healthcare

Article

A Deep Learning Approach for TUG and SPPB Score Prediction of (Pre-) Frail Older Adults on Real-Life IMU Data

Björn Friedrich [1,*], Sandra Lau [2], Lena Elgert [3], Jürgen M. Bauer [4] and Andreas Hein [1]

[1] Assistance Systems and Medical Device Technology, Carl von Ossietzky University, Ammerländer Heerstraße 114-118, 26129 Oldenburg, Germany; andreas.hein@uni-oldenburg.de

[2] Geriatric Medicine, Carl von Ossietzky University, Ammerländer Heerstraße 114-118, 26129 Oldenburg, Germany; sandra.lau@uni-oldenburg.de

[3] Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Carl-Neuberg-Straße 1, 30625 Hannover, Germany; lena.elgert@plri.de

[4] Center for Geriatric Medicine, Agaplesion Bethanien Krankenhaus, University of Heidelberg, Rohrbacher Straße 149, 69117 Heidelberg, Germany; juergen.bauer@bethanien-heidelberg.de

* Correspondence: bjoern.friedrich@uni-oldenburg.de

Abstract: Since older adults are prone to functional decline, using Inertial-Measurement-Units (IMU) for mobility assessment score prediction gives valuable information to physicians to diagnose changes in mobility and physical performance at an early stage and increases the chances of rehabilitation. This research introduces an approach for predicting the score of the Timed Up & Go test and Short-Physical-Performance-Battery assessment using IMU data and deep neural networks. The approach is validated on real-world data of a cohort of 20 frail or (pre-) frail older adults of an average of 84.7 years. The deep neural networks achieve an accuracy of about 95% for both tests for participants known by the network.

Keywords: pre-frail; frail; older adults; mobility assessments; machine learning; supervised learning; decision support

Citation: Friedrich, B.; Lau, S.; Elgert, L.; Bauer, J.M.; Hein, A. A Deep Learning Approach for TUG and SPPB Score Prediction of (Pre-) Frail Older Adults on Real-Life IMU Data. *Healthcare* **2021**, *9*, 149. https://doi.org/10.3390/healthcare9020149

Academic Editor: Tin-Chih Toly Chen

Received: 14 January 2021

Accepted: 25 January 2021

Published: 2 February 2021

1. Introduction

With the demographic change a lot of challenges arise, particularly in the fields of medicine and healthcare. Older adults need frequent medical attention for maintaining their health and physical performance. Functional decline and mobility impairments are symptoms of impending diseases and also for frailty [1]. Early diagnosed diseases can be treated well, impairments can be minimised and there is a good chance of rehabilitation. Additionally, older adults have a high risk of falling and the medical implications can be more serious than the incident itself. Usually, standardised geriatrics assessments are used for monitoring the physical performance and estimating the risk of falling of older adults. Although digital solutions are available, the conventional assessments are still mainly performed under the supervision of a medical professional, therefore it is time consuming in clinical settings. Long-term monitoring on a regular basis assessments exceed the logistic capacities of medical professionals. Another disadvantage of assessments is the test situation itself, because people tend to give their best effort in test situations and studies showed that capacity is not performance [2–4]. Sensor assisted monitoring of the mobility and the physical performance in everyday life situations can give valuable information to health professionals for diagnosis and therapy. IMU sensors are low-cost and unobtrusive sensors for measuring the body's specific force and angular acceleration. The light weight and the small size makes those sensors ideal for carrying around in a pocket or attached to a belt in daily life. Moreover, IMUs are not dependent on external infrastructure like satellites and can be used inside as well as outside without any loss of accuracy.

Machine learning approaches and recently deep learning showed good result in estimating gait parameters and fall risk of older adults [5]. One advantage of deep learning approaches is that the algorithms can approximate arbitrary functions, extract features automatically and the time consuming and difficult step of handcrafting features is not needed [6–9]. This advantage comes at the price of computationally expensive training and hyperparameter optimisation.

In this contribution a machine learning model for predicting the assessment scores for the Short-Physical-Performance-Battery (SPPB) and the Timed Up & Go test (TUG) on IMU data is introduced. The model learned from IMU data collected in everyday life situations from a cohort of older adults (84.75 y, 5.19 y SD). This article is structured as follows, in Section 2 the state of the art of mobility assessments and technology approaches for assisting are described. Technology assistance is divided in the two subgroups technology assisted assessments and unsupervised assessment approaches. In Section 3 the study for data acquisition, steps of preprocessing the data and the machine learning approach are explained. The results are shown in Section 4 and discussed in the following Section 5. In the last section conclusions are drawn and further steps are briefly mentioned.

2. State of the Art

Assessing the mobility of older adults is a common task in geriatrics medicine, validated and well accepted assessments like the SPPB [10] and the TUG [11] test are commonly used.

The TUG assesses the mobility of an older adult by getting up from a standardised chair, walking a distance of three metres, turning around and getting back to the chair to sit down. Assistive devices used for walking are permitted but it must be documented and used for any re-tests. The time from the start command "Go" until the patient's buttocks touches the seat again is measured in seconds. Assessment categories are <10 s = normal, no mobility impairment, 11–19 s = minor mobility impairment not relevant in everyday life, 20–29 s = mobility impairment, >30 s severe mobility impairment, need for intervention [11]. Each category has numeric score, no mobility impairment is 1, minor mobility impairment is 2, mobility impairment is score 3, and severe mobility impairment is 4.

The SPPB consists of three parts assessing balance, gait speed and lower limb strength. During the balance test, the participant stands with the feet side by side, in semi-tandem stance and tandem stance for ten seconds each. Habitual gait speed is measured over a distance of 2.40 m, 3 m or 4 m. The chair rise test assess the muscle strength of the lower extremities by measuring the seconds the participant needs to perform five times from sit to stand. A maximum of 4 points for each task can be achieved. A total SPPB score ≤ 9 points was found as cut-off value for fit and frail people [10,12]. The approaches to support the assessments using technology can be categorised in technology assisted assessments and unsupervised assessments in real-life.

The technology assisted assessments are still performed in a way the assessment is supposed to be in order to use the validated point values for evaluation. Technical devices are used to enhance the measurements and to support the supervisors as well as the participants. In [13] an approach to enhance the TUG test was introduced. The measurements of the values' score were computed and automatically measured using an IMU sensor. The computed values showed a high correlation to the gold standard measurement under supervision of a health professional using a stopwatch. A system for automated SPPB assessment executions has been developed in [14].

The unsupervised assessment approaches are trying to detect motion combinations from assessments in real-life by sensors. Once a motion combination is recognised the recorded sensor data is used to compute the performance. In [15] an approach for automatically detect TUG execution using an IMU sensor is shown and the results were compared to traditional stopwatch measurements. The TUG could be recognised with an accuracy of 96% and the results showed a strong correlation with the conventional method. As an item of the SPPB test and important parameter for functional decline, gait speed is focused as well. Approaches using cameras to measure the gait speed in domestic environments are introduced in [16–18]. Instead of cameras, more privacy respecting ambient sensors can be used as well like in [19–22]. However, all of those sensors are fixed and measure the gait speed in a single location only. To overcome this disadvantage small portable low-cost IMU sensors can be used to measure gait speed as well. The research in [23,24] showed, that gait speed estimations based on IMU data are comparable to gold-standard measurements of a GAITRite walkway. The experiments described in [25] showed the validity and reproducibility of using IMU sensors for gait parameter estimation. Besides the gait parameter measurements IMU data can be used to measure the intensity of activities [26].

