technische universität
dortmund

Person Identification Using Motion Information

Master Thesis

Nilah Ravi Nair

November 9, 2021

Supervisors:
Prof. Dr.-Ing. Gernot A. Fink
Fernando Moya, M.Sc.

Fakultät für Informatik
Technische Universität Dortmund
http://www.cs.uni-dortmund.de

CONTENTS

1 INTRODUCTION

The human gait cycle is a biometric that remains untapped for person identification and verification applications. Unlike its counterparts, fingerprint and retina-scan recognition systems, gait cycles are sensitive to the changes in clothing, terrain, fatigue, environment and idiosyncrasies. Consequently, creating a reference model or template for an individual including all possible variations is not easy; for example, researchers have explored vision-based person identification using gait for security applications. Gait-based person identification for high accuracy applications, such as crime control and detection in high security, public areas, were deemed difficult due to the camera angle, occlusion, obtrusion, clothing, and the realistic scenario of multiple individuals being within a frame. However, low-priority applications, such as identification in a health care system, are presumed to be feasible.

One could consider Inertial Measurement Units (IMU) or infrared sensors as alternative data acquisition methods to overcome the limitations of obtrusion and occlusion of vision-based data acquisition. Interestingly, IMU sensors are extensively used in the field of human activity recognition (HAR). HAR is highly researched due to its possible applications in logistic environments, clinical diagnosis and monitoring systems for elderly care. Furthermore, HAR using IMU sensors is widely used for daily activity monitoring. IMUs in smartwatches and smartphones are used to acquire motion data to log daily activities, such as step count, sleep duration, and running duration to promote a healthier lifestyle. As a result, a massive set of IMU data with varied clothing, terrain and environment is available for each individual.

The gait cycle is the repetitive movement of the body, as seen when performing activities such as walking and running. Given that gait is a biometric, it is interesting to analyse whether the general body movements of an individual would contain a unique signature that could function as an identity. Here, signature refers to movement patterns ingrained in how the individual performs general activities such as holding, waving and walking. Neural networks and a large amount of data available facilitate the exploration of general body motion for person identification. However, the possibility of person identification using motion data collected from IMU sensors raises privacy concerns. For instance, the motion information collected from the IMU sensors of smartphones and smartwatches are often stored in third-party memory locations. The data storage is necessary to facilitate future analysis and developments specific to user requirements. Consequently, given a scenario that the data is hacked and the individuals can be identified based on the motion data, privacy will be compromised.

Similarly, IMU sensors are used for activity recognition in a logistic environment to optimise the order-picking process by ©MotionMiners GMBH [MM] and ©Iterate Labs Inc [IT]. These companies prefer IMU-based data acquisition methods over videos to maintain the employees' privacy. Privacy is mandated by the European Union's General Data Protection Regulation [Cou16]. The Official Journal of European Union Regulations Directive 95/46/EC § 1.24, 1.26, and 1.28 (2016) — state that given a situation where the identity is not given, but there is a possibility of re-identification based on any small publicly available excerpts; data privacy is considered breached. Thus, raising the question, "Is person identification possible with IMU-based motion data created for HAR?".

Research by [GRS+20] and [EBL18] suggests the possibility of general body motion-based person identification. Thus, we need to analyse the data to identify the features that contain identity information. Further, methods to either mask or to delete the features that facilitate identification so that the HAR data is devoid of identity must be designed. In addition, it is of interest to analyse whether soft-biometric characteristics such as age, gender and height can be analysed from the data. The study could lead to an understanding of the effect of soft-biometrics and the subject's individuality on the data created for HAR experiments. Furthermore, developing attribute representation with soft-biometrics can facilitate categorisation and transfer learning of human motion data. Thus, this study can address data privacy concerns and draw out methods to improve HAR data collection.

This thesis aims to be the preliminary work towards identifying the impact of individual motion signature on the HAR dataset and the possibility of masking or deleting identity. As a result, this thesis first explores person identification using general motion information from IMU data. As the experimental results are expected to be dataset-specific, various datasets must be explored to devise a conclusive statement. Further, the thesis will scrutinize the impact of activities on identification accuracy. Next, the possibility to model soft biometrics as attribute representation will be examined. HAR has popularly experimented on Neural Networks such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Following the lead, the experiments of this thesis will be performed on CNN and RNN. A comparison of the methods will be performed, following which the network with better performance will be utilised for further experiments.

The thesis is structured as follows. Chapter: 2 will explain the fundamental concepts of the networks used in this thesis. In addition, the chapter will explain biometrics and

HAR. The previous research on person identification using motion information will be explored in Chapter: 3. The chapter will encase a discussion on the research gaps that were identified in this field. Chapter: 4 will elaborate on the networks and the training methods. Further, attribute representation will be formulated from soft biometrics. The experiments and results will be illustrated in Chapter: 5. Finally, Chapter: 6 will elucidate the conclusions derived from the experiments and results. In addition, future works required in this field following the thesis will be discussed.

2 FUNDAMENTALS

Human activity recognition (HAR) is the process of recognising the physical activities performed by a human using machine learning methods, as explained in Sec: 2.5. HAR is an area of interest for logistics and surveillance companies. Logistics companies use HAR methods to analyse human ergonomics in the order picking process to improve picking time. Ergonomics is the study of human interaction with its environment and is applied by companies to improve the working condition of the employees; for example, the ©Motion-Miner GmbH [MM] uses Inertial Measurement Unit (IMU) data from the sensors attached to the body of employee to analyse the activities and performance time. Then, Motion Miners recommend changes to the work environment to improve employee performance. IMU data is preferred as it is said to maintain the employee's privacy and reduce any source of discomfort of being monitored by another individual. Instead of labelling the data for each activity, a method of automatic recognition of the activity is preferred. Machine Learning (ML) algorithms can be used to facilitate HAR. Often, researchers in this field have considered Artificial Neural Networks (ANN) (see Sec: 2.1) for HAR classifications, as seen in [NRR+20], [ZLC+17], [MGF+18] and [GLR+17].

HAR research requires dedicated datasets. IMU-based HAR datasets such as [CSC+13] and [RS12b] consists of activities of daily living, whereas [NRR+20] consists of activities performed in a logistic environment. From these datasets, it can be seen that walking is a common labelled activity. Interestingly, walking or gait is a biometric (see Sec: 2.4) [JRP04]. As per [GRS+20], person re-identification can be achieved using IMU recordings of gait. Thus, raising the question, can general body motion recordings from IMUs of subjects performing activities, such as cycling, sleeping and handling, be used to identify an individual. Given the possibility that IMU data can be used to identify individuals, privacy would be compromised. Thus, companies such as Motion Miners would have to identify a method to mask or delete the identity information in the IMU data to ensure employee privacy. Consequently, we need to experiment on the possibilities of person identification using motion information from IMU sensors.

[EBL18] has given an overview of how activities of daily living affect person identification. The experiments were conducted on statistical features extracted from the data. Furthermore, the authors focused on classifiers such as Support Vector Machine (SVM), K-Nearest Neighbour (KNN) and Decision Tree. However, it is interesting to explore how features will be learned by Convolution Neural Network (CNN) (see Sec: 2.2) for person identification. CNN variants have been a popular choice in HAR classification. Another variation of ANN is the Recurrent Neural Network (RNN) (see Sec: 2.3), such as Long-Short Term Memory (LSTM) (see Sec: 2.3.1) and Gated Recurrent Units (GRU) (see Sec: 2.3.2). [GRS+20] considered RNNs for person re-identification. However, the author of this thesis explores the possibilities of using CNN for person identification.

This thesis attempts to use the IMU dataset and CNN-based networks created for HAR to explore person identification. Further, the thesis shall investigate the impact of activities on identification accuracy. Consequently, we shall explore the aforementioned terms to facilitate a fundamental understanding of the experiments and the results. In particular, ANN (Sec: 2.1), CNN (Sec: 2.2), RNN (Sec: 2.3), biometric (Sec: 2.4) and HAR (Sec: 2.5).

2.1 ARTIFICIAL NEURAL NETWORK

The ability of a system to acquire sensory information and extract patterns is known as machine learning (ML) [GBC16]. Logistic regression, naive Bayes and ANN are examples of ML algorithms. ANNs are desirable as computational models because of their capability to approximate unknown functions [PP18]. Conceptually, they mimic a Biological Neural Network (BNN).

As shown in Fig: 2.1.1, a unit of ANN can be represented with dendrites, a neuron and an axon. Each dendrite node acts as an input point. The input could be from another artificial neuron or direct sensory input [AH17]. The input can be referred to as x_i, where i is the dendrite in consideration of n dendrites to the artificial neuron. The information from each dendrite may vary in strength or importance and can be represented as synaptic weights or input weights w_i. As a result, the input to a neuron can be represented as $w_1x_1 + w_2x_2 + ... + w_nx_n$ [Gup13]. The neuron is activated based on whether the input to the neuron satisfies a threshold value θ. Therefore, the threshold function, Eq: 2.1.1, is known as an activation function $a(.)$ (see Sec: 2.1.2). The neuron output can be then accessed through the axon.

$$a(x) = \begin{cases} 1, & \text{if } x \geqslant \theta \\ 0, & \text{if } x < \theta \end{cases} \tag{2.1.1}$$

Using the ANN unit, hierarchical multilayered structures called networks can be created to solve complex problems. ANNs can be categorised as Feedforward Neural

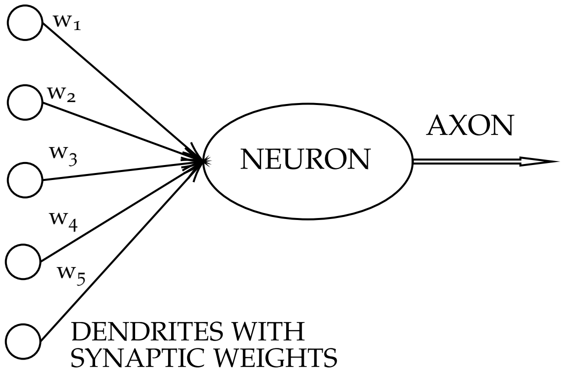


Figure 2.1.1: Artificial Neural Network [Cha18]

Networks or Recurrent Neural Networks (RNN) based on the direction of the flow of information through the network [PP18]. Given that the flow of information is unilateral, the network is referred to as Feedforward Neural Network. When the information from a later layer of the multilayered structure is communicated to a previous layer, forming a loop, the Neural Network is called RNN.

At its inception, one of the significant issues of ANN was the modelling of the input weights for general application-based problems [Roj96], [Gup13]. Such problem-based learning and adaptation of network parameters require a learning algorithm [Roj96]. Consequently, three major learning paradigms were considered: Supervised Learning, Unsupervised Learning, and Reinforcement Learning [Gup13]. For this thesis, we are focused on supervised learning. In supervised learning, we expect the Neural Network to learn a generalised solution to the problem with the help of annotated samples.

2.1.1 *Perceptron and Multilayer Perceptron*

A learning algorithm was first introduced for the perceptron. As shown in Fig: 2.1.2, the perceptron model is designed similar to the ANN unit mentioned in Sec: 2.1. However, it accepts real inputs, $x \in [0, 1] \subset \mathbb{R}$. Further, the weights are real and can be learnt with a numerical algorithm. The inputs are first summated as $w_1x_1 + w_2x_2 + ... + w_5x_5$. Next, the perceptron is activated, provided the summation crosses the threshold value. The threshold activation function is similar to that of an ANN. The output y of the perceptron will be either 0 or 1. As shown in Fig: 2.1.3, the perceptron functions as a classifier by linearly separating the input space [Roj96]. The parameters of the line separating the input space are learned by performing classification trial and error correction on a training set. The training set consists of the inputs x_i to the perceptron and the expected classification y_i^*. The expected classification y_i^* is expressed as 1 to denote a positive example and 0 to denote a negative example. This method of learning is called Perceptron Learning. The classification error is the difference/distance between the prediction y_i of the perceptron and the expected classification y_i^*. Often, the error is quantified as $y_i - y_i^*$. Learning is achieved by updating the weights of the perceptron when the input is wrongly classified. The weight update is achieved by the Eq: 2.1.2. Note that t refers to the iteration, and i refers to a weight out of n weights.

$$w_{(t+1),i} = \begin{cases} w_{t,i} + x, & \text{if } x \text{ belongs to positive example} \\ w_{t,i} - x, & \text{if } x \text{ belongs to a negative example} \end{cases} \tag{2.1.2}$$

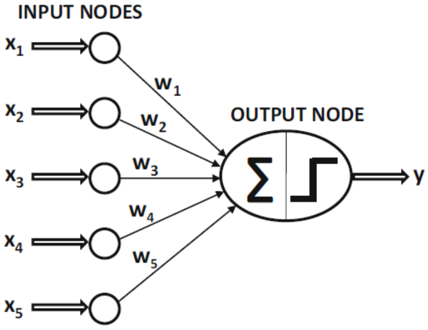


Figure 2.1.2: Perceptron [Roj96]. x_i are the inputs, w_i are the synaptic weights, and y is the output. Σ symbolises summation and the step symbol represents the threshold activation function.

Due to the linearity in classifications, non-linear problems such as the XOR, Fig: 2.1.3, parity, and connectivity problems cannot be solved using a single perceptron [MP88]. However, they can be solved using a multilayer perceptron (MLP). A multilayer perceptron is a fully connected feedforward network, as shown in Fig: 2.1.4. The MLP usually consists of three types of layers; the input layer, the hidden layer and the output layer. The number of neurons in the input layer is dependent on the number of inputs. There can be more than one hidden layer. The number of neurons in the hidden layers depends on the chosen architecture and the problem to be solved. MLP helps to solve problems such as XOR by separating the input space with multiple lines to introduce non-linearity.

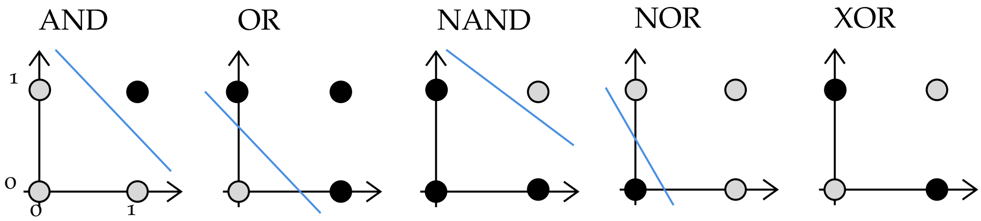


Figure 2.1.3: Boolean operations such as AND, OR, NAND and NOR can be linearly separated using a single perceptron in input space. XOR cannot be linearly separated in input space using a single perceptron [RUD21]

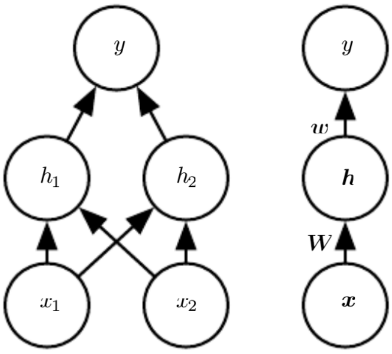


Figure 2.1.4: Multilayer Perceptron [GBC16]. x_i stands for inputs, h_i stands for hidden layer neurons and y represents output. w and W represents the synaptic weights between different layers [GBC16].

Backpropagation algorithms can train MLP. The backpropagation algorithm attempts to find the combination of weights that minimises the classification error [Roj96]. It uses gradient descent to achieve error minimisation (see Sec: 2.1.3). The backpropagation algorithm has two phases. The first phase is called the forward pass. During the forward pass, the MLP performs classification on the input data. Further, the classification error E, Eq: 2.1.3, is calculated from the output [Roj96]. Here, p is the number of data in the training set. The next phase is called the backward pass. In this phase, the error is propagated back into the MLP by calculating the gradient of the error with respect to the weights as $\frac{\partial E}{\partial w_i}$. Next, the weights are updated with the negative gradient (see Sec: 2.1.3). The process is expected to reduce the classification error by re-adjusting the weights of the MLP.

$$E = \frac{1}{2}\sum_{i=1}^{p} \|y_i - y_i^*\|^2 \tag{2.1.3}$$

Gradient descent (see Sec: 2.1.3) is defined only for continuous and differentiable functions [Cha18]. The presence of a threshold activation function in a perceptron implies that the error function is not differentiable. Hence, the threshold activation function in the multilayer perceptron must be replaced with a continuous activation function (see Sec: 2.1.2).

2.1.2 *Activation Function*

The Sigmoid function, $S_c : \mathbb{R} \to (0, 1)$, Eq: 2.1.4, is a continuous function used as an alternative to the threshold function. It is differentiable, as seen in Eq: 2.1.4 and Fig: 2.1.5.

$$a(x) = S_c(x) = \frac{1}{1+e^{-cx}} \qquad a'(x) = S_c'(x) = \frac{e^{-x}}{(1+e^{-x})^2} \tag{2.1.4}$$

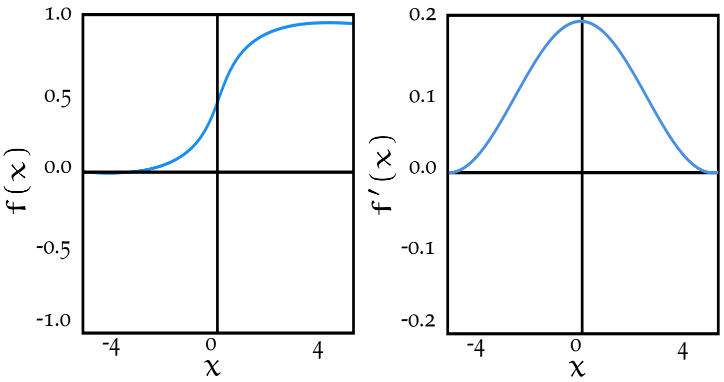


Figure 2.1.5: Sigmoid function and the derivative [AH17]

However, there are a few drawbacks of using the sigmoid function as an activation function. The first drawback is the non-zero centre. When gradient descent is performed on the sigmoid function during backpropagation, the result's sign depends on the neuron output. As a result, the gradient updates could move away from the desired value during some iteration. Thus, the convergence would be slow [AH17]. Furthermore, a zero-centred function is required to maintain the activation within a region of interest to increase convergence speed. Thus, the non-zero center of the sigmoid activation function leads to slow convergence. The next drawback is the issue of vanishing gradient in the multilayered network. As the error propagates back into the early layers of the multilayered network, the gradient tends to become small and eventually vanishes. Consequently, the early layers of the network will fail to learn any useful information about the classification. As shown in Fig: 2.1.4, the vanishing gradient is caused by the saturation of large input values $|x|$ to zero or one [AH17]. Only inputs close to zero (in the x-axis) will have a gradient with a relatively large amplitude, as can be recognised from Fig: 2.1.5.

The hyperbolic tangent was introduced as an activation function, $tanh(x) : \mathbb{R} \to [-1, 1]$ with Eq: 2.1.5, to overcome the drawback of the non-zero center found in the sigmoid activation function. However, the function still suffered from vanishing gradient problem as the amplitude of the derivative was relatively large only for input values close to zero, as seen in Fig: 2.1.6

$$a(x) = tanh(x) = \frac{1-e^{-x}}{1+e^{-x}} \qquad a'(x) = tanh'(x) = 1 - tanh(x)^2 \qquad (2.1.5)$$

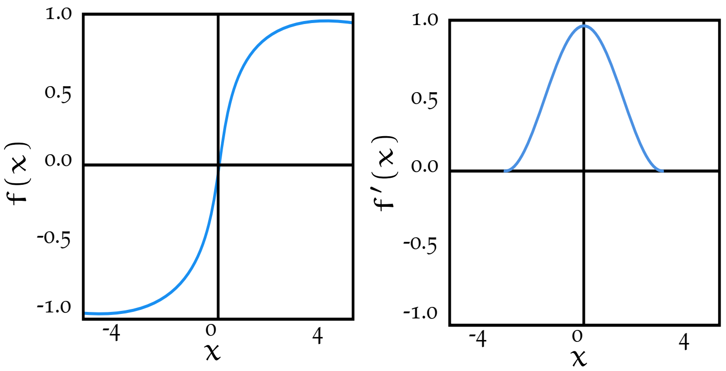


Figure 2.1.6: Hyperbolic tangent and the derivative [AH17]

To mitigate the issue of vanishing gradient, Rectified Linear Unit (ReLU), Eq: 2.1.6, was introduced as an activation function. Since the function does not saturate in the range of $[0, \infty)$, a gradient of larger amplitude can be obtained. As a result, ReLU does not suffer from vanishing gradient [AH17]. However, ReLU may produce dead neurons during training. This property occurs when the weights of a neuron are always negative. Consequently, the neuron is not activated for any sample in the dataset. Although dead neurons may affect the accuracy of the network, removing dead neurons results in a computationally efficient network. Modifications to ReLU, such as leaky ReLU and exponential ReLU, have been identified to overcome dead neurons. However, in practice, ReLU has been an efficient and highly favoured activation function [AH17].

$$a(x) = \mathrm{relu}(x) = \max(0, x) \qquad a'(x) = \mathrm{relu}'(x) = \begin{cases} 0, & \text{if } x < 0 \\ 1, & \text{if } x \geqslant 0 \end{cases} \tag{2.1.6}$$

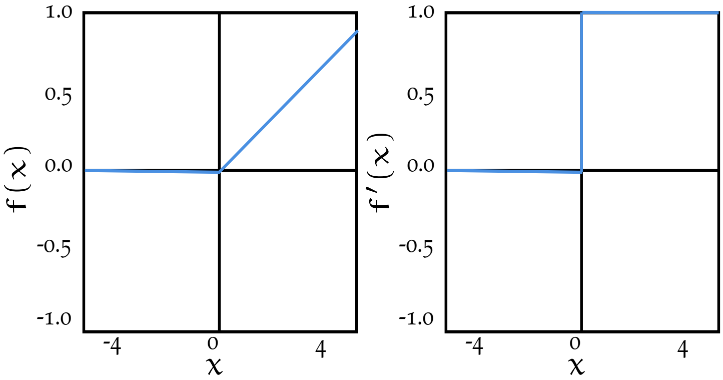


Figure 2.1.7: Rectified Linear Unit and the derivative [AH17]

As can be noticed with the discussion above, choosing an appropriate activation function for the network is an arduous effort in neural network design [Cha18].

2.1.3 *Gradient Descent*

As seen in Sec: 2.1.1, gradient descent is a vital part of the backpropagation algorithm. Gradient descent is an optimisation algorithm, which attempts to find the minimum of the error function E, Eq: 2.1.3 with respect to the weights [Roj96]. The gradient of E can be represented as ∇E as seen in Eq: 2.1.7, where n is the number of weights in the network. Further, weights are updated using the negative of the gradient, as shown in Eq: 2.1.8 [Roj96]. η refers to the learning rate, which defines the step size taken in the minimum gradient direction. In addition, the learning rate can be defined as the constant that controls the extent to which the gradient affects the weight update [Rud16].

$$\nabla E = (\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, ..., \frac{\partial E}{\partial w_n}) \tag{2.1.7}$$

$$w_{(t+1),i} \leftarrow w_{t,i} - \eta \nabla E \tag{2.1.8}$$

There are three types of gradient descent algorithms. The first type, Gradient Descent, computes the gradient of the error function for all the training data and performs a single update. The number of updates will be equal to the number of epochs. Epoch refers to the number of times the neural network sees the complete training dataset. Thus, gradient descent can be slow and inconvenient for large datasets. Further, this method does not allow online updates of the model. The online update refers to the update of the network's weights after the training of a sample or a set of samples.

The second type is called stochastic gradient descent (SGD). Here, an update of the parameters occurs after the gradient of each training sample of the dataset has been calculated. As a result, SGD can perform online learning. Furthermore, the convergence to the minimum is said to be faster [Rud16]. However, the frequent updates imply that the convergence to the minimum is sensitive to noisy data. Although noisy data may prove detrimental to convergence, there exists a possibility that SGD has landed onto a better local minimum. An appropriate learning rate can control the impact of noisy data [Rud16].

The third type, Mini-batch Stochastic Gradient Descent, consists of the best features from gradient descent and SGD. Here, the dataset is split into small batches. An update of the parameters occur after the gradient of the data of a mini-batch has been calculated. Thus, the effect of noisy data, as seen in SGD, is reduced. Furthermore, mini-batch SGD can achieve stable yet fast convergence.

There exists a trade-off between the speed of convergence and the precision of convergence. The appropriate method is chosen based on the amount of data. Various gradient descent optimisation algorithms are present among the Deep Learning community, such as Adagrad, RMSProp, and Adam. These methods attempt to accelerate convergence by introducing momentum and adaptive learning rate during updates [Rud16]. Root Mean Square Propagation (RMSProp) is of interest to this thesis. RMSProp achieves fast convergence by adapting the learning rate of the parameters based on the sign of consecutive gradients.

2.2 CONVOLUTIONAL NEURAL NETWORK

A convolution is a linear mathematical operation where the input signal/image is convolved with a filter, as shown in Fig: 2.2.1. The output of a convolution is referred to as a feature map [GBC16]. Examples of convolution include the Sobel filter for detecting edges in an image and the convolution operation for smoothing signals.

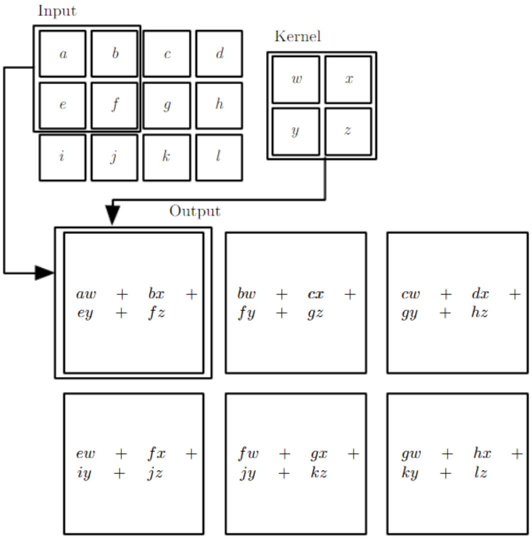


Figure 2.2.1: Example of convolution [GBC16]

The Convolutional Neural Network (CNN) came to popularity when its architecture succeeded at ImageNet challenges in 2011 [Cha18]. A typical CNN consists of convolutional layers, pooling layers and fully connected layers. A convolutional layer consists of convolutional filters that are convolved with the input to the layer. The pooling layer consists of filters that down-sample the feature maps created by the convolutional layer. Two popular pooling layer filters are max-pooling and average-pooling. Different CNN architectures have a varying number of convolutional layers interlaid with pooling layers. Based on the data, some researchers avoid pooling layers in the CNN architecture [NRR+20]. The final layers of the network are usually fully connected. The fully connected layers are the same as the MLP mentioned in Sec: 2.1.1.

