Fusion of Clinical, Self-Reported, and Multisensor Data for Predicting Falls

Joana Silva, Inês Sousa, and Jaime S. Cardoso, *Senior Member, IEEE*

***Abstract*—Falls are among the frequent causes of the loss of mobility and independence in the elderly population. Given the global population aging, new strategies for predicting falls are required to reduce the number of their occurrences. In this study, a multifactorial screening protocol was applied to 281 community-dwelling adults aged over 65, and their 12-month prospective falls were annotated. Clinical and self-reported data, along with data from instrumented functional tests, involving inertial sensors and a pressure platform, were fused using early, late, and slow fusion approaches. For the early and late fusion, a classification pipeline was designed employing stratified sampling for the generation of the training and test sets. Grid search with cross-validation was used to optimize a set of feature selectors and classifiers. According to the slow fusion approach, each data source was mixed in the middle layers of a multilayer perceptron. The three studied fusion approaches yielded similar results for the majority of the metrics. However, if recall is considered to be more important than specificity, then the result of the late fusion approach providing a recall of 78.6% is better compared with the results achieved by the other two approaches.**

***Index Terms*—Fall risk assessment, inertial sensors, machine learning, pressure platform.**

I. Introduction

THE worldwide population aged over 65 is growing rapidly. The consequences of this phenomenon are not only social and health-related, but also economic. The process of aging affects the ability of a person to maintain balance, mobility, and muscle strength and to react properly to unexpected situations such as slipping or stumbling. There are also cross-related factors resulting from health conditions, including loss of auditory and visual capabilities, side effects of medications, dizziness, body pain, depression, and slow walking speed. Aside from these intrinsic risk factors, falls among older people are also associated with extrinsic factors, such as environment hazards, footwear malfunctioning, improper use of assistive devices, and recent hospitalizations [1].

Given a wide range of factors contributing to falls in the context of an aging population, it becomes extremely important to frame strategies that properly evaluate the risk factors of falls in older people. Several scales, questionnaires, functional tests, and protocols have been proposed in the past years to overcome the lack of standardized clinical and medical procedures for assessing the risk of falls [2]. However, in the majority of public sectors, risk factors of falls among the elderly are only assessed after the occurrence of a fall leading to hospitalization or the need for other forms of medical care. When a fall risk assessment is conducted after an occurrence of fall, the collected parameters are altered as a consequence. On the other hand, the majority of the proposed assessment scales and questionnaires are subjective and self-reported and do not consider all major fall risk factors. Proper methods for the objective assessment of individual gait, strength and balance are confined to laboratory settings requiring specialized personnel and equipment, thus leading to higher costs. All these solutions rely on on-time assessments which do not reflect the variation of risk factors over time.

Recently, solutions for the fall risk assessment based on low-cost technologies have been proposed [2], including solutions based on inertial sensors embedded into wearable devices or smartphones. There are also solutions based on force and pressure platforms aiming at assessing multiple factors of balance and correlated fall risks.

This study describes an approach to predicting falls based on a multifactorial screening protocol that combines personal, inertial, and pressure platform data. Three alternative approaches were explored for data fusion: an early approach, that combines all data in a unified feature vector used for optimizing a grid search pipeline; a late fusion approach that combines the predictions of three classification pipelines trained with each of the individual data sources; and a slow fusion approach that uses information from each data source individually in the first layer of a multilayer perceptron (MLP) and then trains this MLP end-to-end using all data sources.

The main contributions of this study are the following: i) the use of multimodal data collected according to a multifactorial screening protocol for predicting falls; ii) the richness of the collected data allowing to infer not only functional capabilities

Manuscript received January 15, 2019; revised May 20, 2019, July 12, 2019, September 2, 2019, and October 9, 2019; accepted October 29, 2019. Date of publication November 8, 2019; date of current version January 6, 2020. This work was supported by the Project Symbiotic Technology for Societal Efficiency Gains: Deus ex Machina (NORTE-01-0145-FEDER-000026) funded by North Portugal Region Operational Programme (NORTE 2020), Portugal 2020, and the European Regional Development Fund (ERDF) from European Union. *(Corresponding author: Joana Silva.)*

J. Silva is with the Fraunhofer Portugal AICOS, 4200-135 Porto, Portugal, and also with the Faculty of Engineering, University of Porto, 4099-002 Porto, Portugal (e-mail: joana.silva@fraunhofer.pt).

I. Sousa is with the Fraunhofer Portugal AICOS, 4200-135 Porto, Portugal (e-mail: ines.sousa@fraunhofer.pt).

J. S. Cardoso is with the INESC TEC, 4200-465 Porto, Portugal, and also with the Faculty of Engineering, University of Porto, 4099-002 Porto, Portugal (e-mail: jaime.cardoso@inesctec.pt).

Digital Object Identifier 10.1109/JBHI.2019.2951230

of a person but also clinical and environmental information; and iii) the exploration of different fusion approaches.

