Inertial measurement unit signal-based machine learning methods for frailty assessment in geriatric health.

Amjad, A., Szczęsna, Agnieszka, Błaszczyszyn, Monika and Anwar, Aamir (2024) Inertial measurement unit signal-based machine learning methods for frailty assessment in geriatric health. Signal, Image and Video Processing, 19 (105).

http://dx.doi.org/10.1007/s11760-024-03719-8

Signal, Image and Video Processing (2025) 19:105
https://doi.org/10.1007/s11760-024-03719-8

ORIGINAL PAPER

Inertial measurement unit signal-based machine learning methods for frailty assessment in geriatric health

Arslan Amjad[1] · Agnieszka Szczęsna[1] · Monika Błaszczyszyn[2] · Aamir Anwar[3]

Received: 4 July 2024 / Revised: 28 August 2024 / Accepted: 6 September 2024

Abstract

Frailty is a geriatric syndrome that may result in poor health outcomes such as hospitalization, disability, psychological distress, and reduced life satisfaction, and it is also associated with higher healthcare costs. The aim of this study is to classify frailty in elders at an early stage (pre-frail) to lower the risk of frailty and, hence, improve the quality of life. The other two classes in the classification task are frail and robust (non-frail). To achieve this, a dataset based on gait was utilized, which was recorded by an Inertial Measurement Unit (IMU) sensor, including gyroscope and accelerometer data. In this study, two approaches were assessed: the first used advanced Deep Learning (DL) algorithms to analyze raw IMU signals, and the second used conventional Machine Learning (ML) methods with hand-crafted features. The DL model, i.e., InceptionTime, beat the other algorithms in the DL approach with a remarkable test accuracy of 98%. On the ML side, Random Forest reported the most successful ML method, which achieved a test accuracy of 63.3%. For a careful assessment of the models, other evaluation metrics like Precision, Recall, and F1-score were also evaluated. The evaluation of both approaches produces research benefits for the classification of frailty in older people and allows for the investigation of new areas, promoting deeper comprehension and well-informed decision-making, particularly in healthcare systems.

Keywords Deep Learning · Frailty · Geriatric · Gait · IMU sensor · Machine Learning

1 Introduction

The number of elderly individuals is rising dramatically on a global scale nowadays. On a chronological basis, the elderly population, typically classified as individuals aged 65 years or older [1]. According to the World Health Organization (WHO) report, a nearly double increase in the percentage of people over 60 globally is estimated, escalating from 12 to 22% between 2015 and 2050 [2, 3]. The dramatic rise in the population of aged people has caused a serious global social issue: frailty among the elderly creates a major concern [4].

Frailty, a medical geriatric syndrome that increases an individual's vulnerability to a decline in muscle mass and quality [5], compromised physiological function, decreased resilience to stress, and pathological and physiological changes affecting various systems, including immunity, muscle, and daily activities [1, 6]. Individuals who are identified as fragile are particularly experiencing unfavorable consequences, such as falls resulting in harm, disability, dementia, long-term care, and death [7, 8]. The rising population of elders experiencing frailty gives rise to significant global challenges in health, social, and financial domains, especially in low-resource settings. This makes frailty analysis an engaging topic for researchers.

To identify the risk of physical frailty, the medical profession commonly used two recognized criteria: Fried's Frailty Phenotype (FFP) and the Frailty Index (FI) [9]. Fried's Frailty Phenotype categorizes patients as frail, pre-frail, or robust based on five measurable criteria: weakness, slowness, poor physical activity, exhaustion, and weight loss. The Frailty Index (FI) assesses the patient's medical history

✉ Arslan Amjad
aabbasi@polsl.pl

Aamir Anwar
anwaaam@uwl.ac.uk

[1] Department of Computer Graphics, Vision and Digital Systems, Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland

[2] Department of Physical Education and Sport, Faculty of Physical Education and Physiotherapy, Opole University of Technology, Prószkowska 76, 45-758 Opole, Poland

[3] School of Computing and Engineering, University of West London, Ealing London W55RF, UK

Published online: 09 December 2024

and cognitive abilities [10, 11]. These assessment methods have some drawbacks, such as being subjective, resource-intensive, inconvenient for the patients to transport for the test, and unable to identify frailty at an early stage [12, 13].

The limitations of subjective methods can be solved by combining wearable sensors with Machine Learning (ML) techniques. Wearable sensors such as Inertial Measurement Unit (IMU) are easy to carry by the elder's, offer real-time monitoring and objectively record an individual's gait metrics such as stride length, cadence, speed, and other features. Whereas ML algorithms enables the identification of complex patterns in the gait data [14, 15]. Facilitating a dynamic and real-time assessment of an individual's physical state and a more precise, individualized classification of frailty in the early stage (pre-frail). As a result, elderly individuals can delay the onset of frailty and could minimize the burden of frailty in an aging population if they are diagnosed in the pre-frail state [16].

This study compares the effectiveness of conventional ML techniques that involve the manual extraction of features from IMU signals with advanced Deep Learning (DL) algorithms capable of performing automatic feature learning from raw IMU data. The proposed strategy involves two phases. The first phase explores the effectiveness of ML approaches with manual feature extraction from IMU signals to classify frailty into three stages (frail, pre-frail, or non-frail/robust). Simultaneously, the second strategy investigates the implementation of DL algorithms capable of automated feature learning from the raw IMU gait recording to classify the frailty.

The research questions addressed in this study are:

1) How effective are traditional ML techniques in classifying frailty using hand-crafted features?
2) How effective are DL algorithms in classifying frailty using raw IMU data?
3) What are the comparative strengths and weaknesses of ML and DL approaches for frailty classification?
4) Empirical experiments to guide the selection of algorithms and parameters that can improve the robustness of frailty classification models.

This study can contribute to improving the efficiency and reliability of frailty assessment models. In clinical settings, it also advances the use of DL and wearable sensor technologies for the early identification and prevention of frailty in older individuals.

The structural arrangement of this paper is as follows: Sect. 2 covers a comprehensive review of relevant prior research; Sect. 3 provides an analysis of the dataset and describes the research methodology. Section 4 discusses the results, whereas the concluding section presents the results and draws overall conclusions.

2 Relevant studies

A study [17] examined six gait features, encompassing intensity, step rate, periodicity, dynamism, and two time-varying representations of gait utilizing wearable sensors for gait analysis in frailty assessment (frail, pre-frail, or robust). The study implemented several ML classifiers, Support Vector Machine (SVM) demonstrated outperformed performance, achieving an average accuracy of 88.5%.

In another study, Kinect sensor was used to extract the features (i.e. weight loss, weakness, poor endurance, slowness, and low physical activity). The results demonstrated that the Support Vector Classifier (SVC) and Multi-layer Perceptron (MLP) were the most effective estimators for predicting Fried's frailty level with median accuracies up to 97.5% [18].

A study classified frailty into three classes (frail, pre-frail, or non-frail) using data from accelerometer and gyroscope sensors. The study extracted seven statistical features and one frequency feature (FFT). KNN outperformed SVM, RF, and NB combined, achieving a 99% higher accuracy [19].