Deep learning approaches showed good results on estimating gait parameters and the risk of falling on wearable sensor data [5]. Yu et al. used IMU data of Parkinson's disease patients collected during TUG assessments for estimating the fall risk and the severity of Parkinson's disease symptoms. The Convolutional Neural Network achieved an F-measure of 94% for estimating the fall risk and a Root-Mean-Squared-Error of 0.06 for severity estimation [27]. Similar approaches showed good results for people suffering from neurological disorders and multiple sclerosis. The networks achieved an accuracy of 92.1% and an Area Under the Curve (AUC) of 0.88 respectively [28,29]. Considering the parameters age and gender in addition to the raw sensor data significantly improves the performance of deep learning models [30]. Convolutional Neural Networks are also able to predict the frailty and cognitive dysfunction in respect to the mini-mental state examination of older adults. The network was trained with spectograms of walking-in-place data collected by IMUs. The network achieved an accuracy of 94.63% and 97.59% for frailty and cognitive dysfunction respectively [31]. An Artificial Neural Network in combination with a pressure sensor was used to detect abnormal foot postures. The pressure sensors were placed inside shoes and the Artificial Neural Network classified abnormal foot postures based on the gait characteristics with an accuracy of 99% [32]. Gait abnormalities can be detected using wrist-worn IMUs as well. A deep neural network trained IMU data collected by smartwatches achieved an accuracy of 88.90%, a sensitivity of 90.60%, and a specificity of 86.20% [33]. Deep Neural Networks also has been used to detect the fall incidents themselves. Using wearable sensor data an accuracy of 97.16% was achieved on the task of fall detection [34]. Using accelerometer data only a sensitivity of 88.20% and a specificity of 96.40% could be achieved in [35].

The mentioned approaches focus on enhancing the execution of geriatrics assessments using technology or detecting motion patterns from assessments in real-life. The approach in the article at hand is different, because the score of an assessment is predicted on IMU data. The participant neither has to execute any special motion patterns nor to complete the assessment in a certain place.

3. Methods and Materials

3.1. Data Acquisition

The data was collected during the observational OTAGO study in 2014 and 2015 over a period of 10 months [36]. The functional performance was assessed every month by the TUG and SPPB tests through conventional stopwatch measurements by a health professional. The cohort consisted of 20 participants (17 female, 3 male) aged 76 to 92 (mean 84.3 y, SD 5.19 y). At baseline, 14 participants (70%) were identified as frail (Frailty-Index $\geq$ 2 pts.) and six participants were pre-frail. The mean scores of the functional performance were 17.9 s for TUG and 5.95 points for SPPB. The baseline characteristics are presented in Table 1. Table 2 shows the characteristics of the cohort at the end of the study. The cohort size is reduced by two, because two participants deceased during the study. Despite dropout the available data of these two participants was considered. Each participant got one IMU sensor for data collection.

Table 1. The baseline characteristics of the study cohort.

n = 20	Age (y)	BMI ($\frac{kg}{m^2}$)	Frailty Index (pts.)	SPPB (pts.)	TUG (s)
Mean	84.75	27.39	1.90	5.95	17.87
SD (±)	5.19	6.10	0.72	2.33	5.33
Range (min-max)	76.00–92.00	17.33–43.09	1.00–3.00	3.00–11.00	11.16 - 31.63

Table 2. The characteristics at the end of the study cohort.

n = 18	Age (y)	BMI ($\frac{kg}{m^2}$)	Frailty Index (pts.)	SPPB (pts.)	TUG (s)
Mean	85.44	28.27	2.00	6.61	16.12
SD (±)	4.92	6.44	0.97	2.85	5.85
Range (min-max)	77.00–93.00	16.89–45.99	0.00–4.00	2.00–12.00	8.15–30.06

The IMU sensor of type Shimmer 3r was capable of measuring force in 9 Degrees of Freedom (9DOF) and was comprised of wide range and low range accelerometer, gyroscope, magnetometer, and pressure sensor [37]. In the first two weeks of the study the sensor was set to 51.2 Hz and in the remaining time of the study the sensor was set to 102.4 Hz. The sensors were given to the participants before the TUG and SPPB were done and collected after around about two weeks. Therefore, the dataset consists test situations and everyday life situations. Participants were asked to wear the the Shimmer3r IMU the whole day on a sensor belt, in a trouser pocket or other small pocket on the right side of the hip. The logo should face the front and the side with the charging port facing down. At night the sensor was supposed to be placed on the charging station. The participants were instructed to store the sensor safely during taking a shower and in the exercise bath. In total, 259 days of IMU data were collected.

3.2. Preprocessing

Before using the data for learning several preprocessing steps were applied. Participants 2, 3, and 4 were excluded, because of unavailable IMU data. The models are tested with two different testing strategies. The first testing strategy is the common testing strategy in machine learning, a test set is separated from the dataset. The second strategy is to test the models on data of participants which have been excluded before processing the data. The data of the excluded participants were neither used for training nor for evaluation of the model during training time. The participants 16 and 19 were randomly chosen for evaluating the performance of the model on unknown participants. This led to an exclusion of the SPPB score 2, because participant 19 was the only participant with a score of 2 in a SPPB assessment. Only two TUG assessments took longer than 30 s (max. +1.6 s). The scores were included in score 3, for not losing the data. The values of the low range accelerometer, wide range accelerometer, gyroscope and magnetometer were chosen as input to the network. The values of the temperature sensor were mostly 0 and contained errors and were not considered. The orientation in relation to the earth coordinate frame of the IMU was unknown and to eliminate the influence of the orientation the magnitude for each sensor modality was computed by

$$m_i = \sqrt{x_i^2 + y_i^2 + z_i^2} \tag{1}$$

where i is the index of the value and x, y, and z are the axis of the sensor. After computing the magnitude the sensor data was filtered by a second order low-pass filter with a cut-off frequency of $\frac{1}{4} \times sampling\ frequency$. The filtered data has been divided into 5 s non-overlapping windows. Since two different sampling frequencies were used in the study the number of values per window differed. The number of values per window were rounded down to 500 values for the sampling frequency of 102.4 Hz and to 250 values for

51.2 Hz. Latter were oversampled to 500 values per window by duplicating each value. The final sets may contain data from assessments as well, but the amount of windows containing assessment data is insignificant. For each participant about 180,018.13 (SD ± 19,565.80) samples are available and a maximum of 1320 samples of each participants contain assessment data. The number of windows per class of the resultant dataset are shown in Table 3. Due to the setting of the data acquisition the classes were imbalanced and for several classes no windows were available at all. To balance the dataset windows of the overrepresented classes were deleted and the score 12 was excluded from the SPPB. Considering the smallest class as lower threshold for balancing, would have led to massive loss of data. The dataset was balanced two times and two different datasets, one for each assessment were derived. The dataset for the SPPB had 148,307 windows per class and the dataset for the TUG had 169,711 windows per class. The sets were divided into subsets for training (75%), validation (15%) and test (10%) in a stratified fashion.

