Weightlifting Exercise Tracker and Classifier

Sohan Arun
dept. Computer Science
Blekinge Institute of Technology
Karlskrona, Sweden
Sohanoffice46@gmail.com

***Abstract*— Strength training, alongside aerobic exercises, is a vital component of a well-rounded fitness regimen. However, the automated tracking of free weight exercises remains an area that requires further exploration. This project investigates the potential of context-aware applications within the strength training domain by analysing accelerometer and gyroscope data collected from wristbands during strength training sessions. The dataset includes recordings from 5 participants performing different barbell exercises. The objective is to develop and evaluate models capable of performing tasks akin to those of human personal trainers, such as tracking exercises, counting repetitions, and identifying improper form. The study adopts a supervised learning approach, training various machine learning algorithms on the collected data. Model performance is assessed and compared to identify the most effective solutions.**

I. Introduction

Over the last decade, constraints associated with wearable sensors like accelerometers, gyroscopes, and GPS receivers have largely been addressed, enabling human activity monitoring and classification using devices such as smartwatches to emerge as a key area in pattern recognition and machine learning. This growth is driven by the commercial potential of context-aware applications and their ability to tackle societal challenges in areas like rehabilitation, sustainability, elderly care, and health [1]. Past efforts have focused on tracking movements and providing feedback through exercise management systems, partly replacing personal trainers. For aerobic exercises like running, swimming, and cycling, devices such as GPS-based pedometers, ECG monitors, and electronic machines have become standard tools for tracking performance [1]. However, free weight exercises remain underexplored, with only one fitness wearable claiming to automatically track exercises and repetitions [4].

Advancements in context-aware applications could eventually enable fully digital personal trainers. These digital trainers must replicate the core responsibilities of human trainers: understanding anatomy and exercise science, designing tailored exercise programs, and tracking workouts to ensure safety and progress [5]. While the first two aspects have seen progress, the third—tracking form and progressive overload—is still inadequately addressed in current wearables.

This projectS explores the potential for context-aware applications in strength training by analysing accelerometer and gyroscope data collected from wristbands during barbell exercises. Data from five participants performing medium- to heavy-weight exercises were used to develop and evaluate supervised machine learning models capable of tracking exercises, counting repetitions, and detecting improper form. Methods build on prior work by Hoogendoorn and Funkuse [2], applying various algorithms to identify the most effective approach.

The paper is structured as follows: Section 2 reviews related work, Section 3 details experiment setup, Section 4 explains methodology, Section 5 covers results and analysis, and Section 7 concludes.

II. Related Work

The first study on activity recognition using wearable sensors was conducted in 2000 by Van Laerhoven [7], where accelerometers attached to pants collected data for activity recognition using Kohonen maps and probabilistic models. Recognizing daily activities remains a widely researched area [8-11] and is now implemented in commercial products like Fitbit [13], Apple Watch [15], and Samsung Gear [14]. Smartphones also track fitness activities such as walking, running, and cycling [16-17], and newer devices recognize sports-specific activities like aerobic workouts and swimming. For instance, Fitbit's SmartTrack records stats like activity duration and calories burned [18]. However, free weight exercises remain underrepresented.

Some studies have explored gym exercise recognition using wearable sensors [19-21]. Chang et al. used accelerometers on gloves and waistbands to classify nine exercises with 90% accuracy for single-user data and 85% for cross-user validation, counting repetitions with a 5% error rate [19]. Koskimäki et al. addressed limited exercise variety by using two accelerometers on a single subject performing 30 exercises, achieving a 96% true positive rate, though accuracy dropped for unseen data [21]. Li et al. employed Dynamic Time Warping on time-series data from a glove accelerometer, yielding promising results across different subjects [19].

These studies demonstrate the feasibility of exercise recognition but neglect key strength training aspects like progressive overload and form. Progressive overload, crucial for strength training success, was oversimplified as subjects performed the same high-repetition sets regardless of weight

[23-24]. Realistic workout scenarios, perceived intensity, and execution quality, critical for preventing injuries with heavy weights, were largely overlooked, limiting the applicability of these approaches for free weight training.

III. Experimental Setup

Previous studies have demonstrated the feasibility of using machine learning algorithms on free weight exercise accelerometer data with promising results. However, these works often overlook the importance of collecting high-quality datasets. This study addresses this by simulating real strength training sessions and focusing on exercises from a specific program, Starting Strength by Mark Rippetoe [24]. Unlike prior works that randomly selected exercises without justification, this approach reduces noise by limiting exercises to those commonly performed together. Starting Strength includes five core barbell exercises: Bench Press, Deadlift, Overhead Press, Row, and Squat [24]. The exercises are shown in Figure 1.