II. Related Work

The related work described here comprises studies that used any type of sensors to retrieve metrics during the execution of fall risk functional tests. The studies focused only on clinical, self-reported, or measurable variables (e.g., [3], [4]), are not discussed in this section. We only noted that the sensitivity achieved in these studies varied from 43% to 100% (median = 80%), whereas the specificity ranged from 38% to 96% (median = 75%). Howcroft *et al.* [2] reviewed previous studies focusing on the fall risk assessment with inertial sensors. The authors concluded that future research should i) consider investigating the relationship between the models' predictive variables and specific fall risk factors and ii) focus on groups with an increased fall risk due to some diseases. As weak points of the previous studies, the authors reported that 50% of them had not used separate datasets for model training and validation, which could have impacted the models' applicability beyond the training set population. Moreover, applying the most commonly used cut-off values in clinical assessment tests could have biased the decisions made since the thresholds typically used to split classes had produced false positives and false negatives, introducing inaccuracies when evaluating sensor-based models. Another aspect to be considered is that clinical assessment thresholds were not used consistently across the research studies included in the review. The prospective fall occurrence rate is considered to be the most reliable criterion for dividing subjects into non-fallers and fallers [2]; however this criterion was only used in 15% of the studies. Regarding the retrospective fall assessment, the most relevant limitations are the inaccurate recording of fall histories most commonly assessed by self-reported questionnaires and the fact that balance, strength, and gait parameters can change due to past falls.

A. Retrospective Studies

Bigelow *et al.* [5] studied posturography for clinical fall risk screening of older adults. They recruited 150 adults aged 65 and above from local senior centers and independent living facilities. The subjects were categorized as recurrent fallers and non-recurrent fallers based on their fall status in the previous year. The participants performed four standing tasks on a force platform. The authors extracted "traditional and fractal measures from the center of pressure data" [5]. Their logistic regression model exhibited a sensitivity (recall) of 75% and a specificity of 94%. The authors highlighted the importance of combining multiple variables rather than using only a single measure to compute the fall risk.

Qiu *et al.* [6] reported a study conducted with multiple wearable inertial sensors for multifactorial fall risk assessment on 196 community-dwelling older women. The sequence included the Timed Up and Go (TUG), Five Times Sit to Stand (5TSTS), and Limits of Stability tests. A model built using inertial sensor data and support vector machine was able to classify between fallers (N = 82) and non-fallers (N = 114) based on fall histories. The model achieved an overall accuracy of 89.4% (92.7% sensitivity and 84.9% specificity). The results of the study support the idea that inertial sensors allow the identification of individuals with a high risk of falls, who should be followed with fall prevention strategies.

Greene *et al.* [7] performed a quantitative estimation of the fall risk using multiple sensors during the standing balance exercise. The authors acquired data from 120 community-dwelling older adults aged over 60 by using a pressure-sensitive platform sensor and attaching a body-worn inertial sensor to the lower back of the participants. The estimation of the fall risk was compared with the Berg Balance Scale (BBS). The results were analyzed by gender using a support vector machine model, which returned a mean classification accuracy of 73.07% for the participants with a self-reported history of falling in the past 5 years. These results compared favorably with those obtained using solely the BBS (with a mean classification accuracy of 59.42%).

B. Prospective Studies

Liu *et al.* [8] reported an accelerometer-based fall prediction model that was trained using wearable inertial sensor data obtained in a routine assessment, including the TUG test, Alternate Step Test (AST), and 5TSTS. The study sample included 68 subjects aged from 72 to 91 from a previous study and a second group of 30 subjects aged from 68 to 92 who were newly recruited. The authors have assessed the prospective falls that occurred in the following 12 months based on fall diaries. The best classification performance allowing to distinguish fallers from non-fallers with a sensitivity of 68% and a specificity of 73% was achieved by a logistic regression model that was trained using only AST data.

Schooten *et al.* performed several studies focusing on the assessment of the ambulatory fall risk. The study participants aged over 65 wore an inertial sensor for one week. The authors extracted metrics related to the amount of physical activity and gait characteristics and reported several approaches ranging from logistic regression to deep learning methods to discriminate between fallers and non-fallers. A logistic regression model trained on accelerometry-derived parameters of gait obtained from 139 participants allowed to substantially improve the area under the curve (AUC) up to a value of 0.82, compared with using questionnaires and functional test scores alone [9]. Deep learning models built using a dataset of 296 older adults achieved an accuracy similar to that of the logistic regression model. Aicha *et al.* [10] highlighted the fact that deep learning models have the advantage of not requiring the implementation of feature extraction methods.On the other hand, deep learning models lack interpretability, which limits their application in medical contexts. The same authors [11] also demonstrated that the gait quality in daily life is "predictive for both time-to-first and time-to-second falls in both univariate and multivariate models" with adequate to good accuracy.

