOPO evaluation methodology calls for non-parametric statistical significance test due to not-normally distributed results. We use the Wilcoxon signed-rank test for paired samples to evaluate the statistical significance at 95% confidence.

$$F_1 = \sum_i 2 \times \frac{precision_i \times recall_i}{precision_i + recall_i} \quad (11)$$

B. Comparative Study

We compare mHAF with following baselines.

1) **Early Fusion:** Raw features from each sensor modality is flattened, concatenated and a shared feature representation is learned.
2) **Mid Fusion:** Modality specific feature representations are followed by a shared feature representation.
3) **mHAF-midHAF:** mHAF + a shared feature representation (2 Dense Layers) between HAF module and the classification layer.
4) **mHAF-noHAF:** mHAF without the HAF module. Modality specific feature representations are concatenated for classification.
5) **mHAF-noSA:** mHAF without the SA module in HAF, only the HA module
6) **mHAF-noHA:** mHAF without the HA module in HAF, only the SA module

The first three baselines evaluate the significance of Modality specific feature representations and the last three baselines evaluate the significance of attention modules in the mHAF architecture (see Figure 3). Therefore this evaluation also act as an ablation study. We re-purpose the models in Table III as the modality specific and shared feature representations for the baselines 1, 2 and 3 by adjusting the layer sizes accordingly.

We do not find appropriate baselines in literature given the heterogeneity of sensor modalities.

TABLE IV
COMPARATIVE EVALUATION OF MHAF

Algorithm	F1-measure(%)
Early Fusion	89.92
Mid Fusion	93.59
mHAF-midHAF	93.89
mHAF-noHAF	94.65
mHAF-noSA	95.25
mHAF-noHA	94.25
mHAF	**95.84**

Table IV presents the results of the comparative study. mHAF outperforms the three baselines with shared feature representations with statistical significance. This confirms our argument towards using late fusion over early or mid fusion where multiple modalities learn a shared feature representation that is degrading for heterogeneous modalities. In addition, mHAF also outperforms its variants, mHAF-noHAF, mHAF-noHA and mHAF-noSA verifying the significance of the HAF module as a whole and the significance of individual modules HA and SA.

C. Comparison of Sensor Modality Combinations

The objective of this evaluation is to identify the best performing minimal modality combinations to cater for user preferences. Using all four modalities can be obtrusive, economically infeasible and discouraging to some users. Accordingly, we aim to identify the best 2 modality and 3 modality combinations in addition to the 4 modality combination suitable for deployment. We create 10 datasets and their respective mHAF architectures by combining 2 modalities (6 combinations) and 3 modalities (4 combinations). We follow the same evaluation methodology and present the F1-measure for each mHAF architecture.

Table V presents the results for the comparison of modality combinations. The 3 modality combination ACT, PM and DC (highlighted in bold text) outperform the 4 modality combination with statistical significance. We refer to Table II which indicates that removing the least performing modality helps to enhance model performance. Closer examination suggests that sensors that are more prone to noise such as the on-body wrist accelerometer, is detrimental for fusion.

We further compare the performance between different modality combinations at exercise class level by visualising the confusion matrices in Figure 4. The confusion matrices are averaged over all 30 folds and normalised. Here we observe the significant difference between the 4 modality mHAF model the and 3 modality mHAF model can be explained by examining the performance of exercise class 6. For instance, exercise class 6 and 7 have the exact same hand movements but differ in their thigh movements, thus including ACW in the 4 modality setting has adversely contributed to overall fusion performance. In contrast, including the DC data in the 3 modality mHAF model has significantly improved recognition of exercises classes 2, 3 and 6 in comparison to the 2 modality mHAF model.

It can be argued that these modality selections can be learnt by a deep architecture. Such deep architecture needs to be "very deep" and have many train-able parameters, which would typically require a large training dataset. As stated in Section I, our goal in this exploratory study is to mitigate this demand on data, and instead maintain the minimalism of mHAF with a manageable collection of data.

We further extend the comparative study from Section V-B for the two best 2 and 3 modality combinations to evaluate mHAF architecture on different modality combinations. The results are presented in Table VI. The results further validate the mHAF architecture for the task of ExR with heterogeneous multi-modal sensors by significantly outperforming all baselines.

D. Learning Feature Importance with Attention

The aim of this evaluation was to observe the capability of the HAF module to learn the feature importance and modality combinations for effective fusion. To this end, we visualise the attention weights learned by the HA and SA modules in the best performing 3 modality mHAF architecture (best overall) in Figure 5. Each row is an exercise class and the weights are grouped, sorted and stacked per modality for clear visualisation. HA weights are on the left and SA weights are on the right in which a darker colour indicates a higher weight.

