An Interpretable Machine Vision Approach to Human Activity Recognition using Photoplethysmograph Sensor Data

Eoin Brophy[0000−0002−6486−5746], José Juan Dominguez Veiga[0000−0002−6634−9606], Zhengwei Wang[0000−0001−7706−553X], Alan F. Smeaton[0000−0003−1028−8389], and Tomás E. Ward[0000−0002−6173−6607]

Insight Centre for Data Analytics,
Dublin City University, Glasnevin, Dublin, Ireland

Abstract. The current gold standard for human activity recognition (HAR) is based on the use of cameras. However, the poor scalability of camera systems renders them impractical in pursuit of the goal of wider adoption of HAR in mobile computing contexts. Consequently, researchers instead rely on wearable sensors and in particular inertial sensors. A particularly prevalent wearable is the smart watch which due to its integrated inertial and optical sensing capabilities holds great potential for realising better HAR in a non-obtrusive way. This paper seeks to simplify the wearable approach to HAR through determining if the wrist-mounted optical sensor alone typically found in a smartwatch or similar device can be used as a useful source of data for activity recognition. The approach has the potential to eliminate the need for the inertial sensing element which would in turn reduce the cost of and complexity of smartwatches and fitness trackers. This could potentially commoditise the hardware requirements for HAR while retaining the functionality of both heart rate monitoring and activity capture all from a single optical sensor. Our approach relies on the adoption of machine vision for activity recognition based on suitably scaled plots of the optical signals. We take this approach so as to produce classifications that are easily explainable and interpretable by non-technical users. More specifically, images of photoplethysmography signal time series are used to retrain the penultimate layer of a convolutional neural network which has initially been trained on the ImageNet database. We then use the 2048 dimensional features from the penultimate layer as input to a support vector machine. Results from the experiment yielded an average classification accuracy of 92.3%. This result outperforms that of an optical and inertial sensor combined (78%) and illustrates the capability of HAR systems using standalone optical sensing elements which also allows for both HAR and heart rate monitoring. Finally, we demonstrate through the use of tools from research in explainable AI how this machine vision approach lends itself to more interpretable machine learning output.

Keywords: deep learning · activity recognition · explainable artificial intelligence.

1 Introduction

Due to the ubiquitous nature of inertial and physiological sensors in phones and fitness trackers, human activity recognition (HAR) studies have become more widespread [13]. The benefits of HAR include rehabilitation for recovering patients, activity monitoring of the elderly and vulnerable people, and advancements in human-centric applications [19].

Photoplethysmography (PPG) is an optical technique used to measure volume changes of blood in the microvascular tissue. PPG is capable of measuring heart rate by detecting the amount of light reflected/absorbed in red blood cells as this varies with the cardiac cycle. Reflected light is read by an ambient light sensor which then has its output conditioned, so a pulse rate can be determined. The pulse rate is obtained from analysis of the small alternating component (which arises from the pulsatile nature of blood flow) superimposed on the larger base signal caused by the constant absorption of light.

For usability reasons the wrist is a common site for wearables used in health and fitness contexts [15]. Most smartwatches are equipped with a PPG sensor capable of measuring pulse rate. Difficulties arising in obtaining a robust physiologically useful signal via PPG can be caused by motion artefacts due to the changes in optical path length associated with disturbance of the source-detector configuration. This disturbance is introduced by haemodynamic effects and gross motor movements of the limbs [20]. This can lead to an incorrect estimation of the pulse rate. Reduction in motion artefacts can be achieved using a range of techniques, for example, through the choice of physical attachment technique or through adaptive signal filtering methods based on an estimation of the artefact source from accelerometer-derived signals [2].

In this study we sought to exploit the motion artefact to infer human activity type from the PPG signals collected at the wrist. Our hypothesis was that there is sufficient information in the disturbance induced in the source-detector path to distinguish different activities through the use of a machine learning approach. In recent years, capabilities of machine learning methods in the field of image recognition has increased dramatically [12]. Building on these advancements in image recognition would allow for simplification of wearables involved in HAR. We chose an image-based approach to the machine learning challenge as this work is part of a larger scope effort to develop easily deployed AI (artificial intelligence) which can be used and interpreted by end users who do not have signal processing expertise.

