Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multilayer Perceptron versus Direct Guideline Rule Application for Four-Class Sarcopenia Classification: A Development and External Validation Study

Si-Hyeong Noh
Wonkwang University
Ji Hee Kim
Wonkwang University School of Medicine and Hospital
Dong-Wook Lim
Wonkwang University
Sung-Kyu Choe
Wonkwang University
Chang Won Jeong
mediblue@wku.ac.kr

Wonkwang University

Research Article

Keywords: sarcopenia, functional sarcopenia, machine learning, external validation, rule-based classification

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11463743/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Multilayer Perceptron versus Direct Guideline Rule Application for Four-Class Sarcopenia Classification: A Development and External Validation Study

Si-Hyeong Noh[1]†, Ji Hee Kim[2]†, Dong-Wook Lim[1], Sung-Kyu Choe[1,3]*, Chang won Jeong[1,4]*

[1] Sarcopenia Total Solution Center, Wonkwang University, Iksan, Jeonbuk, Republic of Korea

[2] Department of Rehabilitation Medicine, Wonkwang University School of Medicine and Hospital, Iksan, Jeonbuk, Republic of Korea

[3] Department of Microbiology, Institute of Wonkwang Medical Science, Wonkwang University School of Medicine, Iksan, Jeonbuk, Republic of Korea

[4] Medical AI Team, Wonkwang University Hospital, Iksan, Jeonbuk, Republic of Korea

†These authors contributed equally to this work and share first authorship.

**Corresponding authors: Sung-Kyu Choi, Sarcopenia Total Solution Center, Wonkwang University / Department of Microbiology, Institute of Wonkwang Medical Science, Wonkwang University School of Medicine, Iksan, Jeonbuk, Republic of Korea. Email: seongkyu642@wku.ac.kr. ORCID: [0000-0002-2102-973X]. Chang-Won Jeong, Sarcopenia Total Solution Center, Wonkwang University / Medical AI Team, Wonkwang University Hospital, Iksan, Jeonbuk, Republic of Korea. Email:mediblue@wku.ac.kr. ORCID: [0000-0002-9305-4686].*

Abstract

Background: Sarcopenia phenotypes are defined by cut-off values for muscle mass, grip strength, and physical performance. When reference labels are generated from the same variables assigned to a model, the aspects contributed by a machine learning classifier to the direct application of cut-off values remain unclear.

Methods: We developed a multilayer perceptron (MLP; 5–36–28–9–4) to classify the states of "normal," "sarcopenia," "functional sarcopenia," and "severe sarcopenia" based on five variables: sex, height,

skeletal muscle index, grip strength (higher of the left- and right-hand means), and Short Physical Performance Battery score. The development cohort comprised 339 participants from one hospital (data collection: September 2023–August 2025) and was evaluated using stratified five-fold cross-validation (out-of-fold predictions). An external cohort of 151 participants from three institutions (148 women) was used for validation. The MLP was compared with the direct application of guideline rules (Rules A and B), logistic regression, a decision tree, and a random forest. Confidence intervals (CIs) were obtained through bootstrap resampling.

Results: Reference labels agreed with Rule A in 93.5% (317/339) of the development cohort and 94.0% (142/151) of the external cohort and with the originally stated definition (Rule B) in 91.7% and 92.1%, respectively. The MLP achieved out-of-fold accuracy, balanced accuracy, and macro F1-score values of 0.870, 0.819, and 0.819, respectively, in the development cohort and 0.901, 0.923, and 0.897, respectively, in the external cohort. Rule A, applied without training, achieved 0.935 (95% CI 0.909–0.959), 0.926, and 0.909 internally and 0.940 (95% CI 0.901–0.974), 0.939, and 0.926 externally, higher than the MLP for every point estimate. The random forest matched Rule A externally (accuracy 0.940) but not internally (0.882); logistic regression and the decision tree performed worse.

Conclusions: The MLP reproduced the reference labels but did not outperform the direct application of the guideline rules. When labels are defined by the cut-off values of the input variables, the model performance largely reflects rule reproduction, and any added value must be tested against independent reference standards or clinical outcomes. As 98% of the external cohort comprised women, generalizability to men could not be assessed.

