Personalised Fitness Training via Workout Difficulty Classification from Biometric and Performance Features

John Joyal, Ahmed O. Basil Al-Mashhadani
University of Northampton
Northampton United Kingdom
joyaljohn3500@gmail.com
ahmed.basil2@northampton.ac.uk

Mohammed Al-Khafajiy
University of Doha for Science and Technology
Doha, Qatar
Mohammed.Alkhafajiy@udst.edu.qa

***Abstract*—This paper presents an AI-enabled fitness application framework that supports personalised training by predicting an individual's work-out difficulty level from biometric and performance measurements. Using a structured dataset containing demographic attributes (e.g., age and gender), anthro-pometrics (height and weight), cardiovascular indica-tors (systolic/diastolic blood pressure), and functional fitness metrics (e.g., grip force, sit-ups, flexibility, and broad jump), we perform data cleaning to remove anomalies and reduce noise, followed by exploratory feature analysis using distribution plots and correlation screening. To ensure an interpretable prediction target, the original class label is operationalised into a four-level workout difficulty recommendation through a rule-based procedure informed by key indicators such as BMI, body fat percentage, and performance tests. We then benchmark multiple supervised learning models under a consistent validation protocol, reporting not only accuracy but also deployment-relevant measures including prediction speed, training time, and compact model size. Experimental results show that a medium feed-forward neural network achieves the highest validation accuracy of 96.4%, with very high discriminative capability reflected by ROC/AUC performance, while also delivering substantially faster inference and a smaller model footprint than the best-performing SVM baseline (medium Gaussian SVM at 81.8%). The findings indicate that neural models provide a strong balance of accuracy and real-time suitability for personalised fitness recommendations.**

I. Introduction

Artificial intelligence (AI) is increasingly embedded in health-facing software systems, where data-driven models support decision-making, monitoring, and personalised recommendations. Recent ad-vances in machine learning have demonstrated that predictive models can deliver robust classification performance when trained on sufficiently representative data and evaluated with rigorous validation protocols [1]–[3]. In parallel, fitness applications have evolved from simple activity tracking tools into interactive platforms that aim to optimise training outcomes through individualised plans, feedback loops, and behavioural support.

Despite this progress, *personalisation* in fitness applications remains challenging in practice. First, fitness-relevant datasets often contain heterogeneous user characteristics (age, sex, anthropometrics, and performance metrics) that require careful feature engineering and modelling choices. Second, real-world fitness data can be imbalanced across demographics or fitness-level classes, which may bias standard learners toward majority groups and reduce reliability for underrepresented users. In the broader health AI landscape, privacy, safety, and trust are also recurring concerns when systems process sensitive biometric data [4].

This paper develops and evaluates an AI-enabled fitness application framework for personalised training recommendations using a supervised learning pipeline trained on user biometrics and performance measures.

TABLE I. Available features within the dataset

Feature	Significance
age	Impacts exercise requirements, metabolic rate, and health considerations across age groups.
gender	Influences body composition, preferences, and health risks.
height_cm	Affects biomechanics, joint stress, and equipment adjustments.
weight_kg	Enables BMI estimation and weight-related risk assessment.
body_fat_%	Adds body composition context beyond weight.
systolic / dias-tolic	Indicates cardiovascular status and exercise safety limits.
gripForce	Proxy for overall strength and functional capacity.
sit & bend_forward	Flexibility measure linked to injury preven-tion.
sit_ups_counts	Core strength and endurance indicator.
broad_jump_cm	Lower-body power/explosiveness metric.
class	Fitness-level label used for supervised learning.

The dataset used in this study contains multiple physiological and functional indicators (summarised in Table I), supporting a modelling approach that maps user inputs to a practical *workout difficulty* recommendation. To address class imbalance and improve minority-class performance, we employ an undersampling-boosting strategy based on RUSBoost [5], and benchmark it against commonly used alternatives, including SVMs, neural networks, and decision-tree families. The key contributions of this work are:

- An AI-driven fitness application framework that operationalises real-time personalisation by mapping user biometrics and performance metrics to workout difficulty recommendations.
- A supervised learning evaluation across multiple model families (SVM, neural networks, decision trees, and ensemble methods) to identify a high-performing and practical classifier for deployment.
- An imbalance-aware modelling approach, leveraging RUSBoost [5], motivated by the observed demographic skew (see Fig. 1) and aimed at improving prediction reliability across user groups.