These exercises target all major muscles with a full range of motion and progressive loading for strength gains. In this program, exercises are performed with heavier weights, typically allowing only about five repetitions per set [24]. This higher load alters execution, affecting bar path and speed compared to lighter weights used in other studies. The study examines whether models can still classify exercises accurately under these conditions and explores methods for detecting improper form.

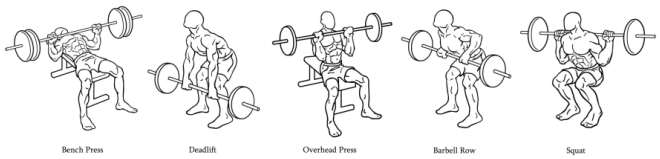

Fig. 1. Basic Barbell Exercises

A. Data Collection

While previous studies primarily focused on accelerometer data, modern smart devices include additional sensors like gyroscopes, providing richer data. Many earlier experiments placed sensors in impractical locations, such as torso bands or workout gloves. To develop a practical commercial product, a smartwatch-like setup is ideal, as it integrates necessary sensors. For this experiment, MbientLab's wristband sensor research kit [12] was used, mimicking smartwatch placement for controlled testing. Data was collected at accelerometer settings of 12.500Hz and gyroscope settings of 25.000Hz. Five participants performed barbell exercises in 3 sets of 5 repetitions, as in the Starting Strength program, and later in 3 sets of 10 to test model generalization across different weights. This resulted in 150 (5x5x6) sets of data, including resting periods where participants stood, walked, or sat, providing state-change data from rest to exercise.

B. Weights

The one rep max (1RM) metric, defined as the maximum weight a person can lift for one repetition [27], was used to determine appropriate weights. 1RM can be calculated directly through maximal testing or indirectly via submaximal estimation, with the latter being safer and faster. Common submaximal formulas include Epley and Brzycki [26]. This experiment used Epley's formula:

$$1RM = w(1 + \frac{r}{30})$$

Here, r represents repetitions performed, and w is the weight used. After calculating 1RM, Epley's formula was also applied to determine weights for 5 or 10 reps, approximating 85% (5 reps) and 75% (10 reps) of the 1RM. This approach ensured participants lifted weights proportionate to their strength.

IV. Methodology

A. Converting Raw Data

The raw dataset comprised 69,677 entries, each containing an epoch timestamp and x, y, and z-values from the wristband sensors. Sensor data was split into separate files, each with unique timestamps, requiring aggregation. A step size of Δt = 0.20 s (five instances per second) was chosen to minimize information loss, using the mean for numerical values and the mode for categorical labels. The aggregated dataset served as a foundation for visualizations.

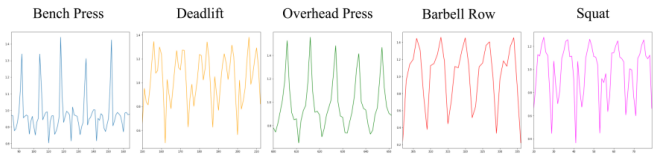

Fig. 2. Accelerometer Data from Exercise

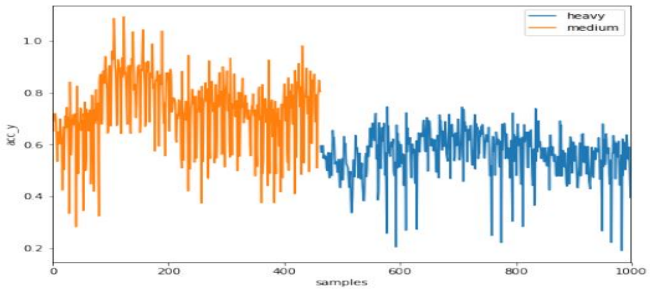

Fig. 3. Medium and Heavy Weight Squats

Figure 2 illustrates unique exercise patterns where repetitions are identifiable by peaks, while Figure 3 highlights differences in y-acceleration between medium and heavy squat sets. Medium sets show higher peaks due to lower resistance,

and heavy sets exhibit deeper drops due to heavier loads. To handle noisy data, this project explores two methods: Low-pass Filtering for individual attributes and Principal Component Analysis (PCA) to identify variance across the dataset.