Keywords: sarcopenia; functional sarcopenia; machine learning; external validation; rule-based classification

Background

Sarcopenia is a clinical condition diagnosed based on assessments of muscle mass, muscle strength, and physical function performed collectively. The European Working Group on Sarcopenia in Older People 2 (EWGSOP2) proposes low muscle strength as a key factor, considers low muscle quantity or quality to confirm the diagnosis, and uses low physical performance to grade severity [1]. Further, the Asian Working Group for Sarcopenia (AWGS) 2019 consensus and Korean Working Group on Sarcopenia (KWGS) guidelines present assessment frameworks integrating these clinical domains [2,3]. Sarcopenia is a muscle disease associated with functional decline in older age [4], and its prevalence varies depending on the diagnostic definition and cutoff values used [5]. Consensus definitions have been revised since the EWGSOP 2010 [6], and the Foundation for the National Institutes of Health (FNIH) project proposed data-derived cutoffs for muscle strength and mass [7].

The Short Physical Performance Battery (SPPB)—a physical function assessment tool—integrates balance, gait, and chair stand performance, and early studies have reported associations with disability, mortality, and nursing home admission [8]. A subsequent meta-analysis revealed that an SPPB score < 10 was associated with all-cause mortality [9], and grip strength is a potential biomarker of health status in older adults [10]. The KWGS guidelines classify functional sarcopenia as low muscle strength and physical performance without low muscle mass [3]. This literature supports the need to assess functional characteristics together with muscle mass.

Machine learning and deep learning models using clinical characteristics and laboratory indices, as well as electronic health record-based detection models, have been employed in sarcopenia research [11,12]. Further, deep learning models using Korean data, machine learning models based on physical activity variables and feature selection [13–15], and approaches combining clinical and biological variables [16] have been proposed. These prior studies demonstrate the potential of approaches integrating diverse clinical information. Yet, as the input variables, diagnostic definitions, and classification targets vary across studies, the performance of the four-class phenotype classification used in the current study requires separate verification.

In real clinical data, the number of participants in each sarcopenia category may not be equal—an imbalance that could affect the training of multiclass models [17]. Specifically, as high overall accuracy driven by the majority class can coincide with a relatively low classification sensitivity for minority classes, it is vital to evaluate the balanced accuracy, macro F1-score, and class-specific performance alongside overall accuracy. Additionally, when evaluating clinical applicability, the calibration of class membership probabilities and clinical utility must be examined together with discriminative performance [18].

Accordingly, in this study, we developed a multilayer perceptron (MLP)-based model using five clinical variables—sex, height, skeletal muscle index (SMI), mean grip strength, and SPPB—to classify four states: “normal,” “sarcopenia,” “functional sarcopenia,” and “severe sarcopenia.” Model performance was internally validated using stratified five-fold cross-validation and out-of-fold classification results, and subsequently evaluated using data from 151 external participants not used in model development. Permutation feature importance was considered to analyze the model’s relative dependence on each input variable. As the reference labels are defined by cut-offs of the same input variables, we also compared performance against the classification obtained by applying the diagnostic rules directly to the input values and against other machine learning models, and additionally evaluated the calibration of class membership probabilities.

Methods

Data and Participants

Model development and internal validation were performed using data from 341 participants collected at Wonkwang University Hospital between September 2023 and August 2025.

Participants were adults aged 20 years or older with suspected sarcopenia who visited Wonkwang University Hospital in response to recruitment notices posted at Wonkwang University Hospital, Wonkwang University, and the Iksan Public Health Center. They were enrolled after receiving an explanation of the study purpose and procedures and providing voluntary written consent. Participation

was not restricted based on race or socioeconomic status. The exclusion criteria were as follows: currently undergoing cancer treatment or less than five years since completion of cancer treatment; current severe psychiatric symptoms; a history of malabsorptive disease or small- or large-bowel resection; pregnancy; and acute bacterial or viral infectious disease.

After reviewing the range and outliers in the raw data, two participants whose grip strength values could not be confirmed to be reliable were excluded, leaving data from 339 participants for the final analysis. The sex and class distributions of the development data are presented in the Results section and Tables 2 and 3.

The existing Result label provided in the dataset was used as the target variable and reference label for the performance evaluation. It was coded as 0, 1, 2, and 3 for normal, sarcopenia, functional sarcopenia, and severe sarcopenia, respectively. Sex was coded as 0 for men and 1 for women.