The data of the two participants reserved for testing were preprocessed in the same way, but were not balanced and not divided into subsets. The data for SPPB score 2 of participant 19 were not considered for evaluation, because the model was not trained for that class. The number of values per class for participants 16 and 19 are shown in Tables 4 and 5 respectively. The data was collected over a certain period of time and the physical performance changed during that time. So, one participant could have achieved different assessment scores.

Before the data was fed to the model the data was scaled to a range of 0 and 1.

Table 3. This table shows the number of windows for each class of the SPPB and the TUG assessments. The range of the TUG score is smaller than the range of the SPPB score.

Score	TUG	SPPB
1	216,624	0
2	2,016,236	83,248
3	467,412	128,331
4	-	371,263
5	-	298,456
6	-	235,227
7	-	522,077
8	-	357,449
9	-	379,761
10	-	145,679
11	-	178,402
12	-	379

Table 4. This table shows the number of windows for participant 16 who was excluded from the training set.

Score	TUG	SPPB
2	88,733	0
3	80	0
4	-	14,505
10	-	43,708

Table 5. This table shows the number of windows for participant 19 who was excluded from the training set. Score 2 of the SPPB were not considered for evaluation, because the model for the SPPB was not trained with class 2.

Score	TUG	SPPB
2	28,336	256
3	82,010	8633
4	-	28,184
5	-	67,403
7	-	5868

3.3. Network Architecture

For this research a deep neural network approach was used. The architecture is shown in Figure 1 and the blocks are shown more detailed in Figure 2. The architecture and window size are adapted from [38]. Two models were trained, one for each assessment. The difference of the models was the number of neurons of the classification layer. For the SPPB model 9 neurons, and for the TUG model 3 neurons were used. In both networks the final classification layer is activated by the softmax function. The trained networks are available in the supplementary materials.

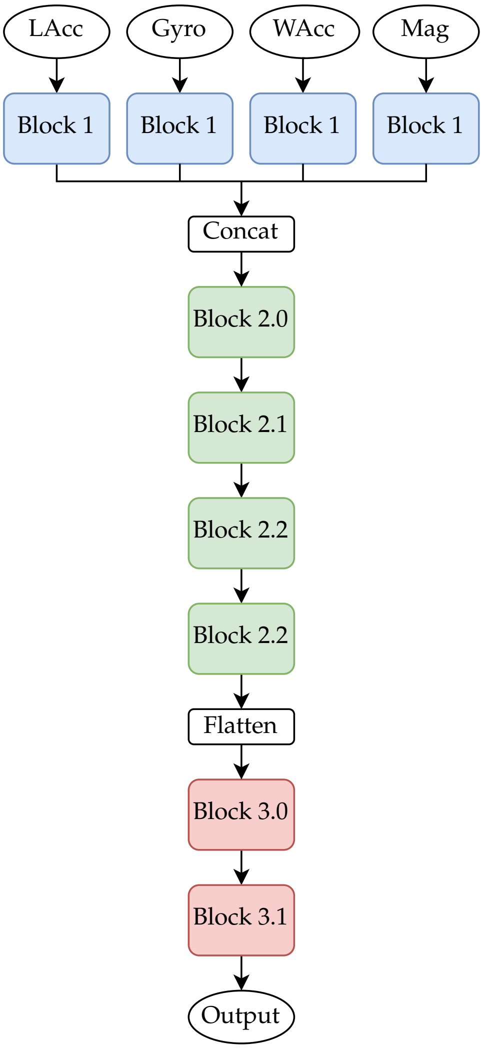

Figure 1. The deep neural network used for this research. Each sensor modality had its own input. The inner structure of the blocks are shown in Figure 2.

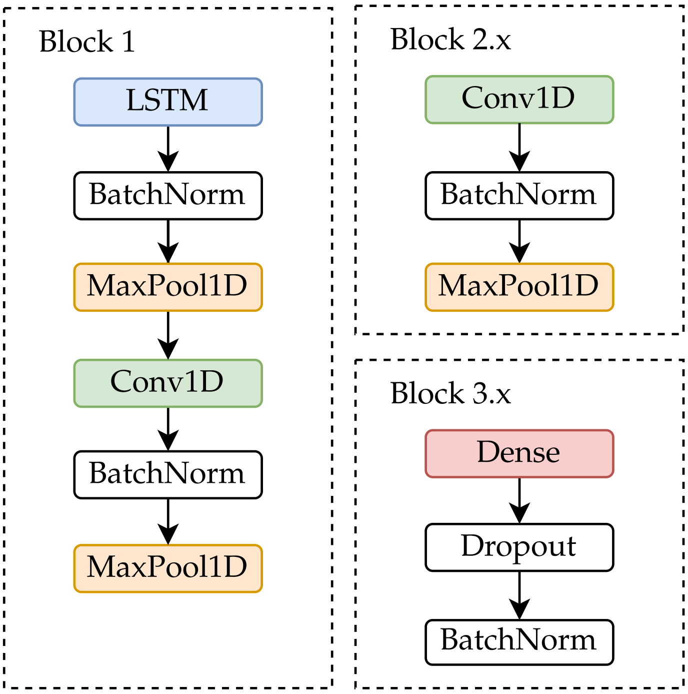

Figure 2. The blocks of the deep neural network. Details about the blocks and the layer parameters can be found in the Tables A1–A6 in Appendix A.

As first layers Long-Short-Term-Memory (LSTM) layers are used for capturing the relation between the time steps. Then a convolutional layer was added to learn features for each sensor modality. All intermediate features were concatenated and forwarded to a sequence of convolutional layers. The final classification was performed by a two layer neural network. Dropout, maximum pooling and batch normalisation layers were added to prevent overfitting.