B. Low-pass Filter

A low-pass filter, such as the Butterworth filter, is suitable for temporal data with periodicity and can reduce high-frequency noise that may hinder learning [2]. In this study, it was applied to all features except the target. Visual inspection revealed a movement frequency of ~2 seconds per repetition, and following trial-and-error as outlined by van den Bogert [3], the cut-off was set at 1.3 Hz. Figure 4 illustrates the y-acceleration for a heavy bench press set before and after smoothing.

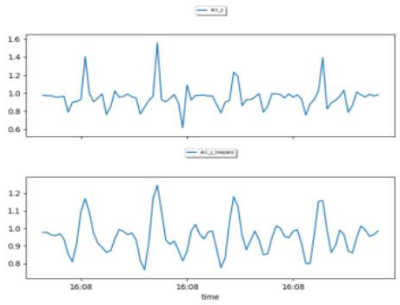

Fig. 4. Low-pass Filter

C. Principal Component Analysis

A principal component analysis (PCA) was conducted to find the features that could explain most of the variance. PCA was applied to all features excluding the target columns. The results are visualized in Figure 5 which shows that the explained variance drastically decreases after 3 components. Therefore, 3 components are selected, and their values are included into the dataset.

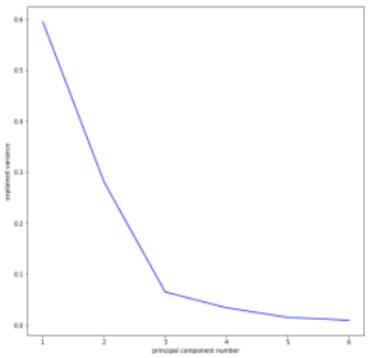

Fig. 5. Principal Component Number

D. Feature Engineering

a) Aggregated Features: To further utilize the data, the scalar magnitude r of the accelerometer and gyroscope was calculated r, derived from the combined x,y, and z data points, offers the advantage of being orientation-independent and capable of managing dynamic re-orientations [29].

b) Time Domain: To leverage the temporal nature of the data, numerical points were aggregated by calculating the standard deviation (sd) and mean of all features, excluding target columns, across varying window sizes. The sd captures data variation over time, expected to be higher during exercises than resting periods. The mean provides general data levels with reduced noise influence. Choosing a window size involves balancing noise effects and predictive power. Results for window sizes of 2, 4, and 6 seconds are shown in Figure 6, with 4 seconds selected for the dataset.

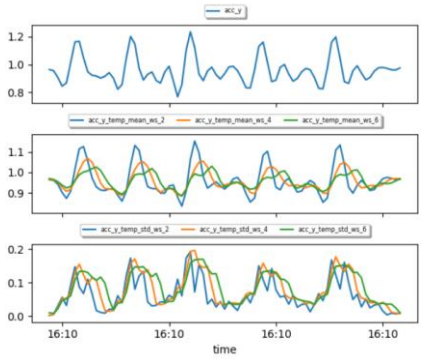

Fig. 6. Numerical temporal aggregation with window sizes of 2, 4, and 6 seconds

c) Frequency Domain: Fourier Transformation: In addition to the time domain, the frequency domain was analyzed using Fourier Transformation, which represents a sequence of measurements as a combination of sinusoidal functions with varying frequencies [2]. A 4-second window was applied to compute frequency features, including maximum frequency, and frequency signal weighted average [2].

d) New dataset: The dataset now includes additional features, but overlapping time windows result in highly correlated attributes. To reduce redundancy and prevent overfitting, a maximum overlap of 50% was set. Instances exceeding this overlap were removed, reducing the dataset to 4505 instances. Although some information was lost, this approach minimizes similar instances that could lead to overfitting [2].

Clustering: Clustering membership can aid in label prediction. The focus was on clustering acceleration data since gyroscope data proved unhelpful. Among tested methods, k-means clustering (k=4) achieved the best silhouette score (0.6478) compared to k-medoids and agglomerative clustering. Although k=2 had a slightly higher score, k=4 was chosen for better label differentiation. Results (Figure 7) and the distribution of measurements and labels (Table 1) show that cluster 1 includes most bench press and overhead press data(This makes sense as both pressing movements are very

similar), cluster 2 captures squats, cluster 3 perfectly captures deadlifts and rows, and cluster 4 partially captures rest data but lacks accuracy.