It is evident that HA has learned skewed weights to highlight only a few features, whilst SA has learned a more normally distributed set of weights. While some exercises classes learn a combination of all sensor modalities for recognition, there exist others that learn to choose a subset of sensor modalities. For instance, exercise 6 has no thigh movement and DC does not capture the hand movements (see Fig 1). Accordingly. exercise 6 relies mostly on PM for recognition. Similarly, exercise 7 includes thigh movement and is captured by the DC which increased the significance of the ACT and DC features compared to exercise 6. Exercise 3, is the Pelvic tilt which is hardly recognisable from the DC and attention weights have learnt that DC should contribute minimally towards the recognition task in that situation. In summary, these observations demonstrate the need for maintaining alternative forms of attention within the fusion architecture.

We also visualised two randomly sampled test instances and observe the weights of the features learned by the two modules HA and SA in Figure 6. These visualisations are obtained with a fixed ordering of the feature set, in order to visualise feature area of comparable weight assignments by both SA and HA (as indicated with a dotted box). Here we can observe that SA and HA consistently learning to attribute the highest importance to the same features for the recognition task. Furthermore, there are additional features identified by SA, thus providing more opportunity for highly accurate modalities like ACT to contribute to the recognition task. Overall, the compatibility

TABLE V
COMPARISON OF MODALITY COMBINATIONS

	ACT and PM	ACT and DC	ACW and PM	ACW and DC	PM and DC	ACT and ACW
F1-measure(%)	*93.54*	92.76	84.66	91.25	90.41	86.88
	ACT, ACW and PM	ACT, ACW and DC	ACW, DC and PM	**ACT, DC and PM**	*ACT, ACW DC and PM*	
F1-measure(%)	94.85	94.21	91.47	**96.24**	*95.84*	

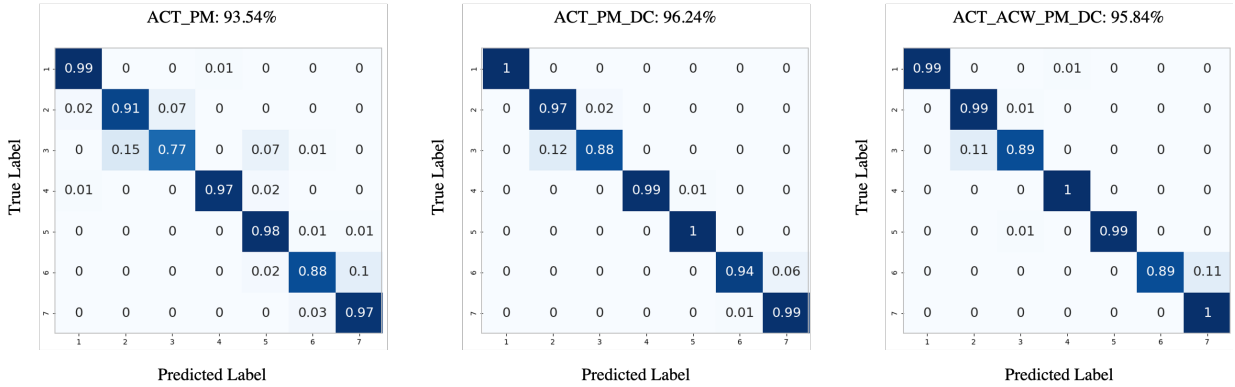

Fig. 4. Confusion matrices for mHAF with the best 2, 3 and 4 modality combinations

TABLE VI
COMPARATIVE EVALUATIONS OF MHAF WITH THE BEST 2 AND 3 MODALITY COMBINATIONS

Algorithm	F1-measure(%)	
	ACT and PM	ACT, PM and DC
Early Fusion	85.33	92.24
Mid Fusion	89.53	92.63
mHAF-midHAF	88.27	93.65
mHAF-noHAF	90.57	93.78
mHAF-noSA	90.75	94.95
mHAF-noHA	90.36	93.45
mHAF	**93.54**	**96.24**

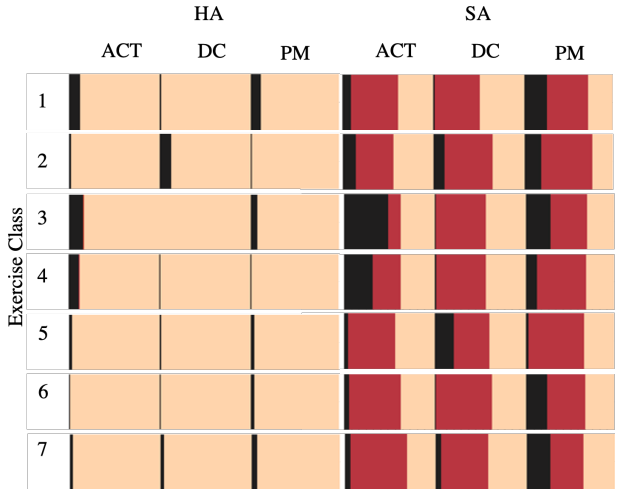

Fig. 5. Visualisation of attention weights

between the interpretation of attention weights and the domain knowledge verify that the mHAF architecture is learning to intelligently reason with heterogeneous multi-modal sensors for the ExR task.