Participants' sarcopenia-related clinical status was classified according to clinical criteria based on SMI, grip strength, and the SPPB. Low muscle mass was defined as SMI < 7.0 kg/m² in men and < 5.7 kg/m² in women. Low muscle strength was defined as grip strength < 28 kg in men and < 18 kg in women, and low physical performance was defined as an SPPB score of ≤ 9.

Based on the combination of these three clinical criteria, the participants were classified as having sarcopenia if they had low SMI together with low grip strength or low SPPB; functional sarcopenia if they had both low grip strength and low SPPB without low SMI; and severe sarcopenia if they had low SMI, grip strength, and SPPB together. Severe sarcopenia was determined preferentially when all three criteria were met. Participants who did not meet any of the above criteria were classified as normal. The final Result label was based on this clinical determination and was coded as normal=0, sarcopenia=1, functional sarcopenia=2, and severe sarcopenia=3 for model training.

The overall analytical workflow proceeded in the following order: construction of training data, selection and standardization of input variables, training of MLP model, internal validation using

stratified five-fold cross-validation, construction of the final model using the entire training dataset, and external validation using data not used in model development (Figure 1).

Model Input Variables

The final input variables used to classify sarcopenia status were five variables in total: sex, height, SMI, mean grip strength, and SPPB score.

The model input order was fixed as sex → height → SMI → mean grip → SPPB.

Height was measured in centimeters. No missing values were observed for the model input variables in either the development data or external data of the 151 participants.

Grip strength was measured three times for each hand. The mean of the three measurements was calculated for each hand, and the higher of the two hand-side means was used as the model input value. Hereafter, this variable is denoted as the mean grip and corresponds to the MaxGrip variable name used in the original model file.

Only the five clinical variables were entered into the model, and the Result class was used as the training target variable. As SMI, grip strength, and SPPB are related to the clinical classification criteria, that no separate binary diagnostic variables were entered does not remove the information overlap between the reference criteria and input variables.

This study evaluated the classification performance for clinical status at each measurement time point; however, the prognostic performance for future onset or progression was not evaluated.

Data Preprocessing and Standardization

Standardization was applied for the input variables to reduce training bias arising from variations in measurement units and value ranges across variables. StandardScaler was used, and each variable was transformed using the mean and standard deviation (SD) calculated from the corresponding training data.

The standardized value was calculated as follows.

Here, x is the original input value, μ (mu) is the mean of the corresponding variable calculated from the training data, and σ (sigma) is its SD.

In internal cross-validation, the scaler was fitted using only the training subset of each fold and then applied to the corresponding validation subset to prevent information from the validation data from being used in training during standardization. When constructing the final deployed model, the scaler was refitted using the full data from all 339 participants, and the resulting mean and SD were stored together with the model in a metadata file.

Neural Network Architecture

An MLP-based neural network was constructed to classify the four sarcopenia states. The model comprised five input nodes, three fully connected hidden layers, and four output nodes (Figure 1).

The network architecture was as follows.

5 -> 36 -> 28 -> 9 -> 4

The first, second, and third hidden layers comprised 36, 28, and nine nodes, respectively. A rectified linear unit (ReLU) activation function was applied for each hidden layer.

The final output layer comprised four output nodes representing the four classes (normal, sarcopenia, functional sarcopenia, and severe sarcopenia), and the class membership probabilities for each class were produced using the softmax function. The final classification was determined as the class with the highest of the four output probabilities.

We did not perform automated hyperparameter optimization through model comparison. The 5–36–28–9–4 network architecture was fixed to reproduce a pre-existing configuration (A-model) and used identically across all internal cross-validation folds and in training the final model. No procedure was used to modify the architecture iteratively based on the validation results.

Model Training

The NAdam optimizer was used for model training, with the learning rate set to 0.001. The loss function was sparse categorical cross-entropy (SparseCategoricalCrossentropy), which directly handles the four classes coded as integers (0–3).

Balanced class weights were applied based on the class frequencies of each training subset to correct class imbalance in the training data. The weight for each class was calculated from the total number of training samples and number of samples in that class, and this weight was reflected in the training loss. In the internal five-fold cross-validation, the class weights were calculated distinctly for the training subset of each fold.

