Monitoring In-Home Emergency Situation and Preserve Privacy using Multi-modal Sensing and Deep Learning

David Andreas Bordvik
Dept. of Informatics, University of Oslo
Oslo, Norway
davidabo@uio.no

Jie Hou
Dept. of Physics, University of Oslo
Dept. of Clinical and Biomedical Engineering, Oslo University Hospital
jieho@fys.uio.no

Farzan M. Noori
Dept. of Informatics, University of Oslo, Norway
farzanmn@ifi.uio.no

Md Zia Uddin
Software and Service Innovation, SINTEF Digital, Oslo, Norway
zia.uddin@sintef.no

Jim Torresen
Dept. of Informatics and RITMO University of Oslo, Norway
jimtoer@ifi.uio.no

***Abstract*—Videos and images are commonly used in home monitoring systems. However, detecting emergencies in-home while preserving privacy is a challenging task concerning Human Activity Recognition (HAR). In recent years, HAR combined with deep learning has drawn much attention from the general public. Besides that, relying entirely on a single sensor modality is not promising. In this paper, depth images and radar presence data were used to investigate if such sensor data can tackle the challenge of a system's ability to detect abnormal and normal situations while preserving privacy. The recurrence plots and wavelet transformations were used to make a two-dimensional representation of the presence radar data. Moreover, we fused data from both sensors using *data-level, feature-level, and decision-level fusions*. The decision-level fusion showed its superiority over the other two techniques. For the decision-level fusion, a combination of the *depth images* and *presence data recurrence plots* trained first on convolutional neural networks (CNN). The output was fed into support vector machines, which yielded the best accuracy of 99.98%.**

***Index Terms*—CNN, LSTM, Multi-modal sensing, Home-care, Sensor fusion, XeThru UWB sensor.**

I. Introduction

According to a 2017 report by the United Nations Department of Economic and Social Affairs, the number of older individuals is growing at a faster rate than that of other age groups [1]. In 2015, one in every eight persons would be 60 years or older. By 2050, the world's population of senior citizens is anticipated to reach over 2.1 billion. With the advancement of technology, people are becoming more independent. The combination of increasing age and independence leads to many social and financial challenges, especially nursing for older people. Autonomous systems to aid in older people's well-being in their homes have been getting increased attention lately. However, analysis based on human behavior has proven to be challenging due to the complexity of human behavior and privacy regulations [2]. As a health monitoring system, HAR has been getting significant attention to tackle the challenges related to behavior analysis, e.g., normal and abnormal, for older people. HAR seeks to classify a person's behavior or actions in real-time from a series of measurements [3], [4]. Even though HAR is looking promising, the task encounters several class challenges, such as the null-class dominance (non-relevant activities), interclass similarity, intraclass variability, and the complexity and diversity of physical activities [5].

When a single sensor is unable to measure all relevant attributes or when the perception is unclear, vulnerability in the sensor arises as a result of occlusions or missing features [6]. Due to uncertainties, relying entirely on a single sensor modality to detect human behaviors is not very practical. Health or activity monitor systems having several sensors can be more useful to discriminate complex activities. Integrating data from various sensors enhance reliability and confidence while reducing ambiguity and uncertainty. The utilization of different sensors data and autonomous reports to the caregiver would be highly beneficial from such a system [7].

Regarding home care of older people, privacy needs to be compromised with a system's ability to notify the caregiver when something seems abnormal. Also, light conditions in a home can vary significantly with sunlight in the daytime and darkness at night-time, hindering video surveillance effectiveness. Rather than transmitting a lot of personal sensor data to the caregiver facility, it is desirable to have an autonomous system that makes qualified decisions based on the sensor data. Today, older people might prefer to live in the comfort of their own homes instead of nursing homes. However, they might be subject to emergencies where they are unable to alarm the authorities. So, to make this a viable living situation, there is an apparent need for a surveillance system. It is desirable that the referred system is autonomous and contributes to reducing

First and second authors contributed equally to this work.

This is the author accepted version of an article published in
2022 International Conference on Electronics, Information, and Communication (ICEIC)
https://doi.org/10.1109/ICEIC54506.2022.9748829

Fig. 1: A schematic setup to detect the abnormal situation when the person is lying on the floor.

the human workload and increasing the person's privacy.