Similarly, other studies explored different temporal-spatial parameters for wearable sensor-based frailty classification using ML. Parameters such as gait speed, velocity, time, stride time, step time, percentage of time in double support, and trunk kinematics of angular velocity are examples of metrics commonly investigated [20–23]. Another metric that varies in assessment within the literature is balance, [24–26] focused on different aspects of balance parameters. In summary, the studies evaluated various temporal-spatial parameters for wearable sensor-based frailty classification using ML methods, highlighting the importance of precise extraction of key gait parameters through precise sensor data.

Previous research articles applied DL techniques (1DCNN, LSTM, Bi-LSTM, RNN, ConvLSTM etc.) to explore the relationship between frailty and gait data. Certain research articles exclusively utilized raw IMU signals, feeding them directly into DL algorithms [27–31]. In contrast, other research efforts adopted a methodology that combined both hand-crafted features and raw signals [32–35]. Whereas many researchers utilized images generated based on raw IMU signals, then fed into DL algorithms, the images discussed in the previous studies were spectrograms and plantar pressure distribution in the foot, gait energy, recurrence plots, and vGRF signals [36–39].

3 Research methodology

This study's methodology combines shallow ML and DL approaches in a two-tiered fashion to classify the frailty stages as depicted in Fig. 1. In the shallow ML category, authors fed different hand-rafted features from a dataset to conventional ML classifiers. On the other hand, raw signals

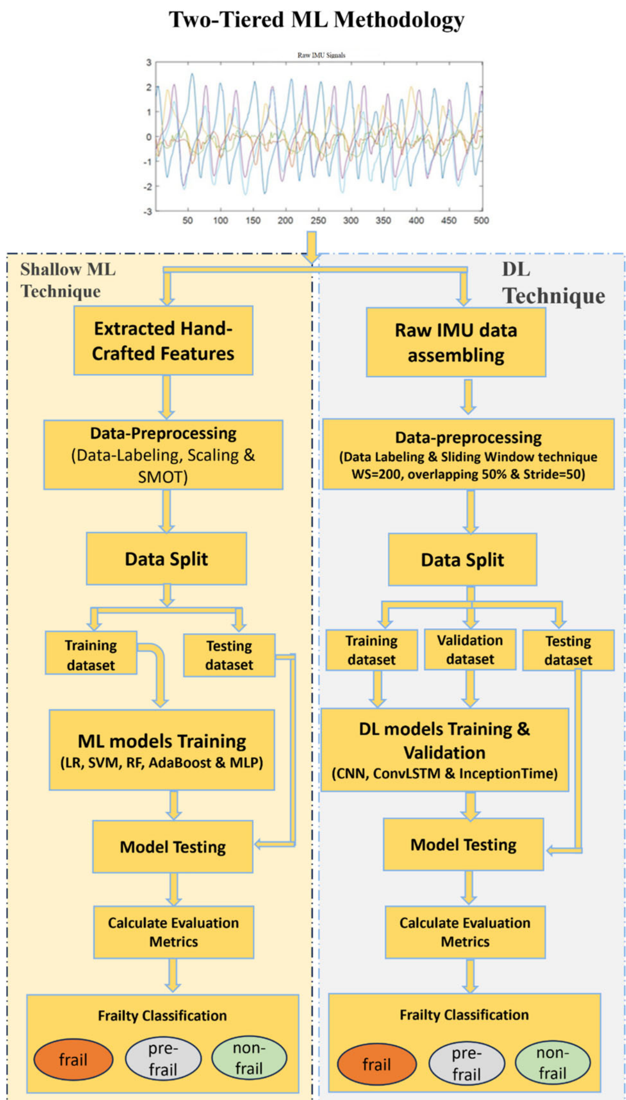

Fig. 1 Two-Tiered research methodology that used both shallow ML techniques and DL techniques for frailty classification

from IMU sensors are analyzed directly by the DL algorithms. This comparative methodology enabled an in-depth analysis of feature-rich shallow ML models, and the DL feature captured structures for frailty analysis. Such an approach provided a detailed understanding of each architecture's contribution to the overall frailty classification task.

3.1 Dataset

In this study, the publicly available GSTRIDE [40] database is utilized for conducting frailty classification tasks. The dataset includes health assessments of 163 elderly individuals (45 men and 118 women) aged between 70 and 98 years, with an average weight of 64.2 ± 13.1 kg and a height of 156.8 ± 10.2 cm offering a comprehensive representation of the aging population.

The database consists of socio-demographic data (i.e., age, gender, and subject's living environment), anatomical, functional, and cognitive variables (i.e., weight, height, Body Mass Index (BMI) and Global Deterioration Scale (GDS) index of the subjects). Authors also outlined the outcomes from tests commonly utilized in elder evaluations, including: the 4-m Gait Speed Test, the Hand Grip Strength, the Timed Up and Go (TUG), the Short Physical Performance Battery (SPPB), and the Short Falls Efficacy Scale International (FES-I) [40].

For gait data, two IMUs sensors i.e., CSIC and Gaitup, were used with frequencies of 104 Hz and 128 Hz, respectively [40]. The authors stated that the varied specifications and sampling frequencies of the sensors have a very little effect on temporal-spatial estimation. However, the estimation accuracy varies slightly [41].

The dataset includes gait parameters obtained through measurements (accelerometer and gyroscope) utilizing an IMU positioned on the subjects' foot. The motivation behind selecting this dataset lies in its diversity and the inclusion of gait related IMU data and parameters, making it a valuable resource for advancing frailty classification methodologies. The use of a dataset with only one kind of sensor is one of the study's limitations. In the future, a more diverse dataset with numerous sensors could be used to improve early frailty identification.

3.2 Hand-crafted features

This study focused on extracting two primary components from the GSTRIDE database. One is raw IMU signals, and the other is gait parameters during a 15-min walk for each subject [42]. Hand-crafted parameters utilized in this research were extracted from an individual's complete gait cycle. They fulfill the need for structured, interpretable representations in ML algorithms, supporting the inherent simplicity and effectiveness of such models. The gait parameters contain metrics like walking distance, total time taken, number of strides, and an array of spatial–temporal gait parameters. Spatial–temporal gait parameters include stride length, stride time duration, step speed, percentage of gait phases (Swing, Stance, Foot-Flat, Push and Load) over the strides, foot angle during Heel Strike and Toe Off events, 3D and 2D paths, cadence, and clearance.

To optimize the model training, physiological parameters (weight, height, and BMI) were also included. This decision is based on the understanding that these parameters demonstrate medium to high correlations with spatial–temporal gait parameters, as supported by previous studies [42, 43]. Hand-crafted features utilized in this study are listed in Table 1 with description.