C. Challenges and Opportunities

One of the most commonly reported limitations of fall risk assessments is the large feature dimensionality relative to the sample size of datasets. With the exception of the work by Schooten *et al.*, the majority of existing studies are based on

data collected from less than 150 participants. The collection of larger datasets is time- and resource-consuming, whereas small-size datasets can impact the quality of the analysis and generalizations retrieved from that data. Moreover, the low incidence of falls (less than 30% in the older population) leads to unbalanced datasets which can negatively impact results. In this study, we employed an oversampling technique to deal with the unbalanced nature of the collected dataset, a procedure that is rarely reported in the literature for fall prediction. Furthermore, we divided the dataset into a training dataset and a hold-out test set for model validation, something that has been lacking in previous research [2].

In this case study, we applied a multifactorial fall risk screening protocol to 403 participants. Only a part of the population aged over 65 was taken into consideration for analysis, resulting in a total of 281 participants. Another challenge presented in previous studies relates to the fall risk parameters being evaluated and data sources used for feature extraction. The majority of the existing studies focusing on fall risk prediction are based on a single source of data, either clinical and self-reported or extracted from inertial sensors. In contrast, we combined several data sources, including not only clinical and self-reported data but also information about functional capabilities, such as mobility, balance and strength, obtained from inertial sensors and a pressure platform during the execution of a multifactorial screening protocol. This protocol combined the most relevant tests for assessing grip strength, balance, mobility and muscle strength. The multifactorial nature of the collected data provided an opportunity to study data fusion approaches, and compare models based on a single source of data with models based on data fusion. While there is a lack of consensus regarding the output metric that should be used to divide population groups into fallers and non-fallers, the 1-year follow-up occurrence of falls has been pointed out as the most reasonable metric [2]. We used monthly follow-up phone calls to record the occurrence of falls over 1-year period and based our analysis on reported fall occurrences instead of automatic detection of falls. while there are tools for automatic detection of falls, such as personal emergency response systems (PERS), these have never been reported to be used during the follow-up period.

III. Methodology

A. Data Collection

1) Subjects: Four hundred and three Portuguese community-dwelling adults aged over 50 (mean age of 69.69 $\pm$ 10.31; 70% women) were recruited from parish councils, physical therapy clinics, senior's universities, and other community facilities. The inclusion criterion was the ability to independently stand and walk with or without walking aids. The excluding criterion was the presence of severe sensory (deafness or blindness) or cognitive impairments [12]. Only adults aged over 65 were considered for the analysis given that many previous studies used this age as a threshold for patient recruitment. The sample used in this study consisted of 281 subjects. The research was approved by the Ethics Committee at the Polytechnic Institute of Coimbra (N°6/2017). All participants gave their written informed consent before the data collection in accordance with the principles of the Declaration of Helsinki [12].

2) Protocol: A multifactorial screening protocol for assessin the risk of falls in community-dwelling adults was defined based on relevant literature. The protocol included demographic and anthropometric data; lifestyle and health behavior data; six functional tests (*handgrip strength test*, *TUG test, 30 s STS*, Step test (*Step*), "modified" 4-Stage Balance test (*4Stage*), and 10-m walking speed test (*10 Meter Walk*) instrumented with inertial sensors and a pressure platform); and questionnaires about environmental home hazards, activity and participation profile related to mobility, and self-efficacy to exercise [12].

3) Data Sources: Several types of data were collected:

- *Clinical data,* including demographic, anthropometric and data such as place of residence, age, sex, medical conditions, and medications taken, as well as functional tests outcomes, such as test timing, number of repetitions, and grip strength.
- *Self-reported data* from questionnaires, such as home hazards, previous number of falls, and fear of falling;
- *Three-dimensional (3D time series* extracted from the 3D accelerometer and 3D gyroscope used in the functional tests, including the time to stand and average acceleration along x, y, and z axes.
- *Two-dimensional (2D) time series* extracted from the pressure platform used in the functional tests, including the center of pressure oscillation in the mediolateral and anteroposterior directions.

Clinical and self-reported data were combined to form one data source named the *personal data.*

4) Prospective Falls After 12 Months: The participants were followed for 12 months via monthly phone calls. "The rate of falls was recorded from the day of inclusion until voluntary dropout, loss of phone contact or the end of the follow-up period" [12]. The participants who reported at least one fall in the 12-month follow-up period were categorized as fallers, whereas those who did not report any falls during this period were categorized as non-fallers. The incidence of fallers in the study sample was 26.3%, which is in accordance with the literature reporting that approximately one-third of people over 65 will fall each year [2].

B. Feature Extraction

During the walking tests (i.e., TUG and 10 Meter Walk), two wearable inertial sensors [14], were placed on the lower back and ankle of the support leg. The sensors were sampled at 50 Hz. For the static tests (i.e., STS Step, and 4Stage) the PhysioSensing pressure platform [15] sampled at 50 Hz was used in addition to the two inertial sensors. The handgrip strength was assessed using a Jamar hydraulic hand dynamometer [12]. Each functional test was divided into phases, e.g., 4Stage was divided into seven balance positions. Several features, as detailed in Table I were extracted from the four sources of data, i.e., clinical, self-reported, inertial sensor, and pressure platform data. Overall, 230 features were extracted.