The model was trained using a batch size of 10 for 300 epochs. Likewise, the 300 epochs were not selected using validation results but were fixed according to a pre-existing training configuration and applied identically across all cross-validation folds and in the final full data training. We did not use early stopping, nor did we apply dropout, additional regularization, or noise augmentation to the final primary model.

For reproducibility, the random seed was fixed at 42, and the same value was applied to the Python random generator, NumPy, and TensorFlow.

[Insert Table 1 here]

Internal Validation

Stratified five-fold cross-validation was performed to evaluate the internal classification performance of the developed model (Figure 1). The full training data were divided into five folds, with stratification applied such that the proportions of the four classes were maintained as closely as possible in each fold.

In each iteration, the model was trained using four folds and the remaining fold was used as the validation dataset. This process was repeated five times, with every fold used once as the validation dataset.

For each fold, the accuracy, balanced accuracy, macro-average F1-score, weighted-average F1-score, and one-vs.-rest macro-average receiver operating characteristic-area under the curve (ROC-AUC) were calculated for the validation dataset. Each metric was summarized as the mean and SD across the five folds.

Out-of-fold (OOF) classification results were generated such that each participant was classified exactly once, by a model trained on a subset not including that participant. The accuracy, balanced accuracy, macro F1-score, and macro ROC-AUC were calculated from the pooled OOF results, and a confusion matrix and class-specific ROC curves were constructed.

Feature Importance Analysis

Permutation feature importance analysis was conducted to evaluate the relative influence of each input variable on classification performance [19].

In each validation fold, the accuracy of the original model was set as the baseline, and the values of one input variable were randomly permuted across participants before reclassification. The resulting decrease in validation accuracy was defined as the importance of this variable.

The permutation for each variable was repeated 10 times in each validation fold, with an independent random permutation in each repetition; the decrease in mean accuracy across the 10 repetitions was the variable importance for this fold. Subsequently, the importance values from the five validation folds were combined to calculate the mean and SD for each variable.

Accordingly, each input variable was evaluated with 50 permutations (10 repeats × five folds) across the five-fold cross-validation. To ensure reproducibility, a fixed random generator based on the SEED + fold number was used for each fold.

Permutation importance [20] is a relative measure of the extent to which a model's classification performance depends on a given input variable and does not imply a causal relationship between that variable and sarcopenia status.

Final Model Construction and Storage

After internal cross-validation, a new StandardScaler was fitted using the full development data of the 339 participants, and the final MLP model was trained for 300 epochs using the same architecture and training conditions.

The final model was saved in H5 format, and the input order, sex coding scheme, class mapping, and the mean and SD of the StandardScaler were stored in a JSON metadata file.

External Validation

The external classification performance of the final model was examined for 151 distinct participants from three institutions—Seulgiroun Rehabilitation Hospital, Dongsan Monastery, and Jungang Monastery—whose data were not used in model development or internal cross-validation. Each participant contributed to a single observation, and the participant was the unit of analysis. The external data were not used for model retraining or hyperparameter tuning.

The same sex coding and input variable order as the development data were applied to the external data. The sex and class distributions are presented in Tables 2 and 3.

The mean and SD of the StandardScaler derived from the full development data of 339 participants were applied unchanged to the external validation data. The standardized inputs were entered into the final H5 model to yield softmax probabilities for the four classes, and the class with the highest probability was defined as the final classification.

The external validation performance was assessed based on the accuracy, balanced accuracy, macro F1-score, weighted F1-score, and one-vs.-rest macro ROC-AUC. We also calculated the class-specific precision, recall, F1-score, and ROC-AUC. The confusion matrix presents the cross-tabulated frequencies of the true and predicted classes, together with the proportion within each true class. ROC curves were constructed from class membership probabilities. Key performance metrics are presented as point estimates with 95% bootstrap confidence intervals (CIs).

Description of Baseline Characteristics

The baseline characteristics of the 339 development cohort participants and 151 external cohort participants were recorded. Sex was summarized as frequency and percentage, and height, SMI, mean grip, and SPPB were summarized as mean ± SD, with the missingness of input variables reported. The absolute standardized mean difference (SMD) for continuous variables was defined as the mean difference divided by the square root of the average sample variances of the two datasets. For sex, the absolute difference in the proportion of women was divided by the square root of the average of $p(1-p)$ for the two datasets. SMD was used for descriptive comparison only, with no significance testing conducted.