Table 1 Hand-crafted features extracted from GSTRIDE database

Features	Description
Weight (kg)	Weight of a subject
Height (m)	Height of a subject
BMI	Body Mass Index (BMI) is the height to weight ratio of a subject
Distance (s)	Total distance travel by a subject
Time (s)	Total walking time of a subject
No. of Strides	Number of steps taken in the total walk
Stride length (m)	Distance between consecutive contacts of one foot with the ground (Avg. and STD values)
Step speed	Foot forward speed during the swing phase (Avg. and STD values)
Swing (%)	Calculated from toe-off to heel strike from total gait cycle time (Avg. and STD values)
Foot-Flat (%)	Calculated as the total time the foot-flat in total gait cycle time (Avg. and STD values)
Load (%)	Calculated as the percentage time from heel-strike to start of foot-flat in total cycle time (Avg. and STD values)
Toe-off-angle (o)	Maximum pitch angle at toe-off (Avg. and STD values)
Heel-strike-angle (o)	Maximum pitch angle at heel strike (Avg. and STD values)
Push (%)	Calculated as the percentage time from the end of foot-flat to toe-off in total cycle time (Avg. and STD values)
Load (%)	Calculated as the percentage time from heel-strike to start of foot-flat in total cycle time (Avg. and STD values)
Cadence (strides/min)	Defined as number of steps per minute (Avg. and STD values)
3D and 2D paths (m)	The distance along the path of the foot in the horizontal plane and 3D space during a step respectively (Avg. and STD values)
Clearance (m)	Most elevated height of the foot during the swing phase in relation to the ground (Avg. and STD values)
Targets	Frailty class label assigned to each subject (either frail, pre-frail or non-frail)

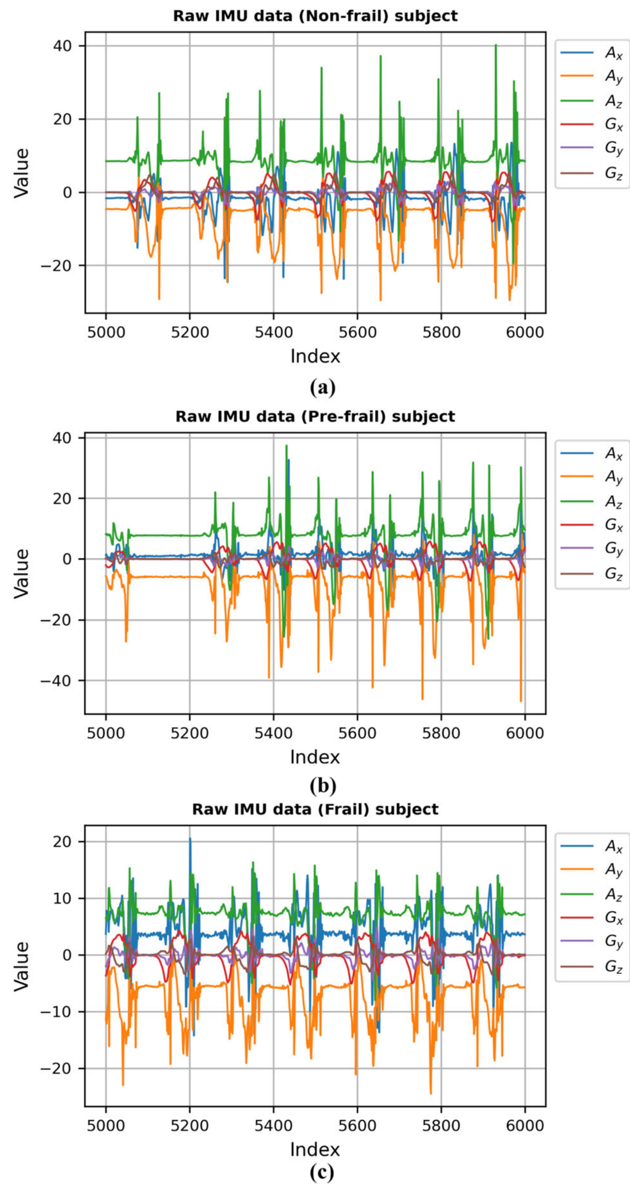

Fig. 2 Raw IMU signals of tri-axial accelerometer (Ax, Ay, Az) and tri-axial gyroscope (Gx, Gy, Gz) data from randomly chosen participants indexed from 5000 to 6000 **a** from the non-frail/robust class, **b** from the pre-frail class **c** from the frail class

3.3 Frailty labeling of participants

The frailty stage of each subject is categorized into three classes: frail, pre-frail, and non-frail. The assessment of the frailty stage for each elderly subject is conducted using the standardized Fried's phenotype test [10]. The Frailty Index (FI) score is computed by summing the values of five Fried's phenotype parameters (assigned a score of 1 for positive responses or 0 for negative) [40]. Subsequently, class labels are assigned to each subject based on their FI score, which ranges from 0 to 5, as in (1). The number of participants categorized as non-frail, pre-frail, and frail classes is 80, 58, and 25, respectively, based on the criteria given in (1). This systematic approach ensures a robust and standardized labeling process for the subsequent supervised classification analyses.

$$T\arg et class = \begin{cases} Non-Frail, \; FI = 0 \\ \Pr e-frail, \; FI = 1 \, or \, 2 \\ frail, \; FI = 3, 4 \, or \, 5 \end{cases} \quad (1)$$

Table 2 Hyperparameters of optimal CNN architecture

Hyperparameter	Value
Learning Rate	0.018550
Regularization Rate	0.00010
Filters	[93, 83, 42, 13, 80, 71, 100, 71]
Fully connected Nodes	1577

Table 3 Hyperparameters of optimal ConvLSTM architecture

Hyperparameter	Value
Learning rate	0.000235
Regularization rate	0.006537
Filters	[91, 25, 54, 89, 48, 35, 99, 12, 72, 71]
Fully connected nodes	[53, 24, 17]

Table 4 Hyperparameters of optimal InceptionTime architecture

Hyperparameter	Value
Learning rate	0.005890
Regularization rate	0.025038
Network depth	6
No. of filters	70
Max. kernel size	23

3.4 Shallow ML techniques

In this phase of the study, five well-known ML models were fed hand-crafted features (shown in Table 1). The initial steps involved data preprocessing, which addressed outliers and ensured that the features were normalized without being overly impacted by extreme values, for this, a robust scaling technique was utilized. A Synthetic Minority Over-Sampling Technique (SMOTE) was also used to resolve the class imbalance in the training data. Following the preprocessing of the data, ML algorithms were implemented using Python with built-in library "*sklearn*". These included Support Vector Machines (SVM) with the radial basis function (RBF) kernel [42, 44, 45] and Logistic Regression (LR) [46] configured with an *L1* penalty and *SAGA* solver.

As the study unfolded, authors chose ensemble approaches because of their capacity to manage complex relationships and improve the general robustness of the classification procedure. The Random Forest (RF) [47] classifier was trained with 250 estimators on resampled data and used the AdaBoost classifier [48] with 300 decision tree base estimators for classification. The Multi-Layer Perceptron (MLP) [49] with activation function 'ReLu' and 50 and 25 neurons in the first and second hidden layers, respectively, gave valuable insights into the complex patterns in the frailty dataset.

An extensive hyperparameter tuning was carried out utilizing random search to ensure the optimal performance of each ML model. In LR, the optimal parameters were *'saga'* solver, an *L1* penalty, and the maximum iterations to *1000*. SVM is fine-tuned with the regularization parameter *(C)* to 1, *RBF* kernel, and set gamma to *'scale'*. The RF model optimizes parameters such as the number of trees *(100),* the minimum sample per leaf *(1)*, and the minimum sample per split *(2)*. AdaBoost was fine-tuned with the base estimator of maximum depth of*3*, the learning rate to *0.001*, and the number of estimators of *250*. Finally, the hyperparameters of the MLP classifier were fine-tuned by utilizing the *'relu'* activation function and hidden layer sizes of *25 and 10* neurons. Each model was then trained with optimal parameters and evaluated by using performance measures.