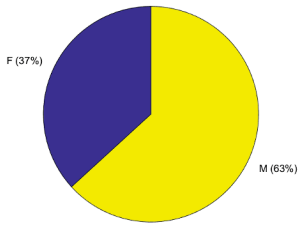

Fig. 1. Gender Distribution (M/F)

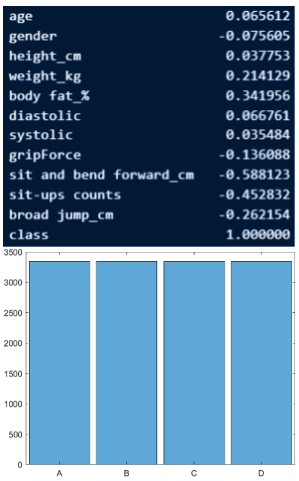

Fig. 2. correlation of Class in comparison with other features in the dataset, and Class distribution

II. Preliminaries and Methodology

A. Problem Formulation

Let $D = \{(x_i, y_i)\}_{i=1}^{N}$ denote a dataset of N participants, where $x_i \in R^d$ is a feature vector containing biometric and performance indicators, and $y_i \in \{1, 2, 3, 4\}$ is the target label represent-ing the recommended *workout difficulty level*. The objective is to learn a classifier $f_\theta(x)$ that maps user inputs to an appropriate difficulty level to support personalised workout recommendations.

B. Dataset Overview

The study uses a structured dataset comprising physiological and functional measurements (Table I), including demographic attributes (e.g., age, gender), anthropometrics (height and weight), cardiovascular indicators (systolic/diastolic blood pressure), and performance measures (e.g., grip force, sit-up counts, broad jump distance). Exploratory analysis highlights non-uniform distributions across several variables (Fig. 11) and a noticeable gender imbalance (Fig. 1), motivating both robust prepro-cessing and imbalance-aware learning.

C. Target Variable Engineering: Workout Difficulty

The original dataset label (*Class*) was found to be institution-defined with unclear grading criteria. To ensure an interpretable and reproducible learning target, we replace it with an engineered label, *Workout Difficulty*, derived from a subset of fitness-relevant indicators: age, body fat percentage, sit-up counts, broad jump distance, and body mass index (BMI). The BMI is computed as:

$$\mathrm{BMI} = \frac{w_{\mathrm{kg}}}{(h_{\mathrm{m}})^2}. \tag{1}$$

A rule-based scoring procedure is then applied to discretise participants into four ordinal difficulty levels. The rule design is guided by established evidence that functional decline may become observable from midlife onward [6], body-fat cut-offs provide meaningful differentiation of adiposity strata [7], and abdominal endurance tests can serve as practical indicators of core capacity [8]. In addition, a minimum sit-up threshold is used during label construction to filter rows likely associated with measurement anomalies or non-performance (see Section II-D).

D. Data Cleaning and Preprocessing

Before training, the dataset is cleaned to mitigate outliers, inconsistencies, and data-entry artefacts. First, extreme anthropometric anomalies are identified via a height–weight scatter analysis (Fig. 13) and removed to

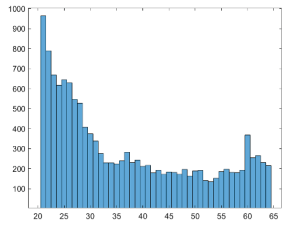

Fig. 3. * Age Distribution

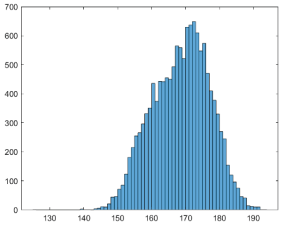

Fig. 4. * Height Distribution (cm)

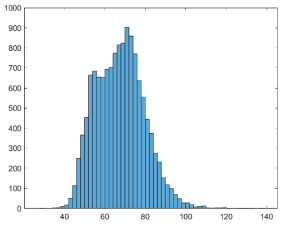

Fig. 5. * Weight Distribution (kg)

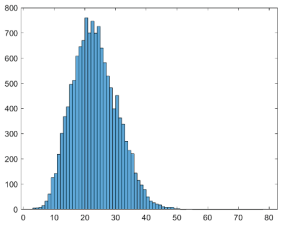

Fig. 6. * Body Fat Percentage Distribution (%)