Smart sensor systems have been demonstrated in human-computer interactions, which help recognize human actions. For instance, recognizing when a person suddenly falls on the floor or has an abnormal heart-rate while the person is resting. With the increasing demand for privacy for older people, this work will demonstrate how a combination of different sensors in a care robot at home can reduce privacy-related concerns compared to surveillance-based system and, at the same time, increase quality in the prediction of emergencies.

Many works have been done related to human activity recognition [8], [9], [10], [11], [12]. Noori et al. [8], [10] investigated the multiple representations of a single sensor data and fusion of multiple representations together with deep convolution neural networks (CNNs) and showed promising results. Xia et al. [13] used a combination of Long-Short term memory (LSTM) and CNN architecture to perform automatic activity feature extraction using very few model parameters. Bulling et al. [14] focused on using on-body inertial sensors and addressed the problems with recognizing different hand gestures from the sensors attached to the arms. Krizhevsky et al. [15] investigated classifying high-resolution images into 1000 classes using CNN, performed using non-saturation neurons and efficient GPU implementation, and showed favorable results. Guo et al. [16] explored the use of multimodal information to classify human activities based on wearable sensors and yield a competitive HAR performance.

The main focus of this work is to monitor the human activities and capture the heart-rate. For instance, if the person does some daily exercise, a sudden increase in heart-rate (HR) is expected, and vice versa. On the other hand, if a change in HR can not easily be explained, it could be a cause of concern and should be reported. To detect the sudden abnormality accurately, this work will use multi-sensor data, together with different sensor fusion techniques to cover a wide range of features. An RGB-D camera, XeThru ultra-wideband (UWB) radar, and a smartwatch (for heart-rate recording) were used to collect the data. A UWB radar is a compact impulse-radio ultra-wideband radar system on a chip UWB sensors [17].

In this work, we explore the potential HAR for providing detection of normal or abnormal behavior. The goal is to explore different sensor fusion methods and machine learning algorithms to make models that can classify normal vs abnormal situations by combining RGB-D images and presence data from the UWB radar. This work focus on using depth images and presence data instead of using the RGB images to preserve privacy. To the best of our knowledge, this is

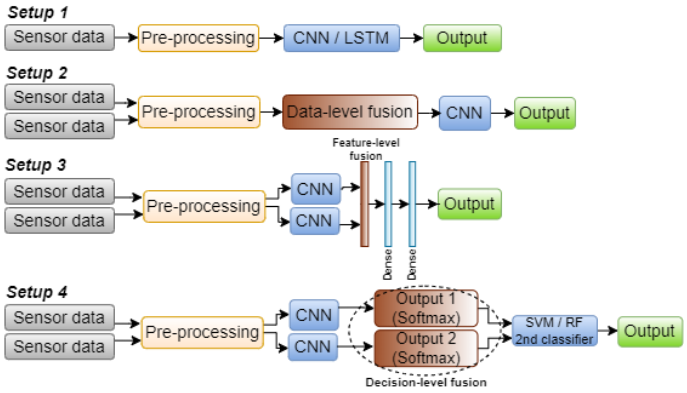


Fig. 2: Experimental setups for different model configurations. Setup 1: single sensor data model; Setup 2: data-level fusion; Setup 3: feature-level fusion; Setup 4: decision-level fusion.

the first work done using only depth images and presence data to predict emergency situations. Deep learning methods have shown promising results in recognizing different human activities. The CNNs and LSTM-based RNNs were used to classify normal and abnormal situations in this work. The contributions of this paper is as follows:

- Different levels of sensor fusion are explored to classify normal or abnormal situation in a home emergency system while preserving the privacy of people.
- The UWB sensor's data was presented in recurrence plots and wavelet transform.
- The use of sensor fusion secures redundancy concerning sensor malfunction.
- Our methods make a more robust system with respect to previous works while also demonstrates good performance.

The paper is organized as follows: Section II presents the methodology. Section III presents results and discussions. The paper is concluded in Section IV.

II. Methodology

The methodology of our proposed approach will be presented in this section.

A. Preprocessing of sensor data

The dataset is collected at the University of Oslo, Norway. During the data recording, the subjects wore a smartwatch to record the heart-rate as the ground truth. The normal activities consist of typical activities, such as lying, sitting, and standing while having normal heart-rate. The abnormal activity consists of situations while lying on the floor and having a heart-rate of more than 90 BPM. For recording the abnormal situation, the participants were asked to exercise until their heart-rates reached 140 BPM. Afterward, they immediately lied down on the floor, and the robot recorded the radar data of their movements.