To ensure the generalizability of the model, a tenfold Cross-Validation (CV) technique was used. The dataset was randomly shuffled and divided into training (75%) and testing (25%) sets with random state of 42. After applying the ML algorithms, the models were evaluated using metrics like precision, recall, and F1-score for each class, and looked at the overall accuracy for each fold. The average accuracy and F1 score of 10-folds were also calculated, allowing an extensive assessment of the models' performance over diverse data subsets.

3.5 Deep learning (DL) techniques

Shallow machine learning has its limitations since it relies on hand-crafted or manual feature selection, which requires domain knowledge [50]. On the other hand, deep learning offers advantages, particularly in frailty classification through gait analysis, as it eliminates the need for manual feature selection by automatically extracting high-level features from raw IMU data through its multiple layers. Raw IMU signals consist of tri-axial accelerometer (A_x, A_y and A_z) and tri-axial gyroscope (G_x, G_y and G_z) data. As the task is to classify frailty into frail, pre-frail or non-frail/robust, some examples of raw IMU signals from all three frailty classes are shown in Fig. 2.

Three deep learning models: 1DCNN, DeepConvLSTM, and InceptionTime were used in this study [51]. Data assembly and class labeling was the first step to performed on the raw IMU data extracted from the GSTRIDE database [40] prior to implementing the DL algorithms into practice. Each participant's accelerometer and gyroscope signals were first normalized with robust scaling technique as in the shallow ML approach, then used to classify frailty. Class labels were added to each subject in accordance with (1). Next data pre-processing stage was data segmentation, in which each subject's raw IMU signals are transformed into the DL

Table 5 Shallow ML methods result for training and testing phase

ML	Training	Testing			
	Avg. accuracy 10-Fold (%)	Avg. precision (%)	Avg. recall (%)	Avg. F1 -score (%)	Test accuracy (%)
LR	61.21	53	56	53	57.14
SVM	70.92	54	60	54	57.14
RF	70.29	59	64	61	63.27
Ada-Boost	61.80	54	53	53	59.18
MLP	70.29	52	54	53	55.10

time-series format using a sliding window technique [52] of window size 200 with 50% overlap and a step size of 50. The input layer size for the model is set at 200 × 6, where 200 is the window size and 6 is the number of features. As, the dataset structured into multiple windows, each with a size of 200 × 6. The dataset was divided randomly into three subsets with random state of 42: training (70%), validation (15%), and testing (15%), as the study suggested [53, 54]. Then the models were trained on the training dataset for 25 epochs using a batch size of 64. To prevent overfitting and ensure generalization, early stopping was implemented with a patience argument of 3 epochs. Finally, the models were evaluated on the respective validation and testing datasets.

DL architectures were implemented using the open-source Python based library McFly [51]. McFly was chosen for its capability to facilitate the creation of DL models for time-series data and conduct hyperparameter optimization. The process involved creating four models for each DL technique. These models were individually trained on the training dataset and assessed on the validation dataset. The selection of the best model for each DL technique was based on criteria such as low training and validation loss and high accuracy. The optimal models, along with their corresponding hyperparameters, were saved after the training process. Finally, the frailty classification results were determined by evaluating each optimal model (1DCNN, DeepConvLSTM, and InceptionTime) on the training and validation dataset. The evaluation of all three DL models was conducted on the test dataset, utilizing metrics including accuracy, F1-score, precision, and recall.

3.5.1 Convolutional neural network (CNN) architecture

CNN architecture consists of eight 1D convolutional layers each followed by batch normalization then a flatten operation, two dense layers, and an output layer, resulting in a total depth of 12 layers. Convolutional layers utilize filters with varying sizes [93, 83, 42, 13, 80, 71, 100, 71]. The output layer has three nodes with '*softmax*' activation for classification. Optimal model's hyperparameters are listed in Table 2.

3.5.2 Convolutional-LSTM Network (ConvLSTM) architecture

ConvLSTM architecture started with batch normalization and reshaping operations, then ten 2D convolutional layers with varying filter sizes [91, 25, 54, 89, 48, 35, 99, 12, 72, 71] were used. Following these convolutional layers, further layers such as batch normalization and activation functions were added before the data was reshaped and fed into three LSTM layers with dimensions [17, 24, 53]. Finally, dropout regularization, time-distributed and activation layers conclude the model. This architecture has 32 layers total, making it an advanced model that can capture complex spatial–temporal features. Hyperparameters of optimal model's on GSTRIDE raw IMU signals are listed in Table 3.

3.5.3 InceptionTime architecture

An input layer is the first step in the InceptionTime architecture, followed by batch normalization. To capture important features, a primary 1D convolutional layer is used, followed by max pooling. The basis of this architecture is a network of inception blocks with a depth of 6, which includes 1D convolutional layers with 70 filters and a maximum kernel size of 23. To capture diverse spatial–temporal features, these pathways are combined. Additional batch normalization and activation are applied to the concatenated features. Every inception block goes through this procedure, which helps the model to capture diverse spatial–temporal features. The last layers are global average pooling, a dense layer, and activation, which results in the model's output. Table 4 depicts the hyperparameters of an optimal InceptionTime model.

4 Results

In the first phase of this study, shallow ML algorithms were trained and assessed on the hand-crafted features as listed in Table 1. The ML models were assessed on the training dataset using the average accuracy obtained over tenfold CV,

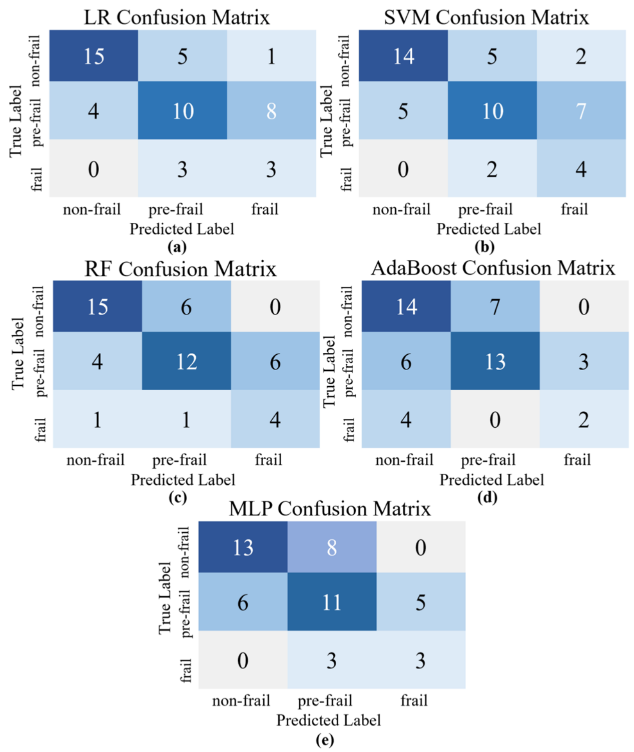

Fig. 3 Confusion matrices for the testing set of ML algorithms: **a** Logistic Regression (LR), **b** Support Vector Machine (SVM), **c** Random Forest (RF), **d** AdaBoost, **e** Multi-Layer Perceptron (MLP)

providing an independent measure of the model's generalization performance. Whereas the overall performance of each ML model was evaluated on the test data using evaluation metrics including precision, recall, F1-score, and accuracy [55, 56].