Uncertainty of Performance Estimates

Uncertainty in the key performance metrics was tested using nonparametric bootstrapping. The accuracy, balanced accuracy, macro F1-score, weighted F1-score, and macro ROC-AUC were recalculated by resampling with replacement, at the participant level and at the original sample size, from the internal OOF data (339 participants) and external validation data (151 participants). Resampling was repeated 5,000 times for each dataset, with the NumPy default_rng seed set to 42 in each case. The 95% CI was the 2.5th and 97.5th percentiles of the bootstrap distribution. All four classes were retained in every one of the 5,000 iterations for both datasets; therefore, no iterations were excluded. In resampling, we used the saved OOF classification results or external classification results from the fixed final model, and the model was not retrained. Therefore, these CIs reflect the sampling variability of the fixed classification results and do not capture the full variability in the model retraining process.

Reference-Label Reproducibility Check and Comparator Models

Sex-specific cut-offs (SMI: < 7.0 kg/m² in men, < 5.7 kg/m² in women; mean grip strength: < 28 kg in men, < 18 kg in women; SPPB: ≤ 9 points) were applied directly to the model input values to examine whether the existing Result label could be reproduced by the criteria described in Section 2.1. Two rules were applied. Under Rule A, participants were classified as having sarcopenia if they had low SMI

together with low grip strength or low SPPB; as having severe sarcopenia if all three criteria were met; as having functional sarcopenia if they had both low grip strength and low SPPB without low SMI; and as normal otherwise. Under Rule B, following the pre-revision wording mentioned in Section 2.1, only participants meeting both the low SMI and grip strength criteria were classified as having sarcopenia. The distribution of existing labels was summarized in a cross-tabulation by each of the eight combinations of low/not low status across the three criteria.

As rule-based classification involves no training, the rules were applied directly and uniformly to the entire development data and to the external data, and evaluated using the same metrics as the MLP. As additional comparator models, multinomial logistic regression (standardized inputs, balanced class weight), a decision tree with maximum depth 4 (balanced class weight), and a random forest [21] (500 trees, minimum leaf size 2, balanced class weight) were trained using the same five input variables. The internal performance of the comparator models was evaluated from the OOF results using the same stratified five-fold split as the MLP (shuffle, random_state=42), and the external performance was evaluated using models fitted on the entire development data. The 95% CIs were calculated using the same bootstrap procedure as reported in Section 2.11 (5,000 iterations, seed 42). When comparing Rule A and the MLP in the same participants, 95% CIs for the difference in metrics were obtained by paired bootstrapping, and the exact McNemar test was applied to pairs differing in correctness [22].

Calibration and Subgroup Analysis

We used the internal OOF probabilities saved from the original analysis (339 participants) and external output of the final H5 model (151 participants) in analyses of misclassification, sex-specific subgroup performance, and calibration. The internal OOF probabilities were matched to the row order of the development data and to the Result labels. The external probabilities were calculated using the saved H5 weights and standardization coefficients in the JSON metadata by applying the same float32 preprocessing and forward pass as the provided inference code. No additional model training or parameter adjustments were performed using the external data.

Calibration was tested using the multiclass Brier score (the mean, across participants, of the sum of squared probability errors across the four classes), a prevalence-only reference Brier score (applying the class proportions of each dataset to all participants), log loss, the expected calibration error (ECE) based on the top-class probability, and the class-specific one-vs.-rest calibration slope. The ECE was calculated by dividing the 0–1 range into 10 equal-width bins and weighting—by the number of participants—the absolute difference between the mean top-class probability and the observed accuracy in each bin. The bins included their left boundary and excluded their right boundary, except that the final bin included 1. Log loss was calculated by clipping probabilities to $[10^{-7}, 1-10^{-7}]$ and renormalizing such that each row summed to 1. The calibration slope was estimated together with an intercept from an unpenalized logistic regression, with the logit of the (similarly clipped) class-specific probability as the predictor and membership in this class as the outcome. Sex-specific subgroup performance was estimated for men and women using the development data and for women using the external data. For the three external men, only the number correctly classified is reported. The near-cut-off group for the misclassification analysis was defined as participants with SMI within ±0.2 kg/m² of the sex-specific cut-off, mean grip strength within ±2 kg of the cut-off, or an SPPB score of 9 or 10, with boundary values included.