Similar to ANN, an activation function is applied on the feature maps created by convolution layers to introduce non-linearity. ReLU is the preferred activation function within the network. CNNs are mainly used for segmentation and classification purposes. Consequently, the final layer of a CNN consists of a softmax activation function, Eq: 2.2.1. z refers to the features from the final fully connected layer. K refers to the number of classes. The result of a softmax activation function layer can be considered the probability that the input belongs to a specific class. As a result, the sum of the output values of each class totals to one. This property can be analysed from the Eq: 2.2.1.

$$\sigma_i(z) = \frac{e^{z_i}}{\Sigma_{j=1}^{K} e^{z_j}} \tag{2.2.1}$$

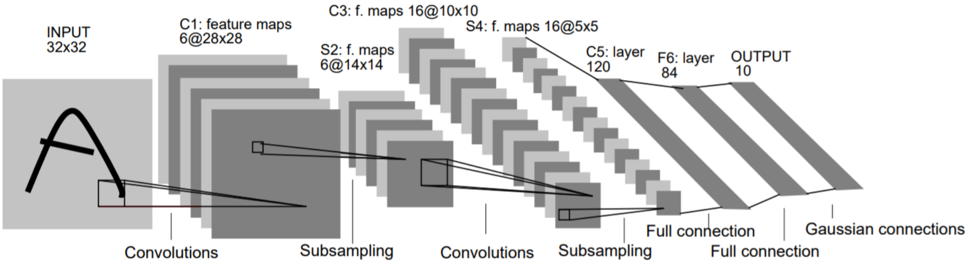


Figure 2.2.2: LeNet architecture [LBBH98]

The Cross-Entropy Loss (CE_{loss}) function is the preferred method to calculate the loss on the outputs of the softmax activation function. The CE_{loss} is also called Log Loss, Eq: 2.2.2. Here, y_i denotes whether the prediction is the same as the label. If the prediction and the label are the same, then $y_i = 1$, and for other labels, the $y_i = 0$. Essentially, the CE_{loss} simplifies the calculation of the derivative of the softmax function $\sigma_i(z)$ as $\sigma_i(z)(1 - \sigma_i(z))$. The simplification can be attributed to their formulation from the probability distribution. As a result, the Cross-Entropy Loss function is preferred when the softmax activation function is in use.

$$CE_{loss} = -\Sigma_{i=1}^{K} y_i \log(\sigma(z)) \tag{2.2.2}$$

A popular example of CNN is the LeNet-5, Fig: 2.2.2. As shown in the figure, the convolutional layers are interlaid with sub-sampling/pooling layers. The final set of layers are fully connected multilayer perceptrons [LBBH98]. This network has the general structure that is followed by most variations of CNN.

A CNN has fewer parameters in comparison to an ANN [Cha18]. This feature of CNN is attributed to the filters. A convolutional layer can have k filters. These filters slide over the input of the layer to create a feature map. Thus, each filter learns a particular feature based on different parts of the input through convolution operation. This process is referred to as parameter sharing and accounts for the reduction in memory consumption [GBC16]. The initial layers of the CNN focus on extracting primitive features of the data. The filters in the later layers attempt to extract complex features based on the primitive features extracted [Cha18].

Though the CNN application examples are more towards image classification, they can be employed on temporal, spatial, or spatio-temporal data [Cha18]. However, the data needs to be shaped into a grid-like topology [GBC16]. For example, time-series data can be considered as a 1-D grid. As a result, a CNN is not constrained by the data's dimensionality or depth/channels [GBC16].

2.2.1 *Convolutional Neural Networks for Time-Series Data*

Time-series data can be defined as recordings of a series of observations in relation to time [ZLC+17]. Examples are Inertial Measurement Unit (IMU) data, electrocardiogram (ECG), and price of stocks. As per [ZLC+17], there are three methods of classification: model-based, distance-based and feature-based. In model-based classification, each class has a model determined from the training set. Further, the time-series data is compared with each of the class models to achieve classification. In the distance-based method, a distance definition has to be identified. The similarity is evaluated with methods such as K-NN or SVM. The third method is the feature-based classification. The main goal is to achieve dimensionality reduction. As a result, a set of features representing the time-series data is extracted. CNN is one such method capable of extracting deep features from raw time-series data [ZLC+17].

There are two types of time-series data: univariate and multivariate [ZLC+14]. In univariate time-series data, only one data point is assigned to a timestamp. An example is the recording of the heartbeat of an individual. Whereas in multivariate time-series data, multiple data points are assigned to a single timestamp. As a result, one can refer to multivariate time-series data as a combination of univariate data with the same timestamp. Because of this feature, the sensors must be synchronised while recording the measurements for a multivariate dataset. An example is the IMU sensor recording, where each timestamp is linked to three data points - Accelerometer, Gyroscope, and Magnetometer measurements. Each of these sensors has axes (x, y and z), referred to as channels.

As mentioned in Sec: 2, this thesis is focused on applying CNNs on multivariate time-series IMU data to perform classification. To perform a convolution, we need to first extract a fixed number of frames from the time-series data. The sliding window approach can be applied to the time-series data to extract frames of fixed temporal length. The temporal length can also be referred to as window size. It is recommended to have overlapping windows [DSGS19]. The overlapping windows ensure that the network will have more data windows to learn from. Fig: 2.2.3 visualises the steps involved in applying CNN on time-series data.

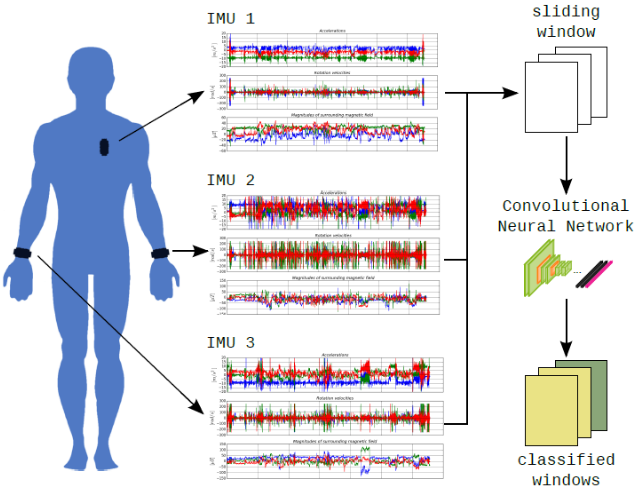


Figure 2.2.3: Example of Time-series Convolution Neural Network process flow [GLR+17]

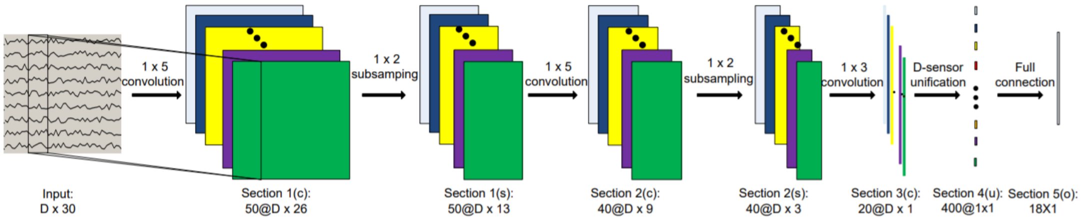


Figure 2.2.4: Example of Time-Series Convolutional Neural Network Architecture[YNS+15]

Fig: 2.2.4 shows an example of a deep CNN for multi-channel time-series data, explored in [YNS+15]. The first step shows the application of a sliding window to extract two-dimensional data of window size r and D channels. Next, the authors convolved the extracted windows with a kernel of size 1x5. This process resulted in 50 feature maps, with each feature map of size Dx26, as shown in the figure. The notation m@Dxn can be read as m feature maps of shape Dxn. The authors use the ReLU activation function. Max-pooling layers were utilised in this network. The features maps of the third convolutional layer, as shown in Fig: 2.2.4, is combined into one dimension in the unification layer. The final layer is a fully connected MLP layer with softmax activation (see Sec: 2.2) for classification [YNS+15].

The architecture can be understood intuitively. The feature maps provide a local representation of the data based on each channel. The fully connected layers give a global view of the data with respect to the local representations.

A convolution can be applied on each sensor separately or all sensors simultaneously [GLR+17]. In the former approach, each sensor data is evaluated separately with independent convolutional filters. Further, fully connected layers attempt to identify the correlations between different sensors. In the latter approach, the same convolutional filters evaluate the data from all the sensors. These filters identify the correlation among sensors. In [GLR+17], an architecture where the data of multiple IMUs are processed separately has been proposed, as shown in Fig: 2.2.5. According to the authors, this method provides robustness against slightly asynchronous IMUs.

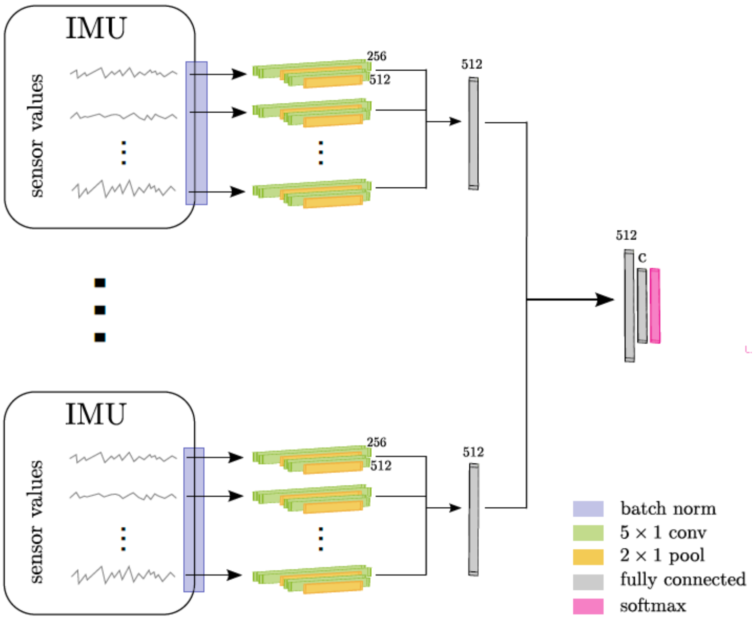
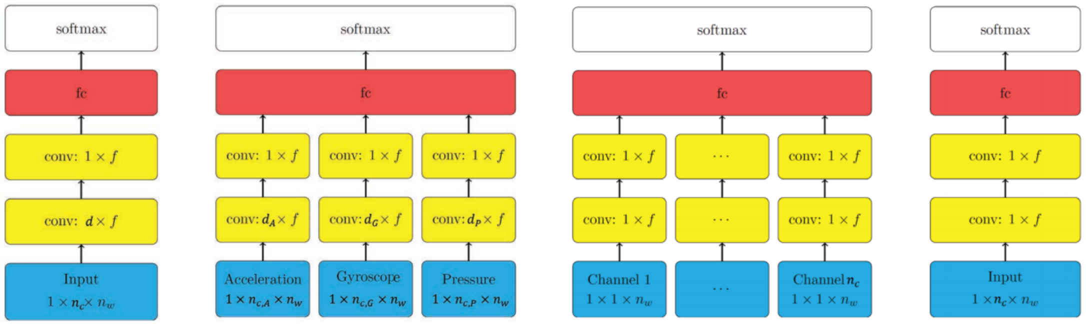


Figure 2.2.5: CNN-IMU [GLR+17]

2.2.2 *Early and Late Fusion Methods*

Human activity recognition (see Sec: 2.5) research often involves multiple sensors placed on the human body. The goal is to find the best method of extracting information from the data to achieve enhanced classification.

According to the authors of [DTC17], two main data fusion methods are used in multiple sensors data: early-fusion and late-fusion. Early-fusion aims to aggregate all the IMU information at the input of the deep CNN. Thus, resulting in high dimensional input data. The early-fusion method cannot process data if one sensor is defective and gives incomplete measurements. However, the late-fusion method trains the network separately with each sensor data. The classification scores generated by the individual networks are then fused to obtained the final classification score. Thus, even if one sensor fails and provides inaccurate classification, the overall classification accuracy is not affected [DTC17].



(a) Early fusion (b) Sensor-based late fusion (c) Channel-based late fusion (d) Shared filters hybrid fusion

Figure 2.2.6: Data Fusion methods [MSR+17]

[DTC17] mentions three levels of abstraction in information fusion. The first level is data level fusion, the second is feature level fusion, and the third is model score fusion. Data fusion can occur at different stages of the network for multivariate data, as given in [MSR+17]. As shown in Fig: 2.2.6, early-fusion, sensor-based late fusion, channel-based late fusion, and shared filters hybrid fusion are the possible multivariate sensor fusion techniques.

The early-fusion method fuses all channels into one dimension in the first convolutional layer of the CNN. As a result, the number of parameters to be learned is small compared to the other fusion methods. Furthermore, this method requires less computation time. In sensor-based late fusion, the data is split based on the sensors.

Each sensor has designated convolutional layers. The feature maps are fused after the end of the convolutional layers. Channel-based late fusion is similar to sensor-based late fusion. However, each channel is handled separately. This method has the highest number of parameters. The final method, shared filters hybrid fusion, uses the same filters for all the channels. Though it looks similar to the early-fusion method, the distinguishing factor is the filter dimensionality and the number of parameters. There are fewer parameters than in the early-fusion method. As per, [MSR+17] and [DTC17], late fusion methods perform better than early-fusion methods as the convolution filters are tailored to a specific sensor or channel. Furthermore, in late-fusion, the filters have a small focus area that enables the extraction of descriptive features [MSR+17]. In conclusion, late-fusion methods are preferred for multi-sensor, multi-channel data types such as IMU.

2.3 RECURRENT NEURAL NETWORK

Recurrent Neural Networks can also be referred to as Auto-Associative or Feedback Networks. As mentioned in Sec: 2.1, RNNs have cyclic internal connections. As a result, RNN supports dynamic-temporal behaviour [PP18] and are applied to sequential data. The cyclic nature supports the storage and reuse of information. As a result, RNNs are said to have memory.

Computational graphs are beneficial to understand the structure of a sequence of computations. In Fig: 2.3.1, the RNN is presented as a computational graph. Let there be a simple RNN structure, with input node x, hidden layer h, and output layer o. L is the loss calculated from the output o of the network and the desired output y for the input x, W is the weight matrix that propagates the hidden layer's output back to the hidden layer for the next iteration. When the network is unfolded for a finite number of iterations, we get the computational graph, Fig: 2.3.1. The unfolded RNN can be considered a feed-forward network for each input in time. Thus, the backpropagation algorithm applied in an RNN is called backpropagation through time (BPTT) [Roj96]. It is to be noted that, because of BPTT, the parameters are shared across the network structure [GBC16].

The RNN BPTT can encounter the vanishing gradient issue. To mitigate the issue, refined variants of RNNs have been identified. As per [GBC16], for practical purposes, gated RNNs have been effective. Two such prominent gated RNNs are the Long-Short Term Memory (LSTM) (see Sec: 2.3.1) and Gated Recurrent Units (GRU) (see Sec: 2.3.2).

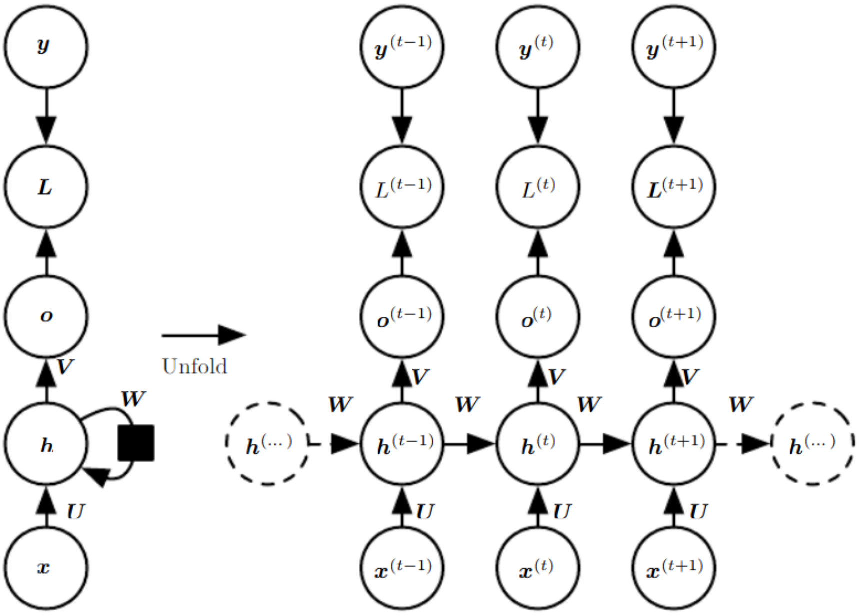


Figure 2.3.1: Unfolding of RNN to form a computational graph [GBC16]

2.3.1 *Long - Short Term Memory*

An LSTM preserves the activations of the hidden nodes from an earlier point in time [CC19]. This property is achieved by introducing self-loops with gated learn-able weights [GBC16]. Thus, the hidden nodes act as memory units that can be updated, erased or read [OR16]. In addition, the gated hidden nodes tackle the vanishing gradient issue.

Here, the gated hidden unit can be called an LSTM cell. An LSTM cell typically consists of a memory cell c_t, an input gate i_t, an output gate o_t, and a forget gate f_t. The three gates use the sigmoid activation function to enable gating [CC19]. The forget gate, Eq: 2.3.2, decides which information must be retained. The decision is dependent on the previous output, current input and the weights associated with the neurons. The input gate, Eq: 2.3.1, decides which new input is relevant to be stored in the memory cell. The output gate, Eq: 2.3.4, is responsible for deciding the output based on the memory cell information. The input and past values bias the output. The memory of past activations, referred to as a hidden value, is denoted as h_{t-1} [OR16]. Let the input from a sequence at a time t be denoted as a_t. Based on Fig: 2.3.2, the following LSTM block equations can be identified for performing an update - Eq: 2.3.1,

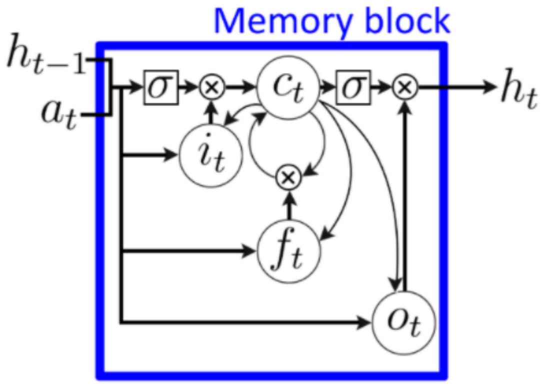


Figure 2.3.2: LSTM cell [OR16]

2.3.2, 2.3.3, 2.3.4, and 2.3.5. The notations in bold represent vectors. The layer notation is avoided to improve readability [OR16].

$$\mathbf{i}_t = \sigma_i(\mathbf{W}_{ai}\mathbf{a}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{W}_{ci}\mathbf{c}_{t-1} + \mathbf{b}_i) \tag{2.3.1}$$

$$\mathbf{f}_t = \sigma_f(\mathbf{W}_{af}\mathbf{a}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{W}_{cf}\mathbf{c}_{t-1} + \mathbf{b}_f) \tag{2.3.2}$$

$$\mathbf{c}_t = f_t\mathbf{c}_{t-1} + \mathbf{i}_t\sigma_c(\mathbf{W}_{ac}\mathbf{a}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1} + \mathbf{b}_c) \tag{2.3.3}$$

$$\mathbf{o}_t = \sigma_o(\mathbf{W}_{ao}\mathbf{a}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{W}_{co}\mathbf{c}_t + \mathbf{b}_o) \tag{2.3.4}$$

$$\mathbf{h}_t = \mathbf{o}_t\sigma_h(\mathbf{c}_t) \tag{2.3.5}$$

Here, **i**, **f**, **o** and **c** represent the input gate, forget gate, output gate and cell activation vectors. As mentioned earlier, the gates are controlled by sigmoid activations (σ). As a result, they are limited to the output value of zero or one. Thus, the gating signals need to be expressed as vector equations [DS17]. These vectors have the same size as vector **h**. Weight matrices are $\mathbf{W}_{ai}, \mathbf{W}_{hi}, \mathbf{W}_{ci}, \mathbf{W}_{af}, \mathbf{W}_{hf}, \mathbf{W}_{cf}, \mathbf{W}_{ac}, \mathbf{W}_{hc}, \mathbf{W}_{ao}, \mathbf{W}_{ho},$

and $\mathbf{W}_{co}$. Their subscripts represent the nodes that they connect and direction of flow. For example, $\mathbf{W}_{ai}$ represents the flow of information from the input to the input gate matrix. $\mathbf{b}_i, \mathbf{b}_f, \mathbf{b}_c$, and $\mathbf{b}_o$ are bias vectors [OR16].

LSTMs are a popular solution for applications with real-world sequences such as handwriting recognition, speech recognition, and machine translation. [GBC16]

2.3.2 *Gated Recurrent Units*

Though similar to LSTMs, GRUs are simple and fast to train. These properties can be credited to the reduced number of parameters by eliminating the memory cell found in the LSTM cell [Roj96]. The gates in GRU are called reset gate and update gate. The update gate replaces the functionality of the input and the forget gate found in the LSTMs [Roj96]. The update gate maintains information from the past that proves helpful for classification. The reset gate decides how much of the past information is retained [GRS+20]. As a result, it replaces the forget gate and output gate functionality of the LSTM [Roj96]. Unlike the LSTM, a GRU does not have a separate internal memory. Essentially, the GRU, similar to the LSTM, is based on the idea of partially resetting hidden states.

The structure of the GRU cell is as shown in Fig: 2.3.3. The reset gate and update gate are represented by the Eq: 2.3.6 and 2.3.7, respectively. Eq: 2.3.8 and 2.3.9 present the previous activation's hidden states and the current hidden activations.

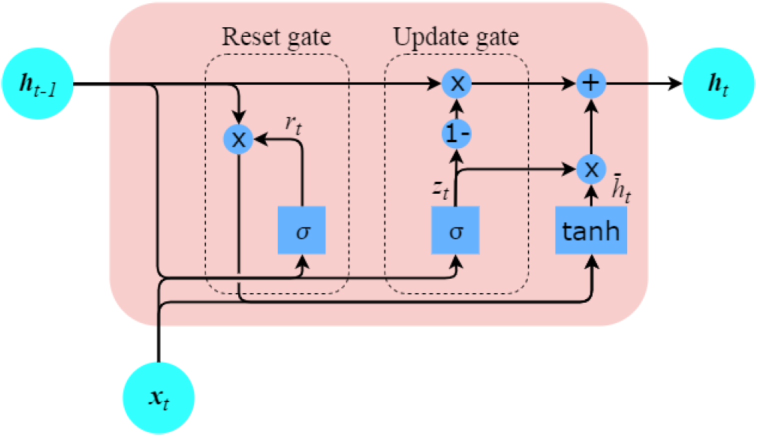


Figure 2.3.3: Gated Recurrent Unit (GRU) cell with marked reset gate and update gate [GRS+20]

$$r_t = \sigma(W_{xr}x_t + W_{hr}h_{t-1} + b_r) \tag{2.3.6}$$

$$z_t = \sigma(W_{xz}x_t + W_{hz}h_{t-1} + b_z) \tag{2.3.7}$$

$$\tilde{h}_t = tanh(W_{xh}x_t + W_{hh}(r_t \odot h_{t-1}) + b_h) \tag{2.3.8}$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t \tag{2.3.9}$$

Here, r stands for reset gate, and z represents update gate. $\odot$ stands for element-wise multiplication. $\tilde{h}_t$ represents previous activation's hidden states, while h represents current hidden activations. $\sigma(.)$ stands for sigmoid activation function. Changes to the gating equation can create variations of GRU [DS17].

2.4 BIOMETRIC

Any physical and/or behavioural quality that can help uniquely identify an individual is called a biometric. When these characteristics are used for automatic recognition, it is referred to as Biometric recognition [JRP04]. Biometric characteristics are human physiological and/or behavioural characteristics that are universal, distinct, permanent, and collectable. Based on these requirements, a few popular biometrics are DNA, face, finger-print, gait, iris, retina, signature, and voice. For robust biometric recognition systems, performance, acceptability, and circumvention need to be considered.

At its core, a biometric system is a pattern recognition system. As a result, it follows a given sequence of functions. Firstly, few samples of the biometric data have to be collected from the individual. Next, features have to be extracted from the samples. A template is created for the individual based on the features extracted from the samples. Templates of multiple individuals will be stored in the template database. Given a test biometric sample for identification, features will be extracted from the test sample and be compared against the template database. Therefore, a biometric system would have four modules: sensor module, feature extraction module, matched module, and the system database module. In case of verification, the biometric system tries to confirm an identity claim. Whereas in identification, the system attempts to find a matching template from the database [JRP04].

For this thesis, gait as a biometric is of interest. Gait is the cyclic movement of our legs while walking [BL05]. It is a spatio-temporal biometric. Though gait may not be very distinct in all cases, it supports verification applications where the security is insignificant. It is difficult to gather a gait template covering footwear, terrain, fatigue, and injury variation [BL05]. Furthermore, gait's long-term relevance as a biometric is questionable because of body weight fluctuations, injuries and changes in the sense of balance. However, gait data can be acquired as images, videos, optical Motion Capture (oMoCap), electromyography, or IMU data. As a result, it is one of the most collectable biometric alongside signature, face thermogram, and hand geometry [JRP04].

2.5 HUMAN ACTIVITY RECOGNITION

Physical activity can be defined as any movement produced by the skeletal muscles that lead to energy expenditure [CPC85]. The process of recognising human activity through computer vision methods can be referred to as HAR. Activities may include human locomotion such as walking and jogging or daily activities such as cooking and cleaning. The activities may include human interaction with the surroundings. Human to human interaction can be handshakes or boxing, while brushing teeth, cutting vegetables, and picking up a box can be examples of human to object interaction.

The activities can be considered as classes. There are two types of class variations: intra and inter-class variation. When the actors perform the same activity with different mannerisms, it leads to intra-class variations. Often in real-world scenarios, intra-class variations are significant [KF18]. In specific scenarios, two different activities may have the same body movements. For example, drinking or eating has the same hand movement as smoking. This type of variation is referred to as inter-class variation. Inter and intra-class variations can often cause confusion while recognising the activity with computer vision [KF18].

Videos, motion capture systems, and IMUs can gather human activity data. Often, the data is labelled with the activity to support supervised learning methods [NRR+20]. The labelling process, also called the annotation process of the HAR dataset, is an arduous process. Some popular labelled human activity datasets are Opportunity [CSC+13], PAMAP2 [RS12b], KTH dataset [SLC04], UT Interaction dataset [RA09], and UCF sports dataset [RAS08].

2.5.1 *Attribute Representation*

Attributes can provide semantic or discriminative information about the data [FEHF09]. For example, the semantic description could be about the scenes or objects in an image or the object's shape, colour, or size for object detection [RF18].

Attribute learning supports intra and inter-class classifications [FEHF09]. As mentioned in Sec: 2.5, inter and intra-class misclassifications are common in HAR. Thus, by representing the classes with attributes, one can verify the correctness of the classification with the help of attribute representation. For example, the activities — pushing a cart and gait cycle — can be differentiated by analysing the hand's position. Given that the hand attribute is positive, the class pushing cart can be selected. Further, given a case where the class is not identifiable or unique, attributes could help with approximate classification. According to [RF18], attributes can reduce annotation requirements in unbalanced or large datasets. In addition, attribute representation supports zero-shot and transfer learning.

The authors in [RF18] implemented an attrCNN-LSTM, attrCNN, and attrCNN-IMU networks to learn HAR attributes, with the final layer consisting of a sigmoid activation function Eq: 2.1.4. The number of neurons in the final layer depends on the number of attributes. Here, the attributes are represented as a vector of 0s and 1s. Consequently, the Binary Cross-Entropy loss (BCE_{loss}) function Eq: 2.5.1 is preferred to calculate the loss [NRR+20]. Here, n is the number of samples, y is the label, and p is the probability of positive prediction.

$$BCE = -\frac{1}{n}\sum_{i=1}^{n} -y^{i}log(p^{i}) - (1-y^{i})log(1-p^{i}) \tag{2.5.1}$$

3 RELATED WORK

Chapter: 1 and Chapter: 2 discuss IMU sensor datasets for HAR (see Sec: 2.5) research. HAR has found application in logistics, as mentioned in Chapter: 2 and in medical care services; for example, Ambient Assisted Living (AAL) and early detection and monitoring of diseases such as Parkinson's. Further, we discussed gait as a biometric and the ease of gait data acquisition. In addition, we discussed the possibility of subjects' identity-based classification using HAR datasets. Having covered the elementary topics related to this thesis, we now answer the question, what are the prominent researches and conclusions derived in gait-based person identification and soft biometrics using vision and IMU data.