RF algorithm outperforms in this shallow ML phase, showing an average CV accuracy of 70.29% and a testing accuracy of 63.27%. RF achieves balanced precision, recall, and F1-score metrics, which are crucial for frailty classification, particularly in identifying pre-frail individuals. Early detection of frailty in pre-frail patients can prevent further progression of frailty, making it a key focus for effective intervention and management. The precision of 63% indicates RF's accuracy in identifying pre-frail cases, while a recall of 55% indicates the identification of true pre-frail instances. The F1-score of 59% confirms that RF reflects a balance between precision and recall. Table 5 shows the overall results of all ML algorithms, whereas confusion matrices are shown in Fig. 3.

In the second phase of our research, DL models were used to automatically extract features from raw IMU signals for frailty classification. In this study, three DL algorithms were utilized, which are CNN, ConvLSTM, and InceptionTime. For each of these DL methods, four models with different hyperparameter setups were built. The training, validation, and testing processes, as well as the metrics used for evaluation, are discussed in the DL techniques section. The focus of this section is to give the results of the best-performing model among the four distinct models developed for each DL technique. Training outcomes of each best DL model are

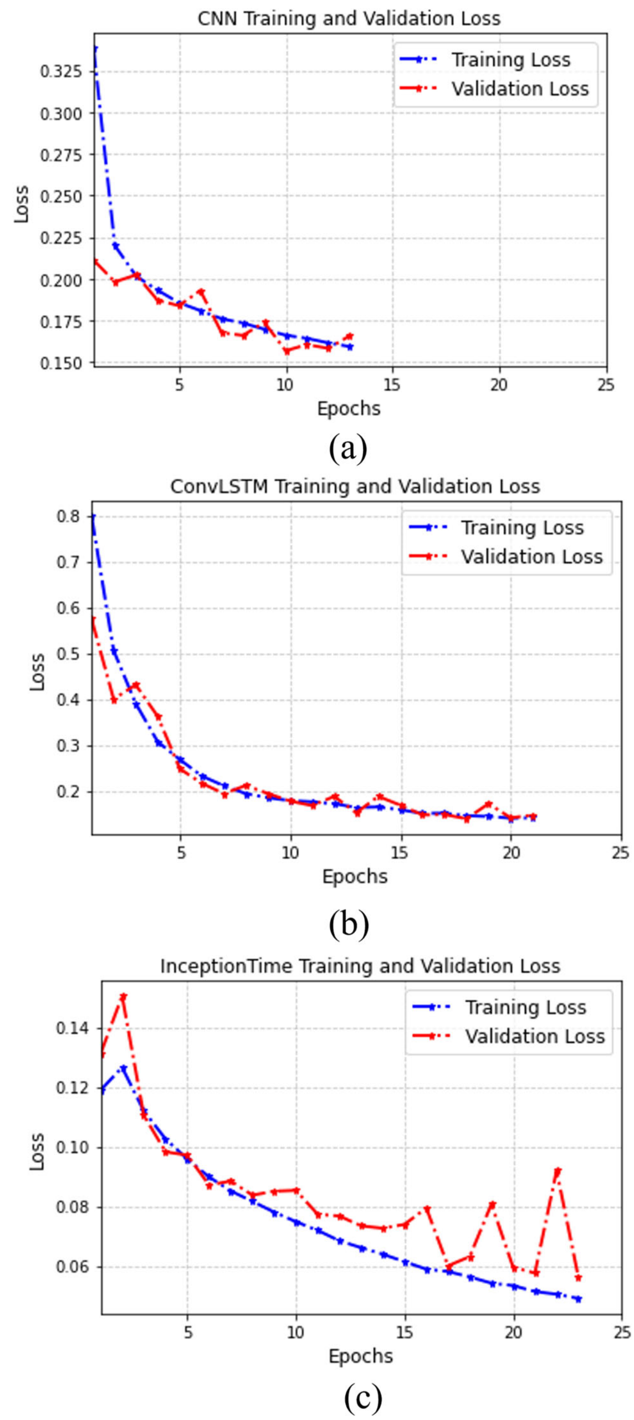

Fig. 4 Training and Validation losses of: **a** CNN, **b** ConvLSTM and **c** InceptionTime algorithms

Table 6 DL algorithms performance on training and validation phases

DL Algorithms	Training accuracy (%)	Validation accuracy (%)
CNN	95	94
ConvLSTM	92	92
InceptionTime	98	98

depicted in the form of training and validation loss, as shown in Fig. 4.

InceptionTime was the best-performing DL approach, with a training loss of 0.0470 and a validation loss of 0.0514, as shown in Fig. 4. The slight fluctuations in the validation loss show the inherent complexity and variability of the time-series data. On the test dataset, the InceptionTime algorithm reported an accuracy of 98%. The other key metrics evaluated for the classification model's performance are precision, recall, and F1-score; these are helpful when there is imbalance across classes. The training and validation results are shown in Table 6. However, the testing phase results reported in Table 7 show a high average precision value, particularly InceptionTime, was useful in decreasing false positives, while recall values show that a significant fraction of true positive cases is effectively identified. The F1-score, which is the harmonic mean of precision and recall, offers a balanced evaluation of a model's overall effectiveness. Confusion matrices of all DL models are shown in Fig. 5.

5 Discussion

The first phase of this study's results, which involved applying conventional ML algorithms to manually extracted features, highlight several important findings. SVM and RF models perform comparatively better, which highlights their strength and efficiency when evaluating structured feature sets. RF benefits especially from its capacity to handle complex feature interactions. However, the moderate performance of other ML methods (i.e., LR, AdaBoost, and MLP) shows the challenges these models encounter with the data set provided due to their limits in capturing intricate patterns and relationships within the features. A significant strength of the proposed approach was the use of the class imbalance strategy, which is critical for improving the model's performance. The models still struggled with the minority class, as seen by lower precision and recall scores for the frail class, as shown in Fig. 3. This shows a significant limitation in traditional ML techniques for efficiently identifying minority classes, which is vital in clinical applications where early and accurate frailty identification is required.

The second phase showed a significant improvement in performance, which was achieved by applying DL algorithms to raw IMU data. The CNN, ConvLSTM, and InceptionTime models all achieved excellent accuracy, with InceptionTime outperforming the others at 98%. This demonstrates the effectiveness of DL techniques in analyzing the raw IMU data for frailty classification and emphasizes its potential for creating reliable frailty classification systems. However, slight fluctuations in validation loss for the InceptionTime algorithm (Fig. 4) suggest that further optimization and possibly larger datasets could enhance stability and performance even further.

These findings suggest that conventional ML techniques serve as a useful benchmark for the frailty classification task. Whereas, the DL models offer significant improvements in accuracy and robustness, especially for complex and real-world frailty assessment clinical settings.