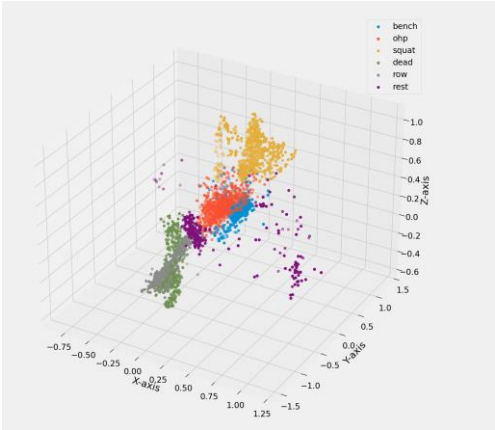

Fig. 7. Clusters

TABLE I. CLUSTER COVERAGE

Label	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Bench Press	99.88 %	0.12 %	0.00 %	0.00 %
Deadlift	0.00 %	0.00 %	100.00 %	0.00 %
OHP	99.28 %	0.72 %	0.00 %	0.00 %
Row	0.00 %	0.00 %	100.00 %	0.00 %
Squat	2.98 %	97.02 %	0.00 %	0.00 %
Rest	4.14 %	3.78 %	50.45 %	41.62 %

E. Modeling

The processed dataset is ready for training, containing 6 basic features, 2 scalar magnitude features, 3 PCA features, 16 time features, 12 frequency features, and 1 cluster feature. This section outlines the development and evaluation of models for classification, and repetition counting.

a) Classification: Due to the temporal nature of the dataset, the training and test sets were divided by exercise sets. The first two sets for each exercise, weight, and participant combination were used for training, while the remaining sets were reserved for testing. This approach ensures the test data consists of unseen sets for model evaluation.

b) Feature selection: Forward feature selection was applied to identify the most impactful features, as irrelevant ones could degrade algorithm performance. Using a decision tree and incrementally adding the best features, it was observed that performance plateaued after 15 features. The top 5 predictive features identified were: pca 1, acc y, pca 3, gyr x temp std ws 4, and acc r pse.

c) Regularization: To penalize complex models, a regularizer was added to the objective functions. Figure 8 illustrates how increasing the regularization parameter initially improves test set accuracy but only up to a certain point, after which accuracy declines for both the training and test sets.

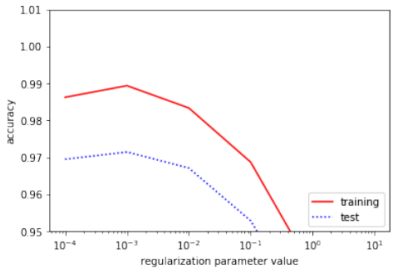

Fig. 8. Regularization

d) Models: An initial test run was conducted to evaluate the performance of selected models and features. The models tested included Random Forest, Support Vector Machine, K-Nearest Neighbors, Decision Tree, and Naive Bayes. Grid search was applied to optimize all models.

e) Counting Repetitions: To count repetitions, a peak counting algorithm was applied to scalar magnitude acceleration data. A low-pass filter with a 0.4 Hz cut-off was used to exclude small local peaks. The method required exercise-specific adjustments for optimal performance. For deadlifts and overhead presses, counting minima produced better results. The overall error rate for repetition counting was approximately 5% for the dataset. An example of 10 deadlift repetitions is shown in Figure 9.

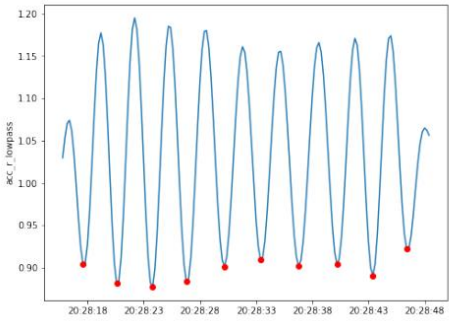

Fig. 9. Counting deadlift repetitions using the minimum values after applying a lowpass filter

f) Evalution Metrics: To evaluate model performance, metrics such as accuracy, precision, recall, and F1-Score were used. Accuracy was used as a primary metric to measure the ratio of correctly predicted instances to the total instances, providing an overall view of model performance. However, its reliability can be limited in cases of class imbalance. Precision and recall were key complementary metrics; precision assessed the reliability of positive classifications, ensuring the model correctly identified and accessed the relevant exercise, while recall measured the ability to capture all such cases, which is especially critical when false negatives have significant consequences. The F1-Score balanced precision and recall, making it particularly valuable for imbalanced datasets.

Additionally, a confusion matrix offered detailed insights into classification performance, highlighting misclassification patterns and providing a deeper understanding of model behavior.