Firstly, we discuss the researches on gait as a biometric (Sec: 3.1). This section highlights the methods for extracting gait data and their limitations. Using IMU data of human body movement, we can either classify individuals based on their soft biometrics (see Sec: 3.2.1) such as age and gender, or classify the IMU data based on subject identity (see Sec: 3.2.2). Thus, in Sec: 3.2, we discuss how IMU data is used to determine soft biometrics and person identity. Body movements have been used to create attribute representations of HAR activities (see Sec: 2.5.1). We explore the research works of attribute representation in HAR in Sec: 3.4. It is expected that using guidelines obtained from attribute representation in HAR, soft biometrics can be modelled as attribute representation for person identification. Concluding the chapter in Sec: 3.5, we discuss the key points, observations and inferences, limitations and future research ideas.

3.1 GAIT AS A BIOMETRIC

Gait as a biometric is in its nascent stage due to complications caused by variation in terrain, footwear, fatigue, and injury [BL05]. Gait has both coordinated and cyclic nature of motion, as shown in Fig: 3.1.1. When the foot is swung forward, as shown in the image, the leg is said to be in the swing phase. Otherwise, when the foot is in contact with the ground, the pose is referred to as stance [GRS+20]. The gait cycle is a combination of swing and stance phases. Gait duration can be split into a half cycle - step or a full cycle - stride. Step refers to the heel strike of one leg followed by the other (swing). Stride refers to the heel strike of one leg, followed by the heel strike of the same leg (swing and stance) [GRS+20].

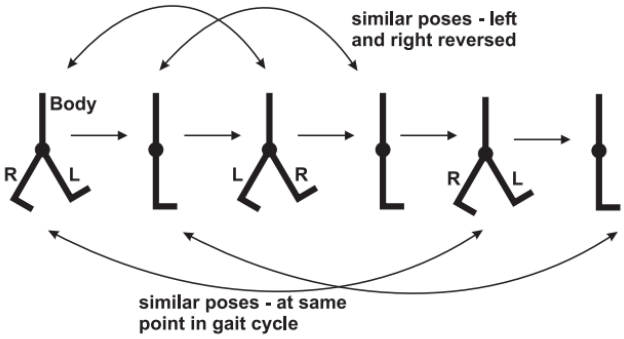

similar poses - left and right reversed
Body
R
L
L
R
R
L
similar poses - at same point in gait cycle

Figure 3.1.1: Gait cycle [BL05]. R refers to right leg and L refers to left leg.

In their study to understand how humans observe gait motion, psychologists found that humans look for frequency entrainment, phase locking, and physical plausibility [BL05]. To elaborate, given images of dots arranged in a particular shape and a particular movement, as shown in Fig: 3.1.2, humans tend to analyse whether there is a frequency found in the movement of the dots reflecting the movement of different parts of the body, i.e., hands and legs, to confirm that the dot image represents a human gait motion. Given a common frequency, the feature is referred to as frequency entrainment. Further, we analyse the pose formed by the position of the dots. If the poses of the dot image repeat cyclically, then the dot image sequence is said to be phase-locked. Humans look for phase-locked movements to identify human gait motion. Physical plausibility indicates whether the cyclic motion viewed is physically possible by the human body. Humans comprehend physical plausibility based on their physical capabilities and previous experiences. In computer vision, physical plausibility cannot be a criterion for gait recognition. However, frequency and phase are features that can be used for recognition. As per [AC99], machines interpret human motion based on: motion analysis, including human body parts, movement tracking from a single view or multiple camera perspectives, and human activity recognition from an image sequence. At its core, these methods consider oscillations of the body shape, joint trajectory, self-similarity, and pixels, to obtain frequency and phase [BL05].

Based on the gait acquisition method, there are three categories, machine vision, floor sensor, and wearable sensors. The authors of [Gaf07] have used the term machine vision to denote video-camera based data acquisition method. The advantage of

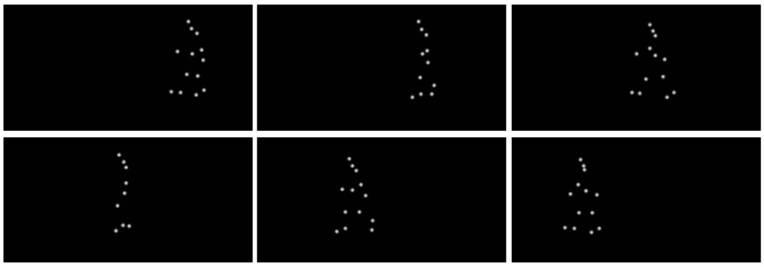


Figure 3.1.2: Experiment conducted by psychologists to understand how humans perceive gait. The dot structure shows a human movement [BL05].

machine vision based gait acquisition is the possibility to acquire data from large distances without requiring the subject's cooperation and awareness [SJAS18]. Thus, the method is of interest in surveillance and forensics [Gaf07], [SJAS18]. According to [SJAS18], vision-based data can be classified as marker-based or marker-free. Marker-free methods refer to video-camera based data acquisition. The Optical Motion Capture (OMoCap) system is an example of a marker-based data acquisition system. Here, the subjects have reflective markers or sensors at particular body joints to facilitate motion capture. Marker-based data acquisition systems are used for clinical gait analysis.

In the floor sensor-based method, force plates are installed to measure gait-related features such as heel strike, stride, and cadence [Gaf07]. Though this method is non-obtrusive and can support localisation within a building, it is limited to laboratory environments. It is not easy to maintain these sensors in home environments due to their sensitive nature. As mentioned in [Yun11], the circuits tend to get damaged when exposed to water.

The final method of gait data acquisition mentioned in [Gaf07] is wearable sensors; for example, smartphones and IMUs. These sensors can be placed on various locations of the human body, for example, on the waist, wrist, chest, and lower part of the legs. Authors in [Gaf07] suggest that wearable sensor-based gait can be utilised for authentication on mobile devices that store financial or private data.

There are two main approaches for interpreting and extracting gait features. They are model-based and model-free approaches. In the model-free approach, the shape and motion of the silhouette extracted from the videos after segmentation are used to obtain gait features. Consequently, the method is also referred to as holistic/appearance-based approach [SJAS18]. In a model-based approach, the subject's physical model is designed from measurable body components, such as joint angle patterns, joint trajectories, height and stride length, prior to data acquisition. The model can be either a 3-D model or a structural model. In the structural model, the geometrical and structural properties of the subject, such as gait period, step length, and stride length, are used for gait recognition. The 3-D model is designed from hip and thigh rotation patterns, motion trajectories, and orientation of limps. The presence of the prior model signifies that the model-based approach can be view-invariant, scale-invariant and unaffected by noise. Consequently, a model-based approach is robust and preferred for practical applications [SJAS18].

The model-free approach can be sub-categorised as statistical and spatio-temporal methods. In the statistical method, the silhouette's shape and motion patterns are used for recognition [SJAS18]; for example, the velocity moments is used to describe the silhouette's motion features and Gait Energy Image (GEI) can be used to analyse the silhouette shape. A spatio-temporal method uses space and time information from video sequences to perform gait recognition [SJAS18]. Space information refers to the appearance of the subject based on clothing variations, while time information refers to the dynamic features of gait such as walking speed. In this method, the motion features, e.g., speed, stride length and stance duration, are extracted and used for recognition using methods such as the Bayesian decision approach. The model-free spatio-temporal approach is affected by camera orientation and appearance variations. However, this method is favourable because of its low computational complexity [SJAS18].

Analysis of gait in itself has various applications. Athletic performance analysis, man-machine interfaces, and content-based image storage and retrieval are few areas of interest [AC99]. Further, gait is used for soft biometric (see Sec: 3.2.1) and clinical analysis. Examples of soft biometrics are age and gender classification or estimation. In clinical analysis, qualitative measures such as cadence, gait speed, and step length are used to analyse ageing and to diagnose diseases. Research on patients with rheumatoid arthritis and Parkinson's disease has shown that gait analysis is an effective diagnostic technique [SJAS18].

3.1.1 *Vision-Based Person Identification*

Vision-based person identification is desirable for surveillance applications. Authors in [SJAS18] mention that gait features can be extracted without the subject's cooperation from a distance of $10\mathrm{m}$ or more. Further, gait feature extraction is possible with low-resolution video sequences.

One of the primary methods considered for identification is silhouette analysis. Silhouettes are affected by shape information. As a result, clothing plays a massive role in same-subject identification [WTNH03]. In addition, silhouette analysis is affected by the viewing angle of the camera. When using a single video camera, the silhouette analysis is limited to one camera viewing angle. However, based on the subject's position with respect to the video camera, the viewing angle can vary. Thus, to facilitate silhouette analysis, a method to generalise the viewing angle is required. The authors in [WTNH03] suggested a multi-camera-based tracking system to overcome the issue.

In [HB05], a method called Gait Energy Image (GEI) was proposed to visualize human motion in a single template or reference image rather than a sequence of templates or reference images. The GEI is the time-normalised accumulative energy image of space-normalised silhouette images of human walking during a complete gait cycle. This spatio-temporal representation considers statistical gait features, such as frequency and phase, from real and synthetic references to generate training templates. The synthetic references are created by distorting the real frequency and phase values. When individual recognition is to be performed, the individual's gait is converted into a GEI and compared with the training template. Benefits of GEI include preservation of template storage memory, reduced computation time, and reduced sensitivity to silhouette noise in each frame. However, GEI still suffered from the issues of multiple viewing angles and missing frames caused by obtrusion.

Another method for gait feature extraction is non-linear machine learning, as shown in [EA07]. Firstly, we extract the binarised silhouette of the subject for few frames. Then, we take four projections of the silhouette and find the correlation of the projections of the frame with the projections of its consecutive frames. Further, the correlated outputs are normalised. Next, a symmetric average filter is applied to smooth the normalised outputs. This output is a 2-D image, referred to as a gait pattern. The gait frequency can be obtained through auto-correlation of the gait pattern. The gait patterns have to be created for the subjects gait motion at different speeds. These frames are used for the training procedure. However, the gait patterns are transformed to the frequency domain to achieve translation invariance before applying the Principal Component Analysis (PCA). PCA extracts the gait features from the gait patterns. These gait features can be used as a template to compare with the features of the test gait data [EA07].

[LJZ09] is a survey on the different gait recognition methods and the future work required in vision-based gait recognition. The authors pointed out that the gait datasets had a limited number of subjects leading to difficulty in generalisation. Most datasets have only 200 subjects. As a result, the performance evaluation is restricted. Furthermore, most datasets consider a single moving subject in the frame. Therefore, evaluation of the methods in real scenarios does not take place. Based on the research trend identified, the authors concluded that 3-D prior modelling of the subject's physical characteristics and multiple-camera employment would be the future research direction. Furthermore, research in the field of spatio-temporal gait features was recommended. Finally, the authors emphasised the need for a more extensive gait database with complex environments.

The authors of [HWZ+12] proposed a method based on optical flow for gait recognition and tracking of subject gait in video-based surveillance. In silhouette-based model-free approaches, background subtraction is the first step towards extracting the gait features. However, extracting silhouettes when the background is cluttered can be tricky. Optical flow method does not require background subtraction. Consequently, optical flow methods are expected to help in gait feature extraction in the case of cluttered background. Local Binary Pattern (LBP) flow was utilised to encode the optical flow information. Further, each individual was assigned a single Hidden Markov Model (HMM) representation of gait dynamics. The recognition could then be achieved either through a model-based approach or an exemplar-based approach. The model-based approach uses a training set to create a statistical model for each subject present in the training set. Further, the method compares the likelihood with the test data [HWZ+12]. In contrast, the exemplar method retains each training data and considers the distance measure between each training data from the set to the test data. Averaging and dynamic time wrapping (DTW) methods are used to find the distance measure. Consequently, the exemplar method is expensive in time and storage. Thus, the model-based approach was said to be efficient [HWZ+12].

The majority of the works in vision-based recognition were focused on using HMM or PCA for classification. However, [WBR16] considered 3-D deep CNN for gait recognition with dataset consisting of multiple views. Thus, implementing the recommendation of multiple camera employment in [LJZ09]. The 3-D deep CNN architecture uses 3x3x3 convolutional filters in each of the seven convolutional layers. Consequently, detection of movements in all directions is expected to be possible [WBR16]. To overcome the challenge of viewing angle, colour, and variation in walking conditions, the 3-D deep CNN was trained with competitive datasets, such as CMU Motion of Body (MoBo), USF Gait-based Human ID Challenge and Casia-B. The datasets had instances of varying clothing, walking speed, and multiple-viewing angles. The authors provided optical flow data as an input along with the grey-scale image to ensure colour invariance. The generalisation of gait features across the different viewing angles was achieved with a CNN. However, the authors proposed experimentation with larger datasets to ensure over-fitting has not occurred.

3.2 IMU BASED CLASSIFICATION

3.2.1 *Soft Biometrics*

In comparison to video-based person identification, IMU-based person identification is a relatively new research area. Interest in this area began with the understanding that gait analysis provides effective disease diagnosis, as mentioned in Sec: 3.1.

Authors in [SNM+12] created a dataset using a biometric suit with wearable sensors, named Intelligent Gait Oscillation Detector (IGOD). The dataset is expected to support gait parameter study, which could be further extended to person identification and walking troubles detection. The suit measured oscillations from eight joints of the human body, specifically knees, hips, elbows, and shoulders. The study analysed the variation in gait oscillation to gait speed as well as gender. The authors identified that analysing the oscillations could determine soft biometric features such as height range and gender.

The focus of [RVKW15] is purely on soft biometrics. 26 subjects were classified based on gender, age, and height using dataset obtained from a single IMU sensor. The IMU data was initially segmented into strides and classified with Random Forest (RF). Here, RF method was preferred based on the results provided by the previous works in this field. An exciting aspect of this research is the dataset. The dataset consists of readings taken with the subject walking on hard surfaces with shoes and without shoes and soft surfaces without shoes. Furthermore, while performing soft biometric classification, the authors brought in restrictions to the training set based on age and height. The restrictions are referred to as sub-groups in Table: 3.2.1. The thresholds selected for creating sub-groups ensures a balanced population in all classes. The authors concluded that a single step recorded from smartphones and smartwatches can be used to reveal personal information such as gender, height and age [RVKW15].

3.2.2 *Person Identification*

IMU sensors support the application of gait-based person identification and verification in human-robot interaction. The authors of [ZKL+13] explored the possibilities of identifying the robot interaction partner through the gait data obtained from a single wearable sensor attached to the pelvis. Bayes classifier was applied to classify the individual. However, the dataset created by the authors consisted of gait data from 20 participants from three IMU sensors. The IMU sensors were placed at the pelvis, right ankle and thorax. [ZKL+13] uses the data from the IMU placed at the pelvis for stride

Classifications	Gender	Age	Height (Hgt in cm)
Classes	Male Female	Age <40 40 <Age <50 50 ⩽ Age	Hgt <170 170 <Hgt <180 180 ⩽ Hgt
Sub-Groups	Male: Age ⩽ 40 Age >40 Female: Age ⩽ 50 Age >50		Male: Hgt ⩽ 180 Hgt >180 Female: Hgt ⩽ 170 Hgt >170

Table 3.2.1: Groups and sub-groups of classifications considered in [RVKW15]

segmentation and person identification. The IMU placed on the ankle was used to validate the stride segmentation. The proposed identification first splits the raw data into a single stride and performs classification using Bayes. The stride segmentation was given high priority in this research. To elaborate, the authors experimented with stride segmentation on data of subjects walking on a straight path, big circular path and small circle. It was found that the stride segmentation algorithm is ineffective on the walking data over a small circular path. Instead of using a single stride for testing, the authors suggested including a voting system by considering three strides. A classification accuracy of 99.3% was achieved on the dataset with the proposed method.

A perspective on cycle extraction, spectro-temporal 2-D expansion and representation of gait cycles, deep CNN, and multi-layer sensor fusion for person identification using gait was provided by authors in [DTC17]. IMU data was collected from five sensors placed on the human body, namely the chest, lower back, right-hand wrist, right knee, and right ankle. As part of data pre-processing, the data was passed through a Butterworth bandpass filter to extract the frequency range of $0.5 - 3.5$Hz. Further, the orientation invariance was achieved by considering the square root of the squared sum of the value along each axis. First, the ankle sensor was used to mark the gait cycles based on peak values to extract gait cycles. Then, the same markers are applied to all sensor data to extract the gait cycles. Next, the input data is mapped onto a time-frequency space using time-frequency distribution. Further, the instantaneous frequency is estimated. The instantaneous frequency can be identified on the time-frequency representation as ridges [DTC17]. Further, the time-frequency representations are given as input to the deep CNN to perform gait classification. The authors used the time-frequency representation to avoid manual feature extraction. Manual feature extraction is said to be prone to error and subjectivity [DTC17]. The authors recommend using the late fusion method of deep CNN, stating better accuracy, given that any of the IMU turns out to be defective or noisy. [DTC17] achieved 91% subject identification accuracy on the proposed method.

[EBL18] provides a different perspective. The authors considered a dataset of 20 daily human activities instead of restricting to gait data; for example, daily activities include stirring, washing dishes, and office-work activities. The dataset was created from six IMU sensors, placed on both wrists, dominant upper arm, thigh, chest and ankle positions of 18 subjects. The experiments were performed on classifiers such as Support Vector Machines (SVM), K-Nearest Neighbour (KNN), Neural Network (NN), Decision Tree and their types. The authors have listed the classifiers that provided the best results in [EBL18]. Unlike in [DTC17], the focus remained on statistical features. Features such as mean, standard deviations, and magnitude were extracted from the data by considering segments extracted using a sliding window approach. The sliding window considered a window size of 2 seconds with 50% overlap. The authors identified that sedentary activities had a higher classification rate. In addition, a combination of accelerometer, magnetometer, and gyroscope provided better results than the sensors individually. A striking result from the experiments is a depreciating relationship between classification success and the number of sensors. Another conclusion is that all subjects are not equally identifiable.

A simple Neural Network was used for classification in [CDHG20]. The authors collected IMU data from the wrist position of 50 subjects while performing 100 seconds of natural walking tests. Gait motions performed by individuals can be very similar. However, variations can be found in the gait characteristics due to the individual's height, weight, arm length, and personal habits [CDHG20]. The authors analysed the characteristics of rotation angle difference, location difference, and the inertial difference of the wrist to obtain a 24-dimensional human wrist gait feature data model [CDHG20]. Next, the selected 24 dimensional feature values are trained on a Neural Network. The method achieved an accuracy rate of 97.65% and was recommended for identity and security authentications [CDHG20].

Another instance where Neural Network was considered is [GRS$^+$20]. Unlike [CDHG20], [GRS$^+$20] does not try to extract features from the data. Instead, the input data is split into either stride or gait data. Further, the split windows are fed to four DNNs: Gated Recurrent Unit (GRU), CuDNNGRU, Long-Short Term Memory (LSTM), and CuDNNLSTM. CuDNN refers to Nvidia's GPU accelerated library for implementing DNNs. GRU and LSTM were implemented on Tensorflow. The data was obtained from the IMU and smartphone sensors placed on the chest of 86 randomly selected subjects. The data were recorded at a sampling rate of 75Hz. The subjects performed gait activity on various surfaces, for example, carpet, grass, tiles and asphalt. The hyperparameters were set for all the models at learning rate 0.001, batch size 32 and epoch 30. The networks were trained with categorical cross-entropy loss function and adam optimizer. The authors found that CuDNNGRU performed better than the other models. The model achieved 87.15% and 86.23% accuracy on step and stride data, respectively. The CNN network achieved 69.05% and 76.04% on step and stride data, respectively. The re-identification was further evaluated by considering age and gender restrictions. To elaborate, the subjects were grouped based on their age and gender. Data of each group were used to train networks. The networks achieved high classification accuracies.

One point of interest is that the step data performed better than the stride data [GRS+20]. A step is defined as the heel strike of one foot followed by the heel strike of the other foot. Stride constitutes two heel strikes from one foot. Previous research concluded that during a normal human walk, the step frequency is between 1-2Hz. Stride frequency is expected to be twice the step frequency. Contrary to the network size paradigm in HAR, the authors noted that increasing the network size and epoch number provided better results. The authors identified the optimal network parameters as 512 neurons, 0.001 learning rate, dropout of 0.5 and epoch 30. A point of interest is that the authors have raised privacy concerns, considering the high accuracy with which identification can be achieved with IMU data.

3.3 IMU BASED ACTIVITY RECOGNITION

Considering how activity recognition applications function successfully using IMU data obtained from smartphones and smart-watches, we can say that stable versions of HAR classifiers are available for applications. Datasets such as Opportunity [CSC+13] and PAMAP2 [RS12b] are now considered trademark datasets in the field of HAR. [HHP16] has utilised the datasets mentioned above and the Daphnet Gait dataset as a trademark to conduct verifiable experimentation. Architectures such as CNNs, RNNs such as LSTMs, and bi-directional LSTMs were experimented upon using these datasets. The authors tabulated the percentage effect of architecture, learning rates, regularisation and iterations on the overall model variance was evaluated. Experiments were conducted on CNN with learning rates ranging from $0.001 - 0.00001$. The authors found that learning rates had an overall effect of about $25 - 50\%$ of model variances. Interestingly, the specific influence of hyperparameters can be found on the dataset; for example, the experiments suggest that PAMAP2 requires correct learning parameters, while, Opportunity is dependent on the architecture. Through experiments, authors identified that for HAR, usage of shallow networks is recommended.

The network of interest for this thesis is the CNN-IMU network. It was conceptualised by [GLR+17] and then further explored by [MGF+18]. CNN-IMU is ideal for processing time-series data from multiple IMUs, as mentioned in Sec: 2.2.1. It follows the late-fusion (see Sec: 2.2.2) method. Initially, each sensor has a branch of convolutional and max-pooling layers (depending on the dataset). These layers extract the temporal-local features of the given data. Further, through a fully connected MLP layer, the local information from the parallel branches is collected and processed to create a global representation. Finally, this information is used for classification using a softmax layer. As a result, CNN-IMU is robust against slightly asynchronous data and is more descriptive. As a proof of concept, the architecture was tested on Opportunity and PAMAP2 datasets. The results obtained were in favour of the CNN-IMU network, in that the architecture outperformed the state-of-the-art on the mentioned datasets [MGF+18].

3.4 HUMAN ACTIVITY-BASED ATTRIBUTE REPRESENTATION

The prospect of utilising attribute representation to denote HAR was explored by authors in [RF18]. Since body movements have particular patterns, the authors reasoned that representing the action classes by the details of coarse human actions would be beneficial; for example, the class handedness can be further explained as left hand, right hand, and both. Furthermore, this method could tackle complications caused by inter-class and intra-class variability and class imbalance, as seen in Sec: 2.5.1. The authors tested the method on three deep network architectures: CNN, deepConvLSTM, and CNN-IMU. The final softmax layer of these networks was replaced with a sigmoid layer to support attribute representation. The sigmoid layer is preferred for attribute representation as each attribute is represented with a binary value. Meaning, each attribute representation can be represented as a string of zeros and ones.

[RSH+18] explains the procedure to create an attribute representation for HAR. Here, the authors have presented attribute representation for HAR in a logistics environment. The authors recommend creating representations that would be semantically understandable for humans. Furthermore, a representation that can be utilised on different HAR datasets was said to be desirable. The authors found that attribute representation could perform at par or even better than classes. Furthermore, representations with a lower number of attributes were found to have a slightly better performance. In addition, it was found that given a semantic relation between the attributes and activities, the performance improved further.

3.5 DISCUSSION

Though person identification with gait information started with vision-based gait surveillance, the lack of an appropriate dataset impeded its advancements. Furthermore, obtrusion and occlusion are dominant hindrances. The presence of smartphones and smartwatches has seen exceptional growth in IMU data compared to vision data. The advancements have facilitated the development of HAR applications with greater accuracy. Similarly, IMU data is expected to boost person identification applications. As a result, the past years have seen more research in this field.

As can be noticed, except for [EBL18], almost all the previous works are focused on gait data. However, data from smartphones and smartwatches are usually a combination of varied activities. It would be ideal to analyse person identification accuracy on datasets that have features of data from smartphones and smartwatches. Experimentation on these datasets will help to observe how varied activities improve or deteriorate person identification. In addition, the privacy concerns associated with the data can be addressed.

The previous works in this field were focused on extracting hand-modelled features for training classifiers such as SVM, KNN and RF. As discussed in [DTC17], hand-modelled features could be erroneous and subjective. As a result, classifiers such as CNN are desirable. The convolutional layers of the CNN are capable of extracting relevant features from the input data to facilitate appropriate classification.

An exception to the trend of using hand-modelled features is [GRS+20], where the training was on segmented gait data without specific feature extraction methods. However, the networks considered were RNNs. Thus, the performance comparison of time-series multi-sensor multi-channel IMU data on time-series CNN-IMU and time-series deepCNNLSTM networks is desirable.

Based on experiments, [GRS+20] had concluded that step data had better accuracy than stride data. The difference between the step and stride data is the segmentation window length. It would be interesting to see whether similar segmentation would facilitate identification on general body movement data.

Ideally, a comparison of the method mentioned above with the dataset of [GRS+20], and [EBL18] would have been an appropriate proof of concept. However, the datasets were unavailable. Hence, the author of this thesis has considered trademarking with OPPORTUNITY and PAMAP2 datasets, as shown in [RF18]. Though the number of subjects is less in these datasets, the subject's identification and the impact of various activities on identity can be evaluated. Furthermore, the data acquisition methods found in these datasets are different. Consequently, the impact of data acquisition methods on identity classification can be explored.

Taking inspiration from [RVKW15], [RF18] and [RSH+18], consideration of soft-biometrics as attributes for person identification would be an intriguing experiment. The groupings and sub-groupings considered in [RVKW15] can be used to develop an attribute representation, as shown in [RSH+18].

Authors of [GRS+20], [EBL18] and [RVKW15] have raised privacy concerns in the case of IMU-based gait recognition. However, research towards understanding the finer aspects that facilitates person identification from IMU data was not found. It would be interesting to analyse which features of the IMU motion data are relevant to the neural network to perform identification.

4 METHOD

As discussed in Sec: 3.5, most previous works on person identification using gait data focus on hand-modelled feature extraction methods and classical classifiers. Authors of [EBL18], have experimented on the impact of activities on person identification using IMU data. However, the authors have opted for hand-modelled features. Furthermore, the data used do not have the properties of a HAR dataset or data extracted from smartwatches or smartphones. Consequently, general body motion-based person identification using IMU data and related privacy concerns are yet to be addressed.

The benefit of using DNNs such as CNN lies in its capability to extract relevant features from the input data to facilitate classification. Therefore, experiments need to be conducted on HAR datasets to analyse person identification using DNNs. Analysing the features that facilitated identification as learned by the models could lead to methods for masking/deleting identity from IMU datasets while maintaining HAR.

From [EBL18] we know that some activities facilitate person identification better than few other activities. However, the authors trained the ML algorithms on data of individuals performing a specific activity. Thus, the impact of activities on identity, with networks trained on data of individuals performing various activities, needs to be analysed. It would be interesting to see whether identity can generalise over activities.

Finally, designing soft biometrics as attribute representation needs to be explored. The designed attribute representation could facilitate an understanding of how certain body characteristics affect identification. In addition, the experiments may give an insight into the impact of an individual signature on HAR.

Research in these areas is expected to function as the preliminary work towards discovering the features that facilitate person identification and the impact of individual motion signatures on the HAR dataset.