1) Inertial Sensors: An analysis and segmentation of the TUG test involving inertial sensors were previously reported

TABLE I
FEATURES EXTRACTED FROM CLINICAL REPORTS, SELF-REPORTED, INERTIAL, AND PRESSURE PLATFORM DATA

Source	Features extracted
Clinical data [12]	**sex**, age, **height**, weight, dwelling place, **benzodiazepines**, **antidepressants**, anti-psychotic, anti-inflammatory analgesics, anti-hypertensive, **total medication**,+ **4 medicines daily**, STS score, TUG score, 4Stage score, Step score, 10 Meter Walk score
Self-reported data [12]	**retrospective falls**, prospective falls, fear of falling, **live alone**, sedentary lifestyle, assistive device, **upper extremities assistance to stand**, **home risks**, not applicable home risks, items answered, risk home entrance, risk stairs out, risk stairs in, risk living areas, risk kitchen, **risk bathroom**, **risk bedroom**, risk outdoor, **index of home risk**, **index of home risk percentage**, self-efficacy score
Inertial sensors [13]	mean, **median**, max, min, rms, std dev, median dev, **iqr**, min avg, max avg, **peak height**, avg peak height, mean cross count, fft max freq, **fft max amp**, **energy**, entropy, skewness, kurtosis, walking steps, walking variability, walking speed, STS power, time to stand
Pressure platform [13]	sway velocity, **sway range**, sum oscillation, std oscillation, **area ellipse**, transfer time, **left foot force**, **right foot force**, left foot higher pressure zone, right foot higher pressure zone, **rising index**, weight symmetry

max: maximum, min: minimum, rms: root mean square, std dev: standard deviation, iqr: interquartile range, avg: average, fft: fast fourier transform, freq: frequency, amp: amplitude.

by Silva and Sousa [16]. Later on, Silva *et al.* [13] presented an analysis of the TUG, STS, and 4Stage tests performed with inertial sensors and a pressure platform, reporting a feature extraction process for both types of sensors. We adopted the analysis procedures reported in these studies. Our analysis of the 10 Meter Walk test was based on a previous work by Aguiar *et al.* [17]. The inertial features presented in Table I were extracted from the magnitude of the accelerometer signal.

2) Pressure Platform: For the Step test, the number of steps was segmented based on the information provided by the pressure platform when a subject raised the leg. If a variation in the number of active cells was detected compared with the initial bipodal position, a step was identified. As the leg was lowered toward the pressure platform, the number of active cells increased, and the end of the segmentation phase was reached. The pressure platform features extracted for the Step and STS tests were the same as previously described for the STS test [13].

C. Classification Pipeline

1) Data Profiling: First, nominal data, such as therapist id, patient id, local id, were removed from the feature vector, and the remaining variables were converted to numerical values. The clinical and self-reported data were converted to numerical values using categorical/dichotomous variables when appropriate. Then, data profiling was performed, and the features with a correlation coefficient above 0.90 were removed. A statistical description of the database was achieved by depicting grouping variables such as the last year falls, follow-up falls, need of walking aid, and need of assistance to stand up in scatter and box plots. We performed an independent samples t-test on each grouping variable, with 95% confidence level. A segmentation of the database for each type of the data source, i.e., personal data (comprising the clinical and self-reported data), inertial sensor data, and pressure platform data, was also considered for testing different fusion approaches.

2) Feature Pre-Processing: Several feature pre-processing methods were employed, mainly for dealing with missing values. As 28 participants were unable to perform at least one of the functional tests due to physical limitations, the data from these tests had missing values. Moreover, missing values were present when participants were unable to reach the last positions among the seven balance positions of the 4Stage test. The last position of the 4Stage test had 80% missing values. The missing values for the remaining features accounted on average for 5.8 $\pm$ 11.8% of all values. Due to time constraints during the data collection, the database also contained some missing answers for participants who filled in the questionnaire. Since the inability to accomplish a functional test could yield valuable information related to functional capabilities, the missing values in such cases were replaced by zero. Removing the participants with missing values would have resulted in a significant reduction of the sample dimension, preventing from accurately representing the target population. All features were normalized by removing the mean and scaling to the unit variance.

3) Data Fusion Approaches: Three approaches to data fusion could be considered using the collected dataset: 1) *data-level fusion*, i.e., combining the data obtained from the inertial sensors and pressure platform to extract features resulting from the joint analysis of both signals; 2) *feature-level fusion*, i.e., extracting features from the three data sources separately and combining all features in the same feature vector; and 3) *decision-level fusion*, i.e., training a model for each data source and combining the predictions of all models. In our study, we experimented with three different data fusion approaches. The first approach, called *early fusion*, involves fusing data after the feature extraction stage and before the classification stage (i.e., feature-level fusion). The second approach, called *late fusion*, involves fusing data after the classification stage (i.e., decision-level fusion. Finally, the third approach, called *slow fusion* is based on combination of the first two approaches. In particular, it gradually fuses multisource information in a neural network, in such a way that higher layers of the network are provided with progressively more information [18].