Initially, pre-processing of the data was done to generate labels and extract relevant and significant features from the dataset. The dataset comprises 137 GB of data stored in 1.6 million files for 20 different users. The images (RGB and depth) takes up the larger part of the dataset. Out of the

1.6 million images, the images that correspond to each user's actual heart-rate acquired at the same time were extracted. The heart-rate is sampled in the dataset once per second, while the images are sampled with 30 images per second. Therefore, a separate program was made to generate a new condensed version of the dataset where all the heart-rate measurements are aligned with exactly one image from the same second in time. We also re-sampled the data in pre-processing, which resulted in fewer samples of class 0, to better balance the classes. Moreover, we decided to work with gray-scale images to make it easier to train for the CNN model. Computational wise it was necessary to reduce the image by automatically remove the regions that were irrelevant background. This removal resulted in the images being reduced from 240×320 down to 200×200. This process enabled us computational wise to train the model with 19777 images while still keeping the relevant information inside the images. The preprocessed data was stored into binary files for easier access. The binary files took up 3.16 GB of storage, and the condensed dataset was more manageable in terms of storage capacity, ease of access, and computational read speed.

B. Data representations

Wavelet transformation was selected as one of the two-dimensional representations of the presence dataset. The wavelet transformation will allow us to know at which frequency our signal oscillates and at what time these oscillations occur. The scaleogram was used to make the two-dimensional spectrum for wavelet transformation. The x-axis in the scaleogram represents the absolute value of the wavelet transformation coefficients of our signal, and the y-axis represents the distance from the sensor.

In addition to the wavelet transformation, recurrence plot [18] was also created. It produces a square matrix, and the matrix elements in the recurrence plot correspond to the times where a state recurs in a dynamical system [19].

$$R_{i,j} = \begin{cases} 1 & if \quad \|x_i - x_j\| \leq \epsilon \\ 0 & otherwise \end{cases} \tag{1}$$

x represents the states, ϵ is a threshold distance, if the difference between x_i and x_j is smaller than the threshold ϵ, $R_{i,j}$ will be 1, otherwise 0.

C. Sensor fusion methods

When one work with multi-modal sensor data, fusing data from different modalities can increase the model performance [16], [20]. In this work, three methods are used in multi-modal sensor fusion, as shown in Fig. 2. Data-level fusion [21] is setup 2, where the depth image with presence dataset were concatenated. The feature-level [22] fusion will combine the features extracted using a CNN model from the depth image and the two-dimensional representation of presence data and then merge the features before the classifiers, as shown in setup 3. In the decision level fusion, meta-classification was used.

D. CNN architecture

CNNs are often used in image recognition due to their ability to automatically extract features, allowing them to differentiate between images accurately. Through convolutional layers, the CNNs can capture the spatial features in the images, and the pooling layer will then reduce the size of the data output so that it is easier to process further [23]. The last dense layer performs the final part in a CNN model; in our case, there are two outputs, 0 for the normal case and "1" for the abnormal case. The CNN architecture used is as follows:

- A kernel size of 3, dilation rate of 2 and padding "same" are used for each of the convolutional layers.
- Max pooling layer of size 2 was used.
- Activation function "RELU" was used for both convolutional and fully connected layers.
- 2 units with the "softmax" activation function was used for the last classification layer, which corresponds to "normal" and "abnormal" activities.

E. LSTM architecture

The LSTM is a special kind of recurrent neural network (RNN). It has an internal mechanism "gates" which regulates the flow of information. These gates decide which information is essential and should be kept and which information should be disregarded [13]. That is to say, the model has memory of its historical data. With this mechanism, the LSTM models do not suffer from the vanishing and exploding gradient problem as traditional RNNs. This method was chosen because the historical information may provide us important features
The LSTM architecture used is as follows:

- Three LSTM layers were used together with a hidden layer
- Batch normalization was added between the layers to perform automatic standardization
- Dropout was used to reduce the overfitting problem
- 2 units with the "softmax" activation function was used for the last classification layer, which corresponds to "normal" and "abnormal" activities.

F. Performance Measures

The metrics accuracy, F_1 score, and recall were chosen to evaluate models performance. They were calculated using True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Accuracy is a measure of the overall performance considering all classes. Precision is a measure for predicted performance for each class specifically. Recall is also a metric specifically for each class, and it takes into account class imbalance between classes. The F_1 score conveys the balance between precision and recall between all classes. F_1 score is, therefore, a good metric for evaluating model performance since it gives deeper insight into how the model performs.