VI. CONCLUSION

Exercise Recognition (ExR) is an essential component of automating digital interventions to effectively provide support and guidance. Towards achieving automated ExR we presented a novel hybrid attention fusion algorithm mHAF that performs ExR by integrating heterogeneous multi-modal sensors. Our algorithm significantly outperforms several baselines and effectively learn to attend to features and learn modality combinations suited to recognise different exercises. In addition, an exploratory study discovered 2, 3 and 4 modality combinations suited for deployment, satisfying user preferences and minimising obtrusiveness. In future, with the MEx dataset and fusion algorithms, we will further explore performance assessment of exercises to assist patients with adherence and provide guidance towards correct exercise execution. This work is contributing towards the implementation of an automated exercise recognition and quality assessment digital behaviour change intervention for self-management of musculoskeletal conditions.

REFERENCES

[1] H. F. Nweke, Y. W. Teh, M. A. Al-Garadi, and U. R. Alo, "Deep learning algorithms for human activity recognition using mobile and

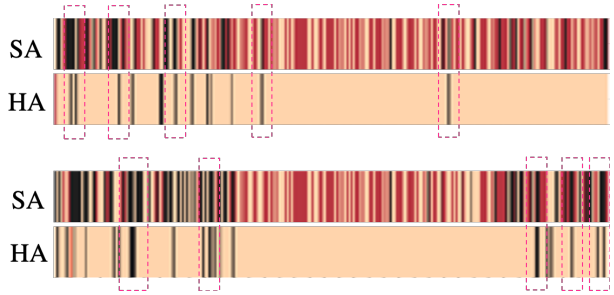

Fig. 6. Two examples of HA and SA learning to attend to features

wearable sensor networks: State of the art and research challenges," *Expert Systems with Applications*, vol. 105, pp. 233–261, 2018.

[2] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognition Letters*, vol. 119, pp. 3–11, 2019.

[3] F. Xiao, J. Chen, X. H. Xie, L. Gui, J. L. Sun, and W. none Ruchuan, "Seare: A system for exercise activity recognition and quality evaluation based on green sensing," *IEEE Transactions on Emerging Topics in Computing*, 2018.

[4] B. Zhou, M. Sundholm, J. Cheng, H. Cruz, and P. Lukowicz, "Never skip leg day: A novel wearable approach to monitoring gym leg exercises," in *Pervasive Computing and Communications (PerCom), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1–9.

[5] J. Qi, P. Yang, M. Hanneghan, A. Waraich, and S. Tang, "A hybrid hierarchical framework for free weight exercise recognition and intensity measurement with accelerometer and ecg data fusion," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 3800–3804.

[6] A. Wijekoon, N. Wiratunga, and K. Cooper, "Mex: Multi-modal exercises dataset for human activity recognition," *arXiv preprint arXiv:1908.08992*, 2019.

[7] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4694–4702.

[8] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[9] E. Velloso, A. Bulling, H. Gellersen, W. Ugulino, and H. Fuks, "Qualitative activity recognition of weight lifting exercises," in *Proceedings of the 4th Augmented Human International Conference*. ACM, 2013, pp. 116–123.

[10] D. M. Burns, N. Leung, M. Hardisty, C. M. Whyne, P. Henry, and S. McLachlin, "Shoulder physiotherapy exercise recognition: machine learning the inertial signals from a smartwatch," *Physiological measurement*, vol. 39, no. 7, p. 075007, 2018.

[11] M. Guo, Z. Wang, and N. Yang, "Aerobic exercise recognition through sparse representation over learned dictionary by using wearable inertial sensors," *Journal of Medical and Biological Engineering*, vol. 38, no. 4, pp. 544–555, 2018.

[12] L. N. N. Nguyen, D. Rodríguez-Martín, A. Català, C. Pérez-López, A. Samà, and A. Cavallaro, "Basketball activity recognition using wearable inertial measurement units," in *Proceedings of the XVI international conference on Human Computer Interaction*. ACM, 2015, p. 60.

[13] M. Sundholm, J. Cheng, B. Zhou, A. Sethi, and P. Lukowicz, "Smart-mat: Recognizing and counting gym exercises with low-cost resistive pressure sensing matrix," in *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*. ACM, 2014, pp. 373–382.

[14] S. Yao, S. Hu, Y. Zhao, A. Zhang, and T. Abdelzaher, "Deepsense: A unified deep learning framework for time-series mobile sensing data processing," in *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017, pp. 351–360.

[15] F. J. Ordóñez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.

[16] S. Sani, S. Massie, N. Wiratunga, and K. Cooper, "Learning deep and shallow features for human activity recognition," in *International Conference on Knowledge Science, Engineering and Management*. Springer, 2017, pp. 469–482.