V. Results and Analysis

The analysis evaluates the performance of multiple machine learning models (Random Forest, K-Nearest Neighbor, Decision Tree, and Naive Bayes) across different feature sets to determine the most effective combination for accurate predictions as shown in figure 10 along with their accuracy resukts in table 2.

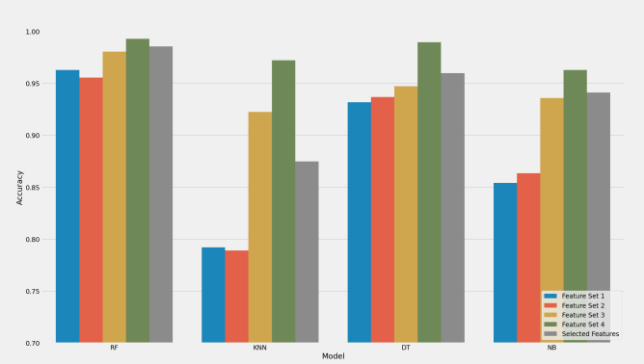

Fig. 10. Model Performances

TABLE II. Accuracy Results

Model	Feature_set	Accuracy
RF	Feature Set 4	0.992761
DT	Feature Set 4	0.989659
RF	Selected Features	0.985522
RF	Feature Set 3	0.980352
KNN	Feature Set 4	0.972079

a) Model Performance: Random Forest (RF) achieved the highest accuracy (99.27%) on Feature Set 4, followed by Decision Tree (DT) with 98.96% accuracy. Selected Features yielded 98.55% accuracy with RF, showcasing its competitiveness. K-Nearest Neighbor (KNN) showed an accuracy of 97.20% on Feature Set 4, indicating its viability for this feature set.

b) Best Model Selection: RF consistently outperformed other classifiers on most feature sets, particularly excelling with Feature Set 4 (accuracy: 99.27%, precision: 99.59%, recall: 99.59%, F1-Score: 99.59%).

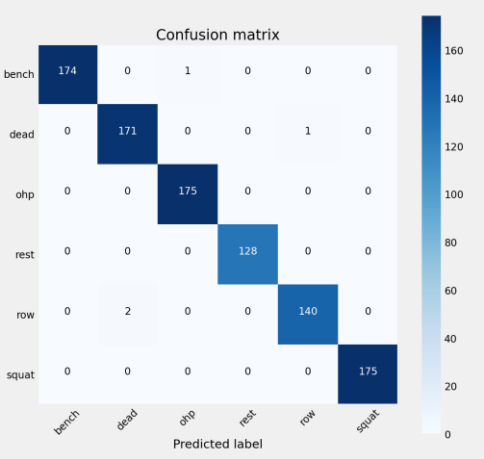

Fig 11. RF Classification Confusion Matrix

c) Participant A Evaluation: When training excluded data from Participant A, the RF model achieved an accuracy of 98.99% on the test data for Participant A. Precision (99.03%), recall(98.99%), and F1-Score(99.00%), indicating consistent performance across metrics. These results demonstrate the model's ability to generalize effectively, as it was trained on data from other participants but still achieved high accuracy and consistency on unseen data from Participant A. This underscores the model's robustness in capturing patterns and its potential to perform well on unseen participants.

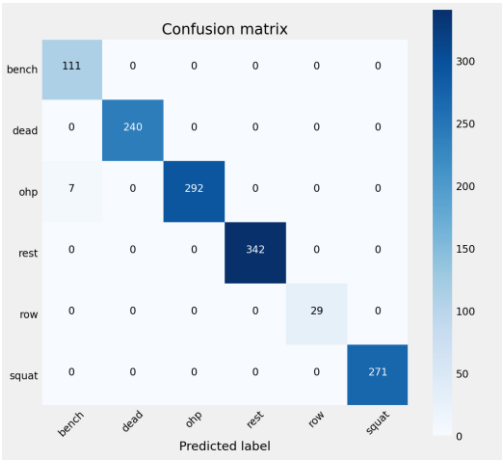

Fig 12. Participant A - RF Confusion Matrix

d) Feature Set Insights: Feature Set 4 demonstrated superior predictive power, producing the highest scores across multiple models, particularly with RF. Selected Features also delivered high accuracy, emphasizing the value of feature engineering.