To solve the problems mentioned above, this thesis has chosen the following method. Firstly, DNNs designed for multi-sensor multi-channel time-series IMU data are used for experimentation. In specific, CNN-IMU and deepCNNLSTM networks. Secondly, HAR datasets will be experimented upon to shift the focus from gait-based person identification to general body motion-based person identification. As a result, the accuracy Acc of identification given a particular activity can be evaluated. Thirdly, to enable the possibility of comparison and bench-marking, only publicly available datasets, such as LARa [NRR+20], Opportunity [CSC+13], and PAMAP2 [RS12b], have been considered. Finally, attribute representations are designed based on soft biometrics. The attribute representation can be used for generalised grouping of the subjects and transfer learning.

Sec: 4.1 of this chapter explains the architecture and training of the DNNs. In Sec: 4.2, the attribute representation will be designed from the soft-biometrics. Further, the modifications to the network for achieving attribute representation will be discussed.

4.1 DEEP LEARNING FOR IDENTIFICATION

This section introduces the deep networks designed for multi-sensor, multi-channel time-series data to achieve person identification using motion information obtained from IMU sensors. The two networks of interest are CNN-IMU and deepCNNLSTM. To facilitate feature extraction, the networks have convolutional layers at their initial layers. Both the networks follow late fusion architecture. Hence, each sensor is adapted with a branch of convolutional layers.

The distinction between the networks lies in the layers following the convolutional layers. If the subsequent layers are fully connected MLP layers, the network is called the CNN-IMU network. The CNN-IMU network was discussed in Sec: 2.2.2. Given, the layers following the convolutional layers are LSTM layers; the network is referred to as deepCNNLSTM. Table: 4.1.1 presents these architectures.

<table>
<tr><td><b>CNN-IMU</b></td><td>||Conv</td><td>||Conv</td><td>||Conv</td><td>||Conv</td><td>||FC</td><td>Concat</td><td>FC</td><td>FC</td><td>Softmax</td></tr>
<tr><td><b>deepCNNLSTM</b></td><td>||Conv</td><td>||Conv</td><td>||Conv</td><td>||Conv</td><td>Concat</td><td>LSTM</td><td>LSTM</td><td>FC</td><td>Softmax</td></tr>
</table>

Table 4.1.1: Comparison of network layers between CNN-IMU and deepCNNLSTM. || denotes that the layers have parallel blocks. Conv refers to the convolutional layer. FC implies a fully connected MLP layer, and Concat stands for concatenation. All the layers consider the ReLU activation function.

In Sec: 2.2.2, it was discussed that late fusion methods are preferred while considering data from multiple sensors. The late fusion method allows convolutional filters to extract descriptive features as the focus area is dimentionsionally less. The networks follow sensor-based late fusion. A branch of convolutional layers processes each IMU sensor. Each branch has four convolutional layers, as visualised in Fig: 4.1.1.

IMU sensors consist of an accelerometer, gyroscope and magnetometer. Each of these devices has multiple channels based on its axes, as discussed in Sec: 2.2.1. These channels are represented as n in Fig: 4.1.1. w stands for the window size, which is dependant on the sliding window process (see Sec: 2.2.1). The convolution filter size

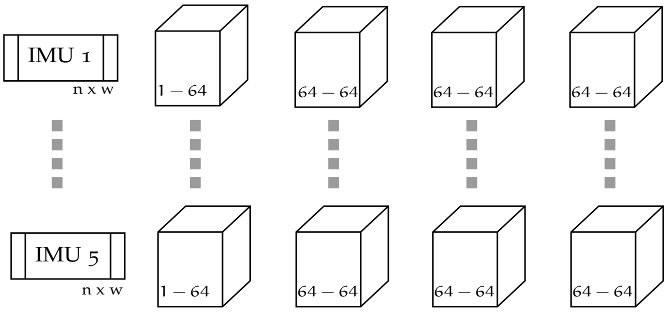


Figure 4.1.1: Visualisation of the parallel blocks of convolutional layers for each IMU. Each IMU input has n channels of w window size. The first convolution layer has one input channel and 64 output channels as depicted. The rest three convolutional layers have 64 input channels and 64 output channels. The filter size (5x1) is constant in all convolutional layers.

is set to 5x1. Similarly, the stride is set to 1x1. The first layer has one input channel and 64 output channels. The rest three layers have 64 input channels and 64 output channels. No pooling layers were considered in these architectures. The rest of the layers are specific to the type of network, as shown in Table: 4.1.1.

The CNN-IMU network has a fully connected layer attached to each block. The outputs from the fully connected layer of the parallel blocks are concatenated. Further, the concatenated outputs are passed through an MLP. ReLU is the preferred activation function of these layers. The final fully connected layer has a softmax activation function to support the classification of identity.

In the deepCNNLSTM network, the output of the parallel convolution blocks are concatenated and passed through two layers of LSTM. Similar to CNN-IMU, the final fully connected layer has a Softmax activation function to facilitate identity classification.

Based on the functionality of the networks, the activation function of the final layer varies. Given that the network is expected to perform identity classification, the final activation function will be a Softmax activation function. The Sigmoid activation function is used in the final layer to facilitate attribute representation.

Training

As mentioned in Sec: 2.2, Softmax, Eq: 2.2.1, output shows the probability that the input belongs to a particular class. As a result, softmax is the preferred activation function for solving classification problems. Loss is calculated with the Cross-Entropy

Loss function, Eq: 2.2.2. Sec: 2.1.3 had introduced gradient descent. Here, the concept of optimisation algorithms was introduced. The two proposed networks use the RMSProp optimisation algorithm to ensure fast convergence. The method adapts the learning rate (see Sec: 2.1.3) according to the variation in gradient in consecutive iterations. When the variation is large, the learning rate is reduced. Consequently, the amount of the parameter update is reduced. Whereas, when the variation is small, the learning rate increases to ensure convergence at a fast rate.

Neural networks have to classify unseen data after being subjected to supervised training on training data. Implying that the network should be able to generalise. When the network fails to classify unseen data but shows good classification on the training data, the issue is referred to as over-fitting. There are various methods to avoid over-fitting, e.g., early stopping, data augmentation and dropout. The data augmentation method used in this thesis is the addition of Gaussian noise (mean $\mu = 0$ and standard deviation $\sigma = 0.01$) to sensor samples. This method was used by [MGF+18] and [GLR+17], to simulate sensor inaccuracies. Furthermore, the networks consider dropout to avoid over-fitting. Dropout is the process of leaving out few neurons of a layer at random during training. The process essentially modifies the network architecture. However, dropout has been effective in improving generalisation.

4.2 ATTRIBUTES REPRESENTATION

Attribute representation was introduced as the semantic description of a scene or an object in Sec: 2.5.1. Sec: 3.2.1 discusses that soft-biometric features can be obtained from human-body movements. Consequently, soft biometrics can be used to describe or categorise an individual [SNM+12]. Thus, it can act as the attribute representation of an individual.

From [RSH+18] and [RVKW15], two sets of attribute representations were designed based on the LARa dataset recording protocol [NRR+20]. A snippet of the LARa dataset recording protocol is presented in Table: 4.2.1. The protocol presents the gender, age, weight, height and handedness of each subject. As all the subjects are right-handed, handedness cannot be considered as an attribute in LARa dataset. Table: 3.2.1 presents an example of sub-categorisations that can be performed on soft biometrics. A similar sub-categorisation on soft biometrics is performed on the LARa subjects (Table: 4.2.1), as shown in Fig: 4.2.1. Type 1 attribute representation splits each soft biometric into two categories. For example, the soft biometric height can be categorised as either $\leqslant 170cm$ or $> 170cm$. However, the Type 2 attribute representation splits the soft biometrics into three categories. It must be ensured

Role	Sex [F/M]	Age	Weight [kg]	Height [cm]	Handedness [L/R]
Subject 07	M	23	65	177	R
Subject 08	F	51	68	168	R
Subject 09	M	35	100	172	R
Subject 10	M	49	97	181	R
Subject 11	F	47	66	175	R
Subject 12	F	23	48	163	R
Subject 13	F	25	54	163	R
Subject 14	M	54	90	177	R

Table 4.2.1: Example of Recording Protocol of LARa dataset [NRR+20]. F/M stands for Female/Male. L/R stands for Left /Right

that the splits are meaningful. To elaborate, care must be given that the sub-category contains variations that the network can learn.

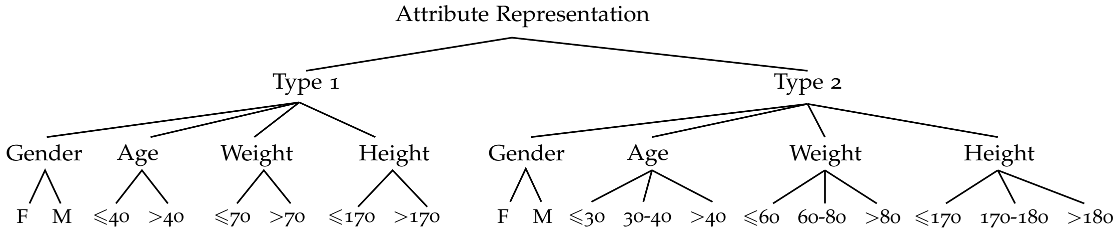


Figure 4.2.1: Soft biometric sub-categorisation for creating attribute representation. F stands for female, and M stands for male. Weight is measured in kg, and Height is measured in cm.

The attribute representations are tabulated, as shown in Table: 4.2.2 and Table: 4.2.3. The tables follow the format of the attribute representation table presented in [RSH+18]. The point of interest is that few subjects have the same set of attribute representations. In Table: 4.2.2, subject 3 and 7 have the same attribute representation. Similarly, subject 5 and 6 have the same attribute representation in Tables: 4.2.2 and 4.2.3. As a result, each attribute representation will be considered a center point, and the subjects will be allocated to each center based on their representation. Fig: 4.2.2 visualises the process. The figure draws the example from Table: 4.2.2.

TYPE 1				
	Gender	**Age**	**Weight**	**Height**
Subject	F/M	$\leqslant 40/>40$	$\leqslant 70/>70$	$\leqslant 170/>170$
0	1	0	0	1
1	0	1	0	0
2	1	0	1	1
3	*1*	*1*	*1*	*1*
4	0	1	0	1
5	*0*	*0*	*0*	*0*
6	*0*	*0*	*0*	*0*
7	*1*	*1*	*1*	*1*

Table 4.2.2: Type 1 Attribute Representation. F/M refers to Female/Male. Weight is measured in kg, and Height is measured in cm.

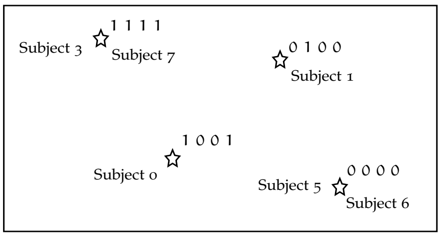


Figure 4.2.2: Visualisation of attribute representation space. Each subject is allocated to the center based on their respective attribute representation. The example is drawn from Type 1 attribute representation, Table: 4.2.2.

To facilitate attribute representation learning, the final layer of the CNN-IMU network has a sigmoid activation function (see Sec: 2.1.2), as shown in [MF18]. Consequently, BCE_{loss} Eq: 2.5.1, is the preferred loss function [NRR+20], as mentioned in Sec: 4.2.

The accuracy Acc, Eq: 5.3.1, of each attribute will be calculated to evaluate the quality of attribute classification. To find the predicted center of the attribute representation, two methods can be considered. The first method is the Nearest Neighbour Approach. Here, the distance between each prediction and the centers are calculated. The prediction is assigned to the center with the least distance. The second method is the Binary Cross-Entropy Loss (BCE_{loss}) approach. In this method, the negative average log of probabilities of the attributes in the representation is calculated, as shown in Eq: 2.5.1. The calculated value is the deviation of the prediction from the

TYPE 2										
	Gender	**Age**			**Weight**			**Height**		
Subject	F/M	$\leqslant 30$	30-40	> 40	$\leqslant 60$	60-80	> 80	$\leqslant 170$	170-180	>180
0	1	1	0	0	0	1	0	0	1	0
1	0	0	0	1	0	1	0	1	0	0
2	1	0	1	0	0	0	1	0	1	0
3	1	0	0	1	0	0	1	0	0	1
4	0	0	0	1	0	1	0	0	1	0
5	*0*	*1*	*0*	*0*	*1*	*0*	*0*	*1*	*0*	*0*
6	*0*	*1*	*0*	*0*	*1*	*0*	*0*	*1*	*0*	*0*
7	1	0	0	1	0	0	1	0	1	0

Table 4.2.3: Type 2 Attribute Representation. F/M refers to Female/Male. Weight is measured in kg, and Height is measured in cm.

expected representation. Thus, the $\mathrm{BCE}_{\mathrm{loss}}$ approach can evaluate the proximity of the predicted center to the desired center. Similar to the Nearest Neighbour Approach, the prediction then represents the center it is most similar to.

5 EXPERIMENTS

In Sec: 3.5 and Sec: 4, we discussed the shortcomings in person identification and the research gaps this thesis attempts to address. This chapter focuses on the experiments and the results. Most of the experiments are focused on the CNN-IMU network (see Sec: 4.1) in combination with the LARa dataset (see Sec: 5.1.1). It is to be noted that the attribute representations designed in Sec: 4.2 is based on the recording protocol of LARa [NRR+20]. Consequently, the concept of attribute representation will be primarily tested on the LARa dataset applied on the CNN-IMU network. Further, based on the results from the primary experiments, the deepCNNLSTM network and datasets such as OPPORTUNITY (see Sec: 5.1.2), PAMAP2 (see Sec: 5.1.3), and Order Picking dataset (see Sec: 5.1.4) will be experimented upon.

We first introduce the datasets in Sec: 5.1. The differences between the datasets will be discussed in Sec: 5.1.5. It is expected that such a comparison, in terms of number of IMUs, IMU placement, number of subjects, and amount of data, will help to understand the results better. Next, the experiments are enlisted in Sec: 5.2. Sec: 5.3 explains the evaluation metrics. Finally, the results will be discussed in Sec: 5.4. The results are categorised into three main sections. The first section focuses on the accuracy of person identification on a given dataset and given network. The results of the impact of activities on person identification are discussed in the second section. Next, the two attribute representations designed in Sec: 4.2, and their performances are evaluated. This section will further analyse the impact of activities on attribute representation. Finally, additional experiments and research will be presented in Sec: 5.5.

5.1 DATASET

Sec: 3.5 discussed the prominence of using publicly available datasets to facilitate benchmarking. Thus, popular HAR datasets are experimented with to analyse the uniqueness of the general body motion of an individual. Each dataset is unique in the type and number of sensors, subjects performing activities, and the activities. It is expected that such varied datasets will illustrate the impact of data acquisition methods on identity within the data. Furthermore, analysing the variations in dataset creation and their impact on identity would help to create a protocol for HAR dataset creation devoid of data privacy concerns. Consequently, this section introduces each dataset based on the aspects mentioned above. Furthermore, the pre-processing steps and their rationale are discussed.

5.1.1 *LARa*

The LARa dataset stands for Logistic Activity Recognition Challenge [NRR+20]. The dataset was created at the Innovationlab Hybrid Services in Logistics at the TU Dortmund University [NRMR+20]. The dataset consists of an Optical Motion Capture System (OMoCap), IMUs, and RGB camera data. Three logistic scenarios were depicted: two picking and one packing scenario. The scenarios were enacted by 14 subjects, resulting in 758 minutes of recordings. The dataset was labelled offline by 12 annotators.

The OMoCap system captures activities by mapping the movement of the reflective markers attached to the body with the help of infrared cameras. As shown in Fig: 5.1.1, the marker suit consists of 39 reflective markers attached to the human body. The arena of data creation has 40 infrared cameras sampled at $200fps$. Consequently, the OMoCap dataset consists of 126 channels of motion information.

Six MbientLab IMUs (MetaMotionRL) were considered with a sampling rate of 100Hz. They were placed on both wrists, chest, waist, and ankles. Each IMU consist of 3-axis Accelerometer (Scale:$\pm 2g - \pm 16g$, Resolution: $16bits$), 3-axis Gyroscope (Scale:$\pm 125^\circ/s - \pm 2000^\circ/s$, Resolution: $16bits$) and 3-axis Magnetometer (Scale: $\pm 1300\mu T(x, y - axis), \pm 2500\mu T(z - axis) range$, Resolution: $16bits$) readings. However, the authors have only provided the accelerometer and gyroscope measurements of five IMUs. The logistics lab is a controlled environment. As a result, no variations in the magnetic field was found. Hence, the authors did not include the magnetometer recordings in the dataset. It was found that the IMU placed at the waist showed a high amount of noise. As a result, the readings of this IMU was not included in the dataset. Consequently, 30 sensor channels of motion information are available for performing experiments.

From the 14 subjects, only eight subjects have IMU-based recorded data, [NRMR+20]. As this thesis is focused on IMU data, only subjects with IMU data will be selected for experimentation. Each subject had participated in a total of 30 recordings of two minutes each. Of the 30 recording, the subjects participated in two recordings of Scenario 1, 14 recordings of Scenario 2, and 14 recordings of Scenario 3. The 30 recordings of an individual were conducted within a day. As a result, the bodysuit

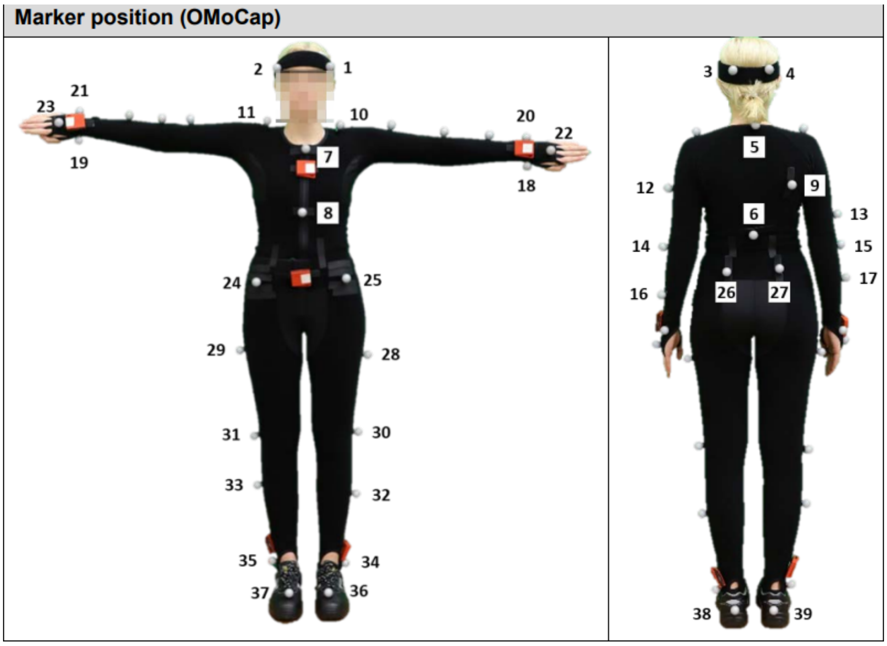


Figure 5.1.1: OMoCap on-body marker placements [NRMR+20]

was not removed from the individual's body once the recording sessions began. Furthermore, the recording protocol consists of before and after recording images of the individuals to map the dislocations of the markers and sensors. To help with synchronisation, each two-minute recording begins with a synchronisation gesture to indicate the beginning of the session. The subjects were given breaks between each recording session. However, they were mandated not to take off the suit during the breaks.

Unfortunately, it was found that a few of the recordings had to be removed because of noise and loss of data. As a result, all the subjects do not have an equal number of recordings for each Scenario. Two extreme cases worth mentioning are that of Subject 11 and Subject 12. The recordings 16-30 are missing for Subject 11. Consequently, Subject 11 do not have any Scenario 3 recordings. Subject 12 do not have recordings 1-10. Therefore, the subject does not have Scenario 1 recordings and only six recordings from Scenario 2.

The Scenarios are mentioned to help identify the type of activities present in the recording. For example, Scenario 1 and 2 are comprised of pushing a cart, handling and walking activities. In terms of body movement, this scenario is composed of activities that require whole-body movements. The difference between the two scenarios is in the layout of the logistic environment [NRR+20]. Scenario 3 is dedicated to packaging activities. The recordings consist of upper body movement and a few walking activities. The scenario does not have any pushing cart activities.

The subjects of interest are shown in Table: 4.2.1. The LARa recording protocol provided the information [NRMR+20].

The authors have considered eight activity classes based on logistics scenarios. Namely, standing, walking, cart, handling (upwards), handling (centered), handling (downwards), synchronization and none [NRR+20].

The activities can be represented as attributes. Attribute representation is expected to provide a coarse-semantic description of the activities [NRR+20]. The authors considered 19 attributes [RSH+18]. However, the activity attributes are not within the interest of this thesis and thus, not explored further.

Pre-processing

Category	Subjects eluded	No: of Classes	T-V-T split	Scenarios
A	12	7	5-1-1	2
B	11	7	6-1-1	2 and 3
C	11 and 12	6	8-2-2	2 and 3
D	10, 11 and 12	5	11-3-3	2 and 3

Table 5.1.1: LARa data is split into sub-categories for experimentation. T-V-T refers to the Train-Validation-Test split of the recordings.

LARa dataset is considered the primary dataset of this thesis. Consequently, the dataset has been extensively experimented. Pre-processing methods are specific to the experiment. The OMoCap and IMU data are used for experimentation. Furthermore, OMoCap experimental results are considered as a benchmark for IMU experimental results.

To tackle the absence of scenarios in few subjects mentioned in Sec: 5.1.1, the dataset was split into four categories, as shown in Table: 5.1.1. This split is implemented on the IMU and OMoCap data.

The IMU and OMoCap readings have a varying range for each sensor channel. For example, the accelerometer reading may range from $-2g$ to $+2g$, whereas the gyroscope reading may range from $-125°/s$ to $+125°/s$. Channel-normalisation is the process where the values are normalised to a value between 0 and 1. Normalisation can be achieved with the Eq: 5.1.1. Here, a refers to the value that needs to be normalised.

min and max refers to the minimum and maximum value of the sensor channel, respectively. Channel-normalisation is recommended for the OMoCap and IMU data while training a neural network, as it accelerates the learning process. This feature can be attributed to the constricted input space.

$$a_{norm} = \frac{a - \mathrm{min}}{\mathrm{max} - \mathrm{min}} \tag{5.1.1}$$

When placing an IMU on an individual, a bias may be induced in the sensor readings based on their physical features or movement. Furthermore, the bias may contribute to person identification. To test this hypothesis, networks were trained with both non-channel-normalised and channel-normalised IMU data. The results of the experiments are presented in Sec: 5.4.1.

To compare the person identification accuracy of IMU data to OMoCap data, the IMU and OMoCap data were both channel-normalised and null labels were removed. The Null labels are the erroneous or irrelevant activities found in the OMoCap data while performing annotation. Furthermore, the OMoCap data was down-sampled to 100Hz to match the frequency of IMU data.

As per [GRS+20], step data of gait gives better person identification than stride data. Step refers to heal strike of one foot followed by the heal strike of the other foot. Stride refers to the consecutive heal strikes of the same foot [GRS+20]. Step typically has a frequency of 1-2Hz. Consequently, the authors of [GRS+20] considered a window size of 100 for IMU data sampled at 75Hz. Similarly, we have considered a sliding window size of 100 and a stride size of 12. As a result, each window overlaps the previous window by 88%.

5.1.2 *OPPORTUNITY*

The OPPORTUNITY dataset [opp10] is focused on achieving human activity recognition from wearables, objects and ambient sensors [RCR+10]. This dataset was created to help benchmark HAR algorithms. Consequently, numerous sensors were used for data creation. In the category of body-worn sensors, seven IMUs, 12 3-axis acceleration sensors, and four 3-axis localization information were used. Further, 12 3-axis accelerations were used as object sensors and eight 3- axis acceleration sensors were used as Ambient sensors. The recording took place within a laboratory environment.

The dataset is comprised of a total of six recordings from each of the four participants. Of the six recordings, five consists of the natural execution of Activities of Daily Living (ADL). The sixth recording is a scripted sequence of activities, referred to as the drill.

The activities were labelled during the recording sessions. Each ADL run typically lasted for 15-25 minutes. The subjects were given a break of 10 to 20 minutes after each run, during which the data was copied, battery levels checked, and appropriate system behaviour ensured. The drill runs were of 20 to 35 minutes duration [RCR+10].

To facilitate context-based learning, the scenarios were annotated at different levels. There are low-level labels for 13 actions to 23 objects. 17 mid-level gesture classes and four high-level activity classes. For this thesis, we are interested in the mid-level gesture classes and high-level activity classes. The high-level activities are stand, walk, sit and lie. The mid-level gestures considered are open/close door1, open/close door2, open/close fridge, open/close dishwasher, open/close drawer1, open/close drawer2, open/close drawer3, clean table, drink cup, and toggle switch.

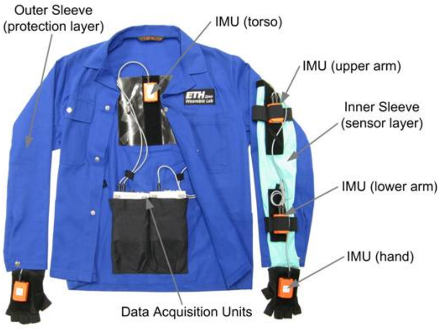


Figure 5.1.2: Motion Jacket [opp10]

The IMU sensors are placed on the subjects with the help of a motion jacket, as shown in Fig: 5.1.2. This jacket has inner sleeves or sensor layers, where the sensors can be placed. Consequently, it ensures unrestricted body movement. The jacket consists of five Xsens inertial units, placed at the subject's mid-back, lower and upper arms. These sensors have a sampling frequency of 30Hz [opp10]. In addition, there are 12 accelerometers placed on the body. These accelerometers have a frequency of 32Hz. Furthermore, two IMU sensors are placed on the toes of both feet, referred to as Inertiacube3. These sensors include a gyroscope, magnetometer, and accelerometer. These IMU sensors were sampled at a frequency of 40Hz. Consequently, a total of 113 sensor channels are considered for experimentation, based on the seven IMUs and 12 accelerometers. Fig: 5.1.3 presents the on-body placement of the IMUs and accelerometers, respectively.

A downside of the OPPORTUNITY dataset is that it does not provide the subject information. Consequently, attempting soft biometrics-based attribute representation is not feasible for this dataset.

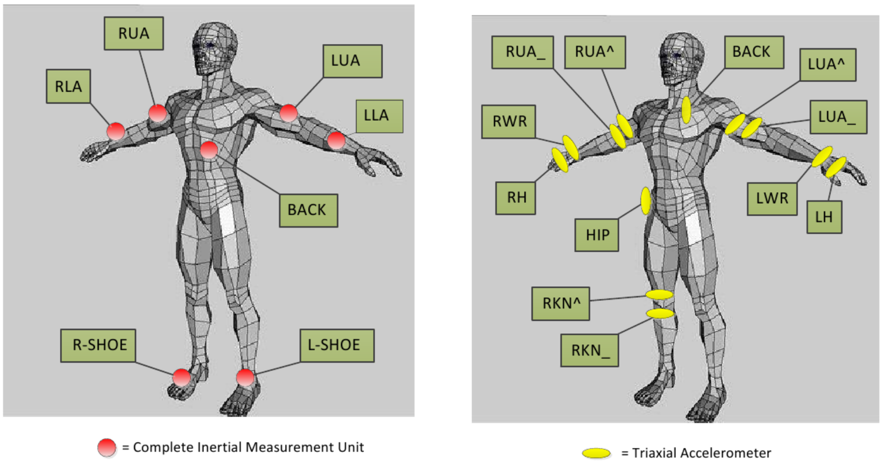


Figure 5.1.3: Sensor placement on the subject body [CSC+13]

Pre-Processing

As mentioned in Sec: 5.1.2, OPPORTUNITY is a sensor abundant dataset. Consequently, one can hypothesis a high accuracy rate of person identification. Furthermore, we are interested in the impact of locomotion and gesture activities on identification. Consequently, we have considered 113 sensor channels for training the network to recognise identity while analysing the activities.