The *late fusion approach* uses the majority voting mechanism, where the predicted class label for a specific instance is assigned based on the class label predicted by the majority of individual classifiers. The *slow fusion approach* combines the information of each data source in the middle layers of a neural network (Fig. 1). For the implementation of the *slow fusion approach*, we employed the Keras library to train a multi-input sequential model, receiving three data sources in a single network. For each data source, we combined three feedforward fully connected (dense) layers with the *ReLu* activation function, intercalated with dropout layers, and with a sequential decrease in the number of layers' nodes. The last layers of each model were concatenated in a stack of two dense layers with *sigmoid* activation. This model was optimized using *binary cross entropy* loss and *Adam* optimization.

4) Classification Pipeline: A randomly stratified train-test split was performed 50 times to ensure the variability between train and test splits, with 33% of the data being selected for

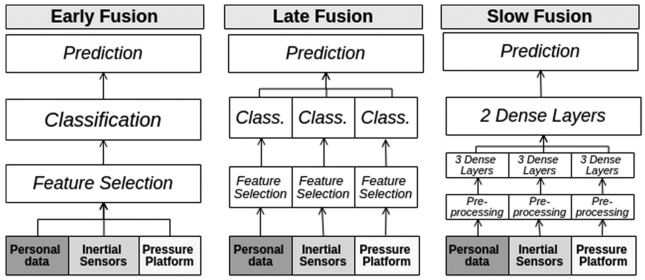

Fig. 1. Early, late, and slow fusion approaches for combining personal, inertial sensor, and pressure platform data, for fall prediction.

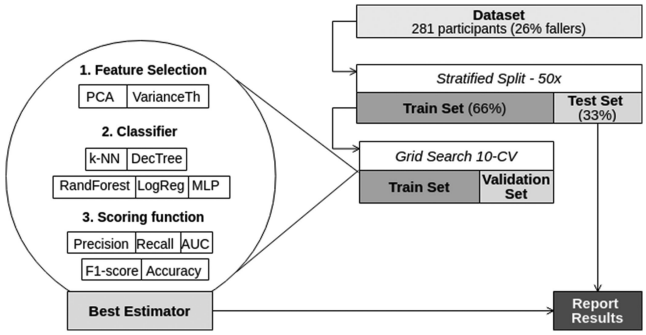

Fig. 2. Classification pipeline for optimizing the of feature selector, classifiers, and scoring function, grid search with CV is applied to the training set, whereas results are reported for the test set.

the test set. Each split yielded a training set of 188 samples (138 non-fallers, 50 fallers) and a test set of 93 samples (69 non-fallers, 24 fallers). Using grid search with cross-validation (CV) over the training set, a classification pipeline was defined to optimize a range of parameters for the three stages: feature selection, classification, and grid search scoring (Fig. 2).

For the feature selection, we optimized the number of components for principal component analysis (PCA) and threshold for the variance threshold method. For the classification stage, we optimized the following hyperparameters of each of the considered classifiers: the variable k and search algorithm of the k-Nearest Neighbors (k-NN) classifier; the maximum depth, number of estimators, and minimum samples to split for the Decision Tree and Random Forest classifiers; and the solver and maximum number of iterations for the Logistic Regression (LogReg) classifier. For the grid search scoring, we considered precision, recall, AUC, F1-score, and accuracy.

Since the incidence of fallers in the database was only 26.33%, we applied an oversampling procedure, namely, the Synthetic Minority Oversampling Technique (SMOTE) [19], to the training set employed in the grid search. In particular, the SMOTE was used to oversample the minority class in the feature space. In this way, the minority class was oversampled by creating synthetic examples rather than oversampling with replacement.

5) Validation: Since the grid search was performed over 50 partitions of the dataset and for the three stages of the pipeline, we obtained several combinations of parameters evaluated with different partitions of the initial dataset. We decided to present the mean and standard deviation across the 50 iterations for each tested classifier combined with different feature selection methods, estimator's hyperparameters, and grid search optimization scores. We report the obtained accuracy, AUC, F1-score, precision, recall, and specificity. To compare the performance metrics across the different fusion approaches, we used ANOVA multiple comparison analysis testing. As a *post-hoc* test we applied Tukey's Honest Significant Difference Test (HSDT) with 95% confidence level to all possible pairs among the three data fusion methods.

IV. Results

A. Descriptive Characteristics

1) Demographic and Anthropometric Information: A total of 281 older people aged over 65 were included in this study. Out of them, 65% were female, 74% were community-dwelling, and 17% used a walking aid. Participants were 75.1 $\pm$ 6.9 years old, 160 $\pm$ 7.9 cm tall and weighed 72.1 $\pm$ 11.1 kg.