VI. Conclusion

This project explored the integration of machine learning with wearable sensor data to address key aspects of strength training, including exercise classification, and repetition counting. By leveraging advanced feature engineering techniques and evaluating multiple models, Random Forest consistently demonstrated superior accuracy and robustness. The findings underline the feasibility of creating intelligent, wrist-worn devices capable of replicating some functions of personal trainers. However, challenges remain, such as the need for more diverse datasets and improved handling of subtle movement variations. With further refinement, this approach has the potential to improve fitness tracking, offering tailored, real-time feedback and promoting safer, more effective strength training practices.

References

[1] The Presidents Council on Physical Fitness and Sports. Fitness fundamentals guidelines for personal exercise programs. online council publications (2003), https://www.hhs.gov/fitness/index.html.

[2] Hoogendoorn, M., & Funk, B. Machine Learning for the Quantified Self.

[3] Van Den Bogert, M. (1996). Practical Guide to Data Smoothing and Filtering.

[4] Atlas Wristband 2, https://atlaswearables.com/

[5] American College of Sports Medicine. (2013). ACSM's Resources for the Personal Trainer. Lippincott Williams Wilkins.

[6] Read, R. Kasparian, M. Li, P. 2015, USD725512S1, https://patents.google.com/patent/USD725512S1

[7] Van Laerhoven, K., & Cakmakci, O. (2000, October). What shall we teach our pants?. In Wearable Computers, The Fourth International Symposium on (pp. 77- 83). IEEE.

[8] Ermes, M., Prkk, J., Mntyjrvi, J., & Korhonen, I. (2008). Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions. IEEE transactions on information technology in biomedicine, 12(1), 20-26.

[9] Banos, O., Damas, M., Pomares, H., Prieto, A., & Rojas, I. (2012). Daily living activity recognition based on statistical feature quality group selection. Expert Systems with Applications, 39(9), 8013-8021.

[10] Zhang, M., & Sawchuk, A. A. (2013). Human daily activity recognition with sparse representation using wearable sensors. IEEE journal of Biomedical and Health Informatics, 17(3), 553-560.

[11] Leutheuser, H., Schuldhaus, D., & Eskofier, B. M. (2013). Hierarchical, multisensor based classification of daily life activities: comparison with state-of-the-art algorithms using a benchmark dataset. PloS one, 8(10), e75196.

[12] MbientLab, https://mbientlab.com/

[13] Fitbit, https://www.fitbit.com/

[14] Samsung Gear, https://www.samsung.com/us/mobile/wearables/

[15] Apple Watch, https://www.apple.com/lae/apple-watch-series-4/

[16] Location API, https://developers.google.com/location-context/"

[17] MotionActivity, https://developer.apple.com/documentation/coremotion/cmmotionactivity"

[18] Fitbit Smarttrack, "https://www.fitbit.com/nl/smarttrack"

[19] Li, C., Fei, M., Hu, H., & Qi, Z. (2012, September). Free weight exercises recognition based on dynamic time warping of acceleration data. In International Conference on Intelligent Computing for Sustainable Energy and Environment (pp. 178-185). Springer, Berlin, Heidelberg.

[20] Chang, K. H., Chen, M. Y., & Canny, J. (2007, September). Tracking freeweight exercises. In International Conference on Ubiquitous Computing (pp. 19-37). Springer, Berlin, Heidelberg.

[21] Koskimki, H., & Siirtola, P. (2014, December). Recognizing gym exercises using acceleration data from wearable sensors. In CIDM (pp. 321-328).

[22] Ward, J. A., Lukowicz, P., & Gellersen, H. W. (2011). Performance metrics for activity recognition. ACM Transactions on Intelligent Systems and Technology (TIST), 2(1), 6.

[23] Kraemer, W. J., Ratamess, N. A., & French, D. N. (2002). Resistance training for health and performance. Current sports medicine reports, 1(3), 165-171.

[24] Rippetoe, M., & Kilgore, L. (2007). Starting strength: Basic barbell training. Wichita Falls, Texas, USA: Aasgaard Company.

[25] One rep max calculator, https://strengthlevel.com/one-rep-max-calculator

[26] Wood, T. M., Maddalozzo, G. F., & Harter, R. A. (2002). Accuracy of seven equations for predicting 1-RM performance of apparently healthy, sedentary older adults. Measurement in physical education and exercise science, 6(2), 67-94.

[27] Marchese, R., & Hill, A. (2011). The essential guide to fitness: for the fitness instructor. Pearson Australia.

[28] Bench Press Mistakes, https://barbend.com/common-bench-press-mistakes/

[29] Arfken, G., & Weber, H. J. Mathematical Methods for Physicists (Academic, San Diego, 1985). Google Scholar, 232-236.