III. Results and Discussions

For the classification of emergencies, we conducted two sets of experiments; one is three classes classification where

the dataset was separated into resting ($\text{Class}_{\text{Resting}}$), lying ($\text{Class}_{\text{Lying}}$) and moving ($\text{Class}_{\text{Moving}}$) classes and binary classification where normal and abnormal situations were detected. For two class classification, which is based on the lying class, the ground truth labels are 0 for abnormal situation and 1 for normal situation, and they were constructed directly from the heart-rate measurements. If a person is lying on the floor, and the heart-rate is above the threshold, it is classified as an abnormal situation ($\text{Class}_{\text{Lying}}$). The goal is illustrated in Fig. 1. For the three class classification, we further classify abnormal and normal situation based upon the lying class. We define an abnormal situation for binary classification to be heart-rate above 90 BPM if the person is lying on the floor as shown in equation (2).

$$\text{Class}_{\text{Lying}} = \begin{cases} \text{Abnormal} & \text{HR} > 90 \\ \text{Normal} & \text{Otherwise} \end{cases} \tag{2}$$

Different models were tested with RGB, depth images, and presence data, both individually and using sensor fusion methods. Both CNN and LSTM were built in Keras with TensorFlow backend. 20 models were made with four different setups shown in Fig. 2.

We used both single sensor data and three different sensor fusion method for the two sets of the experiment, namely data-level fusion, feature-level fusion, and decision level fusion. Experiments were conducted with setup 1,2,3 and 4 shown in Fig. 2 for both two class classification and three class classification. Setup 1 with configuration using RGB images with the CNN model was used as our reference model. By comparing performance from other setups and configurations to the reference model, we were able to find the most optimal method while preserving privacy concerning human identification from RGB images. The following configurations were conducted for our experiment:

- Setup 1
  - CNN trained with RGB images only (base model).
  - CNN trained with depth images only.
  - LSTM trained with presence data using all features (1466), which corresponds to 9.9 meters of radar detection.
  - LSTM trained with presence data using just 400 features, which corresponds to 2.6 meters of radar detection.
- Setup 2 (Data - level fusion)
  - CNN trained with depth images merged with **recurrence plot** or **wavelet** representation of presence dataset.
- Setup 3 (Feature - level fusion)
  - CNN for feature extraction for depth images and **recurrence plots** or **wavelet** representation of presence dataset. The extracted features are concatenated before feeding them into new dense layers and utilizing softmax activation function to discriminate between the classes. At the output, we ultimately choose the most probable class from the distribution given by softmax.
- Setup 4 (Decision - level fusion)
  - Utilizing feature extraction for depth images and the **recurrence plots** or **wavelet** representation of presence dataset using CNN. The extracted features were fed into two separate softmax classifiers. The predicted probability distribution was merged before feeding them as input to the final classifier either Support Vector Machine (SVM) or Random Forest classifier (RF).

Hyperparameter	Values tested
Learning rate η	$[10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}]$
Regularization parameter λ	$[10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}]$
Minibatch size	$[8, 16, 32, 64]$
CNN/LSTM units	$[16, 32, 64]$
Dropout	$[0.2, 0.3, 0.4]$

TABLE I: Hyperparameters tested for both CNN and LSTM models.