This paper extends previous work completed in [4]. The additional contributions of this paper are improved classification through the use of a hybrid classifier approach and more significantly to demonstrate through the use of tools from research in explainable artificial intelligence (XAI) how this machine vision approach lends itself to more interpretable machine learning output.

2 Related Work

Convolutional neural networks (CNNs) have been used since the 1990s and were designed to mimic how the visual cortex of the human brain processes and recognises images [18]. CNNs extract salient features from images at various layers of the network. They allow implementation of high-accuracy image classifiers given the correct training without the need for in-depth feature extraction knowledge.

The current state of the art in activity recognition is based on the use of cameras. Cameras allow direct and easy capture of motion but this output requires significant processing for recognition of specific activities. Inertial sensing is another popular method used in HAR. To achieve the high accuracy of the inertial sensing systems shown in [15], a system consisting of multiple sensors is required, compromising functionality and scalability. The associated signal processing is not trivial and singular value decomposition (SVD), truncated Karhunen-Loève transform (KLT), Random Forest (RF) and Support Vector Machines (SVM) are examples of feature extraction and machine learning methods that have been applied to HAR.

Inertial sensor data paired with PPG are amongst the most suitable sensors for activity monitoring as they offer effective tracking of movement actions as well as relevant physiological parameters such as heart rate. They also have the benefit of being easy to deploy. Mehrang et al. used RF and SVM machine learning methods for a HAR classifier on combined accelerometer and PPG data achieving an average recognition accuracy of 89.2% and 85.6% respectively [16].

The average classification accuracy of the leading modern feature extraction and machine learning methods for singular or multiple inertial sensors range from 80% to 99% [15]. However, this can require up to 5 inertial measurement units located at various positions on the body.

Even with the success of deep learning, similar criticisms that plagued past work on neural networks, mainly in terms of being uninterpretable, are recurring. There is no unanimous definition of interpretability, but works like Doshi-Velez and Kim [7] try to define the term and provide a way of measuring how interpretable a system is. Issues such as algorithmic fairness, accountability and transparency are a source of current research. An example for the need of such work is the European Unions right to explanation [9] in which users can demand to be explained how an algorithmic decision that affects them has been made. The ramifications of these implicitly present issues with governance, policy and privacy demonstrate a need for regulation of explainability in the near future. In this paper the authors use the term interpretability in terms of transparency and accountability.

3 Methodology

3.1 Data Collection

We use a dataset collected by Deleram Jarchi and Alexander J. Casson and is freely available from Physionet [11]. This dataset comprises PPG recordings taken from 8 patients (5 female, 3 male) aged between 22-32 (mean age of 26.5), during controlled exercises on a treadmill and an exercise bike. Data was recorded using a wrist-worn PPG sensor attached to the Shimmer 3 GSR+ unit for an average period of 4-6 minutes with a maximum duration of 10 minutes. A frequency of 256 Hz was used to sample the physiological signal. Each individual was allowed to set the intensity of their exercises as they saw fit and every exercise began from rest. The four exercises were broken down into walk on a treadmill, run on a treadmill, low and high resistance exercise bike. For the walk and run exercises the raw PPG signals required no filtering other than what the Shimmer unit provides. The cycling recordings were low-pass filtered using Matlab with a 15 Hz cut-off frequency to remove any high-frequency noise.

3.2 Data Preparation

The PPG data signal was downloaded using the PhysioBank ATM and plotted in Python. Signals were segmented into smaller time series windows of 8-second intervals chosen to match the time windows used in [3], which acts as a benchmark for this study. A rectangular windowing function was used to step through the data every 2 seconds, also conducted in [3]. It is worth re-emphasising that a machine vision approach is being taken here, the input data to the classifier is not time series vectors but images. These images correspond to simple plots of the 8-second window produced in Python which plots and saves figures and removes all axis labels, legends and grid ticks (removing non-salient features), saving each figure as a 299x299 JPEG file. A total of 3321 images were created, of which 80% (2657) were used for retraining, 10% (332) for validation and 10% (332) for testing.

The image files were stored in a directory hierarchy based on the movement carried out. Four sub-directories of possible classifiers were created; run, walk, high resistance bike and low resistance bike and contained within each were the images of the plotted PPG signal. In Fig 1 an example of a plot for each activity can be seen.