The model was trained using categorical cross-entropy as loss function and accuracy as metric. The accuracy was computed as follows

$$accuracy = \frac{correct\ classified}{all\ samples} \tag{2}$$

The optimiser was AMSGrad version of the *Adaptive Moment Estimation* (Adam) with an initial learning rate of 0.001, a first order derivative momentum of 0.5, a second order derivative momentum of 0.8, and an exponential decay after the first 10 epochs [39].

$$lr(epoch) = 0.001 \times e^{0.1 \times (10 - epoch)} \tag{3}$$

4. Results

The best epochs were epoch 67 for the SPPB model with a validation accuracy of 94.29% and epoch 52 for the TUG model with a validation accuracy of 95.89%. The accuracy on the test %set was 94.28%, and 95.79% for the SPPB and the TUG model respectively. The accuracy for the TUG scores (2, 3) for participant 16 was 98.84%, and 26.15% for participant 19. The accuracy for the SPPB scores (4 and 10) of participant 16 was 6.39%, and for SPPB scores (3, 4, 5, and 7) for participant 19 was 14.13%. The Tables 6–8 give an overview over the results. The Receiver Operating Characteristic (ROC) curves in Figures 3 and 4 show high true positive and low false positive rates at high decision thresholds. The average ROC curves (blue) and the ROC curves for each class are showing similar progress and overlap. The AUCs of the TUG model and scores are 0.99 and the AUCs of the SPPB model are 1 except for the classes 7 and 9, where the AUCs are 0.99. The confusion matrices for the SPPB and TUG models are shown in Tables 9 and 10. The SPPB model classified score 11 best with 573 false classifications (sensitivity: 96.14%, specificity: 99.27%) and score 9 worst with 1381 false classifications (sensitivity: 90.69%, specificity: 98.85%). The TUG model classified score 3 best with 771 (sensitivity: 96.44%, specificity: 97.93%) false classifications and score 2 worst with 1082 false classifications (sensitivity: 95.01%, specificity: 97.25%). The most false classification of the TUG model is for adjacent classes, e.g., 614 samples of class 1 were classified as class 2, but only 268 samples as class 3.

The Figures 5 and 6 are showing the progress of the loss during training and the Figures 7 and 8 are showing the accuracy during training. The validation loss and the validation accuracy are fluctuating in the beginning, but become stabilised after epoch 25. Overall, the loss and accuracy graphs showed the desired behaviour, increasing fast in the early epochs and stabilising during the later epochs. The graphs for both models are similar, but the SPPB model shows a little less performance and little higher loss than the TUG model.

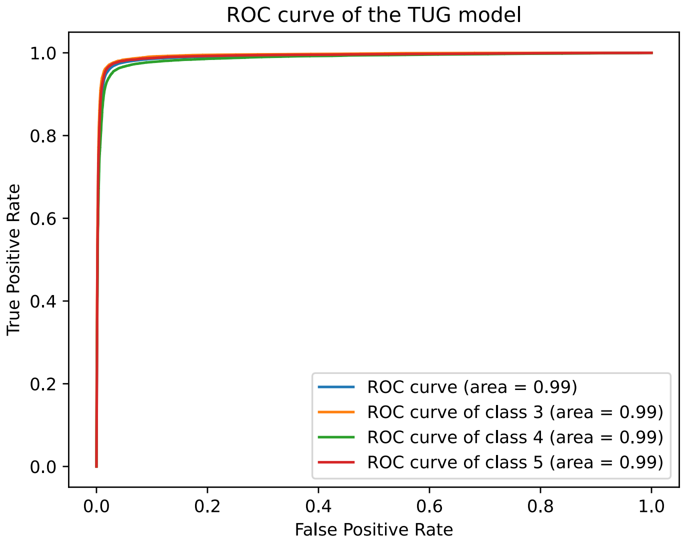

Figure 3. The ROC curve of the TUG model.

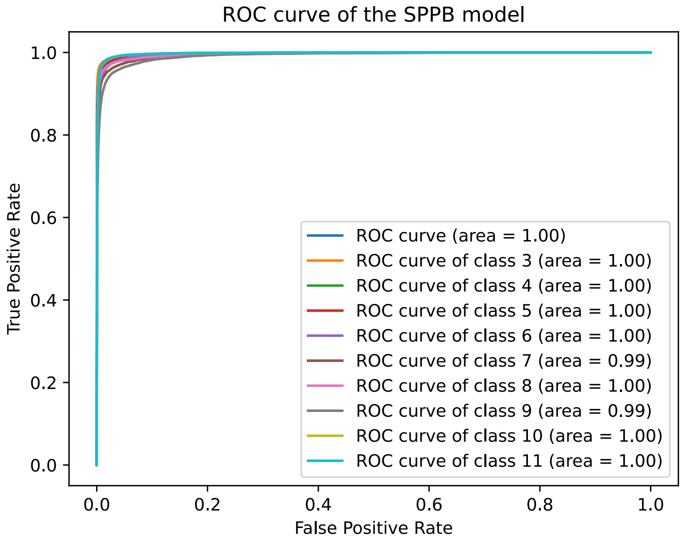

Figure 4. The ROC curve of the SPPB model.

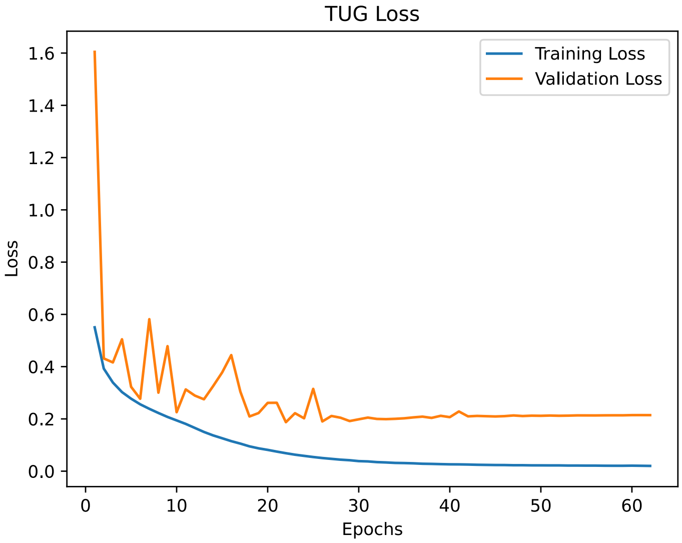

Figure 5. The loss of the TUG model. For the first 25 epochs the loss indicates that the learning rate is slightly too large. From epoch 25 the progress shows an asymptotic behaviour.

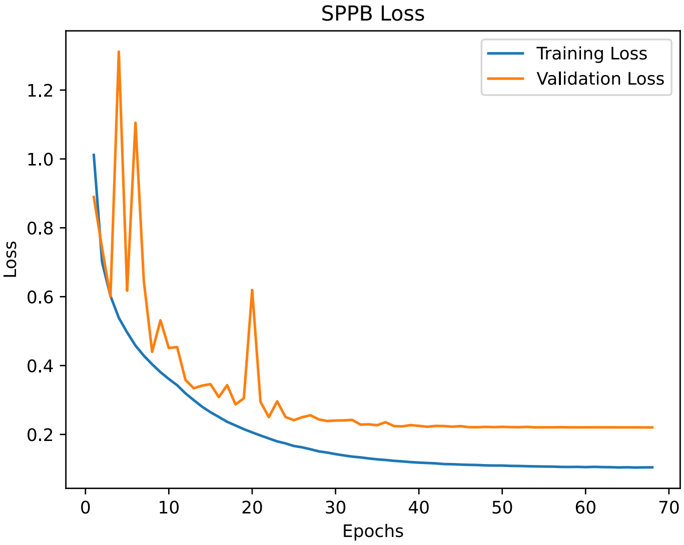

Figure 6. The loss of the SPPB model. For the first 25 epochs the loss indicates that the learning rate is slightly too large. From epoch 25 the progress shows an asymptotic behaviour.