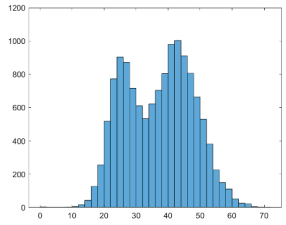

Fig. 7. * Grip Force Distribution

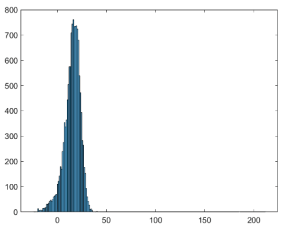

Fig. 8. * Sit and Bend Forward

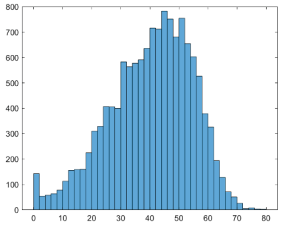

Fig. 9. * Sit-ups Count

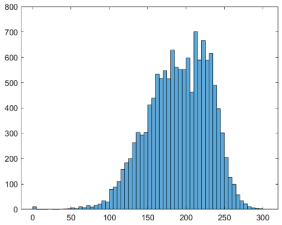

Fig. 10. * Broad Jump (cm)

Fig. 11. Comparison of the features

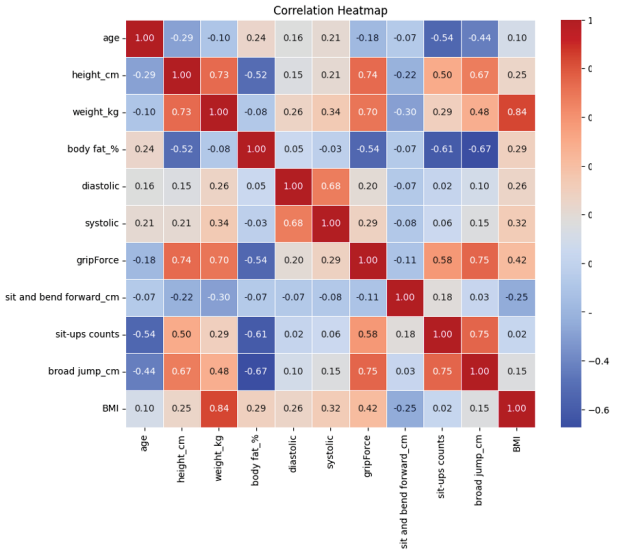

Fig. 12. Correlation Heatmap of Features

prevent distribution skew and unstable decision boundaries. Second, categorical predictors (e.g., gender) are treated as nominal variables to avoid imposing artificial order. Third, observations failing minimum quality criteria (e.g., implausible performance readings such as very low sit-up counts) are filtered to reduce label noise. Finally, numeric predictors are standardised where required by the learner (e.g., distance-based or margin-based models).

E. Feature Analysis and Correlation Screening

To reduce redundancy and improve generalisation, we analyse pairwise feature correlations (Fig. 12). This step highlights strongly coupled predictors (e.g., height–weight) and identifies features with limited predictive contribution. Where appropriate, correlation screening informs feature inclusion to mitigate collinearity and improve model stability.

F. Learning Protocol and Evaluation Metrics

We follow a standard supervised learning pro-tocol using stratified data partitioning and cross-validation to estimate generalisation performance. Specifically, the dataset is split into training and held-out sets, and k-fold

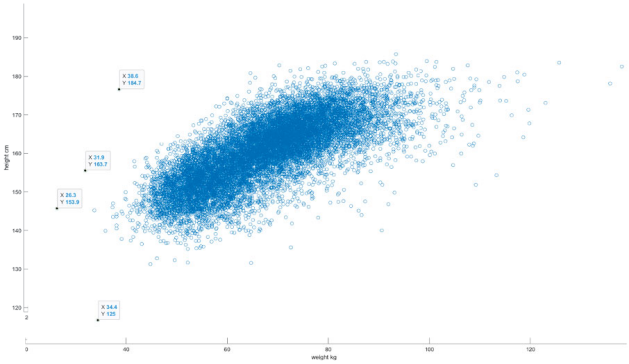

Fig. 13. Scatter plot comparing height and weight, with potential anomalies displayed

cross-validation is used during model selection and hyperparameter tun-ing [9]. Performance is measured using classification accuracy as the primary metric, complemented by confusion matrices and ROC/AUC analysis to assess discriminative behaviour across operating points [10]. Given demographic skew (Fig. 1), we additionally prioritise techniques that are robust to class imbalance.