3.3 The Network Infrastructure

On completion of data preparation, the CNN could then be retrained. Building a neural network from the ground up is not a trivial task, requiring a multilayer implementation for even a simple perceptron [18] which needs optimization of tens of thousands of parameters for even a trivial task such as handwritten digit classification [21]. Also very large amounts of data is required for training. Here we avoided both problems through use of Inception-v3 which can be implemented with the TensorFlow framework and transfer learning.

TensorFlow [1], a deep learning framework, was used for transfer learning which is the concept of using a pretrained CNN and retraining the penultimate layer that does classification before the output. This type of learning is ideal for this study due to our relatively small dataset [22]. The results of the retraining process can be viewed using the suite of visualisation tools on TensorBoard.

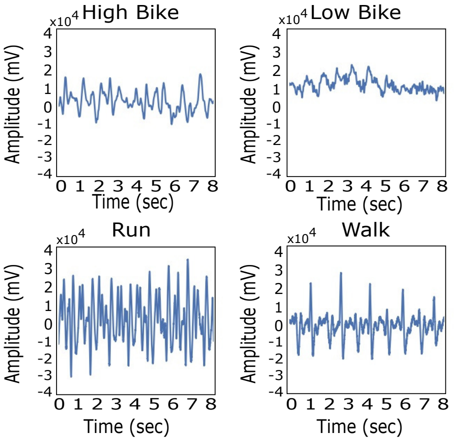

Fig. 1. Sample of PPG image for each activity

Recognition of specific objects from millions of images requires a model with a large learning capacity. CNNs are particularly suitable for image classification because convolution leverages three important properties that can help improve a machine learning system: sparse interaction, parameter sharing and equivariant representations [8]. These properties enable CNNs to detect small, meaningful features from input images and reduce the storage requirements of the model compared to traditional densely connected neural networks. CNNs are tuneable in their depth and breadth meaning they can have much simpler architecture, making them easier to train compared to other feedforward neural networks [12].

3.4 Retraining and Using the Network

A pretrained CNN, Google's Inception-v3, was used which has been trained on ImageNet, a database of over 14 million images and will soon be trained on up to 50 million images [16]. To retrain a network, we use the approach taken by Dominguez Veiga et. al [6]. The experiment was completed again using the features from the retrained Inception-v3 penultimate layer. The 2048 dimensional features were extracted from this layer and used as input into a radial basis function SVM classifier as in [16], a nested cross-validation strategy was applied here to prevent biasing the model evaluations.

The retraining process can be fine-tuned through hyperparameters which allows for optimisation of the training outcome. For the transfer learning approach the default parameters were used except for the number of training steps which were changed from the default of 4,000 steps to 10,000 steps. Selection of this number of iterations allowed for the loss function (cross-entropy) to sufficiently converge, avoiding over-fitting. For the CNN-SVM approach the hyperparameters (kernel, gamma and C) were optimised by grid search[1].

[1] full classifier details: https://github.com/Brophy-E/PPG-Image-Classifier

4 Results

The average classification accuracy of the transfer learning approach was shown to be 88.3% and as can be seen in Fig 2(b) the confusion matrix, demonstrates the accuracy for correctly classifying the test set of PPG images.