6 Conclusion

The increasing elderly population demands an effective frailty analysis system to improve their healthcare quality. A strong research effort in this area has the potential to have a significant socioeconomic impact, such as lower healthcare costs and increased independence for people who are in early stage of frail (pre-frail). Previous studies explored a variety of objective frailty assessment approaches for frailty classification task based on human gait using wearable sensor (IMU) and ML methods. Researchers used both methods such as hand-crafted feature engineering with classic ML methods and DL techniques to extract features from raw IMU signals. This research investigated the classification performance of both ML algorithms and DL algorithms in two phases. In the

Table 7 DL algorithms result on testing phase

DL	Testing phase			
	Avg. precision (%)	Avg. recall (%)	Avg. F1-Score (%)	Accuracy (%)
CNN	94	92	93	95
ConvLSTM	81	84	82	83
InceptionTime	**97**	**97**	**97**	**98**

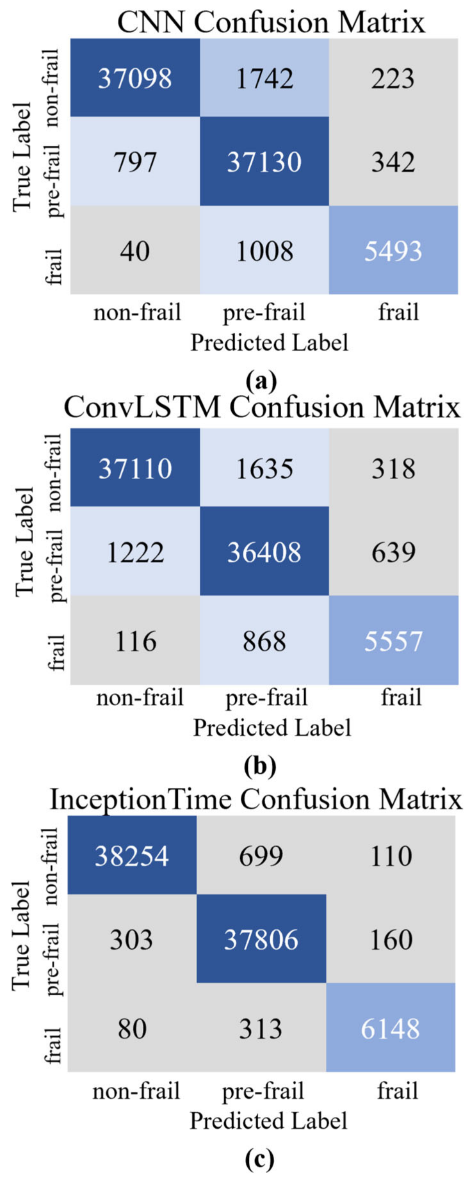

Fig. 5 Confusion matrices for the testing set of DL algorithms: **a** CNN, **b** ConvLSTM and **c** InceptionTime

first phase shallow ML algorithms were utilized with gait-based hand-crafted features. The second phase was to utilize DL techniques on raw IMU signals.

The results showed that DL techniques outperformed shallow ML methods in classifying frailty stages. The GSTRIDE database provided the hand-crafted features and raw IMU (accelerometer and gyroscope) data of elders that were used in this investigation. Among the ML algorithms, RF showed excellent performance, with an average tenfold CV accuracy of 70.29 and 63.27% on training and testing datasets, respectively. Overall, the DL algorithms outperformed; InceptionTime performed exceptionally well, with a test accuracy of 98%. These results demonstrate the effectiveness of DL techniques in the classification of frailty and highlight their potential for accurate and reliable results.

Further efforts should also be directed at finding optimal handcrafted gait features and selecting suitable ML models for effective frailty classification. These steps are required to investigate more diverse dataset across different populations to ensure the development of reliable and accurate models that can handle the challenges involved in detecting frailty in early stage (pre-frail).

Acknowledgements This publication was supported by the Department of Computer Graphics, Vision, and Digital Systems, under the statutory research project for young scientists (Rau6, 2024), Silesian University of Technology, Gliwice, Poland.

Author contributions Conceptualization, A.A. and A.S.; methodology, A.A.; software, A.A. and A.AN.; validation, A.S. and M.B.; formal analysis, A.A. and A.S..; investigation, A.A.; resources, A.AN.; data curation, A.AN.; writing original draft preparation, A.A.; writing review and editing, A.S and M.B.; visualization, A.A. and A.AN.; supervision, A.S.; project administration, M.B. All authors reviewed the final version of the manuscript.

Funding Not applicable.

Data availability The data presented in this study is openly available in [Zenodo] at https://doi.org/https://doi.org/10.5281/zenodo.6883291, reference number [40].

Declarations

Conflict of interest The authors declare that they have no conflict of interest

Ethical approval Not applicable.

References

1. Kojima, G., Liljas, A., Iliffe, S.: Frailty syndrome: implications and challenges for health care policy. Risk Manag. Healthc. Policy **12**, 23–30 (2019). https://doi.org/10.2147/RMHP.S168750
2. World Health, O.: Ageing and health, https://who.int/news-room/fact-sheets/detail/ageing-and-health
3. Kshatri, J.S., Palo, S.K., Bhoi, T., Barik, S.R., Pati, S.: Associations of multimorbidity on frailty and dependence among an elderly rural population: findings from the AHSETS study. Mech. Ageing Dev. **192**, 111384 (2020). https://doi.org/10.1016/j.mad.2020.111384