2) Retrospective and Prospective Falls: Out of the 281 participants, 94 (33.5%) reported at least one fall in the previous year and 74 subjects (26.3%) experienced at least one fall during the 1-year follow-up. Among the 94 subjects that reported previous falls, 35 fell during the follow-up period.

3) Self-Reported Questionnaires: Self-reported questionnaires revealed that 38.8% of the participants required an upper extremity assistance to stand up from a chair. Among all participants, 35.6% reported living alone, 69% reported taking more than four medicines daily, and 50.5% reported having a sedentary lifestyle. When asked if they were afraid of falling, 52.7% answered affirmatively.

4) Functional Tests Scores: The majority of the subjects (253 out of 281) were able to complete all functional tests. Out of the 281 subjects, 13 subjects did not perform the TUG test, 14 subjects were unable to complete the Step test, and 17 subjects were unable to do the 30 s STS test. Only eight subjects were unable to perform any standing position of the 4Stage test, whereas all participants completed the 10 Meter Walk test. Data from the subjects who were only capable of performing one or two tests were still considered for analysis.

5) Individual Predictive Value: We performed a statistical analysis of the individual predictive value of each feature for the prediction of 12 months prospective falls. The differences in the functional test scores between fallers and non-fallers were not statistically significant (p-value $>$ 0.05). The difference between the two groups was statistically significant for the features highlighted in bold in Table I.

B. No Data Fusion - Individual Data Sources

The classification performance metrics using each data source individually were retrieved from the inner loop of the late fusion approach to access the predictive value of each source. The results were grouped by the data source, feature selector, classifier, and grid search score. The average results for the 50 test sets were computed, and the highest recall values were retrieved as listed in Table II.

According to Tukey's HSDT performed for the single-step multiple comparison between the data sources, the averages

TABLE II
AVERAGE RESULTS FOR EACH DATA SOURCE (MEAN AND STANDARD DEVIATION OF THE 50 TEST SETS, IN %)

Source	Personal	Inertial	Platform
Selector	PCA	PCA	PCA
Model	Decision Tree	Decision Tree	Log Reg
Score	Recall	Recall	Recall
Accuracy	40.5 ± 10.2	40.0 ± 7.8	39.8 ± 6.9
AUC	47.8 ± 4.8^{I}	50.5 ± 3.7^{S}	49.8 ± 4.4
F1-score	$34.0 \pm 7.9^{I,P}$	38.0 ± 3.9^{S}	37.1 ± 5.6^{S}
Precision	$23.8 \pm 4.9^{I,P}$	26.1 ± 2.1^{S}	25.5 ± 3.1^{S}
Recall	62.9 ± 22.9^{I}	72.2 ± 15.1^{S}	70.6 ± 16.6
Specificity	32.7 ± 20.6	28.8 ± 14.8	29.1 ± 13.7

S: sig. different from personal; I: sig. different from inertial; P: sig. different from platform.

TABLE III
AVERAGE RESULTS FOR EARLY, LATE, AND SLOW FUSION (MEAN AND STANDARD DEVIATION OF THE 50 TEST SETS, IN %)

Fusion	Early fusion	Late Fusion	Slow Fusion
Selector	PCA	PCA	n.a.
Model	Decision Tree	Decision Tree	MLP
Score	Recall	Recall	Cross Entropy
Accuracy	35.8 ± 10.3^{S}	37.0 ± 9.8^{S}	59.2 ± 4.8
AUC	49.5 ± 3.9	50.5 ± 4.3	50.3 ± 4.4
F1-score	37.1 ± 7.5^{S}	38.2 ± 6.0^{S}	28.5 ± 6.4
Precision	25.2 ± 5.1	26.2 ± 3.2	26.3 ± 5.9
Recall	77.8 ± 24.1^{S}	78.6 ± 22.2^{S}	31.8 ± 8.5
Specificity	21.2 ± 21.6^{S}	22.5 ± 19.9^{S}	68.7 ± 6.9

S: sig. different from slow.

of the accuracy and specificity were not significantly different across all data sources. For the AUC and recall, only the average of the personal data was significantly different from that of the inertial data. For the F1-score and precision, only the averages of the personal data were significantly different from that of the inertial and platform data.

C. Early, Late, and Slow Fusion Approaches

The same classification pipeline was employed for the early and late fusion approaches. For the early fusion approach, we used a combined feature vector with information from the three sources of data and ran it through the classification pipeline illustrated in Fig. 2. For the late fusion approach, the data were split into inertial, pressure platform, and a combination of clinical and self-reported data. The pipeline shown in Fig. 2 was optimized using each source of data individually and then the best estimator for each source of data was combined using voting classification. Finally, the evaluation using the test set was performed.