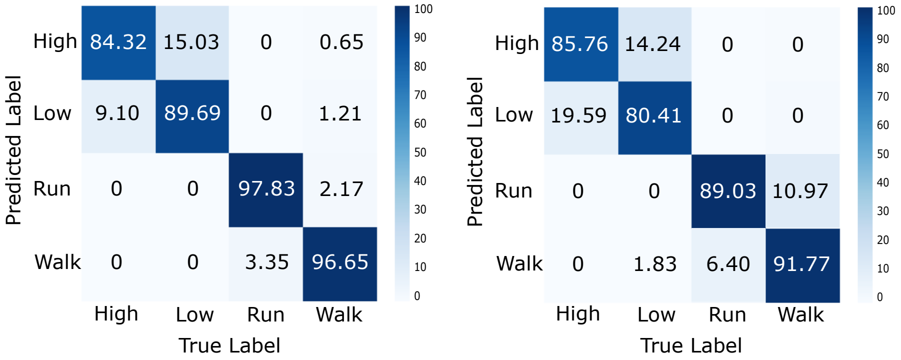

Fig. 2. Confusion matrix for CNN-SVM and for transfer learning approach

The CNN-SVM achieved an average classification accuracy of 92.3%, see Fig 2(a) for the confusion matrix associated with this method. The confusion matrices for these deep learning approaches graphically demonstrates some of the issues classifying the low-resistance bike exercise, where it was misclassified as high-resistance 19.59% of the time. An increase 4.8% classification accuracy was achieved using the CNN-SVM approach with nested cross-validation over the transfer learning approach.

Fig 4 shows two misclassified examples of each class. Based on the plots shown in Fig 4, the errors may have been from a loose wrist strap or excessive movement of the arms. In some circumstances, the signal can be seen to be cut off which indicates gross movement of the limbs for those time instances [20].

A study conducted by Biagetti et al. used feature extraction and reduction with Bayesian classification on time series signals [3]. Their technique focused on using singular value decomposition and truncated Karhunen-Loève transform. Their study used the same time series dataset and was designed to present an efficient technique for HAR using PPG and accelerometer data. Below, in Fig 3 the confusion matrix for their results can be seen, using just the PPG signal for feature extraction and classification.

The feature extraction approach for determining HAR using just the PPG yields an overall accuracy of 44.7%. This shows a reduced classification performance versus the deep learning method employed in this paper. Although our CNN-SVM achieved greater accuracy of over 47 percentage points (92.3% vs. 44.7%), Biagetti et al. in the same paper combined the PPG and accelerometer data to bring their classifier accuracy rate to 78%. We are able to produce

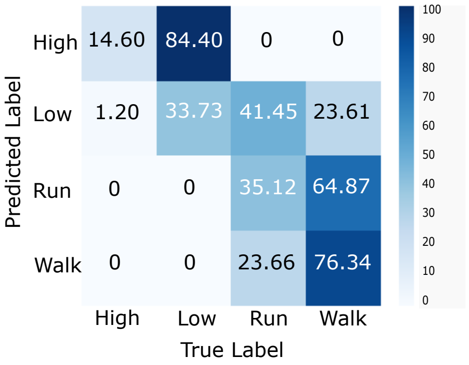

Fig. 3. Confusion matrix for feature extraction approach

very competitive accuracy without the use of an accelerometer, i.e. through the optical signal only.

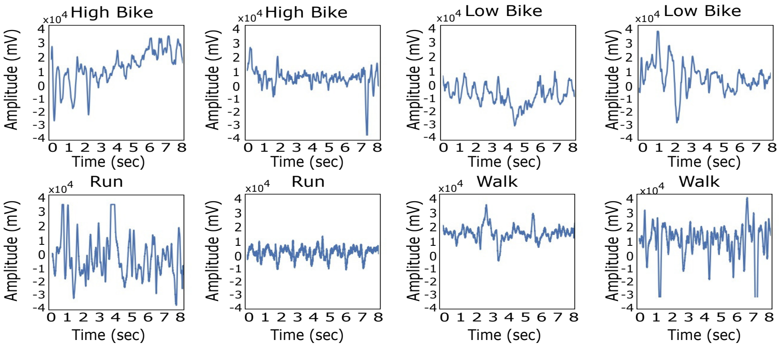

Fig. 4. Sample of eight misclassified images

5 Discussion

Applying the CNN-SVM method to the PPG dataset leads to an accuracy of 92.3%. This result outperforms the combined PPG and accelerometer data for HAR using SVD and KVL (92.3% vs. 78%) and of course is much more accurate than the PPG only result (92.3% vs. 44.7%) [3]. This is a highly competitive result and suggests that simpler wearables based on optical measurements only could yield much of the functionality achievable with more sophisticated, existing multi-modal devices. Of course, the addition of an inertial sensor will always produce more information and therefore more nuanced activity recognition. However, for the types of activity recognition commonly sought in clinical, health and fitness applications a surprisingly good performance can be extracted from a very simple optical measurement.