4. Cohen, R.A., Marsiske, M.M., Smith, G.E.: Neuropsychology of aging. In: Dekosky, S.T. and Asthana, S.B.T.-H. of C.N. (eds.) Geriatric Neurology. pp. 149–180. Elsevier (2019). https://doi.org/10.1016/B978-0-12-804766-8.00010-8
5. Morley, J.E., Vellas, B., van Abellan Kan, G., Anker, S.D., Bauer, J.M., Bernabei, R., Cesari, M., Chumlea, W.C., Doehner, W., Evans, J., Fried, L.P., Guralnik, J.M., Katz, P.R., Malmstrom, T.K., McCarter, R.J., Gutierrez Robledo, L.M., Rockwood, K., von Haehling, S., Vandewoude, M.F., Walston, J.: Frailty consensus: a call to action. J. Am. Med. Dir. Assoc. **14**, 392–397 (2013). https://doi.org/10.1016/j.jamda.2013.03.022
6. Kwak, D., Thompson, L.V.: Frailty: past, present, and future? Sports Med. Health Sci. **3**, 1–10 (2021). https://doi.org/10.1016/j.smhs.2020.11.005
7. Tarekegn, A., Ricceri, F., Costa, G., Ferracin, E., Giacobini, M.: Predictive modeling for frailty conditions in elderly people: machine learning approaches. JMIR Med. Inform. **8**, e16678 (2020). https://doi.org/10.2196/16678
8. Iqbal, M.S., Abbasi, R., Ahmad, W., Akbar, F.S.: Deep Learning and Robotics, Surgical Robot Applications. In: Azar, A.T., Koubaa, A. (eds.) Artificial Intelligence for Robotics and Autonomous Systems Applications, pp. 167–193. Springer International Publishing, Cham (2023)
9. Park, C., Mishra, R., Sharafkhaneh, A., Bryant, M.S., Nguyen, C., Torres, I., Naik, A.D., Najafi, B.: Digital biomarker representing frailty phenotypes: the use of machine learning and sensor-based sit-to-stand test. Sensors. **21**, 3258 (2021). https://doi.org/10.3390/s21093258
10. Fried, L.P., Tangen, C.M., Walston, J., Newman, A.B., Hirsch, C., Gottdiener, J., Seeman, T., Tracy, R., Kop, W.J., Burke, G., McBurnie, M.A.: Frailty in older adults: evidence for a phenotype. J. Gerontol. A Biol. Sci. Med. Sci. **56**, M146–M157 (2001). https://doi.org/10.1093/gerona/56.3.M146
11. Rockwood, K.: A global clinical measure of fitness and frailty in elderly people. Can. Med. Assoc. J. **173**, 489–495 (2005). https://doi.org/10.1503/cmaj.050051
12. Toosizadeh, N., Joseph, B., Heusser, M.R., Jokar, T.O., Mohler, J., Phelan, H.A., Najafi, B.: Assessing upper-extremity motion: an innovative, objective method to identify frailty in older bed-bound trauma patients. J. Am. Coll. Surg. **223**, 240–248 (2016). https://doi.org/10.1016/j.jamcollsurg.2016.03.030
13. Obbia, P., Graham, C., Duffy, F.J.R., Gobbens, R.J.J.: Preventing frailty in older people: an exploration of primary care professionals' experiences. Int. J. Older People Nurs. (2020). https://doi.org/10.1111/opn.12297
14. Amin, F., Majeed, A., Mateen, A., Abbasi, R., Hwang, S.O.: A Systematic survey on the recent advancements in the social internet of things. IEEE Access **10**, 63867–63884 (2022). https://doi.org/10.1109/ACCESS.2022.3183261
15. Osamy, W., Khedr, A.M., El-Sawy, A.A., Salim, A., Vijayan, D.: IPDCA: intelligent proficient data collection approach for IoT-enabled wireless sensor networks in smart environments. Electronics **10**, 997 (2021). https://doi.org/10.3390/electronics10090997
16. Zhang, L., Zeng, X., He, F., Huang, X.: Inflammatory biomarkers of frailty: a review. Exp. Gerontol. **179**, 112253 (2023). https://doi.org/10.1016/j.exger.2023.112253
17. Abbas, M., Le Bouquin Jeannès, R.: Acceleration-based gait analysis for frailty assessment in older adults. Pattern Recogn. Lett. **161**, 45–51 (2022). https://doi.org/10.1016/j.patrec.2022.07.006
18. Akbari, G., Nikkhoo, M., Wang, L., Chen, C.P.C., Han, D.-S., Lin, Y.-H., Chen, H.-B., Cheng, C.-H.: Frailty level classification of the community elderly using Microsoft Kinect-based skeleton pose: a machine learning approach. Sensors **21**, 4017 (2021). https://doi.org/10.3390/s21124017
19. Garcia-Moreno, F.M., Bermudez-Edo, M., Garrido, J.L., Rodríguez-García, E., Pérez-Mármol, J.M., Rodríguez-Fórtiz, M.J.: A microservices e-Health system for ecological frailty assessment using wearables. Sensors **20**, 3427 (2020). https://doi.org/10.3390/s20123427
20. Jansen, C.-P., Toosizadeh, N., Mohler, M.J., Najafi, B., Wendel, C., Schwenk, M.: The association between motor capacity and mobility performance: frailty as a moderator. Eur. Rev. Aging Phys. Act. (2019). https://doi.org/10.1186/s11556-019-0223-4
21. Razjouyan, J., Naik, A., Horstman, M., Kunik, M., Amirmazaheri, M., Zhou, H., Sharafkhaneh, A., Najafi, B.: Wearable sensors and the assessment of frailty among vulnerable older adults: an observational cohort study. Sensors **18**, 1336 (2018). https://doi.org/10.3390/s18051336
22. Parvaneh, S., Mohler, J., Toosizadeh, N., Grewal, G.S., Najafi, B.: Postural transitions during activities of daily living could identify frailty status: application of wearable technology to identify frailty during unsupervised condition. Gerontology **63**, 479–487 (2017). https://doi.org/10.1159/000460292
23. McCullagh, R., Brady, N.M., Dillon, C., Horgan, N.F.: A review of the accuracy and utility of motion sensors to measure physical activity of frail, older hospitalized patients. J. Aging Phys. Act. **24**, 465–475 (2016). https://doi.org/10.1123/japa.2014-0190
24. Schwenk, M., Mohler, J., Wendel, C., D''Huyvetter, K., Fain, M., Taylor-Piliae, R., Najafi, B.: Wearable sensor-based in-home assessment of gait, balance, and physical activity for discrimination of frailty status: baseline results of the Arizona frailty cohort study. Gerontology **61**, 258–267 (2015). https://doi.org/10.1159/000369095
25. Toosizadeh, N., Mohler, J., Wendel, C., Najafi, B.: Influences of frailty syndrome on open-loop and closed-loop postural control strategy. Gerontology **61**, 51–60 (2015). https://doi.org/10.1159/000362549
26. Martínez-Ramírez, A., Lecumberri, P., Gómez, M., Rodriguez-Mañas, L., García, F.J., Izquierdo, M.: Frailty assessment based on wavelet analysis during quiet standing balance test. J. Biomech. **44**, 2213–2220 (2011). https://doi.org/10.1016/j.jbiomech.2011.06.007
27. González, E.G., Villar, J.R., de la Cal, E.: Time Series Data Augmentation and Dropout Roles in Deep Learning Applied to Fall Detection. In: Herrero, Á., Cambra, C., Urda, D., Sedano, J., Quintián, H., Corchado, E. (eds.) 15th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2020). Springer, Cham (2021)
28. Kiprijanovska, I., Gjoreski, H., Gams, M.: Detection of gait abnormalities for fall risk assessment using wrist-worn inertial sensors and deep learning. Sensors **20**, 5373 (2020). https://doi.org/10.3390/s20185373
29. Martinez, M., De Leon, P.L.: Falls risk classification of older adults using deep neural networks and transfer learning. IEEE J. Biomed. Health Inform. **24**, 144–150 (2020). https://doi.org/10.1109/JBHI.2019.2906499
30. Mauldin, T., Canby, M., Metsis, V., Ngu, A., Rivera, C.: SmartFall: a smartwatch-based fall detection system using deep learning. Sensors **18**, 3363 (2018). https://doi.org/10.3390/s18103363
31. 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: 2018 IEEE international conference on healthcare informatics (ICHI). pp. 174–179. IEEE (2018)
32. Tunca, C., Salur, G., Ersoy, C.: Deep learning for fall risk assessment with inertial sensors: utilizing domain knowledge in spatio-temporal gait parameters. IEEE J. Biomed. Health Inform. **24**, 1994–2005 (2020). https://doi.org/10.1109/JBHI.2019.2958879
33. Butt, A.H., Cavallo, F., Maremmani, C., Rovini, E.: biomechanical parameters assessment for the classification of parkinson disease

using bidirectional long short-term memory. In: 2020 42nd annual international conference of the IEEE engineering in medicine & biology society (EMBC). pp. 5761–5764. IEEE (2020)