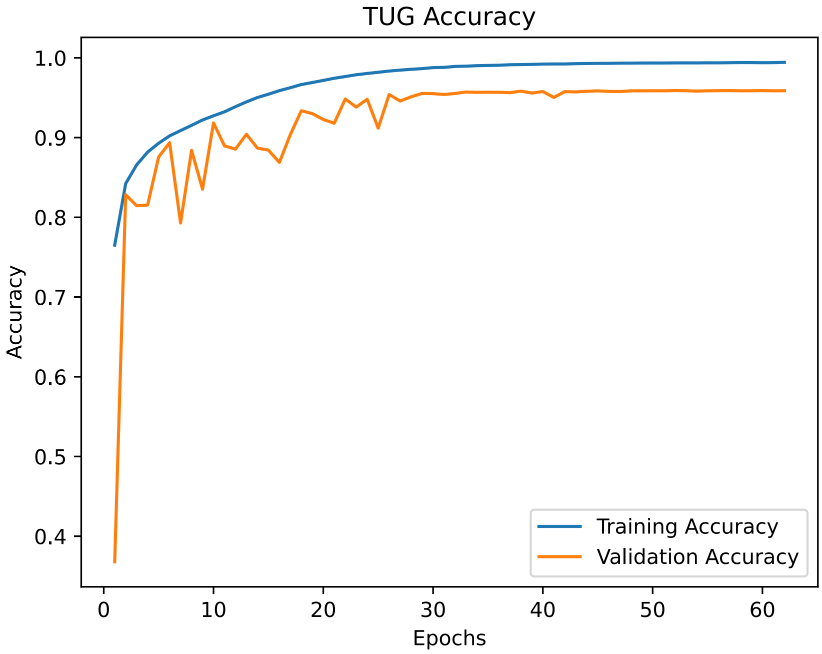

Figure 7. The accuracy of the TUG model. According to the progress of the loss, the accuracy fluctuates in the first 25 epochs and shows an asymptotic behaviour after epoch 25. The best validation accuracy score was 95.89% at epoch 52.

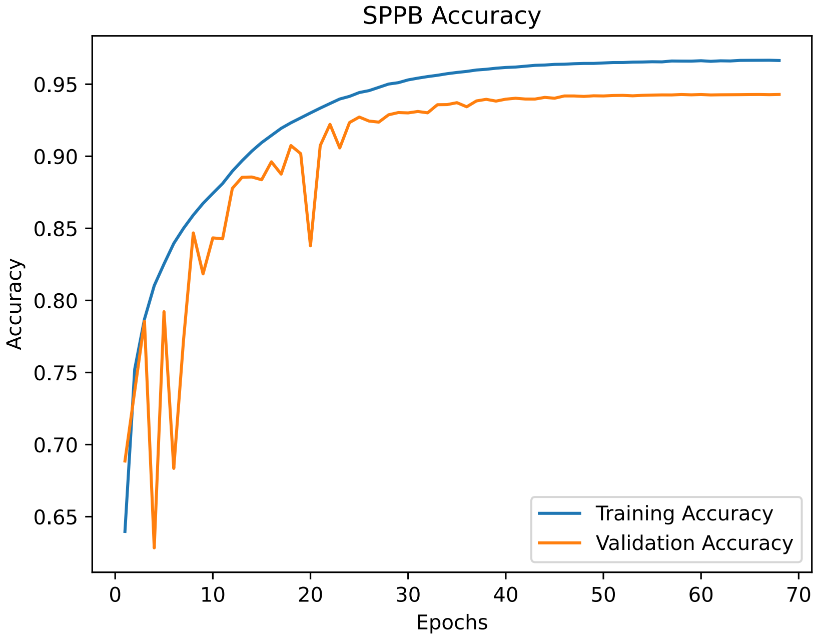

Figure 8. The accuracy of the SPPB model. According to the progress of the loss, the accuracy fluctuates in the first 25 epochs and shows an asymptotic behaviour after epoch 25. The best validation accuracy score was 94.29% at epoch 67.

Table 6. The accuracy on the test set and the two excluded participants.

Assessment	Test Set	P16	P19
SPPB	94.27%	6.39%	14.13%
TUG	95.79%	98.84%	26.15%

Table 7. The specificity and sensitivity for each TUG score on the test set.

Score	Sensitivity	Specificity
1	95.93%	98.50%
2	95.01%	97.25%
3	96.44%	97.93%

Table 8. The specificity and sensitivity for each considered SPPB score on the test set.

Score	Sensitivity	Specificity
3	95.72%	99.67%
4	95.62%	99.38%
5	94.48%	99.45%
6	93.61%	99.06%
7	92.23%	99.25%
8	94.09%	99.31%
9	90.69%	98.85%
10	95.93%	99.31%
11	96.14%	99.27%

The confusion matrices for participant 16 and 19 are shown in Tables 11–14, respectively. The accuracy of the model is lower for the unknown participants. The accuracy for the class 2 of the TUG assessment is 98.93% and 99.49% for participant 16 and 19 respectively. The accuracy for the class 3 of the TUG assessment is 2.50% and 0.81% for participant 16 and 19 respectively. The class 1 is not available for the two participants. The best class for the SPPB score of participant 16 is class 4 with an accuracy of 12.51% and the best class for participant 19 is class 6 with an accuracy of 17.23%. The worst classes are 10 for participant 16 and 3 for participant 19 with an accuracy of 0.000641% and 0.00% respectively.

Participant 16 (76 y) was undergoing chemotherapy two months after the study began and deceased two months before the study ended. Therefore, values of only three assessment dates were available. Facing a severe loss of physical condition following inactivity, functional performance decreased very fast - especially in gait speed, but not in total scores of the SPPB. The interesting part is that, the intra-individual range varied from 4 to 10 points within the short period of time. TUG stopwatch measurements showed a more linear decline and a category change from 2 to 3.

The TUG results for participant 19 (90 y) were very close to the decision boundary of classes 2 and 3. The time of 3 of 6 TUG assessments were of an average of 0.39 s slower than the maximum of 19 s needed for scoring 2 points. The participant scored 3 in those assessments.