5.1 Limitiations

Better understanding of the hyperparameters of the transfer learning prior to the CNN-SVM may lead to higher average classification accuracy than the one achieved in this paper (92.3%). The results generated in this paper are based on data from a small sample size which contained a low number of individuals and this may affect the classifier if implemented of new, previously unseen data.

5.2 Visualising the Network

Due to the increase in the use of machine learning and neural networks in the last number of years there has been a growing demand to make their working more transparent and understandable by end users. Often, neural network models are opaque and difficult to visualise. A lack of insight into the model results in diminished trust on the part of the end user. People want to know why and when the model succeeds and how to correct errors [10].

CNNs can appear to operate as a black box. In our case, the input are images, the layers process this and intermediate layers to do their work; extracting patterns, edges and more abstract features. Then a decision is output to the user foregoing any explanation as to why the particular decision was reached e.g. "This image is a dog". Consequently there is emerging and growing efforts to make such AI systems more explainable. For example, the US Defense Advanced Research Projects Agency (DARPA) are currently working on an explainable model with an explainable interface, i.e. "This is a dog because of the snout, ears, body shape etc." This explainable interface can help end-users better accept the decisions made by the AI system.

A number of works have been able to produce high quality feature visualisation techniques, with the goal of making CNNs more explainable. The work done in [17] has the capability to demonstrate the function of the layers within a CNN, how it builds up extracting edges, textures and patterns to extracting parts of objects. The work done by Olah, et al., shows us what a CNN 'sees' and how it builds up its layers as we go deeper into the network.

However, end users are still unable to intuitively determine why the input caused the output. The information retaining to what feature in the input image is a significant discriminating factor in the classification decision is unavailable. The issue here is to be able to explain the decision making process rather than visually representing what each layer extracts. A simple, visual representation of the decision can be a very effective method of explainable artificial intelligence.

Class activation maps (CAM) are one such visual method capable of aiding in the explanation of the decision making process. CAMs indicate particularly discriminate regions of an image that are used by a trained CNN to determine the prediction of the image for a given layer [23], in other words what region in the input image is relevant to the classification. The CAM for this paper uses the global average pooling layer of the retrained Inception-v3 model.

Fig 5 shows the CAM for the retrained network on all of the input classes; high, low, run and walk respectively, they are representative of the average CAMs for their class. Red areas indicate high activations while the dark blue areas indicate little to no activation. Activations across all four classes demonstrate that the model is correctly discriminating the plotted signal from the background of the image. The location of the activations can be used to indicate the inter-class variations. Each of the classes produces a distinct CAM from one another, the difference in these activations can be used to indicate how a class was classified. For a human observer we can say the class is such due to it's amplitude or frequency. The inter-class variation in the CAM across these plots may provide an explainable interface into the decision making process taken by the CNN regarding the image classification problem. "The activity was classified as a walk due to the location of the content activations not present in other classes."