34. San-Segundo, R., Torres-Sánchez, R., Hodgins, J., De la Torre, F.: Increasing robustness in the detection of freezing of gait in Parkinson's disease. Electronics **8**, 119 (2019). https://doi.org/10.3390/electronics8020119
35. Sánchez-DelaCruz, E., Weber, R., Biswal, R.R., Mejía, J., Hernández-Chan, G., Gómez-Pozos, H.: Gait biomarkers classification by combining assembled algorithms and deep learning: results of a local study. Comput. Math. Methods Med. **2019**, 1–14 (2019). https://doi.org/10.1155/2019/3515268
36. Arshad, M.Z., Jung, D., Park, M., Shin, H., Kim, J., Mun, K.-R.: Gait-based frailty assessment using Image Representation of IMU Signals and Deep CNN. In: 2021 43rd annual international conference of the IEEE engineering in medicine & biology society (EMBC). pp. 1874–1879. IEEE (2021)
37. El-ziaat, H., El-Bendary, N., Moawad, R.: Hand-Crafted and Learned Features Fusion for Predicting Freezing of Gait Events in Patients with Parkinson's Disease. In: Abraham, A., Siarry, P., Ma, K., Kaklauskas, A. (eds.) 19th International Conference on Intelligent Systems Design and Applications (ISDA 2019) held December 3–5, 2019. Springer, Cham (2021)
38. Liu, Y., He, X., Wang, R., Teng, Q., Hu, R., Qing, L., Wang, Z., He, X., Yin, B., Mou, Y., Du, Y., Li, X., Wang, H., Liu, X., Zhou, L., Deng, L., Xu, Z., Xiao, C., Ge, M., Sun, X., Jiang, J., Chen, J., Lin, X., Xia, L., Gong, H., Yu, H., Dong, B.: Application of machine vision in classifying gait frailty among older adults. Front. Aging Neurosci. (2021). https://doi.org/10.3389/fnagi.2021.757823
39. Lin, C.-W., Wen, T.-C., Setiawan, F.: Evaluation of vertical ground reaction forces pattern visualization in neurodegenerative diseases identification using deep learning and recurrence plot image feature extraction. Sensors **20**, 3857 (2020). https://doi.org/10.3390/s20143857
40. García-de-Villa, S., Neira, G.G.-V., Álvarez, M.N., Huertas-Hoyas, E., Ruiz, L.R., Del-Ama, A.J., Sánchez, M.C.R., Jiménez, A.R.: A database with frailty, functional and inertial gait metrics for the research of fall causes in older adults. Sci. Data **10**, 566 (2023). https://doi.org/10.1038/s41597-023-02428-0
41. García-Villamil, G., Ruiz, L., Jiménez, A.R., Granja, F.S., Rodríguez-Sánchez, M.C.: Influence of IMU's Measurement Noise on the Accuracy of Stride-Length Estimation for Gait Analysis. In: IPIN-WiP (2021)
42. Álvarez, M.N., Ruiz, A.R.J., Neira, G.G.-V., Huertas-Hoyas, E., Cerda, M.T.E., Delgado, L.P., Robles, E.R., Del-Ama, A.J., Ruiz-Ruiz, L., García-de-Villa, S., Rodriguez-Sanchez, C.: Assessing falls in the elderly population using G-STRIDE foot-mounted inertial sensor. Sci. Rep. **13**, 9208 (2023). https://doi.org/10.1038/s41598-023-36241-x
43. Xu, L., Zhang, J., Shen, S., Hong, X., Zeng, X., Yang, Y., Liu, Z., Chen, L., Chen, X.: Association between body composition and frailty in elder inpatients. Clin. Interv. Aging **15**, 313–320 (2020). https://doi.org/10.2147/CIA.S243211
44. Reches, T., Dagan, M., Herman, T., Gazit, E., Gouskova, N., Giladi, N., Manor, B., Hausdorff, J.: Using wearable sensors and machine learning to automatically detect freezing of gait during a fog-provoking test. Sensors. **20**, 4474 (2020). https://doi.org/10.3390/s20164474
45. Suykens, J.A.K., Vandewalle, J.: Least squares support vector machine classifiers. Neural. Process. Lett. **9**, 293–300 (1999). https://doi.org/10.1023/A:1018628609742
46. Hosmer, D.W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression. Wiley, Hoboken (2013)
47. Lockhart, T.E., Soangra, R., Yoon, H., Wu, T., Frames, C.W., Weaver, R., Roberto, K.A.: Prediction of fall risk among community-dwelling older adults using a wearable system. Sci. Rep. **11**, 20976 (2021). https://doi.org/10.1038/s41598-021-00458-5
48. Ghosh, S., Fleiner, T., Giannouli, E., Jaekel, U., Mellone, S., Häussermann, P., Zijlstra, W.: Statistical learning of mobility patterns from long-term monitoring of locomotor behaviour with body-worn sensors. Sci. Rep. **8**, 7079 (2018). https://doi.org/10.1038/s41598-018-25523-4
49. Thinh, L.V., Thanh, N.L. Van, Huan, T.T., Nha, N.T.: Human Gait Classification Model Based on Data of IMU Sensor and Multilayer Perceptron Neural Network Model. BT-Proceedings of the International Conference on Advanced Mechanical Engineering, Automation, and Sustainable Development 2021 (AMAS2021). Presented at the (2022). https://doi.org/10.1007/978-3-030-99666-6_121
50. Dargan, S., Kumar, M., Ayyagari, M.R., Kumar, G.: A Survey of deep learning and its applications: a new paradigm to machine learning. Arch. Comput. Methods Eng. **27**, 1071–1092 (2020). https://doi.org/10.1007/s11831-019-09344-w
51. van Kuppevelt, D., Meijer, C., Huber, F., van der Ploeg, A., Georgievska, S., van Hees, V.T.: Mcfly: automated deep learning on time series. SoftwareX. **12**, 100548 (2020). https://doi.org/10.1016/j.softx.2020.100548
52. Jaén-Vargas, M., Reyes Leiva, K.M., Fernandes, F., Barroso Gonçalves, S., Tavares Silva, M., Lopes, D.S., Serrano Olmedo, J.J.: Effects of sliding window variation in the performance of acceleration-based human activity recognition using deep learning models. PeerJ Comput. Sci. **8**, e1052 (2022). https://doi.org/10.7717/peerj-cs.1052
53. Gupta, S.: Deep learning based human activity recognition (HAR) using wearable sensor data. Int. J. Inform. Manag. Data Insights **1**, 100046 (2021). https://doi.org/10.1016/j.jjimei.2021.100046
54. Akber, S.M.A., Kazmi, S.N., Mohsin, S.M., Szczęsna, A.: Deep learning-based motion style transfer tools. Tech. Future Chall. Sens. **23**, 2597 (2023). https://doi.org/10.3390/s23052597
55. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Inf. Process. Manag. **45**, 427–437 (2009). https://doi.org/10.1016/j.ipm.2009.03.002
56. Ali, M., Mazhar, T., Shahzad, T., Ghadi, Y.Y., Mohsin, S.M., Akber, S.M.A., Ali, M.: Analysis of Feature selection methods in software defect prediction models. IEEE Access **11**, 145954–145974 (2023). https://doi.org/10.1109/ACCESS.2023.3343249