As part of pre-processing, the dataset was normalised, Eq: 5.1.1. A sliding window size of 100 with a stride of 12 was considered. Though the sliding window size is not per [GRS+20], it was hypothesised that the sensor channel abundance would compensate for any lapses caused by the window size. To confirm the hypothesis, the dataset experimented with a window size of 24 and stride of 12.

5.1.3 *Pamap2*

PAMAP stands for Physical Activity Monitoring dataset [pam12]. The dataset consists of 18 activities of daily living and postures. The recordings were obtained from nine subjects wearing three IMUs and a heart rate monitor. The IMU sensors are Colibri Wireless IMUs from Trivisio. The IMU sensors are placed on three body locations: the chest, wrist of the dominant arm, and ankle of the dominant side. Each IMU sensor consists of two 3-axis MEMS accelerometers (Scale: $\pm 16g/\pm 6g$, Resolution: 13bit), a 3-axis MEMS gyroscope (Scale: $\pm 1500^\circ/s$, Resolution: 13bit), a 3-axis magnetoresistive magnetic sensor (Scale: ±$400\mu T$, Resolution: $12bit$), and a temperature sensor ($^\circ C$). All sensors were sampled at $100Hz$. The heart rate information is obtained from BM-CS5SR HR-monitor and sampled at $9Hz$. Consequently, there are 40 sensor channels for experimentation.

The dataset was not created in a laboratory environment to accommodate activities such as running and ascending stairs. Consequently, a battery pack with a battery life of 6 hours and a data collection companion unit - Viliv S5 UMPC (Intel Atom Z520 $1.33GHz$ CPU and $1GB$ RAM) was made use of. These were stored in a custom bag attached to the subject's body. The labelling of the activities was performed online using an application [RS12a].

Of the nine participants, eight are male, and one participant is female, of age $27.22 \pm 3.31 years$ and BMI of $25.11 \pm 2.62 kgm^{-2}$. One left-handed individual was present amongst the subjects. Similar to the LARa dataset recording protocol, the PAMAP2 has a subject information table facilitating attribute representation, as shown in Table: 5.1.2.

Subject ID	**Sex [F/M]**	**Age**	**Weight [kg]**	**Height [cm]**	**Handedness [L/R]**
101	M	27	83	182	R
102	F	25	78	169	R
103	M	31	92	187	R
104	M	24	95	194	R
105	M	26	73	180	R
106	M	26	69	183	R
107	M	23	86	173	R
108	M	32	87	179	L
109	M	31	65	168	R

Table 5.1.2: PAMAP2 Subject Information [pam12]. F/M stands for Female/Male. L/R stands for Left/Right.

The protocol includes basic activities such as walking, running, cycling, and Nordic walking, postures such as lying, sitting and standing, and everyday activities including ascending and descending stairs, ironing, vacuum cleaning, and rope jumping. The optional activities that the subjects can choose to perform are watching TV, computer work, driving a car, folding laundry, house cleaning, and playing soccer. It was noticed that all the subjects did not mandatorily perform the 12 basic activities. For example, the recording of Subject 109 mainly consists of optional activities. Rope jumping was the lone activity performed amongst the 12 basic activities.

Altogether, there are over ten hours of data, of which eight hours are labelled. The performed activities summary lists the activities performed by each subject and the duration of each activity in seconds.

A major issue found in the PAMAP recording is extensive data loss. Long data recording duration and the complexity of the activities can be attributed to the data loss. However, the authors of [RS12a] have specifically associated data loss with two main reasons: wireless data loss and fragile set up. The Colibri IMUs are wireless sensors, thus, accounting for wireless data loss. The activities such as jumping ropes and running causes mechanical stress on the sensor connections. Consequently, the wired connections may become loose or disconnected and lead to data loss.

Pre-processing

The major difference between PAMAP and datasets acquired in laboratory environments such as LARa and OPPORTUNITY lies in the continuity of data collection. The PAMAP2 dataset did not stop recording during transient actions like waiting for equipment setup or movement from one activity location to another. These activities were classified as activityID 0 as per the recording protocol. Consequently, it is recommended to remove data of this label. Furthermore, data associated with watching TV, computer work, driving car, folding laundry, house cleaning, and playing soccer has been removed. These activities are removed as most of the subjects do not perform these activities. For example, looking into the performed activities summary, one can see that except for subject 1, no other subject has performed the same activity. As a result, 12 activities are in focus. Namely, rope jumping, lying, sitting, standing, walking, running, cycling, Nordic walk, ascending stairs, descending stairs, vacuum cleaning, and ironing.

A matter of concern is the amount of data on the subject 109. Of the considered activities, subject 109 only performs the activity jumping rope. Thus, the person identification results of subject 109 is interesting.

According to the recording protocol, the data dropping was indicated with NaN. Consequently, to fill in the data, the data points are rewritten as 0. Furthermore, the data were normalised based on the Eq: 5.1.1. As the recording frequency was 100, the sliding window size was fixed at 100 with a stride of 12 as mentioned in Sec: 5.1.1.

5.1.4 *Order Picking Dataset*

The Order Picking dataset is a HAR dataset for logistics activities created at two different warehouses. The scenario of warehouse A requires the subject's interaction with a paper list for order-picking process guidance, whereas Warehouse B deals with handheld devices for order-picking process guidance. Each scenario has three subjects performing seven main activities. Namely, walking, searching, picking, scanning, info, carrying and acknowledge. In addition, there are two background activities, unknown-representing irrelevant actions- and sensor flip- marking the beginning and end of an order line [GLR+17].

Each subject has three sensors—one sensor on each wrist and one on the chest. The number of sensors is restricted to ensure unrestricted body movement. Data is sampled at a rate of 100Hz. Each IMU consist of an accelerometer, gyroscope and magnetometer. Consequently, there are 27 sensor value readings. There are 10 minutes of data for warehouse A and 23.30 minutes for warehouse B.

Though the dataset is not publicly available, it can be accessed on request. The details of the dataset can be found in [MGF+18] and [GLR+17]. A disadvantage of the Order Picking dataset is the lack of recording protocol. Consequently, understanding the activities and processing methodology was difficult. Furthermore, data is created with a fixed window size of 100 and a stride of 1. Thus, experiments with varying window sizes cannot be conducted on this dataset.

Pre-processing

The pre-processing steps for the Order-Picking dataset are quite limited. The dataset was normalised based on Eq: 5.1.1. Further, the null labels were removed. As mentioned earlier, the sliding window size is fixed at 100. In addition, due to the dataset's structure, the stride size is fixed at 1. The authors of [GLR+17] have considered data augmentation for using the Order Picking dataset. One of the data augmentation techniques recommended was the re-sampling of sensor values within a window at random. This process was not performed on the data used for this thesis.

5.1.5 *Discussions*

Table: 5.1.3 presents a comparison between the datasets discussed.

The LARa and OPPORTUNITY were created in a laboratory environment. Consequently, after each recording, the sensor placement, connections and battery charge were verified. As a result, these two datasets did not have continuous data acquisition.

Dataset	No: of Subjects	No: of IMUs	No: of sensor channels	Additional sensor	No: of activities	IMU placement	Location
LARa	8	5	30	OMoCap	7	Chest, both wrist & legs	Lab
OPPORTUNITY	4	7	113	Accelerometers	L = 4 G = 17	Chest, both wrist, arms & legs	Lab - Kitchen
PAMAP2	9	3	40	Heart monitor	12	Chest, dominant wrist & ankle	Outdoor
Order-Picking	6	3	27	-	7	Chest & both wrist	Warehouses

Table 5.1.3: Comparison of the dataset. L stands for Locomotion, and G stands for Gestures
.

Furthermore, both dataset creators gave sufficient break periods to their subjects. The duration of breaks was mentioned in the recording protocol. The authors of LARa and OPPORTUNITY have not explicitly mentioned data loss issues. Furthermore, recordings with extreme noise content were removed in LARa. As a result, the pre-processing steps did not require filling in the missing data.

The sequence of activities in the recordings of LARa and OPPORTUNITY repeats. As a result, any recording taken for testing or validation would consist of all the activities the dataset was trained on. However, Scenario 3 of LARa does not have Cart activity, but the rest of the activities of Scenario 3 were similar to Scenario 2.

Unlike the LARa and OPPORTUNITY datasets, the PAMAP2 recording protocol does not register any break period for the subjects performing the activities. Furthermore, most of the activities do not repeat. The authors were focused on gathering the maximum amount of realistic data. Consequently, the sensors recorded transitive activities such as waiting periods and change of location. However, the transitive data was not used in the experiments of this thesis as we are not aware of the type of activities or sensor disturbances that may have taken place during the transitive periods. In an ideal situation, the transitive data could be considered identity rich data.

In the Order Picking dataset, it was found that most of the activities were not present equally in all the recordings. Due to the absence of a recording protocol, further information regarding the data acquisition process and structure was unavailable. In addition, it was found that the relation between sensor channel to sensor placement was not explicitly mentioned. Consequently, a simple CNN network was used to process the Order Picking data obtained from three IMUs. The CNN network has four convolutional layers, no pooling layers and two fully connected MLP layers.

LARa has more IMUs but less number of sensor channels in comparison to PAMAP2. The reduced number of sensor channels in LARa is attributed to the removal of the magnetometer readings. Furthermore, the number of subjects in these datasets are comparative. As a result, it would be interesting to compare the performance of these datasets specifically. Furthermore, these two datasets provide detailed subject information, which facilitates attribute representation. The application of the attribute representations created for LARa onto PAMAP2 was considered but was found not feasible. In Type 1 attribute representation, the majority of the subjects were found to have the same attribute representation. Type 2 attribute representation had more variation in centers; however, there were attributes with no variations. Thus, an attribute representation with thresholds specific to the PAMAP2 needs to be created to perform experiments. The issue of transferability of attribute representation onto a different dataset can be associated with the number of subjects present and the variation in their characteristics. To facilitate transferability, attribute thresholds need to be generalised. Furthermore, experimentation on larger datasets is required. The research can be part of the future work of this thesis.

PAMAP2 and Order-Picking make use of only three IMUs. However, the placement of the IMUs varies in that the third IMU of PAMAP is placed on the leg, and Order-picking places it on the non-dominant-hand wrist. Consequently, it is of interest to compare these two datasets.

Although LARa and Order-picking are associated with logistic activities, the activity labels vary drastically. The activity and attribute labels of LARa is oriented towards the postures the subject would be in while performing the activity. In comparison, the activity labels of Order-picking is focused on the activity that takes place during order-picking. As a result, LARa labels can be considered a more generalised activity label for the logistic environment.

Unfortunately, all the datasets do not provide detailed sensor information. Thus, a comparison of the technology cannot be performed uniformly.

5.2 EXPERIMENT SUMMARY

Exepriment 1 - LARa dataset

Firstly, the LARa dataset was trained on the CNN-IMU network. A major task of network training is to find the right set of hyperparameters (HP). Hyperparameters (HP) are a set of variables relating to the neural network model, which can be modified to achieve learning. HPs can be classified as mini-batch gradient descent and model HPs. Examples of model HPs are the number of hidden units, weight initialisation and activation function. Learning rate (Lr), mini-batch size ($mBsize$) and epochs are examples of mini-batch gradient descent HPs. The model HPs are fixed for the CNN-IMU network. Consequently, the experiments of this thesis are focused on varying the gradient descent HP. We have considered the $Lr = \{10^{-4}, 10^{-5}, 10^{-6}\}$ and $mBsize = \{50, 100, 200\}$. Though the epoch was experimented upon, it was maintained at 10 epochs for most experiments.

Table: 5.1.1 shows the sub-categorisation of the LARa dataset based on the scenarios and subjects. These sub-categorisations were equally applied on the IMU and OMoCap datasets. It is expected that the sub-categorisation may help to analyse the effect of the activities on identification; for example, category C was trained on Scenario 2 and tested with Scenario 3. Categories B and D were trained on Scenario 2 and 3. However, the trained network was tested with just Scenario 3. Thus, the experiment is expected to give an insight into the generalisation of identity over activities.

As mentioned in Sec: 5.1.1, the pre-processing steps may include channel normalisation and Null label data removal. The various types of pre-processing steps performed on the dataset can be seen in Fig: 5.2.1. Data types that are compared during an experiment are colour coded in the figure. Sub-categories of channel normalised OMoCap data which do not include Null label data will be represented as $OMoCap_S^{C,NN}$, where the sub-category label replaces S. C stands for channel-normalised, and NN means that no Null label data are present in the data. If the sub-script is W, then the usage of the whole dataset is indicated. Similarly, sub-categorised, non-channel normalised IMU data, which include Null label data, will be represented as $IMU_S^{NC,N}$, where S can be replaced with the category label. Here, NC represents non-channel normalised, and N indicate that the data set include data with Null label.

The following are the list of experiments on the sub-categorisation of the LARa dataset:

- Experiment 1A: Comparison of $IMU_S^{NC,N}$ and $IMU_S^{C,N}$.
- Experiment 1B: Comparison of $IMU_S^{C,NN}$ and $OMoCap_S^{C,NN}$.
- Experiment 1C: Comparison of $IMU_W^{C,NN}$ and $OMoCap_W^{C,NN}$

Experiment 1A is expected to give an insight on the contribution of the sensor placement on subject identity, as discussed in Sec: 5.1.1. The data labelled as Null were not removed for the IMU sensor data because the label indicate corrupted OMoCap data. In LARa, to facilitate annotation, the OMoCap data was viewed as a figure. The annotators labeled the frames based on the OMoCap figure movement. Consequently, if the annotators were not able to understand the activity being performed by the subjects during annotation due to glitching, crumbling or absence of the OMoCap figure, the frames were labelled Null. The OMoCap and IMU data are synchronised. Thus, the labels and markings on OMoCap are applied on IMU data. While training OMoCap data, the corrupted data have to be deleted. However, corrupted OMoCap data does not imply that the IMU data is corrupted for the same time frames. Thus, the data associated with the label was maintained for experiments on IMU data.

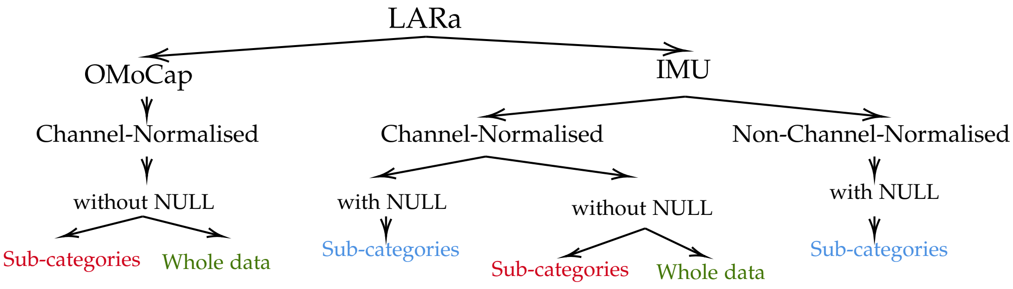


Figure 5.2.1: Dataset sub-categories. The data that are compared are colour coded.

In Experiment $1C_{IMU_W^{C,NN}/OMoCap_W^{C,NN}}$, the network was trained on eight subjects' data.

The results from Experiments $1B_{IMU_S^{C,NN}/OMoCap_S^{C,NN}}$ and $1C_{IMU_W^{C,NN}/OMoCap_W^{C,NN}}$ will be compared with the results of the deepCNNLSTM network. To elaborate:

- Experiment $1D$: Comparison of $OMoCap_S^{C,NN}$ and $IMU_S^{C,NN}$ trained on deepCNNLSTM with CNN-IMU for the same HPs.
- Experiment $1E$: Comparison of $OMoCap_W^{C,NN}$ and $IMU_W^{C,NN}$ trained on deepCNNLSTM and CNN-IMU of the same HPs.

These comparisons intend to evaluate the performance of the two networks on the IMU and OMoCap datasets for similar window size, stride, epochs, $mBsize$ and Lr.

Experiment 2 - Additional Datasets

The focus of Experiment 2 is to train the CNN-IMU network on OPPORTUNITY, PAMAP2 and Order Picking dataset. The experiments are focused on training the network with various combinations of Lr and $mBsizes$ to analyse which combination provides the best classification accuracy. For initial experimentation, $Lr =$

$\{10^{-4}, 10^{-5}, 10^{-6}\}$ and $mBsizes = \{50, 100, 200\}$ will be considered for 10 epochs. Given the scenario where the mentioned values do not provide good classification, alternatives will be explored.

Experiment 2A is the training of CNN-IMU network with OPPORTUNITY dataset. PAMAP2 dataset training on CNN-IMU is called Experiment 2B, and training on CNN-IMU with Order-Picking dataset is called Experiment 2C.

Experiment 3 - Impact of Activities

To identify the impact of activities on the identification accuracy during testing, the ratio of correctly classified windows were recorded with respect to the activity label of the window. The ratio associated with the positive classifications are denoted as IOA^{+}_{al}, Eq: 5.2.1, while negative classifications are denoted as IOA^{-}_{al}, Eq: 5.2.2. al refers to the activity label. $+/-$ denotes correct or wrong classifications, respectively. n^{+}_{al} refers to the number of windows belonging to al and was correctly classified. n^{-}_{al} refers to the number of windows belonging to al but misclassified. The hypothesis was that if a particular activity inversely affected the training of the network, classification accuracy during testing will be poor for that particular activity. This experiment was conducted on the datasets based on their respective activity labels.

$$IOA^{+}_{al} = \frac{n^{+}_{al}}{n^{+}_{al} + n^{-}_{al}} \tag{5.2.1}$$

$$IOA^{-}_{al} = \frac{n^{-}_{al}}{n^{+}_{al} + n^{-}_{al}} \tag{5.2.2}$$

Experiment 3A is associated with LARa activities. OPPORTUNITY activities impact will be presented in Experiment 3B. Experiment 3C and 3D will present the impact of activities on the PAMAP2 and Order Picking dataset, respectively.

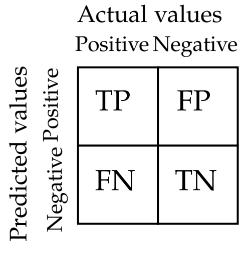
Experiment 4 - Attribute Representation

The attribute representations were designed for the LARa dataset in Sec: 4.2. The two types of attribute representation will be tested on the OMoCap and IMU data of LARa in Experiment 4A. The focus of the experiment is to identify which representation would perform better. There are two methods for finding centers, namely the Nearest Neighbour and BCE_{loss} method. Experiments will be conducted on both methods of finding centers. Next, an experiment by leaving out two subjects while training would be performed ($Experiment4A_{Leave-out2subjects}$). Subject 6 and 7 were left out during the training of the network for this experiment. The testing phase of the network was with the test data of subject 6 and 7. Furthermore, leave one out cross-validation (LOOCV) was performed ($Experiment4A_{LOOCV}$). The average performance over the eight subjects will be considered the final result of how well the attribute representation types perform. In addition, the average performance of each attribute can be analysed.

The impact of activities on the attribute representation is to be analysed in Experiment 4B. It would be interesting to analyse whether the impact of activities on attributes will be similar to that found with identity. The experimental results were collected at random to get an overall effect of activities on the attribute representation.

5.3 EVALUATION METRICS

The most intuitive method to evaluate the network's performance is calculating the accuracy Acc, Eq: 5.3.1. Here, n_{y_c} refers to the number of correct predictions and n_y represents the total number of predictions. However, Acc fails to account for the unbalanced training dataset. Consequently, researchers prefer to quantify and visualise the classifier's performance using the Confusion Matrix, Fig: 5.3.1. As mentioned in the figure, TP stands for true positive, where the prediction y_i and the label y_i^* are equal and shows a positive value. When y_i and y_i^* are equal and negative valued, it is called true negative (TN). FP stands for false positive, where y_i^* and y_i are not equal, and the value of y_i^* is negative. The opposite scenario is represented as a false negative (FN).



		Actual values	
		Positive	Negative
Predicted values	Positive	TP	FP
	Negative	FN	TN

Figure 5.3.1: Confusion Matrix. TP stands for True Positive. FP for False Positive. FN implies False Negative. TN stands for True Negative.

$$\begin{aligned} \mathrm{Acc} &= \frac{n_{y_c}}{n_y} \\ &= \frac{\mathsf{TP}+\mathsf{TN}}{\mathsf{TP}+\mathsf{TN}+\mathsf{FP}+\mathsf{FN}} \end{aligned} \tag{5.3.1}$$

Evaluation metrics such as Precision P, Recall R, and F1-score $F1$ can be derived from the Confusion matrix. Precision P, Eq: 5.3.2, is the ratio of the number of classifications that belong to the class and the number of classifications predicted to be of the class. Recall R, Eq: 5.3.3, is the ratio of the number of correct classifications of the class and the number of classifications that belong to the class. Precision and recall are calculated for each class. Evaluating these metrics help to analyse the bias of the model towards a particular classification. Calculating the $F1$, Eq: 5.3.4, is ideal when a balance between precision and recall is expected. $F1$ is the harmonic mean of precision and recall. Furthermore, the $F1$ accounts for the unbalanced dataset. A variation of $F1$, called the weighted F1 ($wF1$), Eq: 5.3.5, is used as an evaluation metric in this thesis. The difference between $F1$ and $wF1$ is the class average calculation in $wF1$, Eq: 5.3.5. Here, n_i refers to the number of samples of each class, and N is the total number of samples [MGF+18].

$$P = \frac{\mathsf{TP}}{\mathsf{TP}+\mathsf{FP}} \tag{5.3.2}$$

$$R = \frac{\mathsf{TP}}{\mathsf{TP}+\mathsf{FN}} \tag{5.3.3}$$

$$F1 = \frac{2}{\frac{1}{P}+\frac{1}{R}} \tag{5.3.4}$$

$$wF1 = \sum_i 2\frac{n_i}{N}\frac{P_i \times R_i}{P_i + R_i} \tag{5.3.5}$$

5.4 RESULTS

In this section, we analyse the results of the experiments in Sec: 5.2. Further, we shall attempt to derive plausible reasons for the experimental outcome. The experimental results are categorised into three sections: person identification, the impact of activities, and attribute representation. Under person identification, we are interested in the network's performance and identification accuracy. Results from the four datasets and the sub-categories will be presented.

The next aspect is the impact of activities on person identification. The experiments were performed on the datasets based on their respective activity label set. It is of interest to analyse whether the results will be following the findings presented in [EBL18]. Further, the attribute representation results for the LARa dataset will be presented.

5.4.1 *Person Identification*

LARa

Results of Experiment $1A_{IMU_S^{NC,N}/IMU_S^{C,N}}$*:*

The first experimental result is the comparison between $IMU_S^{NC,N}$ and $IMU_S^{C,N}$ on the CNN-IMU network. The comparison was performed on the four sub-categories (see Table: 5.1.1). The sub-categories A, B and C performed the best at 10 epochs, $Lr = 0.0001$ and $mBsize = 50$. An average Acc of 85.41% and $wF1$ of 85.63% was achieved. The following observations were made on sub-categories A, B and C:

- $IMU_A^{-,N}$ had the least number of training windows - 34318 windows.
- Network trained on $IMU_A^{C,N}$ performed slightly better than the network trained on $IMU_A^{NC,N}$ by about 2%.
- In sub-category B, the network trained on $IMU_B^{NC,N}$ performed better than network trained on $IMU_B^{C,N}$ by about 1.2%.
- Networks trained on sub-category C had the worst performance compared to sub-categories A and B.
- The network trained on $IMU_C^{C,N}$ was the sole experiment that gave an average Acc greater than 80% at 82.75% in the case of sub-category C.

Networks trained on sub-category C showed poor performance in comparison to sub-categories A and B. The result was unexpected as sub-category C has more training windows than sub-categories A and B. On analysing the recordings used for training and testing, we recognised that sub-category C was trained on Scenario 2, while the validation and test were performed on recordings of Scenario 3. In the case of sub-category A, training, validation and testing were conducted on recordings of Scenario 2. Similarly, the train-validation-test of sub-category B were on Scenario 3. Consequently, the poor performance of sub-category C was attributed to the recording Scenarios of the train-validation-test set. The experiment indicates that the model cannot generalise identity over activities that were not present in the training set.

Sub-category D showed good performance on the mini-batch sizes and learning rates. The performance could be attributed to the larger number of training windows. The network was trained on 52752 training windows. The epoch was fixed at 10. $\mathrm{IMU}_D^{C,N}$ performed slightly better than $\mathrm{IMU}_D^{NC,N}$, as shown in Table: 5.4.1. The Acc and $wF1$ are averaged over five train-validation-test cycles. From the table, it can be identified that the network's performance is sensitive to the learning rates. While an average Acc of 94.11% can be achieved with $\mathrm{mBsize} = 50$ and $\mathrm{Lr} = 0.0001$, the performance degrades to 82.27% of average Acc for $\mathrm{Lr} = 0.00001$.

Norm	mBsize	Lr	Avg Acc (x5)	Avg $wF1$(x5)
No	**50**	10^{-4}	**93.09 ± 0.51**	**93.12 ± 0.51**
No	**100**	10^{-4}	**90.04 ± 0.41**	**90.03 ± 0.41**
No	200	10^{-4}	85.60 ± 0.48	85.55 ± 0.48
Yes	**50**	10^{-4}	**94.11 ± 0.17**	**94.09 ± 0.17**
Yes	**100**	10^{-4}	**92.68 ± 0.15**	**92.64 ± 0.15**
Yes	200	10^{-4}	89.11 ± 0.43	89.09 ± 0.43
Yes	50	10^{-5}	82.27 ± 0.35	82.39 ± 0.34

Table 5.4.1: Comparison of channel-normalised and non-channel-normalised IMU data of sub-category 4. Acc and $wF1$ are presented in percentage. (x5) indicate that the presented values are averaged over five iterations of the experiment.

Based on the experiments, we can conclude that channel-normalised data performs better than non-channel-normalised data. Consequently, the rest of the experiments will be performed on channel-normalised data.

Results of Experiment $1\mathrm{B}_{\mathrm{OMoCap}_S^{C,NN}/\mathrm{IMU}_S^{C,NN}}$*:*

The performance of the networks trained on subcategories in Experiment 1B showed similarity to Experiment 1A. To elaborate, the networks trained on sub-categories A, B and C, performed best at epoch 10, $\mathrm{Lr} = 0.0001$ and $\mathrm{mBsize} = 50$. The average Acc was at 83.5%. In sub-categories A and B, IMU data performed better than OMoCap.

However, in sub-category C, OMoCap performed better than IMU data, with about a 5% difference in accuracy. The performance variation could be attributed to the training conducted on Scenario 2 while testing performed on Scenario 3. The drastic difference between OMoCap and IMU performance could be associated with the greater number of sensor channels present in OMoCap than IMU.

Experimental results of sub-category D were interesting, as the IMU data was found to perform better than OMoCap with a difference of about 5% for the same HPs. Furthermore, the results showed that OMoCap data capped at 85.62% average Acc, while IMU data reached 93.96% of average Acc for the same HPs. Table: 5.4.2 presents the results of sub-category D, averaged over five train-validation-test cycles. The table is composed of results that gave an average Acc greater than 80%.

Data	Lr	$mBsize$	**Avg Acc (x5)**	**Avg $wF1$ (x5)**
MoCap	10^{-4}	50	85.62 ± 0.23	83.52 ± 0.39
MoCap	10^{-4}	100	85.49 ± 0.21	83.37 ± 0.31
MoCap	10^{-5}	50	83.90 ± 0.76	81.52 ± 1.35
MoCap	10^{-5}	100	82.13 ± 0.93	80.44 ± 0.87
IMU	10^{-4}	**50**	**93.96 ± 0.03**	**93.84 ± 0.03**
IMU	10^{-4}	**100**	**92.43 ± 0.14**	**92.38 ± 0.15**
IMU	10^{-4}	200	88.41 ± 0.67	88.36 ± 0.66
IMU	10^{-5}	50	81.78 ± 0.72	81.93 ± 0.72

Table 5.4.2: Comparison between CNN-IMU network trained on sub-category D of OMoCap and IMU data. The Acc and $wF1$ are presented as percentages. (x5) indicate that the presented values are averaged over five iterations of the experiment.