G. Imbalance-Aware Learning

To address imbalance during training, we employ RUSBoost, which combines random undersampling with boosting to improve minority-class recognition while maintaining efficient training dynamics [5]. Baseline model families considered in this work include support vector machines (SVMs) [11], neural networks, and tree-based learners; comparative results are presented in the subsequent section.

III. Experimental Setup and Model Benchmarking

A. Toolchain and Dataset Instance

All experiments were executed using the cleaned dataset (bpcleanedfull.mat) loaded into the MATLAB workspace, where the training pipeline was invoked through a trainClassifier function that outputs the trained model and validation accuracy. The objective is to benchmark multiple classifier families under consistent preprocessing and evaluation settings to identify an accurate and deployable predictor for workout difficulty classification.

B. Data Partitioning and Validation Protocol

The dataset was partitioned into training (80%) and held-out test data (20%), with a validation process used during model selection [?]. To reduce optimistic bias and mitigate overfitting, *k*-fold cross-validation was applied during model training, enabling each observation to contribute to both training and validation across folds [12]. Model selection was guided primarily by validation accu-racy, complemented by error analysis using con-fusion matrices and discrimination analysis using ROC/AUC [10].

C. Model Families and Hyperparameter Search

We evaluate representative classifiers that com-monly appear in supervised learning pipelines for structured biometric data, including kernel-based methods and neural architectures. Specifically: (i) Support Vector Machines (SVMs) with linear, polynomial, and Gaussian kernels; and (ii) feed-forward neural networks with varying widths and depths. For each family, multiple configurations were trained and compared under the same protocol, with tuning performed using systematic search strategies (grid/random search) as implemented in the experimental script [**?**]. To support deployment-oriented decisions, we additionally record *predic-tion speed*, *training time*, and *compact model size* (Tables II–III).

D. SVM Benchmarking

SVMs were evaluated across several kernel configurations (Table II). In addition to overall validation accuracy, we inspect the confusion matrix of the best-performing SVM configuration to identify systematic class confusions (Fig. 15). ROC curves and the corresponding AUC values are used to quantify separability across operating points (Fig. 14) [10]. The SVM formulation follows the standard maximum-margin learning principle [11].

TABLE II
SVM Model Training

Model Type	Linear SVM	Quadratic SVM	Fine Gaussian SVM	Medium Gaussian SVM	Coarse Gaussian SVM
Accuracy (Validation)	69.4%	78.1%	78.2%	81.8%	72.6%
Total Cost (Validation)	3273	2341	2334	1945	2926
Error Rate (Validation)	30.6%	21.9%	21.8%	18.2%	27.4%
Prediction Speed	∼8400 obs/sec	∼2900 obs/sec	∼1200 obs/sec	∼2300 obs/sec	∼3000 obs/sec
Training Time (s)	217.61	1465.8	408.3	356	471.02
Model Size (Compact)	∼82kB	∼675kB	∼3MB	∼802kB	∼1MB

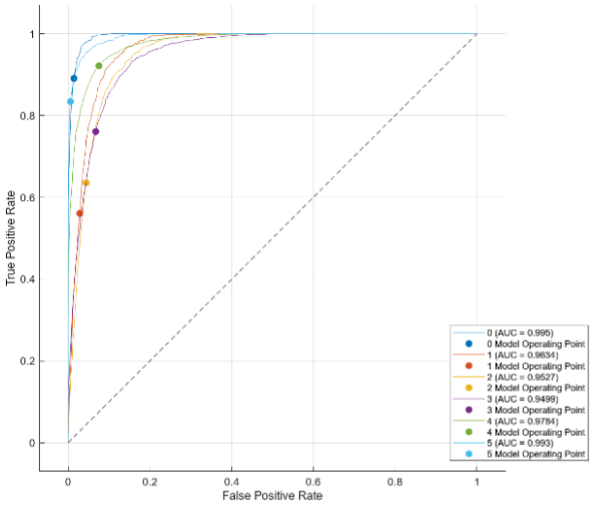

Fig. 14. ROC Curve of Medium Gaussian SVM

E. Neural Network Benchmarking

We further benchmark multiple feed-forward neural network designs (Table III) to evaluate non-linear decision boundaries and feature interactions. The confusion matrix

of the selected neural model (Fig. 16) provides a class-wise error profile, while the ROC curve (Fig. 17) summarises discrimination performance via AUC [10]. The training objective is optimized through backpropagation-based learning [13]. In addition to accuracy, we emphasise runtime constraints by reporting prediction speed and model size, since these factors directly affect real-time recommendation responsiveness in the fitness application.