Table 9. The confusion matrix of the TUG model. The class with the least false classifications is 1 and the class with the most false classifications is 3. t = true label, p = predicted label.

t\p	1	2	3
1	20,780	614	268
2	455	20,580	627
3	193	578	20,891

Table 10. The confusion matrix of the SPPB model. The class with the least false classifications is 3 and the class with the most false classifications is 9. t = true label, p = predicted label.

t\p	3	4	5	6	7	8	9	10	11
3	14,195	98	35	147	80	95	123	47	10
4	39	14,181	44	163	111	40	157	69	26
5	27	67	14,012	205	72	76	187	155	29
6	99	153	136	13,883	107	116	140	107	89
7	61	154	70	118	13,677	123	229	86	312
8	73	43	52	154	121	13,954	188	97	148
9	62	144	169	173	211	189	13,449	221	212
10	13	46	117	88	47	73	180	14,226	40
11	12	25	24	73	136	106	155	42	14,257

Table 11. The confusion matrix of the TUG for participant 16. The participant scores 2 and 3, but not 1. The most values are in class 2. Due to the imbalance the accuracy is high, but the performance is low. t = true label, p = predicted label.

t\p	1	2	3
1	0	0	0
2	8	87,782	943
3	1	77	2

Table 12. The confusion matrix for SPPB of participant 16. The only class with correct predictions is class 4. t = true label, p = predicted label.

t\p	3	4	5	6	7	8	9	10	11
3	0	0	0	0	0	0	0	0	0
4	1814	5643	7484	4245	3043	9421	9322	4041	92
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0
10	3921	15	2977	814	7112	22,327	4832	28	1682
11	0	0	0	0	0	0	0	0	0

Table 13. The confusion matrix for TUG of participant 19. The model is not able to distinguish between score 3 and score 2 for this participant. t = true label, p = predicted label.

t\p	1	2	3
1	0	0	0
2	3	28,191	142
3	86	81,259	665

Limitations

The cohort does not represent the complete scale of the SPPB and the scores 1, 2, and 12 are not learned by the model and hence the validated assessment is not completely represented by the model. Another point is the IMU, which was used by the participants independently. Using filtering approaches some invalid data was filtered, but certainly not all, e.g., if a participant would have given the IMU to another person, the invalid data of that person remains in the dataset. Inactivity covered by noise in the sensor signal will also remain in the dataset.

Table 14. The confusion matrix for SPPB of participant 19. This participant was the only one with a SPPB score of 2. Since the model was not trained to classify this score, the values for score 2 were not considered for classification. t = true label, p = predicted label.

t\p	3	4	5	6	7	8	9	10	11
3	197	1	37	1	13	3	0	1	3
4	39	1,756	11	235	1753	328	403	99	4009
5	28	14	19,027	299	6143	444	1255	778	198
6	2276	219	15,315	974	363	2388	14,474	30,534	860
7	0	0	0	0	0	0	0	0	0
8	8	594	70	267	220	398	4234	75	2
9	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0

5. Discussion

The results showed that the models were performing well on all classes on known participants, but significantly worse on unknown participants. Finding one patient the network is not performing well for already proofs that the model cannot be used for unknown participants without further adjustments like fine tuning. The reason why the SPPB model performs slightly worse than the TUG model is that assessment has more classes and is more complex in execution than the TUG. Moreover, the score of balancing test of the SPPB is subject to the impression of the supervising professional, rather than an objective measurement. The models learned the variety amongst the participants of the OTAGO study, even though the cohort was very heterogeneous. The heterogeneity is expected to be the reason for the low model performance for the unknown participants 16 and 19. The TUG assessment results of participant 16 were very close to the decision boundary of two classes. This made it even more difficult for the model to distinguish the data of those classes for an unknown participant correctly. Considering Table 9, this seems to be a general problem of the TUG model. That shows the limitations of the assessment scores and boundaries. The scores are defined using full seconds, but with today's technology much more precise measurements up to milliseconds are possible. The low accuracy on the SPPB data is mainly due to the exclusion from the training set, i.e. the participant is unknown to the network.

Even though the ROC curves, the AUCs and the accuracy show a high performance of the model, inferring reasons for score changes are not possible. A change in one item of SPPB changes the score of the assessment, so the change could be due to decreasing gait speed, a declining balance or decreasing lower limb strength. The same holds for the TUG test, because the same aspects are implicitly assessed. Standing up from a chair is dependent on the balance and lower limb strength, and the walking part assesses the gait speed. So, the TUG score incorporates the same dimensions like the SPPB.

The results showed that the models were performing well on all classes on known participants, but significantly worse on unknown participants. The loss and accuracy graphs showed the desired behaviour of increasing fast in the early epochs and stabilising during the later epochs. The graphs for both models were similar, but the SPPB model showed a slightly lesser performance and slightly higher loss than the TUG model. Since the SPPB assessment has more classes and is more complex in execution than the TUG, this is a reasonable finding. Moreover, the score of balancing test of the SPPB is subject to the impression of the supervising professional, rather than an objective measurement. The models learned the variety amongst the participants of the OTAGO study, even though the cohort was very heterogeneous. However, this heterogeneity is expected to be the reason for the low model performance for the unknown participants 16 and 19 as well.

Participant 16 (76 y) was undergoing chemotherapy two months after the study began and deceased two months before the study ended. Therefore, values of only three assessment dates were available. Facing a severe loss of physical condition following inactivity, functional performance decreased very fast - especially in gait speed, but not in total scores of the SPPB. The interesting part is that, the intra-individual range varied from 4 to 10 points within the short period of time. TUG stopwatch measurements showed a more linear decline and a category change from 2 to 3.

The TUG results for participant 19 (90 y) were very close to the decision boundary of classes 2 and 3. The time of 3 of 6 TUG assessments were on an average of 0.39 s slower than the maximum of 19 s needed for scoring 2 points. The participant scored 3 in those assessments. This made it even more difficult for the model to distinguish the data of those classes for an unknown participant correctly. The low accuracy on the SPPB data is mainly due to the exclusion from the training set, i.e. the participant is unknown to the network. The results show as well that the age is not important itself. Participant 16 is 14 y younger than participant 19, but the physical condition is much worse. The physical condition of participant 19 is clinically constant and varies slightly between two classes. The latter shows the limitations of the assessment categories and boundaries. The categories are defined using full seconds, but with today's technology much more precise measurements up to milliseconds are possible.

6. Conclusions and Future Work

The results showed that it is possible to use machine learning to predict the geriatrics mobility assessment scores on real-life IMU data, even though the cohort was very heterogeneous and the IMUs were not rigidly attached to the body. The models performed well on known participants and were able to predict the scores of SPPB and TUG with an accuracy of 95.79% and 94.27% for the TUG and SPPB assessment respectively. The ROC curves, the specificities and the sensitvities for each class show, that the models performing well enough to be used by professionals. On the downside, the results showed that the models are very inaccurate on data of unknown participants.

The most promising approach is to use deep unsupervised learning. In the first step the network could be trained to approximate the underlying probability distribution of the training data. In the next step fine tuning for the new participant could be applied. Since one participant is likely to show only a certain subset of all available scores, the output of the network is the probability whether the sample belongs the class the network was fine tuned with. So, changes could be detected.

Another important point is the investigation of the performance on a different cohort. The cohort used for this research was very special, e.g., study inclusion criteria was being pre-frail at least. For healthier cohort of older adults an early detection of physical decline is important for early treatment and prevention as well. Long-term monitoring would be useful for those cohorts. So the models must be evaluated on data of a healthier cohort. Moreover, measures to simplify the model without loss of accuracy should be taken, because the data acquisition is difficult and costly. A less complex model needs less training data.

Supplementary Materials: The following are available online at https://www.mdpi.com/2227-9032/9/2/149/s1, trained TUG model, trained SPPB model.