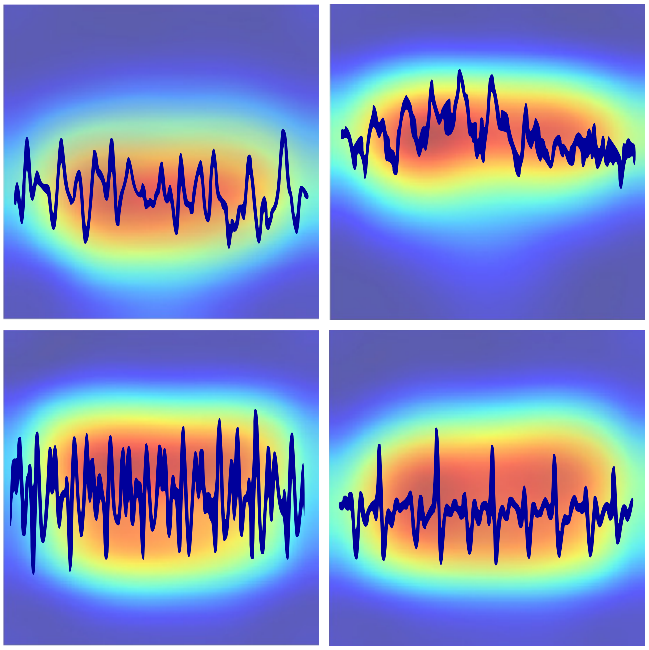

Fig. 5. Average CAMs for High, Low, Run and Walk classes

Another XAI approach that can be adopted is the use of t-Distributed Stochastic Neighbor Embedding (t-SNE) [14]. This dimensionality reduction technique is well suited to the visualisation of high-dimensional datasets such as the 2048 dimensional feature vector describing an image in the penultimate layer of Inception-v3. Using t-SNE on Inception-v3 global average pooling layer allows visualisation of the dimensionally reduced feature vectors for all classes. Fig 6 demonstrates a clustering effect of the data, close groupings of the same class demonstrate the similarity between features available in that class. This allows a user to see the where features for one class appears in relation to the other classes in the trained model. t-SNE produces visualisations that are interpretable by the user in relating the output to the input. One way to help the user comprehend the visualisation of this technique is to place a decision boundary over the output. A feature found within the decision boundary for one class in particular would explain why the input generates the output. Conversely, it might raise further questions if the feature appears outside the boundary. It is also possible to graph one feature against another for a simple dataset (sepal vs petal length [5]). However, this has not been investigated for this dataset yet and this will form part of our future work.

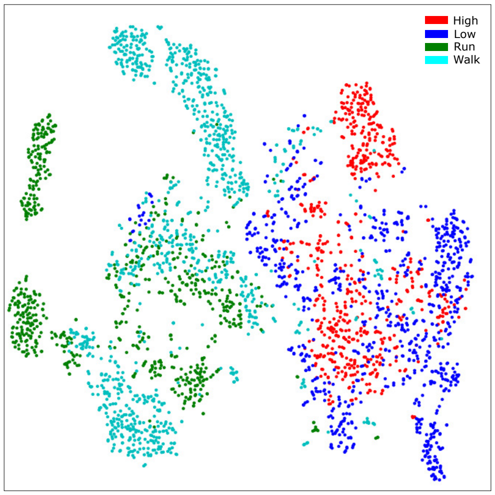

Fig. 6. t-SNE visualisations of CNN codes

6 Conclusion

The application of transfer learning in computer vision for human activity recognition is a novel approach to extracting new useful information from wrist-worn PPG sensors which are conventionally used for heart rate monitoring. We show how CNN suitably configured can create powerful classifiers for HAR applications based on simple images of PPG time series plots [12].

A great benefit of the deep learning approach adopted here is its performance and relative simplicity. Users of this system need not possess a signal processing background to understand the approach and this opens up the possibility that non-experts can develop their own HAR classification applications more readily. Pathways for HAR using deep learning are beginning to be explored on a larger scale thanks to the simplicity of the transfer learning approach, significantly reducing the development time of a suitable CNN. Furthermore this new approach to HAR classification will allow for the easier testing of hypotheses relating to HAR with wearable sensors. The development of AI applications and platforms in the last number of years have led to significant progressions particularly in the domain of image classification problems. These black box AI techniques have a need for more explainable methods to reinforce the machines decision making process. The activation maps (Fig 5) in conjunction with the t-SNE features (Fig 6) have the potential to help with this explainable artificial intelligence issue. Inter-class variations in activations have the capability to provide an explainable interface to the end user; "The exercise was classified as a run because of the high amplitude, high frequency components present in your signal".

The presented process allows activity classification models to be constructed using PPG sensors only, potentially eliminating the need for an inertial sensor set and simplifying the overall design of wearable devices.

Acknowledgements This work was part-funded by Science Foundation Ireland under grant number SFI/12/RC/2289 and by SAP SE.