1) Early Fusion Approach: According to the early data fusion approach, clinical, self-reported, and multisensor data were fused using feature fusion prior to the classification pipeline. In addition to the clinical and self-reported data retrieved mainly from questionnaires (categorical data) and measured variables (e.g., timed tests or anthropometric characteristics), we employed features engineered from the raw signals of the inertial sensors and pressure platform. The initial analysis was performed individually for each type of sensor data. After retrieving features from the inertial sensor and pressure platform signals, they were combined in a unified feature vector together with clinical and self-reported data. This feature vector was then used for the optimization of the grid search pipeline and for retrieval of the best estimator. The resulting feature vector included 229 features extracted from the three data sources for 281 participants (aged over 65). The results were grouped by the data source, feature selector, classifier, and grid search score. The average results for the 50 test sets were computed and the highest recall values were retrieved as listed in Table III. The best combination was PCA, Decision Tree and recall as grid search score function.

2) Late Fusion Approach: In the case of the late fusion approach, the same procedure as described for the early fusion was applied; however individual data sources were used in this case. We combined the predictions of three different estimators (based on individual inertial, pressure platform, and clinical/self-reported data) using a voting classifier. The feature vector constructed based on the inertial data comprised 125 features. In particular, we extracted 59 features from the pressure platform data and 44 features from the clinical/self-reported data. The model selection method was the same as described for the early fusion. The best combination was PCA, Decision Tree, and recall as the grid search score (Table III).

3) Slow Fusion Approach: The slow fusion approach slows the process of fusing estimations by using a MLP to combine multiple data sources. In this case, we mixed information from each data source in the middle layers of the MLP, where the output from each individual stack of layers for each data source was concatenated in the last layers of the MLP. The three branches operated independently from each other until they were concatenated. In this way, we designed a network with three inputs and one output. The average results for the 50 partitions of the dataset are reported in Table III.

According to the Tukey's HSDT performed for the single-step multiple comparisons between the fusion methods, the averages of the AUC and precision were not significantly different across the fusion methods. For the remaining performance metrics (accuracy, F1-score, recall and specificity), only the difference between the averages of the early and late fusion approaches was not significantly different.

V. DISCUSSION AND CONCLUSION

We tested three approaches for multisource data fusion, namely, early, late, and slow fusion, using the procedure illustrated in Fig. 1. We investigated the impact of fusing data at different stages of the pipeline on the obtained results. In this study looking at predicting falls in elderly, similar results were found for the majority of the considered performance metrics. Nevertheless, it should be noted that the late and slow fusion approaches can provide a set of advantages regarding the deployment of a prediction system. For example, a system capable of dealing with fewer sources of information can be designed and trained when a certain data source is not available. Moreover, we found that recall was more important than specificity, for the predictive system considered in this study, since the fall risk screening was used to select elderly with a higher risk of fall that should be considered for fall prevention. It would be preferable to minimize the error of losing a potential faller (i.e.,

maximizing recall) instead of losing a potential non-faller (i.e., maximizing specificity). This rationale was used to select the best models among all possible combinations in the optimization pipeline.

We compared each data source, regarding their predictive value individually. The inertial and platform data alone revealed a higher F1-score and precision compared with those of the personal data. Furthermore, the inertial data alone allowed to achieve a higher AUC and recall compared with those obtained when considering only the personal data. These results reinforce the added value of sensor instrumentation in fall risk screening protocols.

The average results obtained for the early and late fusion approaches were not statistically different from each other, which may indicate that the different data sources were highly correlated. The early fusion approach can be preferred given its lower computational requirements. The slow fusion approach obtained a higher accuracy score but a lower F1-score. The standard deviation of all scores achieved by this fusion approach was lower compared with those of the other approaches because the pipeline for slow fusion was only optimized for one loss function. By optimizing for cross entropy, the model with slow fusion retrieved a higher specificity and lower recall compared with the early and late fusion approaches. The slow fusion approach can also be useful in scenarios where specificity is more important than recall.

To the best of our knowledge, no previously published work has attempted to study different approaches to data fusion using multiple sources of data for prospective fall prediction. However, we found one previous work that reported a late fusion approach with clinical and inertial data for retrospective fall prediction validated using nested CV [20]. The authors reported a significant added value of data fusion compared with analyzing individual data sources. The majority of previous studies report the use of a combination of personal (clinical and self-reported) data and one source of sensor data (either inertial sensors, pressure platform, or other types of sensor-based data) in an early fusion approach.

Furthermore, the classification and validation pipeline used in this study covers different stages of optimization.

We reported the results for a test set that was not used during the training of the proposed grid search pipeline. The lack of an external test set, which is considered to be essential for the evaluation of trained models to avoid overfitting, has been considered as one of the main disadvantages of previous studies [2]. Moreover, few studies have used neural networks for the prediction of falls or employed slow fusion approaches, which are more common for video classification [18].