Author Contributions: Conceptualization, B.F.; methodology, B.F.; software, B.F.; validation, B.F.; formal analysis, B.F.; investigation, B.F. and S.L. and L.E.; resources, J.M.B. and A.H.; data curation, B.F. and S.L. and L.E.; writing—original draft preparation, B.F. and S.L. and L.E.; writing—review and editing, J.M.B. and A.H.; visualization, B.F.; supervision, A.H.; project administration, A.H.; funding acquisition, J.M.B. and A.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board of the Carl von Ossietzky University (protocol code: Drs.27/2014, date of approval: 30.04.2014).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data are not publicly available due to privacy concerns.

Acknowledgments: We acknowledge Enno-Edzard Steen (University of Oldenburg) for setting up and maintaining the sensor systems in the flats and Bianca Sahlmann (University of Oldenburg) for performing the assessments. The OTAGO study has been funded by an internal funding of the Carl von Ossietzky University of Oldenburg. The experiments were performed at the HPC Cluster CARL, located at the University of Oldenburg (Germany) and funded by the DFG through its Major Research Instrumentation Programme (INST 184/157-1 FUGG) and the Ministry of Science and Culture (MWK) of the Lower Saxony State. Moreover, we gratefully acknowledge the support of the NVIDIA Corporation with the donation of the TITAN V GPU used for this research.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AUC	Area Under the Curve
BMI	Body Mass Index
IMU	Inertial Measurement Unit
LSTM	Long-Short-Term-Memory
ReLU	Rectified Linear Unit
ROC	Receiver Operating Characteristic
SD	Standard Deviation
SPPB	Short-Physical-Performance-Battery
TUG	Timed Up & Go

Appendix A

Table A1. The layer details of Block 1.

Layer	Neurons/Kernel Size	Filters	Strides	Activation	Additional Parameters
LSTM	64	-	-	LeakyReLU	LeakyReLU $\alpha = 0.001$
Batch Normalisation	-	-	-	-	-
Maximum Pooling 1D	-	-	2	-	-
Convolution 1D	8	64	2	LeakyReLU	LeakyReLU $\alpha = 0.001$
Batch Normalisation	-	-	-	-	-
Maximum Pooling 1D	-	-	2	-	-

Table A2. The layer details of Block 2.0.

Layer	Kernel Size	Filters	Strides	Activation	Additional Parameters
Convolution 1D	16	64	2	LeakyReLU	LeakyReLU $\alpha = 0.001$
Batch Normalisation	-	-	-	-	-
Maximum Pooling 1D	-	-	2	-	-

Table A3. The layer details of Block 2.1.

Layer	Kernel Size	Filters	Strides	Activation	Additional Parameters
Convolution 1D	32	64	2	LeakyReLU	LeakyReLU $\alpha = 0.001$
Batch Normalisation	-	-	-	-	-
Maximum Pooling 1D	-	-	2	-	-

Table A4. The layer details of Block 2.2.

Layer	Kernel Size	Filters	Strides	Activation	Additional Parameters
Convolution 1D	64	128	2	LeakyReLU	LeakyReLU $\alpha = 0.001$
Batch Normalisation	-	-	-	-	-
Maximum Pooling 1D	-	-	2	-	-

Table A5. The layer details of Block 3.0.

Layer	Neurons	Dropout	Activation
Dense	48	0.3	Sigmoid
Batch Normalisation	-	-	-

Table A6. The layer details of Block 3.1.

Layer	Neurons	Dropout	Activation
Dense	48	0.0	Sigmoid
Batch Normalisation	-	-	-
Dense	9 (SPPB) /3 (TUG)	0.0	Softmax

References

1. Searle, S.D.; Mitnitski, A.; Gahbauer, E.A.; Gill, T.M.; Rockwood, K. A standard procedure for creating a frailty index. *BMC Geriatr.* **2008**, *8*, 1–10. [CrossRef] [PubMed]
2. Giannouli, E.; Bock, O.; Mellone, S.; Zijlstra, W. Mobility in Old Age: Capacity Is Not Performance. *Biomed Res. Int.* **2016**, *2016*. [CrossRef] [PubMed]
3. Peel, N.M.; Kuys, S.S.; Klein, K. Gait Speed as a Measure in Geriatric Assessment in Clinical Settings: A Systematic Review. *J. Gerontol. Ser. A* **2013**, *68*, 39–46. [CrossRef] [PubMed]
4. Middleton, A.; Fulk, G.D.; Beets, M.W.; Herter, T.M.; Fritz, S.L. Self-Selected Walking Speed is Predictive of Daily Ambulatory Activity in Older Adults. *J. Aging Phys. Act.* **2016**, *24*, 214–222. [CrossRef]
5. Nouredanesh, M.; Godfrey, A.; Howcroft, J.; Lemaire, E.D.; Tung, J. Fall risk assessment in the wild: A critical examination of wearable sensors use in free-living conditions. *Gait Posture* **2020**. [CrossRef]
6. Cybenko, G. Approximation by superpositions of a sigmoidal function. *Math. Control. Signals Syst.* **1989**, *2*, 303–314. [CrossRef]
7. Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Netw.* **1991**, *4*, 251–257. [CrossRef]
8. Leshno, M.; Lin, V.Y.; Pinkus, A.; Schocken, S. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Netw.* **1993**, *6*, 861–867. [CrossRef]
9. Pinkus, A. Approximation theory of the MLP model in neural networks. *Acta Numer.* **1999**, *8*, 143–195. [CrossRef]
10. Guralnik, J.M.; Simonsick, E.M.; Ferrucci, L.; Glynn, R.J.; Berkman, L.F.; Blazer, D.G.; Scherr, P.A.; Wallace, R.B. A Short Physical Performance Battery Assessing Lower Extremity Function: Association With Self-Reported Disability and Prediction of Mortality and Nursing Home Admission. *J. Gerontol.* **1994**, *49*, M85–M94. [CrossRef]
11. Podsiadlo, D.; Richardson, S. The Timed "Up & Go": A Test of basic functional mobility for frail elderly persons. *J. Am. Geriatr. Soc.* **1991**, *32*, 142–148. [CrossRef]
12. da Câmara, S.M.A.; Alvarado, B.E.; Guralnik, J.M.; Guerra, R.O.; Maciel, A.C.C. Using the Short Physical Performance Battery to screen for frailty in young-old adults with distinct socioeconomic conditions. *Geriatr. Gerontol. Int.* **2013**, *13*, 421–428. [CrossRef] [PubMed]
13. Fudickar, S.; Kiselev, J.; Frenken, T.; Wegel, S.; Dimitrowska, S.; Steinhagen-Thiessen, E.; Hein, A. Validation of the ambient TUG chair with light barriers and force sensors in a clinical trial. *Assist. Technol. Off. J. RESNA* **2020**, *32*, 1–8. [CrossRef] [PubMed]
14. Jung, H.W.; Roh, H.; Cho, Y.; Jeong, J.; Shin, Y.S.; Lim, J.Y.; Guralnik, J.M.; Park, J. Validation of a Multi—Sensor-Based Kiosk for Short Physical Performance Battery. *J. Am. Geriatr. Soc.* **2019**, *67*, 2605–2609. [CrossRef]
15. Hellmers, S.; Izadpanah, B.; Dasenbrock, L.; Diekmann, R.; Bauer, J.; Hein, A.; Fudickar, S. Towards an Automated Unsupervised Mobility Assessment for Older People Based on Inertial TUG Measurements. *Sensors* **2018**, *18*, 3310. [CrossRef]
16. Kamnardsiri, T.; Khuwuthyakorn, P.; Boripuntakul, S. The Development of a Gait Speed Detection System for Older Adults Using Video-based Processing. In Proceedings of the 2019 4th International Conference on Biomedical Imaging, Signal Processing, Nagoya, Japan, 17–19 October 2019; pp. 1–6.
17. Goffredo, M.; Bouchrika, I.; Carter, J.N.; Nixon, M.S. Performance analysis for gait in camera networks. In Proceedings of the 1st ACM workshop on Analysis and Retrieval of Events/Actions and Workflows in Video Streams, Vancouver, BC, Canada, 31 October 2008; pp. 73–80.