From Table: 5.4.2, the drop in IMU performance at $Lr = 10^{-5}$ can be observed. Furthermore, variation of $mBsize$ does not seem to cause large variations in performance. However, it is interesting to note that OMoCap performance is comparatively stable for different $mBsizes$ and Lr. The experiments of $mBsize = 200$ were not conducted for OMoCap due to memory issues during training.

From the experiment, it can be concluded that identity classification on IMU data performs better for small datasets than OMoCap.

Results of Experiment $1C_{OMoCap^{CNN}_{W}/IMU^{CNN}_{W}}$*:*

The following results compare the performance of the CNN-IMU network trained on OMoCap and IMU data of eight subjects without sub-categorisation. Compared to the data considered for the sub-categories experiments, this experiment is conducted on a larger but unbalanced dataset. It was found that for a epoch 10, $\mathrm{Lr} = 10^{-4}$ and $w\mathrm{Bsize} = 100$, the CNN-IMU network trained OMoCap was able to achieve an Acc of 97% and $w\mathrm{F1}$ of 96.89%. However, the network trained on IMU data showed relatively poor performance. The network attained an Acc of 92.07% and $w\mathrm{F1}$ of 91.99%. The difference in performance between OMoCap and IMU can be attributed to the larger number of training windows available for OMoCap. Furthermore, OMoCap data has comparatively more sensor channels that the network can learn from.

From the results from CNN-IMU network, it can be concluded that the network performs competitively for epoch 10, $\mathrm{Lr} = 10^{-4}$ and $m\mathrm{Bsize} = 50$. In addition, identification accuracy on OMoCap exceeds IMU data with the increase in data size.

Results of Experiment $1\mathrm{D}_{\mathrm{OMoCap}_S^{C,NN}/\mathrm{IMU}_S^{C,NN}}$*:*

The experiments attempt to compare the performance of deepCNNLSTM to CNN-IMU. Consequently, the network was trained on the sub-categories and whole data of OMoCap and IMU. The hyper-parameters were, epoch 10, $\mathrm{Lr} = \{10^{-4}, 10^{-5}, 10^{-6}\}$ and $m\mathrm{Bsize} = \{50, 100, 200\}$.

The performance of the deepCNNLSTM network was found to be poor in comparison to that of the CNN-IMU network. Sub-category A did not yield Acc greater than 80% for any hyper-parameter combination of the network trained on OMoCap data. However, for IMU data, the deepCNNLSTM network performed exceptionally for epoch 10, $\mathrm{Lr} = 0.0001$ and $m\mathrm{Bsizes} = \{50, 100\}$. Average Acc of 84.09% and 81.13% were achieved respectively for $m\mathrm{Bsize} = \{50, 100\}$ for five train-validation-test cycles.

The epoch 10, $m\mathrm{Bsize} = 50$ and $\mathrm{Lr} = 0.0001$ yielded average Acc of 80.33% and 81.55% for Sub-categories B and C of OMoCap data, respectively. The accuracy presented is an average over five iterations. However, the network trained on IMU data did not result in any Acc greater than 80% for sub-categories B and C. No definite conclusions could be derived on the results.

Networks trained on sub-category D performed better than the other three sub-categories for the OMoCap and IMU data. Table: 5.4.3 shows the network performance, averaged over five train-test cycles. OMoCap and IMU data does not show much variation in performance on the deepCNNLSTM network. However, similar to the CNN-IMU, the IMU data performs slightly better than OMoCap data for the same HPs in category 4.

Data	Lr	mBsize	**Avg Acc (x5)**	**Avg** $wF1$ **(x5)**
MoCap	10^{-4}	50	88.76 ± 3.75	87.57 ± 4.66
MoCap	10^{-4}	100	87.49 ± 2.26	85.90 ± 2.93
MoCap	10^{-5}	50	87.49 ± 2.87	86.33 ± 3.76
MoCap	10^{-5}	100	87.22 ± 4.03	86.03 ± 4.92
IMU	10^{-4}	50	89.15 ± 0.77	89.20 ± 0.77
IMU	10^{-4}	100	86.19 ± 0.52	86.26 ± 0.50
IMU	10^{-4}	200	81.69 ± 0.79	82.02 ± 0.77

Table 5.4.3: Comparison of OMoCap and IMU sub-category D data performance on deepCNNLSTM network. The Acc and $wF1$ are presented as percentages. (x5) indicate that the values presented are average of five iterations of the experiment.

Results of Experiment $1E_{\mathrm{OMoCap}^{C,NN}_{W}/\mathrm{IMU}^{C,NN}_{W}}$:

The deepCNNLSTM was trained on the OMoCap and IMU data without sub-categorisation for epochs 10, $Lr = 0.0001$ and $mBsize = 100$. OMoCap data obtained an average Acc of 93.88%, while IMU achieved 87.40%. Consequently, it can be concluded that the overall performance of the CNN-IMU network on the datasets and sub-categories is better than the performance of deepCNNLSTM. The consecutive experiments are conducted on the CNN-IMU network.

Results of Experiment $2A_{\mathrm{OPPORTUNITY}}$:

It was hypothesised in Sec: 5.1.2, that due to the sensor-rich nature of the OPPORTUNITY dataset, there is a possibility of achieving high Acc and $wF1$ rates. The results confirmed the hypothesis. The CNN-IMU was trained with epochs $= \{5, 10\}$, $mBsize = \{25, 100\}$, and $Lr = \{10^{-4}, 10^{-5}, 10^{-6}\}$. An Acc and $wF1$ of 99% was achieved for $mBsizes = \{25, 100\}$, of $Lr = 0.0001$ with epochs 5 and 10. Similar to LARa IMU data, it was noticed that reducing the Lr to 10^{-6} deteriorated the performance of the CNN-IMU network. However, the Acc remained greater than 90%. Experiments were conducted with a sliding window size $w = \{24, 100\}$ and stride 12. It was found that the window size did not bring forth any difference. 30 experiments were conducted in all. An average Acc of 96.03% and average $wF1$ of 95.84% was achieved.

Results of Experiment $2\mathrm{B}_{\mathrm{PAMAP2}}$*:*

Initially, each recording of the individuals was split into the train-validation-test set without considering their activities. The results on the dataset were abysmal irrespective of the Lr, mBsize, window size, stride, and removal of subject 9. The maximum Acc was 57.2% for $\mathrm{Lr} = 0.0001$, $\mathrm{mBsize} = 50$, epoch 50 and without subject 109.

A new set was created to investigate whether the poor identification performance was related to the distribution of activities throughout the train-validation-test set. The new set first sorted the frames with respect to the activity labels and then split the frames into train-validation-test set at $64\% - 18\% - 18\%$. It was found that the new set could perform better and provide better accuracy, as shown in Table: 5.4.4. The experiments are for epoch 10. Subject 109 was included in the experiment. The average over five experiments are presented.

Lr	mBsize	**Avg** Acc **(x5)**	**Avg** $w\mathrm{F1}$ **(x5)**
10^{-3}	**50**	**91.01 ± 1.11**	**90.98 ± 1.14**
10^{-3}	100	88.03 ± 1.46	87.92 ± 1.49
10^{-3}	200	84.47 ± 1.29	84.51 ± 1.32
10^{-4}	**50**	**90.35 ± 0.61**	**90.36 ± 0.61**
10^{-4}	100	85.03 ± 0.43	84.98 ± 0.43

Table 5.4.4: Person Identification results of CNN-IMU network on PAMAP2. The Acc and $w\mathrm{F1}$ are presented as percentages. (x5) indicate that the values are averaged over five iterations of the experiment.

The data was found to be extremely sensitive to the Lr. When experimented with $\mathrm{Lr} = 0.00001$, the Acc dropped to 51.74%, for $\mathrm{mBsize} = 50$ and epoch 10.

After comparing the PAMAP2 results with the IMU results of LARa (Experiment $1\mathrm{C}_{\mathrm{OMoCap}_W^{\mathrm{C,NN}}/\mathrm{IMU}_W^{\mathrm{C,NN}}}$), it can be said that the performance of LARa is better than PAMAP2 at $\mathrm{Lr} = 0.0001$, $\mathrm{mBsize} = 100$ and epoch 10. However, considering that the PAMAP2 has only three IMU sensors attached to the body, in contrast to the five IMU sensors in LARa, PAMAP2 shows favourable results. In addition, the experiment re-enforces the fact that the model cannot generalise over activities it was not trained on.

Results of Experiment $2C_{\mathrm{OrderPicking}}$:

Similar to PAMAP2, the dataset was initially split into a train-validation-test set without considering the activities. As the results were poor, the train-validation-test split was performed by considering the activities. However, unlike PAMAP2, no improvement was observed irrespective of the Lr, $mBsize$ and epoch. The maximum Acc was 52.23% at $Lr = 0.0001$, $mBsize = 200$ and epoch 5.

No conclusive statements can be made on the poor results obtained. However, it can be hypothesised that the small dataset and IMU placement may be the reasons for the poor performance.

To ensure that the CNN architecture did not cause the performance degradation, PAMAP2 was trained on CNN. As PAMAP2 and Order Picking have a similar number of IMUs and data structure, it was expected that if CNN architecture performed poorly compared to CNN-IMU for the same HPs, the architecture could be indicted. However, PAMAP2 trained on CNN performed similarly to CNN-IMU. Thus, the architecture can be ruled out from being the cause of the poor performance.

5.4.2 *Impact of Activities*

Results of Experiment $3A_{\mathrm{LARa}}$:

The LARa dataset has seven main activities, excluding the Null label. The frames labelled as Null are removed from the dataset due to the erroneous OMoCap data. Ten experiments of the CNN-IMU network were selected at random. The networks were trained on four sub-categories and whole OMoCap and IMU data. By selecting the experiments at random, we expect to analyse the overall effect of activities on the identity. The average IOA_{al} over ten experimental results is presented in Table: 5.4.5.

It was interesting to note that the activities that contain gait cycles performed better than activities with upper body movement. Pushing cart activity achieved an average Acc of 93.37% of 10 experiments. Handling downwards has the lowest classification Acc at 66.74%. Handling center and handling up have an average Acc of 83.08% and 82.18% respectively.

Results of Experiment $3B_{\mathrm{OPPORTUNITY}}$:

OPPORTUNITY has two sets of activity labels: locomotion and gesture. Ten classification experiments were chosen at random where the IOA_{al} was calculated. The averaged IOA_{al}s for locomotion and gesture activities are presented in Table: 5.4.6

Activity	IOA^{+}_{al}	IOA^{-}_{al}
Walking	87.67 ± 12.23	12.31 ± 12.23
Cart	93.37 ± 6.37	6.62 ± 6.37
Handling cen	83.08 ± 8.02	16.90 ± 8.02
Handling down	66.74 ± 22.58	33.24 ± 22.58
Handling up	82.187 ± 13.27	17.803 ± 13.27
Standing	89.21 ± 8.90	10.77 ± 8.90
Synch	85.02 ± 9.48	14.97 ± 9.48

Table 5.4.5: Impact of activities on LARa dataset. The values are presented as a percentage.

and 5.4.7, respectively. Table: 5.4.6 shows the activity labels related to locomotion. Interestingly, the windows with minimal body movements obtain a higher rate of correct identification, similar to the experimental results of [EBL18]. The windows with stand locomotion had the least correct window classification.

Activities	None	Stand	Walk	Lie	Sit
IOA^{+}_{al}	98.64 ± 1.68	92.42 ± 3.93	97.69 ± 2.83	98.04 ± 5.72	98.82 ± 2.53
IOA^{-}_{al}	1.36 ± 1.68	7.58 ± 3.93	2.31 ± 2.83	1.96 ± 5.72	1.18 ± 2.53

Table 5.4.6: Impact of locomotion activities on OPPORTUNITY dataset. Values are presented as percentages.

The impact of gestures on classification is presented in Table: 5.4.7. The presented values are an average of ten experiments. Almost all the gestures show good accuracy. CloseDishWasher and CloseDoor1 are the two gestures that have performed relatively poor. No conclusive reason can be hypothesised for their poor performances without further experimentation. During this thesis, the videos of the activities were not analysed. Hence, examining the video recordings of the data might help to identify the reason for the relatively weak performance of CloseDishWasher and CloseDoor1 activities.

The data labelled as Null were not removed from the OPPORTUNITY dataset for both gesture and locomotion sets. In the OPPORTUNITY dataset, the Null label refers to activities that do not belong to the labels explicitly mentioned. Consequently, it is interesting to note that the Null label shows high classification accuracy in both scenarios.

Activities	**OD1**	**OD2**	**ODW**	**ODw1**	**ODw2**	**ODw3**
IOA_{al}^{+}	96.7 ± 2.73	99.74 ± 0.49	98.74 ± 1.57	91.84 ± 8.70	98.81 ± 1.26	97.28 ± 3.35
IOA_{al}^{-}	3.29 ± 2.73	0.26 ± 0.49	1.26 ± 1.57	8.16 ± 8.70	1.18 ± 1.26	2.71 ± 3.35
Activities	**CDW**	**CD1**	**CD2**	**CDw1**	**CDw2**	**CDw3**
IOA_{al}^{+}	78.12 ± 30.05	89.37 ± 31.41	98.96 ± 1.17	95.92 ± 4.32	96.96 ± 3.23	92.24 ± 9.09
IOA_{al}^{-}	21.88 ± 30.05	10.63 ± 31.41	1.04 ± 1.17	4.08 ± 4.32	3.04 ± 3.23	7.76 ± 9.09
Activities	**CF**	**None**	**CT**	**OF**	**Toggle**	**DC**
IOA_{al}^{+}	99.25 ± 0.87	98.75 ± 3.48	95.19 ± 6.24	97.86 ± 2.52	93.89 ± 6.45	91.97 ± 7.45
IOA_{al}^{-}	0.75 ± 0.87	6.09 ± 3.48	4.81 ± 6.24	2.14 ± 2.52	6.09 ± 6.45	8.31 ± 7.35

Table 5.4.7: Impact of gesture activities on OPPORTUNITY dataset. Values are represented as percentages. ODx refers to the OpenDoor gesture, and x stands for the door number. CDx represents CloseDoor. ODW and CDW stand for OpenDishWasher and CloseDishWasher, respectively. ODwx and CDwx, respectively, represent OpenDrawer and CloseDrawer. OF and CF denote OpenFridge and CloseFridge gestures. Gesture CleanTable is abbreviated as CT. DrinkCup gesture is denoted as DC.

Results of Experiment $3\mathrm{C}_{\mathrm{PAMAP2}}$*:*

The PAMAP2 dataset has 18 activity types available in its recordings. Six of the activities in this dataset are optional. Seven experiments with good identification accuracies were selected. Table: 5.4.8 presents the impact of the 12 main activities on the identity classification, averaged over seven experimental results.

Of the four basic activities - walking, running, cycling and Nordic walk - cycling activity shows a poor classification rate. Nordic walk and walking, which are essentially gait activities, perform better than all other activity classes. Lying, sitting and standing are classified as postures. In general, the classification rate of postures is low. However, standing posture has a relatively higher positive classification rate. This finding negates the conclusion of [EBL18], that activities with little body movement have high identification accuracy.

Identical to the findings in [EBL18], the classification accuracy of vacuuming activity was the worst performance. It was mentioned that activities with tools that make noise or cause vibrations could impact the sensors. Thus, leading to poor classification accuracy. The difference between the type of tools can be seen in vacuuming and ironing activities. Ironing has better classification accuracy.

Activity	$\text{IOA}^{+}_{\text{al}}$	$\text{IOA}^{-}_{\text{al}}$
Rope Jumping	83.74 ± 6.81	16.26 ± 6.81
Lying	74.72 ± 26.85	25.27 ± 26.85
Sitting	75.58 ± 13.97	24.42 ± 13.97
Standing	78.73 ± 14.46	21.27 ± 14.46
Walking	85.65 ± 19.15	14.35 ± 19.16
Running	77.41 ± 12.76	22.58 ± 12.76
Cycling	68.64 ± 22.18	31.36 ± 22.18
Nordic Walk	87.23 ± 13.98	12.77 ± 13.98
Ascending Stairs	74.09 ± 22.46	25.9 ± 22.46
Descending Stairs	66.64 ± 24.77	33.36 ± 24.77
Vaccuming	49.39 ± 19.84	50.59 ± 19.84
Ironing	74.91 ± 22.71	25.09 ± 22.71

Table 5.4.8: Impact of activities on PAMAP2 dataset. Values are presented as a percentage.

Interestingly, rope jumping has shown good performance. It was expected that the large body movements produced while jumping could induce noise and misplacement in the sensors. Thus, deteriorating the identification accuracy. However, Rope jumping has shown accuracy rates equivalent to the gait activities. One anomaly of the entire dataset is descending stairs activities. Though ascending stairs has an average performance, descending stairs has a similar performance to that of cycling. It is unclear why two activities that show cyclic body movement showed accuracy rates worse than that of posture activities.

Results of Experiment $3\text{D}_{\text{OrderPicking}}$*:*

Though, in general, the performance of identity classification is unsatisfactory, the impact of activities were recorded for five experiments. As the creators of the dataset recommended, data labelled as Null were not considered in these experiments. The average accuracy of identity classification concerning the activities is presented in Table: 5.4.9.

Activities classified as Unknown has shown the highest performance, followed by Flip. Though the performance of a walking activity is above average, it is exceeded

Activity	IOA^{+}_{al}	IOA^{-}_{al}
Unknown	97.51 ± 5.57	2.49 ± 5.57
Flip	69.52 ± 3.79	30.48 ± 3.79
Walk	51.51 ± 0.98	48.39 ± 0.98
Search	14.81 ± 6.69	85.19 ± 6.69
Pick	47.38 ± 1.95	52.63 ± 1.95
Scan	67.2 ± 2.71	32.79 ± 2.71
Info	68.08 ± 0.82	31.92 ± 0.82
Carry	42.8 ± 5.09	57.19 ± 5.09
Ack	8.84 ± 2.53	91.15 ± 2.53

Table 5.4.9: Impact of activities of Order Picking dataset. Values are presented as percentages. Ack stands for Acknowledge.

by scanning and information checking activities. Acknowledge activity has the worst performance.

As mentioned, the person identification on the Order Picking dataset was exorbitantly poor with respect to the performance of the other three datasets. As a result, it may not be advisable to make conclusions on the impact of activities with the results given in Table: 5.4.9.

It is to be noted that the Null label and Unknown label are not the same. Null label in Order picking refers to erroneous data, while Unknown label refers to activities that are not relevant for logistic activities. The activities under the Unknown label were not explored further due to the absence of a recording protocol.

5.4.3 *Attribute Representation*

Results of Experiment 4A:

The attribute representations designed in Sec: 4.2 is evaluated in this section. The centers were found with the Nearest Neighbour method and BCE_{loss} method. The experiments were conducted on the whole data of eight subjects of the OMoCap and IMU. The HPs used for training the CNN-IMU network were: epoch 10, $Lr = 0.0001$ and $mBsize = 100$.

Results of Attribute Representation Type 1*:*

The result of the Type 1 attribute representation on the OMoCap and IMU data, with the Nearest Neighbour method for finding centers, is presented in Table: 5.4.10. It is interesting to note that though the individual attribute classification is accurate, the attribute representation classification to the center has poor performance.

Data	Acc	wF1	**Gender**	**Age**	**Weight**	**Height**
			M/F	$\leqslant 40/ > 40$	$\leqslant 70/ > 70$	$\leqslant 170/ > 170$
MoCap	60.03	53.11	99.30	98.40	98.13	98.97
IMU	47.53	33.14	92.70	92.33	93.19	92.52

Table 5.4.10: Type 1 attribute representation results with the Nearest Neighbour method. Values are presented as percentages. M/F represents Male/Female.

Each individual attribute label y_i^* is presented in a binary form $\{0, 1\}$. However, the classification prediction y_i is a value between 0 and 1. The threshold is set at 0.5. Consequently, the y_i has to be rounded. However, while considering the Nearest Neighbour method, we do not round the individual attributes. The predictions y_i are considered to be points in a multi-dimensional space. Thus, making the classification to the right center a difficult process.

To overcome this limitation, the Nearest Neighbour method can be replaced with BCE_{loss}. The results are presented in Table: 5.4.11. As expected, a huge variation is seen in Acc and $wF1$ of BCE_{loss} method in comparison to the Nearest Neighbour method. The accuracy in finding the centers is now comparable to the attribute accuracy.

Data	Acc	wF1	**Gender**	**Age**	**Weight**	**Height**
			M/F	$\leqslant 40/ > 40$	$\leqslant 70/ > 70$	$\leqslant 170/ > 170$
MoCap	96.69	96.68	99.57	97.86	98.31	99.05
IMU	89.53	89.37	93.58	93.20	94.07	93.04

Table 5.4.11: Type 1 attribute representation results with BCE_{loss} method. Values are presented as percentages. M/F represents Male/Female.

In both methods, it can be seen that OMoCap consistently performs better than IMU data. The difference in performance can be attributed to the number of sensor channels. From Tables: 5.4.10 and 5.4.11, the performance of gender attribute is seen to be consistently high. The height attribute shows the next best performance.

Results of Attribute Representation Type 2:

The major difference between the attribute representations, Type 1 and Type 2, lies in the dimensionality of the attribute representations. The impact of dimensionality can be seen while using the Nearest Neighbour method. Table: 5.4.12 shows the Type 2 attribute representation results with the Nearest Neighbour method for OMoCap and IMU data. Compared to the results found in Type 1 attribute representation with Nearest Neighbour, it can be seen that the accuracy in finding the centers has increased. This performance can be credited to the increased dimensionality of the Type 2 attribute representation.

Data	Acc	wF1	**Gender**	**Age**			**Weight**			**Height**		
			M/F	$\leqslant 30$	**30-40**	> 40	**40-60**	**60-80**	> 80	$\leqslant 170$	**170-180**	**>180**
MoCap	95.81	95.83	98.63	98.68	99.97	98.46	99.24	97.61	97.68	98.48	97.94	99.11
IMU	89.05	88.95	93.40	93.52	98.10	93.31	93.77	94.65	93.90	93.27	92.10	96.17

Table 5.4.12: Type 2 attribute representation results with the Nearest Neighbour method. Values are presented as percentages. M/F represents Male/Female

OMoCap data has performed better than IMU data. No specific trend could be identified from the attribute results presented.

The results of the $\mathrm{BCE}_{\mathrm{loss}}$ method is presented in Table: 5.4.13. The Acc and wF1 of $\mathrm{BCE}_{\mathrm{loss}}$ and Nearest Neighbour methods were found to be comparable. The performance of the individual attributes are similar to the performance found in Table: 5.4.12.

Data	Acc	wF1	**Gender**	**Age**			**Weight**			**Height**		
			M/F	$\leqslant 30$	**30-40**	> 40	**40-60**	**60-80**	> 80	$\leqslant 170$	**170-180**	**>180**
MoCap	95.32	95.29	98.28	98.54	1	98.22	99.28	96.65	96.86	98.57	97.97	98.89
IMU	88.76	88.58	93.34	93.35	98.10	93.20	93.62	94.77	93.81	93.08	92.05	96.38

Table 5.4.13: Type 2 attribute representation results with $\mathrm{BCE}_{\mathrm{loss}}$ method. Values are presented as percentages. M/F represents Male/Female.

On close examination of the attributes, it was noticed that attributes such as $\mathrm{Age}_{(30-40)}$ and $\mathrm{Height}_{(>180)}$, which showcased high performance, do not contain many variations in training data. To elaborate, in the case of $\mathrm{Age}_{(30-40)}$, of the eight individuals, only one individual was of the range $30 - 40$. As a result, the network does not have much variation to learn from. To understand any underlying trend, it is recommended to reiterate the experiments with a larger dataset, with more variations.

Attribute Representation Experiment Leaving Out Two Subjects:

It can be noticed from Table: 4.2.2 and 4.2.3 representations that few subjects share the same attribute representation. Consequently, it is interesting to know whether the network can correctly place the attributes on individuals it has not seen before. As the focus is on individual attributes and not the center, we experimented using the Nearest Neighbour method.

The recordings of subjects 6 and 7 were removed from the training set and were only used for testing. The data of this format was created using the OMoCap and IMU data. The CNN-IMU network was trained with HPs: epoch 10, $Lr = 0.0001$ and $mBsize = 100$. The result for Type 1 and Type 2 attribute representations are presented in Table:5.4.14 and 5.4.15, respectively.

Data	Acc	$wF1$	**Gender**	**Age**	**Weight**	**Height**
			M/F	$\leqslant 40 / > 40$	$\leqslant 70 / > 70$	$\leqslant 170 / > 170$
MoCap	0.0	0.0	95.28	15.14	48.5	95.44
IMU	00.15	00.31	59.07	53.18	60.29	52.91

Table 5.4.14: Type 1 attribute representation leaving out two subjects. Experimental result of testing on subject 6 and 7. Values are presented as percentage. M/F represents Male/Female.

The attribute representations of both subjects 6 and 7 are present in the training set of Type 1 attribute representation. From Table: 5.4.14, it can be seen that OMoCap performs better than IMU data. Further, the attributes, Gender and Height, perform the best in the case of OMoCap. In contrast, Gender and Weight perform best in the case of IMU data.

In Type 2 attribute representation, only subject 6's representation are present in the training set. Similar to the results of Type 1, OMoCap data performed better than the IMU, Table: 5.4.15. The performance difference can be linked to the number of sensor channels. Further, Height and Gender attributes perform the best in the case of OMoCap data. Meanwhile, Weight and Gender attributes perform best in the case of IMU data.

Data	Acc	$wF1$	Gender	Age			Weight			Height		
			M/F	$\leqslant 30$	30-40	> 40	40-60	60-80	> 80	$\leqslant 170$	170-180	>180
MoCap	4.66	8.49	86.94	22.69	100	24.26	57.70	3.09	48.53	95.17	95.28	100
IMU	9.60	16.63	56.68	58.12	70.38	54.88	55.46	86.73	60.45	52.33	51.48	64.33

Table 5.4.15: Type 2 attribute representation leaving out two subjects. Experimental result of testing on subject 6 and 7. Values are presented as percentage. M/F represents Male/Female.

Attribute Representation Experiment Leave One Out Cross-Validation Method:

To achieve a comprehensive understanding of the attributes performance, Leave One Out Cross Validation (LOOCV) was performed. The CNN-IMU network was trained eight times by leaving out one subject and testing the network with the left-out subject. The average of eight iterations will be presented here. The HPs used were: epoch 10, $\mathrm{Lr} = 0.0001$ and $\mathrm{mBsize} = 100$. The experiments were conducted using the Nearest Neighbour and $\mathrm{BCE}_{\mathrm{loss}}$ method.

Attributes	Data	MoCap	IMU
	Acc	2.57 ± 5.47	2.31 ± 4.18
	$wF1$	4.58 ± 9.42	4.24 ± 7.49
Gender	M/F	77.51 ± 34.47	54.48 ± 21.63
Age	$\leqslant 40/ > 40$	34.02 ± 41.81	32.51 ± 25.61
Weight	$\leqslant 70/ > 70$	68.53 ± 39.36	50.67 ± 18.02
Height	$\leqslant 170/ > 170$	71.25 ± 33.97	55.85 ± 33.49

Table 5.4.16: LOOCV on Type 1 attribute representation using the Nearest Neighbour method. The values presented are average of eight experimental results. The values are presented as percentages. M/F refers to Male or Female.