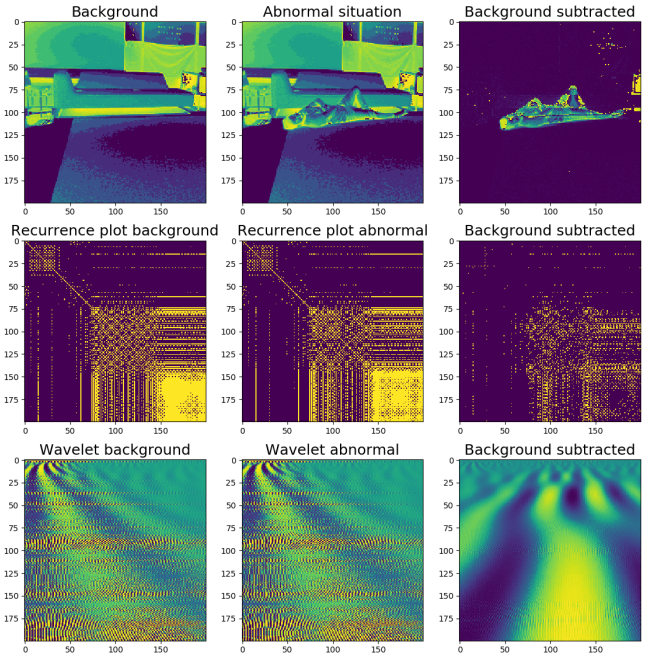
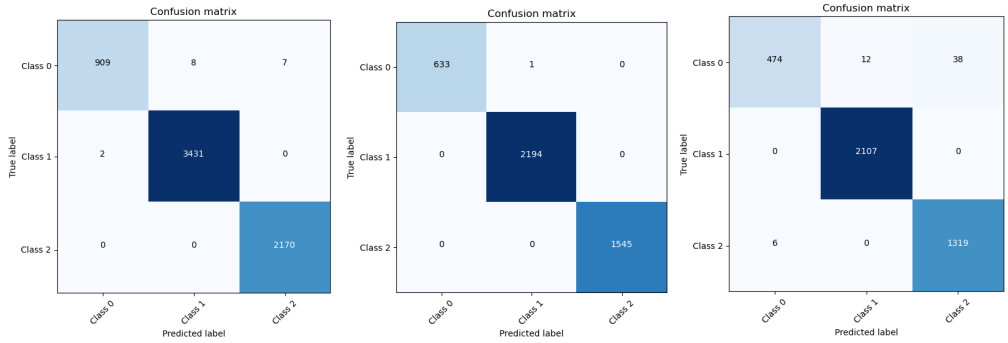


Fig. 3: The three plots on the first row represents the background, abnormal situation where the person is lying on the ground and the person alone where the background has been subtracted. The second and third rows represent the corresponding two dimensional representation of both the recurrence plot and the wavelet transformation using the first 200 features.

The recurrence plot and wavelet transformation were used to represent the presence data in two-dimensional space. Using this data representation, it was possible to merge the image grid of presence data with depth images for our data fusion while still keeping the temporal and spacial dimensions in the presence data. Fig. 3 shows the result of pre-processed data concerning RGB images and presence (Xethru) data.



(a) (b) (c)

Fig. 4: (a) shows the Confusion matrix of setup 1, CNN model trained with depth images only, (b) shows the confusion matrix of setup 4, CNN model trained with depth images, and recurrence plot representation of presence data (Decision - level fusion), (c) shows the confusion matrix for setup 1, LSTM model trained with presence data only.

Model performance for three classes classification						
Setup	**ANN**	**Sensor data**	**Sensor fusion**	**Accuracy**	**F1-Score**	**Recall**
1 (reference)	CNN	RGB	N/A	0.9969	0.9968	0.9969
1	CNN	Depth	N/A	0.9973	0.9959	0.9943
2	CNN	Depth & Recurrence plots	Data-level	**0.9981**	0.9981	0.9981
3	CNN	Depth & Recurrence plots	Feature-level	0.9959	0.9958	0.9959
4	CNN	Depth & Recurrence plots	Decision-level (SVM)	**0.9998**	0.9997	0.9995
2	CNN	Depth & Wavelet transform	Data-level	**0.9974**	0.9973	0.9974
3	CNN	Depth & Wavelet transform	Feature-level	0.9956	0.9955	0.9956
4	CNN	Depth & Wavelet transform	Decision-level (RF)	**0.9973**	0.9972	0.9973
1	LSTM	Presence (all features)	N/A	0.9846	0.9755	0.9666
1	LSTM	Presence (200 features)	N/A	0.9590	0.9589	0.9590

TABLE II: Model performance for three classes classification. Highlighted cases indicate better performance compared to the reference case.

The effect of both normalizing the data and subtracting the background was significant concerning model performance. We got a good portion of increased accuracy by working with normalized presence data that contained relative change with respect to a fixed baseline (background).

In total, 19,777 images and measurements from 20 persons were used for RGB, depth images, and the presence data, respectively. From the total dataset, 44%, 22%, and 33% of the data are chosen for training, validation, and testing, respectively. For all the models, the training process was performed for 100 epochs. The best hyperparameters were determined by repeated grid search procedure. Table I shows a complete list of the hyperparameters used for all the experiments. Confusion matrix was used to describe the performance of our classifier. Several confusion matrices were generated among the different setups in order to visualize the model performances. Fig. 4 shows selected confusing matrix for three classes classification.