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X.: Tensorflow: A system for large-scale machine learning. In: Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation. pp. 265–283. OSDI'16, USENIX Association, CA, USA (2016)
2. Allen, J.: Photoplethysmography and its application in clinical physiological measurement. Physiological Measurement **28**(3) (2007). https://doi.org/10.1088/0967-3334/28/3/R01
3. Biagetti, G., B, P.C., Falaschetti, L., Orcioni, S., Turchetti, C., Dipartimento, D.I.I.: Human Activity Recognition Using Accelerometer and Photoplethysmographic Signals. In: Intelligent Decision Technologies. vol. 73 (2018). https://doi.org/10.1007/978-3-319-59424-8
4. Brophy, E., Wang, Z., Dominguez Veiga, J.J., Ward, T.: A machine vision approach to human activity recognition using photoplethysmograph sensor data. In: 29th Irish Signals and Systems Conference 2018. Belfast (2018)
5. Dheeru, D., Karra Taniskidou, E.: Uci machine learning repository (2017), http://archive.ics.uci.edu/ml
6. Dominguez Veiga, J.J., O'Reilly, M., Whelan, D., Caulfield, B., Ward, T.E.: Feature-Free Activity Classification of Inertial Sensor Data With Machine Vision Techniques: Method, Development, and Evaluation. JMIR mHealth and uHealth **5**(8), e115 (2017). https://doi.org/10.2196/mhealth.7521
7. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)

8. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning, vol. 22. MIT Press (2016). https://doi.org/10.1038/nature14539
9. Goodman, B., Flaxman, S.: European union regulations on algorithmic decision-making and a “right to explanation”. arXiv preprint arXiv:1606.08813 (2016)
10. Gunning, D.: Explainable Artificial Intelligence (XAI) (2016). https://doi.org/DARPA, http://www.darpa.mil/program/explainable-artificial-intelligence
11. Jarchi, D., Casson, A.: Description of a Database Containing Wrist PPG Signals Recorded during Physical Exercise with Both Accelerometer and Gyroscope Measures of Motion. Data **2**(1), 1 (2016). https://doi.org/10.3390/data2010001
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet Classification with Deep Convolutional Neural Networks. Advances In Neural Information Processing Systems pp. 1–9 (2012). https://doi.org/http://dx.doi.org/10.1016/j.protcy.2014.09.007
13. Lara, O.D., Labrador, M.A.: A Survey on Human Activity Recognition using Wearable Sensors. IEEE Communications Surveys & Tutorials **15**(3), 1192–1209 (2013). https://doi.org/10.1109/SURV.2012.110112.00192
14. van der Maaten, L.J.P.: Accelerating t-{SNE} using Tree-Based Algorithms. Journal of Machine Learning Research (JMLR) **15**, 3221–3245 (2014). https://doi.org/10.1007/s10479-011-0841-3
15. Mannini, A., Sabatini, A.M.: Machine learning methods for classifying human physical activity from on-body accelerometers. Sensors **10**(2), 1154–1175 (2010). https://doi.org/10.3390/s100201154
16. Mehrang, S., Pietila, J., Tolonen, J., Helander, E., Jimison, H., Pavel, M., Korhonen, I.: Human Activity Recognition Using A Single Optical Heart Rate Monitoring Wristband Equipped with Triaxial Accelerometer. In: European Medical and Biological Engineering Conference. pp. 587–590 (2017). https://doi.org/10.1007/978-981-10-5122-7_147, http://link.springer.com/10.1007/978-981-10-5122-7_147
17. Olah, C., Mordvintsev, A., Schubert, L.: Feature Visualization (2017). https://doi.org/10.23915/distill.00007
18. Raschka, S., Mirjalili, V.: Python Machine Learning. Packt Publishing (2017)
19. Sazonov, E.S., Fulk, G., Sazonova, N., Schuckers, S.: Automatic recognition of postures and activities in stroke patients. Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Engineering the Future of Biomedicine, EMBC 2009 pp. 2200–2203 (2009). https://doi.org/10.1109/IEMBS.2009.5334908
20. Tautan, A.M., Young, A., Wentink, E., Wieringa, F.: Characterization and reduction of motion artifacts in photoplethysmographic signals from a wrist-worn device. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 6146–6149 (Aug 2015). https://doi.org/10.1109/EMBC.2015.7319795
21. TensorFlow: A Guide to TF layers Building a Convolutional Neural Network, https://www.tensorflow.org/tutorials/layers
22. Wang, T., Chen, Y., Zhang, M., Chen, J.I.E., Snoussi, H.: Internal Transfer Learning for Improving Performance in Human Action Recognition for Small Datasets. IEEE Access **5**, 17627–17633 (2017). https://doi.org/10.1109/ACCESS.2017.2746095
23. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning Deep Features for Discriminative Localization. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015). https://doi.org/10.1109/CVPR.2016.319, http://arxiv.org/abs/1512.04150