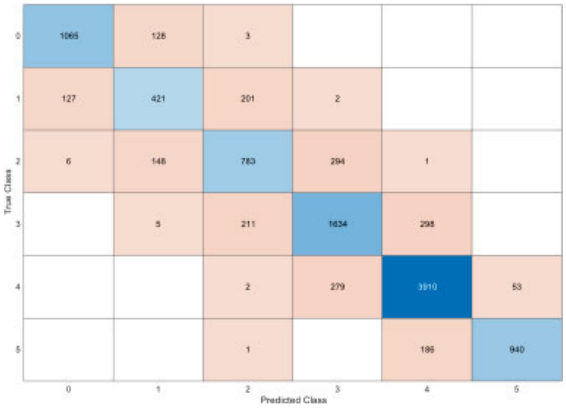

Fig. 15. Validation Confusion Matrix of Medium Gaussian SVM Model

TABLE III. NEURAL NETWORK MODEL TRAINING

Model Type	Narrow Neural Network	Medium Neural Network	Wide Neural Network	Bi-layered Neural Network
Accuracy (Validation)	86.7%	96.4%	94%	93.5%
Total Cost (Validation)	1421	380	643	696
Error Rate (Validation)	13.3%	3.6%	6%	6.5%
Prediction Speed	∼31000 obs/sec	∼35000 obs/sec	∼38000 obs/sec	∼33000 obs/sec
Training Time (s)	354.1	573	970.7	438.05
Model Size (Compact)	∼8kB	∼9kB	∼19kB	∼9kB

IV. RESULTS AND DISCUSSION

A. SVM Results

Table II summarises validation performance across multiple SVM kernels. Among the evaluated configurations, the *Medium Gaussian SVM* achieved the highest validation accuracy of 81.8% with an error rate of 18.2%, while maintaining a compact model size of approximately 802 kB and a prediction throughput of ∼ 2300 observations per second. Although the linear SVM attained substantially higher throughput (8400 obs/sec), its validation accuracy was lower (69.4%), suggesting that non-linear decision boundaries are beneficial for this task.

To understand error structure, Fig. 15 presents the validation confusion matrix for the Medium Gaussian SVM. The matrix indicates noticeable confusion between specific difficulty levels, particularly between classes 1 and 3, suggesting overlapping feature signatures for adjacent or conceptually similar fitness categories. Discrimination is further quantified via the ROC curve (Fig. 14), where the AUC values across operating points range from 0.9499 to 0.995, indicating strong separability de-spite the observed class-level confusions [10].