Table III shows the accuracy for two classes classification. Using only RGB images gives the best accuracy at 0.9998. Comparing different sensor fusion methods, decision-level fusion outperformed both the data-level fusion and the feature-level fusion, which is in agreement with results from earlier studies [8], [10]. Both SVM and RF models were tested as the final classifier; in most cases, the SVM performs better than the RF classifier. Comparing Table III and Table II with the reference model, results from different sensor fusion methods are a bit worse than 0.9998 from the reference model for two classes classification. However, 4 of the models with sensor fusion methods outperformed the reference model for three classes classification. Among those models, the CNN model trained on a combination of depth images and the recurrence plot representation of presence data stands out with an accuracy of 99.98 %. Some of the sensor fusion methods we experimented with performed worse than models using only single sensor data. We believe that this is stemming from the most optimal hyperparameters not taking part in the grid search. In addition, deep learning methods are becoming more and more complex, which results in black-box models where the interpretations of the relation between different variables are not very explainable [24]. In this regard, the most optimal parameters were not found, and better performance may still be possible.

IV. Conclusions

In this work, we evaluated different sensor fusion (data, feature and decision) levels and different deep learning algorithms to detect normal and abnormal situations based on heart-rate. Depth images and presence data from XeThru ultra-wideband (UWB) radar together with a smartwatch for heart-rate detection were used to validate our sensor fusion approaches. Considering a person's privacy, we chose to use depth images instead of RGB images. We found that the decision-level fusion outperformed the other approaches in classifying normal and abnormal situations with 99.98 %

Model performance for two classes classification						
Setup	**ANN**	**Sensor data**	**Sensor fusion**	**Accuracy**	**F1-Score**	**Recall**
1 (reference)	CNN	RGB	N/A	0.9998	0.9987	0.9989
1	CNN	Depth	N/A	0.9218	0.9216	0.9217
2	CNN	Depth & Recurrence plots	Data-level	0.9212	0.9209	0.9211
3	CNN	Depth & Recurrence plots	Feature-level	0.9613	0.9612	0.9612
4	CNN	Depth & Recurrence plots	Decision-level (SVM)	0.9863	0.9859	0.9863
2	CNN	Depth & Wavelet transform	Data-level	0.9225	0.9221	0.9224
3	CNN	Depth & Wavelet transform	Feature-level	0.9633	0.9630	0.9633
4	CNN	Depth & Wavelet transform	Decision-level (RF)	0.9973	0.9979	0.9971
1	LSTM	Presence (all features)	N/A	0.9431	0.9430	0.9431
1	LSTM	Presence (200 features)	N/A	0.9270	0.9268	0.9270

TABLE III: Model performance for classification of normal and abnormal situations for $\text{Class}_{\text{Lying}}$.

accuracy, which is even better than using RGB images directly. This gives us another method to supervise the health status of older people and protect their private lives.

In the future, we plan to predict future heart-rate with regression approach using LSTM and the transformers model. Furthermore, we will apply CNN-LSTM model for classification where we use the CNN model to extract features from the images and LSTM for historical context.

V. Acknowledgment

This work is partially supported by The Research Council of Norway (RCN) as a part of the Multimodal Elderly Care systems (MECS) project under Grant 247697, Collaboration on Intelligent Machines (COINMAC) project, under Grant 261645 and 309869, Vulnerability in the Robot Society (VIROS) under Grant 288285, Predictive and Intuitive Robot Companion (PIRC) under Grant 312333 and through its Centres of Excellence scheme, RITMO with Grant 262762.

References

[1] *Department of Economic and Social Affairs, Population Division (2017). World Population Ageing*. United Nations, 2017.

[2] Farzan Majeed Noori, Zia Uddin, and Jim Torresen. Robot-care for the older people: Ethically justified or not? In *2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 43–47. IEEE, 2019.

[3] Emilio Sansano, Raúl Montoliu, and Óscar Belmonte Fernández. A study of deep neural networks for human activity recognition. *Computational Intelligence*, 36(3):1113–1139, 2020. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/coin.12318.

[4] Farzan Majeed Noori, Benedikte Wallace, Md Zia Uddin, and Jim Torresen. A robust human activity recognition approach using openpose, motion features, and deep recurrent neural network. In *Scandinavian conference on image analysis*, pages 299–310. Springer, 2019.