Providing the results of this study, as our future work we consider studying different methods for feature processing and training different types of classifiers that are more suitable for each data source. Furthermore, it is possible that the nature of falls is not completely covered by the screening protocol used in this study. For example, once an elderly person with poor functional capabilities and clinical history of associated fall risk factors is institutionalized, the fall probability is reduced due to the resulting movement restriction. Adding strategies for data pre-processing and variables that better describe the unexpected nature of fall occurrences should be considered.

Acknowledgment

Authors would like to thank all recruited participants, physiotherapists, and partners of FallSensing project. Authors would also like to thank A. F. Sequeira, R. Marques, and J. Marques for the English proofreading.

References

[1] A. F. Ambrose, G. Paul, and J. M. Hausdorff, "Risk factors for falls among older adults: A review of the literature," *Maturitas*, vol. 75, no. 1, pp. 51–61, 2013.

[2] J. Howcroft, J. Kofman, and E. D. Lemaire, "Review of fall risk assessment in geriatric populations using inertial sensors," *J. Neuro Eng. Rehabil.*, vol. 10, Aug. 2013, Art. no. 91.

[3] A. Nelson, K. L. Perell, L. Z. Rubenstein, N. Prieto-Lewis, R. L. Goldman, and S. L. Luther, "Fall risk assessment measures: An analytic review," *J. Gerontol., Series A*, vol. 56, pp. M761–M766, 2001.

[4] M. Vassallo, L. Poynter, J. C. Sharma, J. Kwan, and S. C. Allen, "Fall risk-assessment tools compared with clinical judgment: An evaluation in a rehabilitation ward," *Age Ageing*, vol. 37, pp. 277–281, 2008.

[5] K. E. Bigelow and N. Berme, "Development of a protocol for improving the clinical utility of posturography as a fall-risk screening tool," *J. Gerontol. Series A, Biol. Sci. Med. Sci.*, vol. 66, no. 2, pp. 228–33, 2011.

[6] H. Qiu, R. Z. U. Rehman, X. J. Yu, and S. Xiong, "Application of wearable inertial sensors and a new test battery for distinguishing retrospective fallers from non-fallers among community-dwelling older people," *Scientific Rep.*, vol. 8, 2018, Art. no. 16349.

[7] B. R. Greene *et al.*, "Quantitative falls risk estimation through multi-sensor assessment of standing balance," *Physiol. Meas.*, vol. 33, no. 12, pp. 2049–2063, 2012.

[8] Y. Liu *et al.*, "Validation of an accelerometer-based fall prediction model," in *Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2014, pp. 4531–4534.

[9] K. S. van Schooten, M. Pijnappels, S. M. Rispens, P. J. M. Elders, P. T. A. M. Lips, and J. H. van Dieën, "Ambulatory fall-risk assessment: Amount and quality of daily-life gait predict falls in older adults," *J. Gerontol. Series A, Biol. Sci. Med. Sci.*, vol. 70, no. 5, pp. 608–15, 2015.

[10] A. N. Aicha, G. Englebienne, K. S. van Schooten, M. Pijnappels, and B. J. A. Kröse, "Deep learning to predict falls in older adults based on daily-life trunk accelerometry," *Sensors*, vol. 18, 2018, Art. no. 1654.

[11] K. S. van Schooten *et al.*, "Daily-life gait quality as predictor of falls in older people: A 1-year prospective cohort study," *PLoS ONE*, vol. 11, 2016, Art. no. e0158623.

[12] A. C. Martins *et al.*, "Multifactorial screening tool for determining fall risk in community-dwelling adults aged 50 years or over (fallsensing): Protocol for a prospective study," *JMIR Res. Protocols*, vol. 7, Aug. 2018, Art. no. e10304.

[13] J. Silva *et al.*, "Comparing machine learning approaches for fall risk assessment," in *Proc. 10th Int. Joint Conf. Biomed. Eng. Syst. Technol.*, 2017, pp. 223–230.

[14] Fraunhofer Portugal AICOS, "A day with pandlets," White Paper, 2016.

[15] "PhysioSensing balance and pressure plate," 2019. [Online]. Available: https://www.physiosensing.net/

[16] J. Silva and I. Sousa, "Instrumented timed up and go: Fall risk assessment based on inertial wearable sensors," in *Proc. IEEE Int. Symp. Med. Meas. Appl.*, 2016, pp. 1–6.

[17] B. Aguiar, J. Silva, T. Rocha, S. Carneiro, and I. Sousa, "Monitoring physical activity and energy expenditure with smartphones," in *Proc. IEEE-EMBS Int. Conf. Biomed. Health Inform.*, Jun. 2014, pp. 664–667.

[18] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 1725–1732.

[19] N. Chawla, K. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[20] B. Greene, K. Mcmanus, and B. Caulfield, "Automatic fusion of inertial sensors and clinical risk factors for accurate fall risk assessment during balance assessment," in *Proc. IEEE Conf. Biomed. Health Inform.*, 2018.