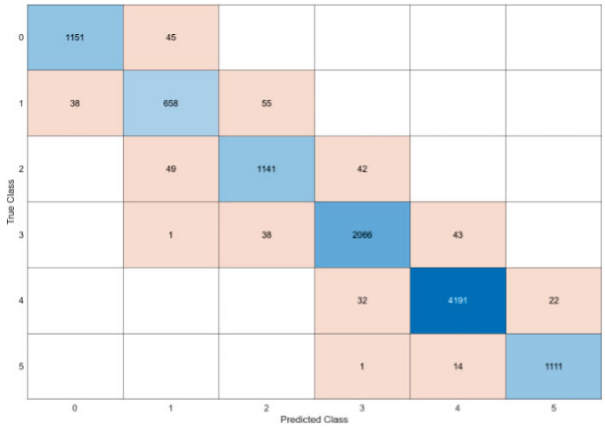

Fig. 16. Validation Confusion Matrix of Medium Neural Net-work Model

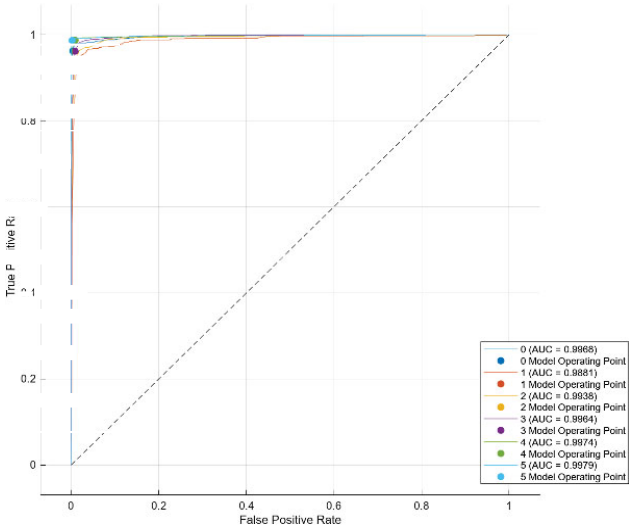

Fig. 17. ROC Curve of Neural Network Model

B. Neural Network Results

Neural models consistently outperformed SVM variants in validation accuracy (Table III). The *Medium Neural Network* achieved the highest validation accuracy of 96.4% with an error rate of 3.6%, alongside a prediction throughput of ∼ 35000 observations per second and a compact model size of approximately 9 kB. The trilayered network pro-vided comparable accuracy (96.1%) and the highest prediction throughput (∼ 48000 obs/sec), but with increased training time relative to the medium network.

Fig. 16 shows the confusion matrix for the Medium Neural Network, where the dominant diagonal structure indicates substantially fewer misclassifications than the SVM case. The ROC analysis (Fig. 17) yields very high AUC values, ranging from 0.9881 to 0.9979, further supporting strong class discrimination capability [10]. These results are consistent with the ability of neural architectures to learn non-linear feature interactions through backpropagation-based optimisation [13].

C. Comparative Discussion and Deployment Implications

Comparing the best-performing representatives of each family, the Medium Neural Network improves validation accuracy by 14.6 percentage points over the Medium Gaussian SVM (96.4%vs. 81.8%). From a deployment perspective, the neural model also provides substantially higher inference throughput ($\sim$ 35000 vs. $\sim$ 2300 obs/ sec, i.e., $\approx$ 15×) and a significantly smaller compact model size (9 kB vs. 802 kB, i.e., $\approx$ 89×). The primary trade-off is training overhead: the Medium Neural Network requires longer training time (573 s) than the Medium Gaussian SVM (356 s), which is acceptable when training is per-formed offline and inference is executed in real time.

The confusion-matrix evidence suggests that remaining errors are primarily driven by boundary cases where adjacent difficulty levels exhibit overlapping physiological and performance profiles. This motivates future work on (i) more explicit ordinal modelling for difficulty levels, and (ii) en-richment of the feature space with longitudinal workout adherence and wearable-derived signals, which may reduce ambiguity between neighbouring classes.

REFERENCES

[1] K. Zhang, X. Liu, J. Shen *et al.*, "Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography," *Cell*, vol. 181, no. 6, pp. 1423–1433.e11, 2020.

[2] V. Gulshan, L. Peng, M. Coram *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016.

[3] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, pp. 24–29, 2019.

[4] S. A. Alowais, S. S. Alghamdi, N. Alsuhebany *et al.*, "Revolutionizing healthcare: The role of artificial intelligence in clinical practice," *BMC Medical Education*, vol. 23, no. 1, p. 689, 2023.

[5] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "RUSBoost: A hybrid approach to alleviating class imbalance," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 40, no. 1, pp. 185–197, 2010.

[6] K. S. Hall *et al.*, "Physical performance across the adult life span: Correlates with age and physical activity," *The Journals of Gerontology: Series A*, vol. 72, no. 4, pp. 572–578, 2017.

[7] B. H. M. Branco, M. P. Bernuci *et al.*, "Proposal of a normative table for body fat percentages of brazilian young adults through bioimpedanciometry," *Journal of Exercise Rehabilitation*, vol. 14, no. 6, pp. 974–979, 2018.

[8] M. H. Diener, L. A. Golding, and D. Diener, "Validity and reliability of a one-minute half sit-up test of abdominal strength and endurance," *Sports Medicine, Training and Rehabilitation*, vol. 6, pp. 105–119, 1995.

[9] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, 1995, pp. 1137–1143.

[10] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[11] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[12] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics Surveys*, vol. 4, pp. 40–79, 2010.

[13] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986.