18. Stone, E.; Skubic, M.; Rantz, M.; Abbott, C.; Miller, S. Average in-home gait speed: Investigation of a new metric for mobility and fall risk assessment of elders. *Gait Posture* **2015**, *41*, 57–62. [CrossRef]
19. Aicha, A.N.; Englebienne, G.; Kröse, B. Continuous measuring of the indoor walking speed of older adults living alone. *J. Ambient. Intell. Humaniz. Comput.* **2017**, *9*, 589–599. [CrossRef]
20. Frenken, T.; Steen, E.E.; Brell, M.; Nebel, W.; Hein, A. Motion Pattern Generation and Recognition for Mobility Assessments in Domestic Environments. In Proceedings of the 1st International Living Usability Lab Workshop on AAL Latest Solutions, Trends and Applications, Rome, Italy, 28–29 January 2011; pp. 3–12.
21. Chapron, K.; Bouchard, K.; Gaboury, S. Real-time Gait Speed Evaluation at Home. In Proceedings of the 5th EAI International Conference on Smart Objects and Technologies for Social Good, Valencia, Spain, 25–27 September 2019; pp. 55–60.
22. Hsu, C.Y.; Liu, Y.; Kabelac, Z.; Hristov, R.; Katabi, D.; Liu, C. Extracting Gait Velocity and Stride Length from Surrounding Radio Signals. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, 6–11 May 2017; pp. 2116–2126.
23. Byun, S.; Lee, H.J.; Han, J.W.; Kim, J.S.; Choi, E.; Kim, K.W. Walking-speed estimation using a single inertial measurement unit for the older adults. *PLoS ONE* **2019**, *14*. [CrossRef]
24. Yeo, S.S.; Park, G.Y. Accuracy Verification of Spatio-Temporal and Kinematic Parameters for Gait Using Inertial Measurement Unit System. *Sensors* **2020**, *20*, 1343. [CrossRef]
25. Washabaugh, E.P.; Kalyanaraman, T.; Adamczyk, P.G.; Claflin, E.S.; Krishnan, C. Validity and Repeatability of Inertial Measurement Units for Measuring Gait Parameters. *Gait Posture* **2017**, *55*, 87–93. [CrossRef]
26. Hellmers, S.; Peng, L.; Lau, S.; Diekmann, R.; Elgert, L.; Bauer, J.; Hein, A.; Fudickar, S. Activity Scores of Older Adults based on Inertial Measurement Unit Data in Everyday Life. In Proceedings of the HEALTHINF, Valletta, Malta, 24–26 February 2020; pp. 579–585.
27. Yu, S.; Chen, H.; Brown, R.; Sherman, S. Motion Sensor-Based Assessment on Fall Risk and Parkinson's Disease Severity: A Deep Multi-Source Multi-Task Learning (DMML) Approach. In Proceedings of the 2018 IEEE International Conference on Healthcare Informatics (ICHI), New York, NY, USA, 4–7 June 2018; pp. 174–179.
28. Tunca, C.; Salur, G.; Eroy, C. Deep Learning for Fall Risk Assessment With Inertial Sensors: Utilizing Domain Knowledge in Spatio-Temporal Gait Parameters. *IEEE J. Biomed. Health Inform.* **2019**, *24*, 1994–2005. [CrossRef] [PubMed]
29. Meyer, B.M.; Tulipani, L.J.; Gurchiek, R.D.; Allen, D.A.; Adamowicz, L.; Larie, D.; Solomon, A.J.; Cheney, N.; McGinnis, R. Wearables and Deep Learning Classify Fall Risk from Gait in Multiple Sclerosis. *IEEE J. Biomed. Health Inform.* **2020**. [CrossRef] [PubMed]
30. Aicha, A.N.; Englebienne, G.; van Schooten, K.S.; Pijnappels, M.; Kröse, B. Deep Learning to Predict Falls in Older Adults Based on Daily-Life Trunk Accelerometry. *Sensors* **2018**, *18*, 1654. [CrossRef] [PubMed]
31. Jung, D.; Dung Nguyen, M.; Park, M.; Kim, M.; Won Won, C.; Jinwook, K.; Mun, K.R. Walking-in-Place Characteristics-Based Geriatric Assessment Using Deep Convolutional Neural Networks. In Proceedings of the 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC), Montreal, QC, Canada, 20–24 July 2020; pp. 3931–3935.
32. Luna-Perejón, F.; Domínguez-Morales, M.; Gutiérrez-Galán, D.; Civit-Balcells, A. Low-Power Embedded System for Gait Classification Using Neural Networks. *J. Low Power Electron. Appl.* **2020**, *10*, 14. [CrossRef]
33. Kiprijanovska, I.; Gjoreski, H.; Gams, M. Detection of Gait Abnormalities for Fall Risk Assessment Using Wrist-Worn Inertial Sensors and Deep Learning. *Sensors* **2020**, *20*, 5373. [CrossRef] [PubMed]
34. Musci, M.; De Martini, D.; Blago, N.; Facchinetti, T.; Piastra, M. Online Fall Detection using Recurrent Neural Networks. *arXiv* **2018**, arXiv:1804.04976.
35. Luna-Perejón, F.; Domínguez-Morales, M.J.; Civit-Balcells, A. Wearable Fall Detector Using Recurrent Neural Networks. *Sensors* **2019**, *19*, 4885. [CrossRef]
36. Carl von Ossietzky Universität Oldenburg. OTAGO. Available online: https://uol.de/en/amt/research/projects/otago (accessed on 20 December 2020).
37. Research, S. Shimmer3 IMU Unit. Available online: http://www.shimmersensing.com/products/shimmer3-imu-sensor (accessed on 20 December 2020).
38. Friedrich, B.; Lübbe, C.; Hein, A. Combining LSTM and CNN for Mode of Transportation Classification from Smartphone Sensors. In Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers (UbiComp/ISWC '20 Adjunct), Virtual Event, Mexico, 12–16 September 2020.
39. Reddi, S.; Kale, S.; Kumar, S. On the Convergence of Adam and Beyond. *arXiv* **2018**, arXiv:1904.09237.