Table: 5.4.16 presents the LOOCV average performance of Type 1 attribute representation with the centers evaluated using the Nearest Neighbour method. As expected, the Acc and $wF1$ averages are poor. Consistently, OMoCap performs better than the IMU data. Gender and Height attributes perform the best with OMoCap data. Interestingly, IMU follows the same trend as OMoCap and shows a better classification average Acc for Gender and Height attributes.

Contrary to expectation, the $\mathrm{BCE}_{\mathrm{loss}}$ method for finding centers performed poorly. The individual attribute representation accuracy values were similar to that of Nearest Neighbour. Consequently, the results are not presented here.

Attributes	**Data**	**MoCap**	**IMU**
	Acc	13.15 ± 31.95	3.06 ± 4.91
	wF1	15.0 ± 33.5	5.57 ± 8.83
Gender	**M/F**	87.21 ± 14.71	53.72 ± 24.36
Age	$\leqslant 30$	55.63 ± 43.3	49.87 ± 26.12
	30-40	76.99 ± 42.39	69.95 ± 30.46
	> 40	33.37 ± 42.62	36.07 ± 23.55
Weight	**40-60**	75.28 ± 38.49	52.09 ± 36.02
	60-80	47.22 ± 44.38	48.47 ± 41.02
	> 80	72.37 ± 40.89	50.93 ± 15.85
Height	$\leqslant 170$	83.35 ± 23.79	58.92 ± 32.79
	170-180	58.33 ± 41.97	46.32 ± 22.41
	>180	74.85 ± 46.2	74.89 ± 32.1

Table 5.4.17: LOOCV on Type 2 attribute representation using the Nearest Neighbour method. The values presented are average of eight experimental results. Values are presented as percentages. M/F represents Male/Female.

Table: 5.4.17 presents the LOOCV results on Type 2 attribute representation with the Nearest Neighbour method for finding centers. The average Acc of centers was unsatisfactory. Performance on the OMoCap was better than that on the IMU data. In Type 2 attribute representation, the performance of Gender and Height attributes outperforms that of Weight and Age for the OMoCap and IMU data. Similar results were obtained from the $\mathrm{BCE}_{\mathrm{loss}}$ method for finding center.

Some interesting points that were analysed during the experimentation are:

- The more the dimensionality of the attribute representation, the more the accuracy can be achieved in finding the centers
- If different subjects have the same representation, the classifier merged the two subjects classifications into one.
- Attribute accuracy was found to be better for OMoCap data than IMU data.

Results of Experiment 4B*:*

This section discusses the impact of activities on attribute representation. The experiments follow the format explained in Sec: 5.2 (Experiment 3_{IOA}). We consider five experiments on IMU data for four sub-categories and the whole data. A distinction between Type 1 and Type 2 attribute representation was not enforced. The result is presented in Table: 5.4.18.

Activity	IOA^{+}_{al}	IOA^{-}_{al}
Walking	97.94 ± 2.81	2.05 ± 2.81
Cart	97.60 ± 3.14	2.39 ± 3.14
Handling cen	94.14 ± 5.87	5.85 ± 5.87
Handling down	77.42 ± 8.67	22.56 ± 8.67
Handling up	97.21 ± 3.62	2.78 ± 3.62
Standing	93.42 ± 3.78	6.56 ± 3.78
Synch	96.99 ± 4.22	3.0 ± 4.22

Table 5.4.18: Impact of activity on attribute representation. Values are presented as percentages.

Walking and cart activities have high classification accuracy. Interestingly, handling up activity has a similarly high classification ratio. Similar to the result of Experiment $3A_{LARa}$, handling down has the least classification accuracy.

Based on the results, the activities with upper body movements, such as synchronisation, handling up and handling center, shows better classification results for attribute representation. This better performance of upper body activities is a deviation from the performance shown in identity classification, where the activities did not perform as well as the activities with gait.

5.5 LAYER-WISE RELEVANCE PROPAGATION

From the experimental results of Sec: 5.4, we concluded that person identification is feasible with IMU-based motion information. Furthermore, we analysed the impact of activities and the dataset on the identification accuracy. Though these experiments gave an overview of how the number of sensors and type of activities affect person identification, we were not able to identify data specific features or motion signatures that help the network to identify the individual. A CNN has the ability to extract innate features from the data to facilitate classification with the help of the convolutional filters. However, accessing these features of the data learnt by the convolutional filters are difficult. The branch of research that facilitates human-understandable interpretation and explanation of non-linear ML algorithm's behaviour is called Explainable Artificial Intelligence (XAI) [SML+21]. XAI methods applied on Neural Networks trained on human motion datasets can help identify the data features that contribute to person identification.

According to [SML+21], there are various techniques for explaining the predictions of the network based on the input data, namely, gradient-based techniques, occlusion analysis and layer-wise relevance propagation (LRP). This thesis attempted at layer-wise relevance propagation (LRP) [MLB+17]. LRP can be derived from deep Taylor decomposition. Here, the contribution of the neurons to the solution/classification is traced back into the network. Each neuron of a layer receives a relevance score. The relevance score accounts for the contribution of the neuron to the final solution. The total relevance score of a layer will be equal to that of the previous layer, following Kirchhoff's law of conservation. Fig: 5.5.1 visually explains the steps followed by the LRP method.

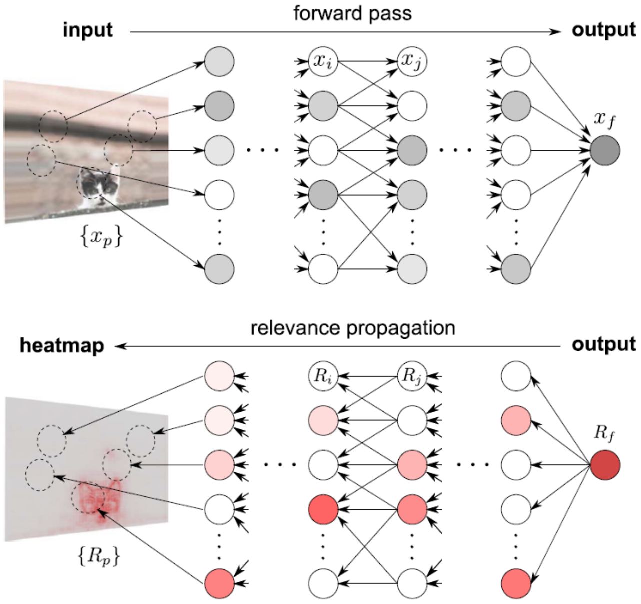


Figure 5.5.1: Visualisation of layer-wise relevance propagation [MLB+17].

LRP is mainly applied on Neural Networks [MBL+19] after they have completed training and can achieve good classification accuracy. Firstly, the trained network is tested with a sample. During the forward pass of the test sample, the activation of each neuron of every layer is recorded. In the relevance propagation phase, the activation obtained for the class of interest in the final layer is maintained as the relevance score. The relevance score of the final layer is propagated back into the network proportional to the neurons' activation value [MLB+17]. The backpropagation of relevance is achieved with deep Taylor decomposition. Following the Kirchhoff's law of conservation, the relevance of a neuron j (R_j) in the lower layer will be the summation of the proportional neuron activation that contributed information to the neuron k in the consecutive layer. Thus, propagation of relevance from neurons in the top layer to a neuron in the lower layer will follow the Eq: 5.5.1. Here, j and k refers to two neurons in consecutive layers, R refers to relevance, and z_{jk} decides the contribution of neuron j that facilitated in the activation of neuron k. The denominator helps to achieve conservation as per Kirchoff's law [MBL+19]. Finally, the relevance is mapped onto the input image and represented as a heat map.

$$R_j = \sum_k \frac{z_{jk}}{\Sigma_j z_{jk}} R_k \tag{5.5.1}$$

Based on the type of layer, various propagation rules can be applied. Few prominent propagation rules are: Basic rule (LRP-0) Eq: 5.5.2, Epsilon rule (LRP-ϵ) Eq: 5.5.3, and Gamma Rule (LRP-γ) Eq: 5.5.4 [MBL+19]. Where, a_j refers to the lower layer activation, w_{jk} refers to the synaptic weights and $\Sigma_{0,j}$ implies that the summation is over all the lower layer activations.

The LRP-0, Eq: 5.5.2, conducts redistribution of relevance proportional to the neuron's contribution, which is controlled by the synaptic weights. The rule can be applied uniformly on the network. However, the rule is susceptible to gradient noise.

$$R_j = \sum_k \frac{a_j w_{jk}}{\Sigma_{0,j} a_j w_{jk}} R_k \tag{5.5.2}$$

A modification to the LRP-0 is the LRP-ϵ rule Eq: 5.5.3. Here, a small positive term ϵ is added to the denominator of the function, calculating the relevance score. The ϵ term helps to filter gradient noise by absorbing the relevance of weak activations of neuron k. Thus, increasing the ϵ value would imply that only strong activations would be maintained [MBL+19]. The LRP-ϵ was used as the propagation rule for this thesis.

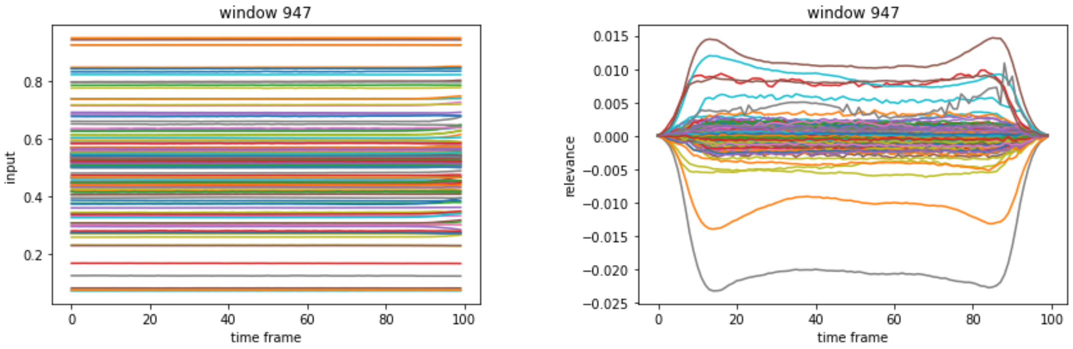
$$R_j = \sum_k \frac{a_j w_{jk}}{\epsilon + \Sigma_{0,j} a_j w_{jk}} R_k \tag{5.5.3}$$

The LRP-γ, Eq: 5.5.4, favours positive contributions of the neurons over negative contributions. The term γ controls the amount of favouritism to positive contributions. Consequently, this rule achieves a stable explanation of the features contributing to the classification [MBL+19].

$$R_j = \sum_k \frac{a_j(w_{jk} + \gamma w_{jk}^+)}{\Sigma_{0,j} a_j(w_{jk} + \gamma w_{jk}^+)} R_k \tag{5.5.4}$$

The relevance score can be a positive, zero or negative value. A positive relevance score is associated with a feature that has contributed to the selection of the class. Zero relevance score imply that the features do not contribute to the class selection. A negative relevance score implies that the feature present in the data is indicating a different class. Features with negative relevance scores lead to misclassification [MBL+19].

Here, we applied LRP on $\mathrm{OMoCap}_W^{C,NN}$ of LARa dataset trained on a CNN with HPs: $Lr = 10^{-4}$, epoch 10 and $mBsize = 100$, for the purpose of person identification. The LRP-ϵ rule was uniformly applied on the CNN layers. The ϵ value was fixed at 10^{-9}. Fig: 5.5.2 shows the input data and relevance score of each sensor channel of Subject 1. The frame was correctly classified as Subject 1. The subject was performing the activity "Standing" in this frame. As mentioned in Sec: 5.1.1, the sliding window size is 100. As the input data is normalised, their values range from 0 to 1.

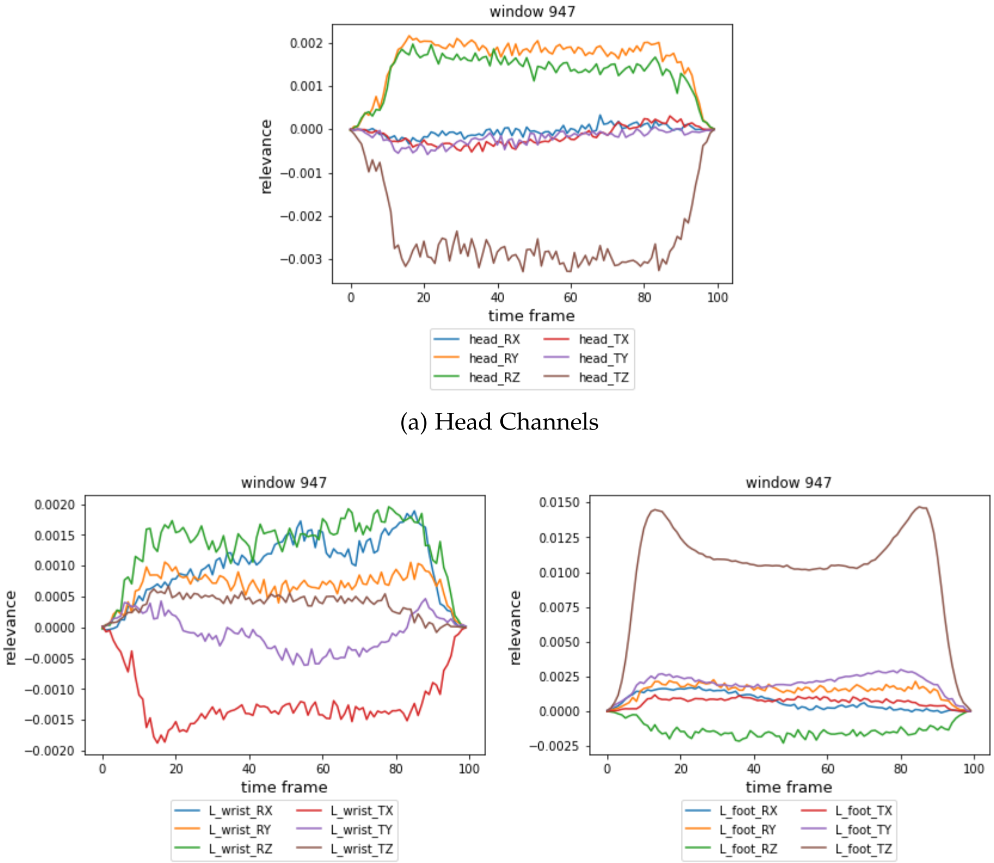


(a) Normalised sensor values

(b) Relevance graph

Figure 5.5.2: Input and relevance graph of Subject 1. The subject was correctly identified in this frame. The activity label of the frame was "Standing".

From the Fig: 5.5.2 (b) Relevance graph, it can be seen that many of the sensor channels have relevance scores close to zero. ϵ needs to be experimented upon to ensure that the relevance of the channels is not gradient noise. The negative relevance value presented is larger than the positive relevance. Consequently, LRP-γ might be the appropriate choice of propagation rule to be applied on the network trained with normalised OMoCap data. Thus, future work using LRP on person identification needs to experiment on various propagation rules to identify the appropriate relevance graph.



(a) Head Channels

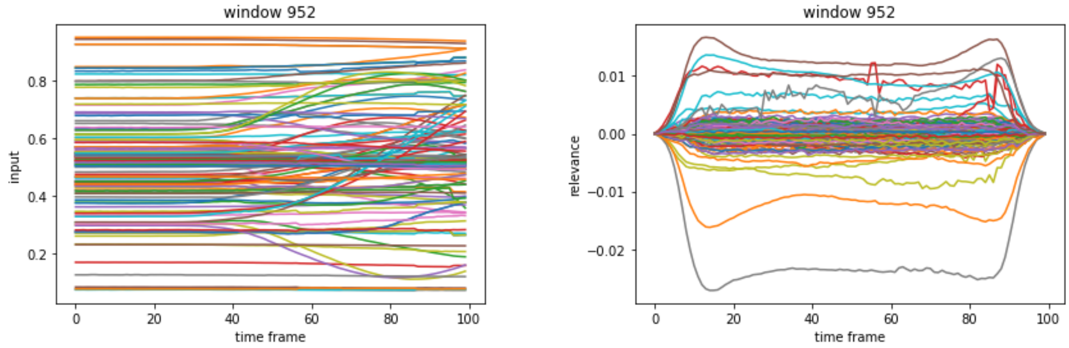
(b) Left Wrist Channels

(c) Left foot Channels

Figure 5.5.3: LRP on subject 1 of LARa dataset. The window was correctly classified as subject 1. The plots show the relevance of the sensor channels of Subject 1 performing the activity "Standing". (a) presents the relevance of the sensor channels related to the head. (b) and (c) represent the sensor channels with respect to the left wrist and left foot.

OMoCap data has 126 sensor channels. Due to the impracticality of representing all the sensor channels in a legend, we have selected six sensor channels related to the head of the subject, left wrist and left foot, and presented their relevances in Fig: 5.5.3. The relevance presented is from window 947 of the test dataset, Fig: 5.5.2(b). The highest noticeable relevance present in Fig: 5.5.2 is close to 0.015. The relevance is associated with the left leg's translation axis Z, as shown in Fig: 5.5.3(c). This implies that one of the features or sensor channels that helps the network to predict subject 1 is the movement of the subject along the translation axis Z. The rotational axes Y and Z of the head have positive relevance on person identification, whereas the translation axis Z has a negative relevance, Fig: 5.5.3(a). The relevance of the wrist channels are presented in Fig: 5.5.3(b). The translational axis X shows high negative relevance in comparison to the other sensor channels.

The input and relevance score of Subject 1 performing the activity "Synchronisation" is presented in Fig: 5.5.4. The window was correctly classified as subject 1 by the network. Compared to the activity "Standing", the sensor channels show variation of sensor values with respect to time in the input graph. As the activity "Synchronisation" is the waving motion of the hands, the sensor channels associated with the hand are shown to vary in time. The subject is expected to be standing while performing the wave. We can see that few of the sensor channel values remain constant in time. These sensor channels show the coordinates of the leg.



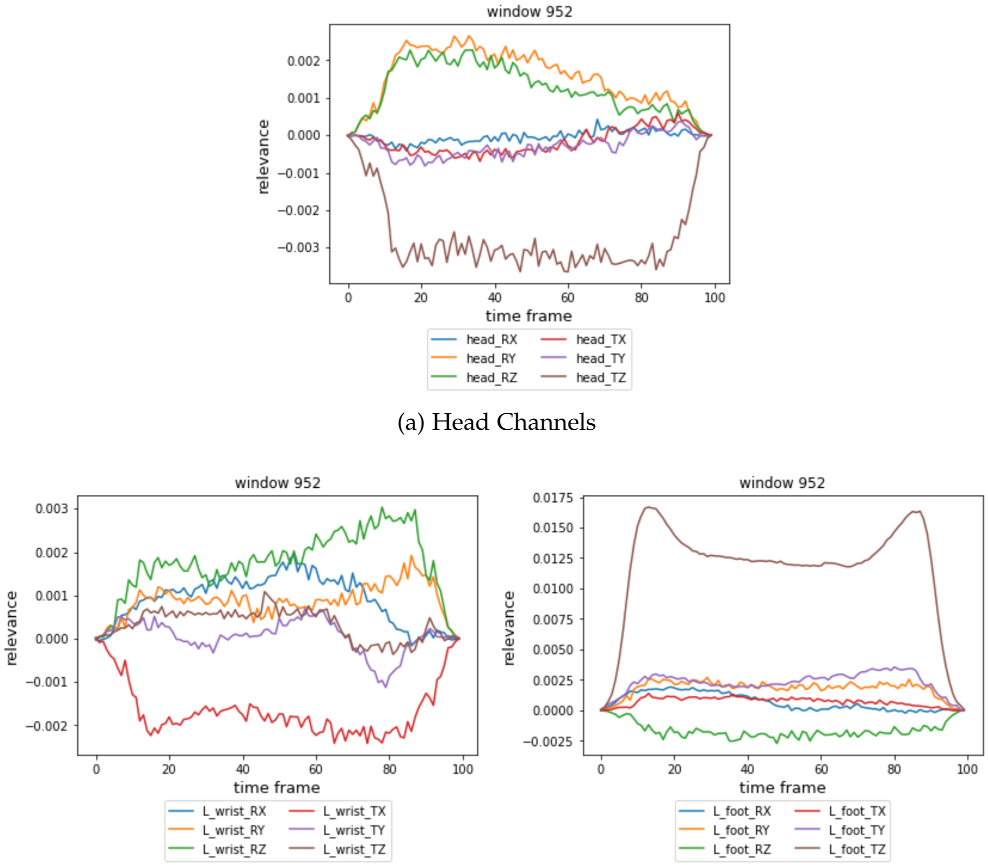
(a) Normalised sensor values

(b) Relevance graph

Figure 5.5.4: Input and relevance graph of Subject 1. The subject was correctly identified in this frame. The activity label of the frames was "Synchronisation"

.

Similar to Fig: 5.5.3, we have extracted the rotational and translational axes relevances of the head, left wrist and left foot of the subject 1 for the window 952, as shown in Fig: 5.5.5. As mentioned, the network was able to classify the subject for this window.



(a) Head Channels

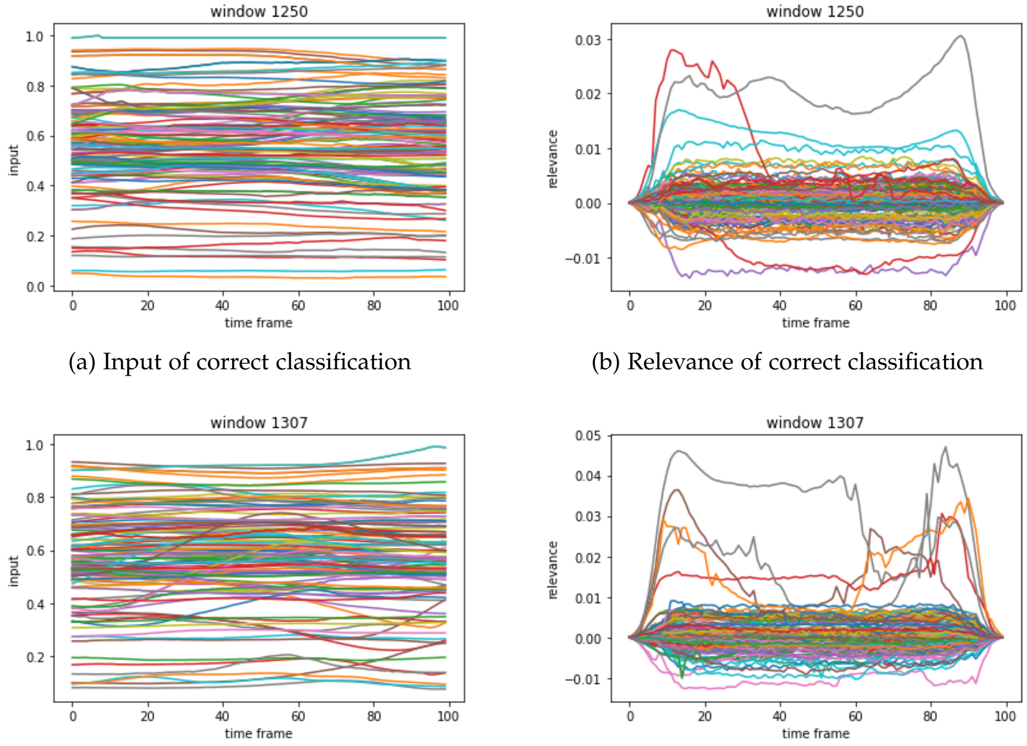
(b) Left Wrist Channels

(c) Left foot Channels

Figure 5.5.5: LRP on subject 1 of LARa dataset. The window was correctly classified as subject 1. The plots show the relevance of the sensor channels of Subject 1 performing the activity "Synchronisation". (a) presents the relevance of the sensor channels related to the head. (b) and (c) represent the sensor channels with respect to the left wrist and left foot.

One interesting aspect to notice from the graphs (a) and (b) of Fig: 5.5.5 is that, as the subject raises the hands to show the synchronisation action, the relevance score of the wrist rotational axis Z increases. The relevance score of the rotational axes Y and Z associated with the head decreases simultaneously. This observation could confirm the conclusions derived from the Sec: 5.4. That is, activities performed by the subject influences the person identification and that the network does not generalise identity over activities. More experiments need to be conducted in this direction to confirm the understanding.

Though Fig: 5.5.2 and Fig: 5.5.4 represent two different activity classes, a similarity in the relevance of the sensor channels can be noticed; for example, there are three sensor channels showing prominent negative relevance in the graphs. Analysing such trends with respect to the individuals and activity can help draw conclusive statements on the impact of activity and the sensor positions on the identity classification.



(a) Input of correct classification

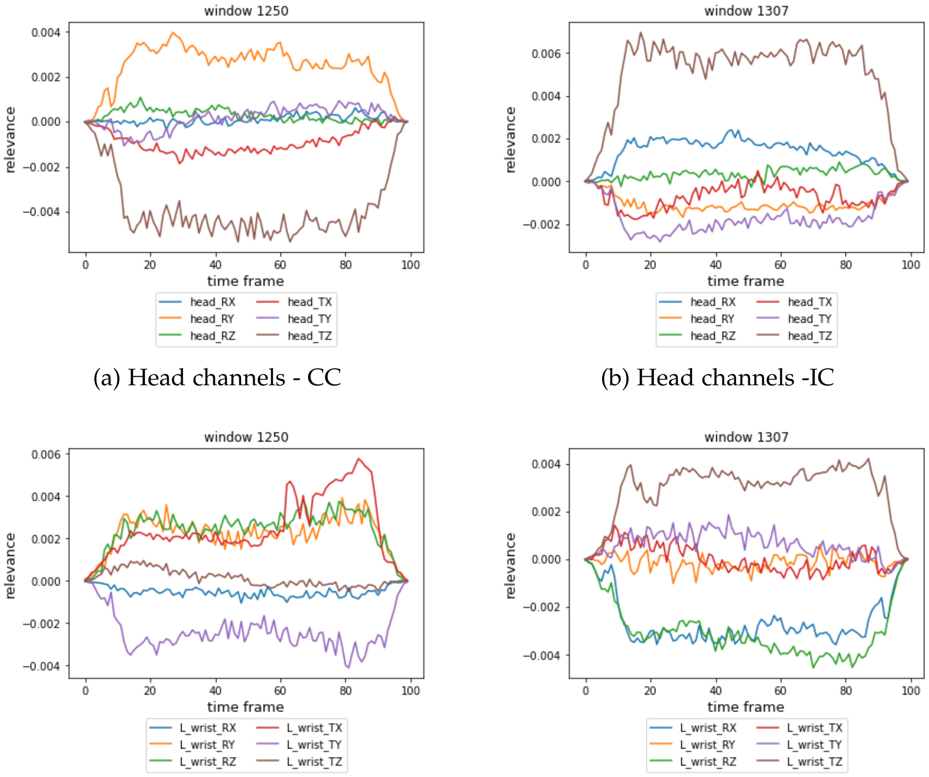
(b) Relevance of correct classification

(c) Input of incorrect classification

(d) Relevance of incorrect classification

Figure 5.5.6: LRP on subject 1 of LARa dataset. The plots show the relevance of the sensor channels of Subject 1 performing the activity "Cart". (a) and (b) present the input and relevance of the frame that was correctly classified as subject 1. (c) and (d) represent the input and relevance of the sensor channels of the frame that was misclassified as subject 6, though it belonged to subject 1.

Next, a comparison between a correctly classified window and a misclassified window is presented in Fig: 5.5.6. The windows belong to subject 1 performing cart activity. Window 1250 was correctly classified as subject 1; however, window 1307 was classified as subject 6.