[5] Jianbo Yang, M. Nguyen, P. P. San, X. Li, and S. Krishnaswamy. Deep Convolutional Neural Networks on Multichannel Time Series for Human Activity Recognition. In *IJCAI*, 2015.

[6] W Elmenreich. An introduction to sensor fusion. vienna university of technology; vienna. Technical report, Austria: 2002. Research report.[Google Scholar].

[7] Lisa Schrader, Agustín Vargas Toro, Sebastian Konietzny, Stefan Rüping, Barbara Schäpers, Martina Steinböck, Carmen Krewer, Friedemann Müller, Jörg Güttler, and Thomas Bock. Advanced Sensing and Human Activity Recognition in Early Intervention and Rehabilitation of Elderly People. *Journal of Population Ageing*, 13(2):139–165, June 2020.

[8] Farzan Majeed Noori, Michael Riegler, Md Zia Uddin, and Jim Torresen. Human Activity Recognition from Multiple Sensors Data Using Multi-fusion Representations and CNNs. *ACM Trans. on Multimedia Computing, Communic., and Applications*, 16(2):45:1–45:19, May 2020.

[9] Mandar Gogate, Ahsan Adeel, and Amir Hussain. Deep learning driven multimodal fusion for automated deception detection. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–6, November 2017.

[10] Farzan Majeed Noori, Enrique Garcia-Ceja, Md. Zia Uddin, Michael Riegler, and Jim Tørresen. Fusion of Multiple Representations Extracted from a Single Sensor's Data for Activity Recognition Using CNNs. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6, July 2019. ISSN: 2161-4407.

[11] Henry Friday Nweke, Ying Wah Teh, Ghulam Mujtaba, Uzoma Rita Alo, and Mohammed Ali Al-garadi. Multi-sensor fusion based on multiple classifier systems for human activity identification. *Human-centric Computing and Information Sciences*, 9(1):34, September 2019.

[12] Farzan M Noori, Md Zia Uddin, and Jim Torressen. Ultra-wideband radar-based activity recognition using deep learning. *IEEE Access*, 2021.

[13] Kun Xia, Jianguang Huang, and Wang Hanyu. LSTM-CNN Architecture for Human Activity Recognition - IEEE Journals & Magazine. 8, March 2020.

[14] Andreas Bulling, Ulf Blanke, and Bernt Schiele. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys*, 46(3):33:1–33:33, January 2014.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, May 2017.

[16] Haodong Guo, Ling Chen, Liangying Peng, and Gencai Chen. Wearable sensor based multimodal human activity recognition exploiting the diversity of classifier ensemble. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, pages 1112–1123, New York, NY, USA, September 2016. Association for Computing Machinery.

[17] N. Andersen, K. Granhaug, J. A. Michaelsen, S. Bagga, H. A. Hjortland, M. R. Knutsen, T. S. Lande, and D. T. Wisland. A 118-mW Pulse-Based Radar SoC in 55-nm CMOS for Non-Contact Human Vital Signs Detection. *IEEE Journal of Solid-State Circuits*, 52(12):3421–3433, December 2017. Conference Name: IEEE Journal of Solid-State Circuits.

[18] J.-P. Eckmann, S. Oliffson Kamphorst, and D. Ruelle. Recurrence Plots of Dynamical Systems. *Europhysics Letters (EPL)*, 4(9):973–977, November 1987. Publisher: IOP Publishing.

[19] Norbert Marwan, M. Carmen Romano, Marco Thiel, and Jürgen Kurths. Recurrence plots for the analysis of complex systems. *Physics Reports*, 438(5):237–329, January 2007.

[20] Zeeshan Ahmad and Naimul Khan. Towards Improved Human Action Recognition Using Convolutional Neural Networks and Multimodal Fusion of Depth and Inertial Sensor Data. *arXiv:2008.09747 [cs]*, August 2020. arXiv: 2008.09747.

[21] Rao Muhammad Anwer, Fahad Shahbaz Khan, Joost van de Weijer, Matthieu Molinier, and Jorma Laaksonen. Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 138:74–85, April 2018.

[22] Omid Dehzangi, Mojtaba Taherisadr, and Raghvendar ChangalVala. IMU-Based Gait Recognition Using Convolutional Neural Networks and Multi-Sensor Fusion. *Sensors*, 17(12):2735, December 2017. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

[23] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, and Tsuhan Chen. Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354–377, May 2018.

[24] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery*, 9(4):e1312, 2019. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1312.