(a) Head channels - CC

(b) Head channels -IC

(c) Left Wrist channels - CC

(d) Left Wrist channels - IC

Figure 5.5.7: LRP on subject 1 of LARa dataset. The plots show the relevance of the sensor channels of Subject 1 performing the activity "Cart". A comparison between a correctly identified window and an incorrectly classified window is presented. CC refers to correct classification, and IC refers to incorrect classification. (a) and (b) shows the head channels' relevance, and (c) and (d) shows the left wrist channels' relevance.

When calculating the relevance, we take into account the class with the maximum final layer activation. The propagation of relevance score takes place with respect to this class. In Fig: 5.5.6(d), the graph presents the relevance of the sensor that contributed to the classification of the window as subject 6. Thus, the figure represents the relevance of the sensor channels that help the network identify subject 6. Compared to Fig: 5.5.6(b), the magnitude of relevance values in Fig: 5.5.6(d) is larger. Furthermore, the sequence of relevance seems to be different. To facilitate better comparison, Fig: 5.5.7 shows the relevance values of the head and wrist sensor channels.

From Fig: 5.5.7, we can analyse the drastic deviation in the relevance of the two windows of subject 1. An interesting trend was noticed with respect to the translational Z axis of the head. Comparing the Fig: 5.5.3(a), Fig: 5.5.5(a), Fig: 5.5.7(a), and (b), we can see that when the windows were correctly classified as subject 1, the translational axis Z had a negative relevance. However, in the case of incorrect window classification as subject 6, translation Z has high positive relevance. This may suggest that the subject 1 and 6 have similar head movement along the Z axis, but the movement is a strong indicator of subject 6's identity. Drawing out such similarities may help to tackle misclassification and achieve high person identification accuracy. To confirm the observation, more examples have to be analysed and experimented.

From these preliminary experiments and analyses, we can conclude that by applying LRP on human motion data, we can derive interesting observations and proofs that support the observations of this thesis. As a result, this thesis recommends LRP as part of the future work of this thesis.

6 CONCLUSIONS

The main goal of this thesis was to explore the possibility of person identification using motion information obtained from IMU data by training Deep Neural Networks, such as CNN and RNN. Further, it aimed to analyze the impact of activities contained in the data on the identification process. In addition, attribute representations designed from soft biometrics are intended.

The networks of interest were the CNN-IMU and deepCNNLSTM networks. The networks are designed based on the sensor-based late-fusion method. Consequently, each IMU sensor is associated with a branch of convolutional layers to extract descriptive features from the sensor channels (e.g., x, y, z axes of accelerometer) to facilitate identification. Each branch has four convolutional layers. Pooling layers were not used in these architectures. In the CNN-IMU network, the last layer of the branches is a fully connected layer. The features of the branches are then concatenated and passed through an MLP. The classifier layer is activated by the softmax activation function to facilitate classification. In contrast, the features from the branches of the deepCNNLSTM are directly concatenated, and passed through two LSTM layers. However, similar to CNN-IMU, the classifier layer is activated by the softmax activation function. LARa, OPPORTUNITY, PAMAP2 and Order Pickings are the four datasets considered for experimentation.

An individual can be described by their soft biometrics, such as age, gender, height and handedness. An attribute representation method provides semantic information of data. Consequently, soft biometrics can be designed to be an attribute representation of individuals. This thesis designed two datasets specific attribute representation. The soft biometric details of the subjects were obtained from the LARa dataset's recording protocol. The Type 1 attribute representation considered a binary split of soft biometrics; for example, height soft biometric can be classified as $\leqslant 180$ or > 180. The second type of attribute representation has considered more splits; for example, soft biometric height was split as $\leqslant 170$, 170-180 and > 180. As a result, the Type 1 attribute representation has four attributes, while the Type 2 attribute representation has ten attributes. To facilitate attribute representation classification, the final layer of the CNN-IMU network was activated by the sigmoid activation function.

The attribute representation assigned to an individual is referred to as center. To calculate the center, two methods were followed; the Nearest Neighbour method and the Binary Cross-Entropy Loss method. In the Nearest Neighbour method, the predicted attribute representation is assigned to the nearest center in the attribute space. In the Binary Cross Entropy loss method, the deviation of the prediction from the expected representation is calculated. The attribute representation with the most proximity is assigned as the center.

To identify the better network, the CNN-IMU and deepCNNLSTM networks were trained on the LARa dataset's OMoCap and IMU data for similar hyperparameters. The CNN-IMU network was found to perform classifications with better accuracy in comparison to deepCNNLSTM for similar hyperparameters. As a result, CNN-IMU network was used for the rest of the experiments.

The CNN-IMU networks were trained on channel-normalised, and non-channel normalised IMU data of the LARa dataset. The results pointed that channel-normalised data performed better than non-channel normalised data. Thus, a conclusion was derived that the placement biases did not significantly affect identification and that channel normalisation improved identification accuracy.

The LARa dataset set recordings were grouped based on Scenarios, recordings and individuals. Four sub-categories were created. The sub-category A excluded the subject 12 of the LARa dataset. Only recordings of Scenario 2 were present in this group. Similarly, the sub-category B excluded the subject 11. However, the group consisted of Scenarios 2 and 3 recordings of the LARa dataset. Both subjects 11 and 12 were excluded from the sub-category C. The sub-category D excluded the subjects 10, 11 and 12. The category had the maximum number of recordings for training. The sub-categories B, C and D consists of recordings from the Scenario 2 and 3. The sub-categories were applied on the OMoCap and IMU data, and were trained using the CNN-IMU network. In few cases, networks trained on the sub-categorised IMU data performed better than the sub-categorised OMoCap data; for example, the networks trained with the same hyperparameters on IMU and OMoCap were able to achieve an average accuracy of 93.96% and 85.62%, respectively. However, when the whole dataset was considered for training, the network trained on OMoCap performed better than IMU. The network trained on OMoCap data achieved an average accuracy of 97%, while the network trained on IMU achieved 92.07% average accuracy.

From the experiments performed on the sub-categorised data, it was observed that networks do not learn to generalise identity over different activities. To elaborate, the CNN-IMU network was trained with recording of Scenario 2 of the LARa dataset and tested on recordings of Scenario 3, the networks' performance dropped compared to the networks trained on less number of data. Thus, the network cannot identify an individual present in the training set, if the activity performed by the individual in the test sample is not present in the training set.

To benchmark the possibility of person identification with the IMU dataset created for HAR, identification was experimented on OPPORTUNITY, PAMAP2 and Order Picking datasets. OPPORTUNITY is a sensor-rich dataset. A CNN-IMU network trained on the OPPORTUNITY was able to achieve an averaged person identification accuracy of 96.03% over all experiments. Similarly, a network trained on PAMAP2 was able to achieve an average accuracy of 91.01%. Furthermore, experiments on PAMAP2 confirmed the observation that networks can not generalise identity over activities. Person identification on the Order Picking dataset performed poorly. The maximum accuracy achieved was 52.23% on six subjects. The performance could be attributed to the low amount of data, and placement of IMUs.

Due to the absence of a recording protocol for the Order Picking dataset, the author of the thesis was unable to identify the relation between the IMU sensor channel to the placement of the sensor. Consequently, a CNN network was used to train the dataset. To ensure that the network chosen has not affected the identification accuracy, the PAMAP2 was experimented on a CNN. Similar to the Order Picking dataset, the PAMAP2 consisted of three IMUs. However, unlike CNN trained on the Order Picking dataset, the CNN trained on the PAMAP2 showed performance similar to the CNN-IMU network. Hence, ruling out the effect of the network architecture on identification. The major difference between the PAMAP2 and Order picking is the placement of IMUs. The IMUs of the Order Picking dataset were placed on both wrists and chest, unlike in PAMAP2, where the IMUs were placed on the chest, dominant hand wrist and ankle. Due to the absence of the recording protocol, experiments to analyse the effect of the sensor placement on person identification were not feasible.

The impact of activities on person identification was tested with the four datasets, namely, LARa. OPPORTUNITY, PAMAP2 and Order Picking datasets, trained on the CNN-IMU network. Consistent with the gait-based person identification research, identification accuracy was favourable for test windows with activities such as walking, Nordic walk, and pushing cart. This behaviour was observed on the LARa and PAMAP2 datasets. From the research by [EBL18], activities with less body movements were observed to have high classification accuracy. OPPORUTNITY dataset's experimental results had windows with sedentary activities with high person identification accuracy similar to [EBL18]. Interestingly, experiments on the impact of activities on attribute representation on LARa dataset showed results similar to [EBL18]. That is, activities with small body movements were found to have higher attribute representation accuracy than activities with large body movements.

Of the two types of attribute representations, the attribute representation with larger dimensionality had better performance. In particular, the Type 2 attribute representation had higher accuracy while calculating the centers of the individual.

The Binary Cross-Entropy Loss method for finding the centers performed better than the Nearest Neighbour approach. Gender and height soft biometrics consistently provided exceptional results during experiments with both OMoCap and IMU data. The performance can account for the swing movement of the body during gait. The experiments collectively pointed towards the necessity of a larger dataset for attribute representation.

The soft biometric split for creating attribute representations was specific to the LARa dataset. Attempting to apply the same attribute representation on another dataset may not be feasible. To elaborate, when applied on different datasets, attribute representation may have attributes with no variation. Consequently, the network will not be able to learn any useful information with respect to the attribute. Thus, research is needed to create attribute representations transferable to different datasets.

To obtain an interpretable explanation of the features learnt by the network, the Explainable Artificial Intelligence (XAI) method, layer-wise relevance propagation (LRP) was explored. LRP is capable of visualising the features the neural network expect within the input data to support classification. LRP was implemented on a CNN network, trained on the OMoCap data of LARa dataset. The LRP-ϵ rule was applied on all layers of the CNN. The LRP was applied on the test windows of subject 1 of the LARa dataset. The value of negative relevances were found to be greater than the positive relevances in few frames. Furthermore, the relevance graph was found to be noisy. Consequently, future experiments are recommended by varying the ϵ value of the LRP-ϵ. In addition, the authors recommend experiments on different propagation rule or combination of propagation rules as part of future work. From the brief analysis of the subject 1 frames, it was identified that, experiments on LRP can deliver proofs for the observations and conclusions derived in this thesis with respect to person identification and impact of activities on identification accuracy.

6.1 FUTURE WORK

The thesis established the possibility of person identification using motion information. The future work is the analysis of the features learned by the Neural Network. Identifying the features that facilitated the network to perform person identification, would help in masking or deleting the feature to ensure privacy or in identifying the impact of subject identity on HAR. This thesis implemented the LRP on a CNN trained on OMoCap data of LARa, using the LRP-ϵ rule on all layers of the CNN. The relevance of few windows were presented and analysed. However, analysing few individual windows is not sufficient to have a comprehensive understanding of the features. Consequently, methods to facilitate group analysis, such as relevance pooling and Spectral Relevance Analysis (SpRAy) [SML+21], need to be implemented. Furthermore, an implementation of LRP on a CNN-IMU network with the help of [ANS+21] needs to be attempted. This step would help to analyse and cross-verify the experimental results obtained in this thesis. In addition, the LRP propagation rules have to be experimented upon. This thesis applied LRP-ϵ rule on all the layers, however, [MBL+19] has recommended attempting combination of propagation rules to achieve the optimal feature explanation. [ANS+21] provides frameworks for implementing LRP on custom networks and analysing LRP on datasets. These frameworks are expected to help carry out the relevance analysis in a structured manner. Consequently, experimentation on these frameworks is a direction this thesis would recommend.

[EBL18] mentions that few subjects are easily identifiable than others. That is, few subjects have prominent action signatures. The uniqueness and identifiability of these individuals can be associated with their idiosyncrasies. The motion signatures can be further explored by analysing the features that support identification using the LRP method. In addition, the impact of individual motion signatures on HAR can be researched.

Soft biometrics-based attribute representation needs to be explored further. The primary effort required in the direction of attribute representation using soft biometrics is the creation of a larger dataset. The dataset has to be inclusive of larger variations in soft biometric characteristics and ranges. To elaborate, when considering the soft biometric age, the dataset must ideally have individuals from a wider age range. Thus, the attribute thresholds for creating soft biometric splits must be revisited. A method for obtaining transferable attribute representation based on soft biometrics have to be designed.

The thesis confirmed the impact of activities on identities and soft biometric attribute representations. Future research may constitute the impact of the identities on activities, in specific, on HAR datasets.

BIBLIOGRAPHY

[AC99] Aggarwal, Jake K. ; Cai, Quin: Human motion analysis: A review. In: *Computer vision and image understanding* 73 (1999), Nr. 3, S. 428–440

[AH17] Aghdam, Hamed H. ; Heravi, Elnaz J.: Guide to convolutional neural networks. In: *New York, NY: Springer* 10 (2017), S. 978–973

[ANS+21] Anders, Christopher J. ; Neumann, David ; Samek, Wojciech ; Müller, Klaus-Robert ; Lapuschkin, Sebastian: Software for Dataset-wide XAI: From Local Explanations to Global Insights with Zennit, CoRelAy, and ViRelAy. In: *CoRR* abs/2106.13200 (2021). https://arxiv.org/abs/2106.13200

[BL05] Boyd, Jeffrey E. ; Little, James J.: Biometric gait recognition. In: *Advanced Studies in Biometrics*. Springer, 2005, S. 19–42

[CC19] Chien, JT ; Chien, JT: Deep neural network. In: *Source Separation and Machine Learning* (2019), S. 259–320

[CDHG20] Chunsheng, Hu ; De, Wang ; Huidong, Zhao ; Guoli, Li: Human Gait Feature Data Analysis and Person Identification Based on IMU. In: *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)* IEEE, 2020, S. 437–442

[Cha18] Charu, C A.: *Neural Networks and Deep Learning*. 2018

[Cou16] Council, European: *REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL*. Official Journal of the European Union, 2016

[CPC85] Caspersen, Carl J. ; Powell, Kenneth E. ; Christenson, Gregory M.: Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research. In: *Public health reports* 100 (1985), Nr. 2, S. 126

[CSC+13] Chavarriaga, Ricardo ; Sagha, Hesam ; Calatroni, Alberto ; Digumarti, Sundara T. ; Tröster, Gerhard ; Millán, José del R ; Roggen, Daniel: The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition. In: *Pattern Recognition Letters* 34 (2013), Nr. 15, S. 2033–2042

[DS17] Dey, Rahul ; Salem, Fathi M.: Gate-variants of gated recurrent unit (GRU) neural networks. In: *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)* IEEE, 2017, S. 1597–1600

[DSGS19] Dehghani, Akbar ; Sarbishei, Omid ; Glatard, Tristan ; Shihab, Emad: A quantitative comparison of overlapping and non-overlapping sliding windows for human activity recognition using inertial sensors. In: *Sensors* 19 (2019), Nr. 22, S. 5026

[DTC17] Dehzangi, Omid ; Taherisadr, Mojtaba ; ChangalVala, Raghvendar: IMU-based gait recognition using convolutional neural networks and multi-sensor fusion. In: *Sensors* 17 (2017), Nr. 12, S. 2735

[EA07] Ekinci, Murat ; Aykut, Murat: Human gait recognition based on kernel PCA using projections. In: *Journal of Computer Science and Technology* 22 (2007), Nr. 6, S. 867–876

[EBL18] Elkader, Seham A. ; Barlow, Michael ; Lakshika, Erandi: Wearable sensors for recognizing individuals undertaking daily activities. In: *Proceedings of the 2018 ACM International Symposium on Wearable Computers*, 2018, S. 64–67

[FEHF09] Farhadi, Ali ; Endres, Ian ; Hoiem, Derek ; Forsyth, David: Describing objects by their attributes. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition* IEEE, 2009, S. 1778–1785

[Gaf07] Gafurov, Davrondzhon: A survey of biometric gait recognition: Approaches, security and challenges. In: *Annual Norwegian computer science conference* Annual Norwegian Computer Science Conference Norway, 2007, S. 19–21

[GBC16] Goodfellow, Ian ; Bengio, Yoshua ; Courville, Aaron: *Deep Learning*. MIT Press, 2016. – `http://www.deeplearningbook.org`

[GLR+17] Grzeszick, Rene ; Lenk, Jan M. ; Rueda, Fernando M. ; Fink, Gernot A. ; Feldhorst, Sascha ; Hompel, Michael ten: Deep neural network based human activity recognition for the order picking process. In: *Proceedings* *of the 4th international Workshop on Sensor-based Activity Recognition and Interaction*, 2017, S. 1–6

[GRS+20] Gohar, Imad ; Riaz, Qaiser ; Shahzad, Muhammad ; Ul Hasnain Hashmi, Muhammad Z. ; Tahir, Hasan ; Ehsan Ul Haq, Muhammad: Person re-identification using deep modeling of temporally correlated inertial motion patterns. In: *Sensors* 20 (2020), Nr. 3, S. 949

[Gup13] Gupta, Neha: Artificial neural network. In: *Network and Complex Systems* 3 (2013), Nr. 1, S. 24–28

[HB05] Han, Jinguang ; Bhanu, Bir: Individual recognition using gait energy image. In: *IEEE transactions on pattern analysis and machine intelligence* 28 (2005), Nr. 2, S. 316–322

[HHP16] Hammerla, Nils Y. ; Halloran, Shane ; Plötz, Thomas: Deep, convolutional, and recurrent models for human activity recognition using wearables. In: *arXiv preprint arXiv:1604.08880* (2016)

[HWZ+12] Hu, Maodi ; Wang, Yunhong ; Zhang, Zhaoxiang ; Zhang, De ; Little, James J.: Incremental learning for video-based gait recognition with LBP flow. In: *IEEE transactions on cybernetics* 43 (2012), Nr. 1, S. 77–89

[IT] *Iterate Labs Inc.* https://iteratelabs.co/, . – Accessed: 2021-06-29

[JRP04] Jain, A. K. ; Ross, A. ; Prabhakar, S.: An introduction to biometric recognition. In: *IEEE Transactions on Circuits and Systems for Video Technology* 14 (2004), Nr. 1, S. 4–20. http://dx.doi.org/10.1109/TCSVT.2003.818349. – DOI 10.1109/TCSVT.2003.818349

[KF18] Kong, Yu ; Fu, Yun: Human action recognition and prediction: A survey. In: *arXiv preprint arXiv:1806.11230* (2018)

[LBBH98] LeCun, Yann ; Bottou, Léon ; Bengio, Yoshua ; Haffner, Patrick: Gradient-based learning applied to document recognition. In: *Proceedings of the IEEE* 86 (1998), Nr. 11, S. 2278–2324

[LJZ09] Liu, Ling-Feng ; Jia, Wei ; Zhu, Yi-Hai: Survey of gait recognition. In: *International Conference on Intelligent Computing* Springer, 2009, S. 652–659

[MBL+19] Montavon, Grégoire ; Binder, Alexander ; Lapuschkin, Sebastian ; Samek, Wojciech ; Müller, Klaus-Robert: Layer-wise relevance propagation: an overview. In: *Explainable AI: interpreting, explaining and visualizing deep learning* (2019), S. 193–209

[MF18] Moya Rueda, Fernando ; Fink, Gernot A.: Learning attribute representation for human activity recognition. In: *2018 24th International Conference on Pattern Recognition (ICPR)* IEEE, 2018, S. 523–528

[MGF+18] Moya Rueda, Fernando ; Grzeszick, René ; Fink, Gernot A. ; Feldhorst, Sascha ; Ten Hompel, Michael: Convolutional neural networks for human activity recognition using body-worn sensors. In: *Informatics* Bd. 5 Multidisciplinary Digital Publishing Institute, 2018, S. 26

[MLB+17] Montavon, Grégoire ; Lapuschkin, Sebastian ; Binder, Alexander ; Samek, Wojciech ; Müller, Klaus-Robert: Explaining nonlinear classification decisions with deep taylor decomposition. In: *Pattern Recognition* 65 (2017), S. 211–222

[MM] *Motion-Miners GmbH*. https://www.motionminers.com/?lang=en, . – Accessed: 2021-06-29

[MP88] Minsky, Marvin L. ; Papert, Seymour A.: *Perceptrons: expanded edition*. 1988

[MSR+17] Münzner, Sebastian ; Schmidt, Philip ; Reiss, Attila ; Hanselmann, Michael ; Stiefelhagen, Rainer ; Dürichen, Robert: CNN-based sensor fusion techniques for multimodal human activity recognition. In: *Proceedings of the 2017 ACM International Symposium on Wearable Computers*, 2017, S. 158–165

[NRMR+20] Niemann, Friedrich ; Reining, Christopher ; Moya Rueda, Fernando ; Altermann, Erik ; Nair, Nilah R. ; Steffens, Janine A. ; Fink, Gernot A. ; Hompel, Michael ten: *Logistic Activity Recognition Challenge (LARa) – A Motion Capture and Inertial Measurement Dataset*. http://dx.doi.org/10.5281/zenodo.3862782. Version: Mai 2020. – Acknowledgement: The work on this publication was supported by Deutsche Forschungsgemeinschaft (DFG) in the context of the project Fi799/10-2, HO2403/14-2 "Transfer Learning for Human Activity Recognition in Logistics".

[NRR+20] Niemann, Friedrich ; Reining, Christopher ; Rueda, Fernando M. ; Nair, Nilah R. ; Steffens, Janine A. ; Fink, Gernot A. ; Hompel, Michael t.: LARa: Creating a Dataset for Human Activity Recognition in Logistics Using Semantic Attributes. In: *Sensors* 20 (2020), Nr. 15, S. 4083

[opp10] *OPPORTUNITY Activity Recognition Data Set*. https://archive.ics.uci.edu/ml/datasets/OPPORTUNITY+Activity+Recognition#, 2010. – Accessed: 2021-07-27

[OR16] Ordóñez, Francisco J. ; Roggen, Daniel: Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. In: *Sensors* 16 (2016), Nr. 1, S. 115

[pam12] *PAMAP2 Physical Activity Monitoring Data Set*. https://archive.ics.uci.edu/ml/datasets/PAMAP2+Physical+Activity+Monitoring, 2012

[PP18] Poznyak, Tatyana ; Poznyak, Alex: *Ozonation and biodegradation in environmental engineering: Dynamic neural network approach*. Elsevier, 2018

[RA09] Ryoo, Michael S. ; Aggarwal, Jake K.: Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In: *2009 IEEE 12th international conference on computer vision* IEEE, 2009, S. 1593–1600

[RAS08] Rodriguez, Mikel D. ; Ahmed, Javed ; Shah, Mubarak: Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: *2008 IEEE conference on computer vision and pattern recognition* IEEE, 2008, S. 1–8

[RCR+10] Roggen, Daniel ; Calatroni, Alberto ; Rossi, Mirco ; Holleczek, Thomas ; Förster, Kilian ; Tröster, Gerhard ; Lukowicz, Paul ; Bannach, David ; Pirkl, Gerald ; Ferscha, Alois u. a.: Collecting complex activity datasets in highly rich networked sensor environments. In: *2010 Seventh international conference on networked sensing systems (INSS)* IEEE, 2010, S. 233–240

[RF18] Rueda, Fernando M. ; Fink, Gernot A.: Learning Attribute Representation for Human Activity Recognition. In: *Proc. Int. Conf. on Pattern Recognition*. Bejing, China : IEEE, 2018. – ISSN 1051–4651, S. 523–528

[Roj96] Rojas, Raul: Statistics and neural networks. In: *Neural Networks*. Springer, 1996, S. 227–261

[RS12a] Reiss, Attila ; Stricker, Didier: Creating and benchmarking a new dataset for physical activity monitoring. In: *Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments*, 2012, S. 1–8

[RS12b] Reiss, Attila ; Stricker, Didier: Introducing a new benchmarked dataset for activity monitoring. In: *2012 16th International Symposium on Wearable Computers* IEEE, 2012, S. 108–109

[RSH⁺18] Reining, Christopher ; Schlangen, Michelle ; Hissmann, Leon ; Hompel, Michael ten ; Moya, Fernando ; Fink, Gernot A.: Attribute representation for human activity recognition of manual order picking activities. In: *Proceedings of the 5th international Workshop on Sensor-based Activity Recognition and Interaction*, 2018, S. 1–10

[Rud16] Ruder, Sebastian: An overview of gradient descent optimization algorithms. In: *arXiv preprint arXiv:1609.04747* (2016)

[RUD21] RUDOLPH, Günther: *Lecture notes from Introduction to Computational Intelligence.* https://ls11-www.cs.tu-dortmund.de/people/rudolph/teaching/lectures/CI/WS2020-21/lec09.pdf, 2021. – Version: March 2021

[RVKW15] Riaz, Qaiser ; Vögele, Anna ; Krüger, Björn ; Weber, Andreas: One small step for a man: Estimation of gender, age and height from recordings of one step by a single inertial sensor. In: *Sensors* 15 (2015), Nr. 12, S. 31999–32019

[SJAS18] Singh, Jasvinder P. ; Jain, Sanjeev ; Arora, Sakshi ; Singh, Uday P.: Vision-based gait recognition: A survey. In: *IEEE Access* 6 (2018), S. 70497–70527

[SLC04] Schuldt, Christian ; Laptev, Ivan ; Caputo, Barbara: Recognizing human actions: a local SVM approach. In: *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* Bd. 3 IEEE, 2004, S. 32–36

[SML⁺21] Samek, Wojciech ; Montavon, Grégoire ; Lapuschkin, Sebastian ; Anders, Christopher J. ; Müller, Klaus-Robert: Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. In:

Proceedings of the IEEE 109 (2021), Nr. 3, S. 247–278. http://dx.doi.org/10.1109/JPROC.2021.3060483. – DOI 10.1109/JPROC.2021.3060483

[SNM+12] Shahid, Saman ; Nandy, Anup ; Mondal, Soumik ; Ahamad, Maksud ; Chakraborty, Pavan ; Nandi, Gora C.: A study on human gait analysis. In: *Proceedings of the Second International Conference on Computational Science, Engineering and Information Technology*, 2012, S. 358–364

[WBR16] Wolf, Thomas ; Babaee, Mohammadreza ; Rigoll, Gerhard: Multi-view gait recognition using 3D convolutional neural networks. In: *2016 IEEE International Conference on Image Processing (ICIP)* IEEE, 2016, S. 4165–4169

[WTNH03] Wang, Liang ; Tan, Tieniu ; Ning, Huazhong ; Hu, Weiming: Silhouette analysis-based gait recognition for human identification. In: *IEEE transactions on pattern analysis and machine intelligence* 25 (2003), Nr. 12, S. 1505–1518

[YNS+15] Yang, Jianbo ; Nguyen, Minh N. ; San, Phyo P. ; Li, Xiaoli ; Krishnaswamy, Shonali: Deep convolutional neural networks on multichannel time series for human activity recognition. In: *Ijcai* Bd. 15 Buenos Aires, Argentina, 2015, S. 3995–4001

[Yun11] Yun, Jaeseok: User identification using gait patterns on UbiFloorII. In: *Sensors* 11 (2011), Nr. 3, S. 2611–2639

[ZKL+13] Zhang, Tianxiang ; Karg, Michelle ; Lin, Jonathan Feng-Shun ; Kulic, Dana ; Venture, Gentiane: Imu based single stride identification of humans. In: *2013 IEEE RO-MAN* IEEE, 2013, S. 220–225

[ZLC+14] Zheng, Yi ; Liu, Qi ; Chen, Enhong ; Ge, Yong ; Zhao, J L.: Time series classification using multi-channels deep convolutional neural networks. In: *International conference on web-age information management* Springer, 2014, S. 298–310

[ZLC+17] Zhao, Bendong ; Lu, Huanzhang ; Chen, Shangfeng ; Liu, Junliang ; Wu, Dongya: Convolutional neural networks for time series classification. In: *Journal of Systems Engineering and Electronics* 28 (2017), Nr. 1, S. 162–169