Research Ethics

This study was conducted after review and approval by the Institutional Review Board (IRB) of Wonkwang University Hospital (IRB no. WKUH 2023-08-031; approval date: September 8, 2023). All participants received a full explanation of the study purpose and content, the benefits and risks of participation, and personal information protection. Data collection commenced only after voluntary consent was obtained. Separate consent for human-derived material research was also obtained for research using blood specimens, including proteomic and metabolomic analyses.

Results

Participants and Data Composition

The development data were from 339 participants in total: 193 (56.9%) with normal status, 36 (10.6%) with sarcopenia, 48 (14.2%) with functional sarcopenia, and 62 (18.3%) with severe sarcopenia (Table 3 and Figure 2).

The baseline characteristics of the development data and 151 external participants are presented in Table 2. The proportions of women were 63.1% (214/339) and 98.0% (148/151), respectively. Mean height, SMI, and mean grip were lower in the external data than in the development data, with absolute SMDs of 0.841, 0.660, and 0.546, respectively; the absolute SMD for SPPB was 0.028. No missing values were noted for model input variables. Table 3 presents the class distribution.

Table 2. Characteristics of the development cohort and external participants

Characteristic	Development cohort (N=339)	External validation (151 participants)	Absolute SMD
Female, n (%)	214 (63.1%)	**148 (98.0%)**	**0.982**
Male, n (%)	125 (36.9%)	**3 (2.0%)**	—
Height, cm	158.78 ± 8.86	**152.49 ± 5.77**	**0.841**
SMI, kg/m²	6.41 ± 1.07	**5.81 ± 0.70**	**0.660**
Mean grip strength, kg	19.63 ± 9.04	**15.73 ± 4.50**	**0.546**
SPPB score, points	8.81 ± 3.58	**8.73 ± 2.51**	**0.028**
Missing values in model inputs, n (%)	0 (0%)	**0 (0%)**	—

Values are n (%) or mean ± SD. The external cohort comprised 151 participants, with one observation per participant.

Table 3. Clinical class distribution in the development cohort and external participants

Clinical class	Development cohort (N=339)	External validation (151 participants)
Normal	193 (56.9%)	**69 (45.7%)**
Sarcopenia	36 (10.6%)	**19 (12.6%)**
Functional sarcopenia	48 (14.2%)	**31 (20.5%)**
Severe sarcopenia	62 (18.3%)	**32 (21.2%)**

Internal Model Classification Performance

Stratified five-fold cross-validation was performed to evaluate the internal performance of the developed MLP model. The mean across the five validation folds was 0.8702 ± 0.0457 for accuracy, 0.8196 ± 0.0506 for balanced accuracy, 0.8184 ± 0.0546 for the macro F1-score, and 0.8696 ± 0.0451 for the weighted F1-score. The mean macro ROC-AUC was 0.9748 ± 0.0124.

In the pooled OOF results, accuracy was 0.8702 (95% CI 0.8319–0.9056), balanced accuracy 0.8188 (95% CI 0.7630–0.8696), the macro F1-score 0.8188 (95% CI 0.7646–0.8654), the weighted F1-score 0.8702 (95% CI 0.8329–0.9045), and macro ROC-AUC 0.9733 (95% CI 0.9610–0.9837).

In the OOF-based confusion matrix, the following participants were correctly classified: 179 of 193 (92.7%) for normal, 26 of 36 (72.2%) for sarcopenia, 37 of 48 (77.1%) for functional sarcopenia, and 53 of 62 (85.5%) for severe sarcopenia (Figure 3).

The normal group had the highest recall among the four classes. In the sarcopenia group, five cases each were misclassified as normal or severe sarcopenia. In the functional sarcopenia group, misclassification as normal was the most common (nine cases), and in the severe sarcopenia group, misclassification as sarcopenia (five cases) or functional sarcopenia (four cases) was the most common.

A high classification performance was also observed in multiclass ROC analysis. The one-vs.-rest class-specific ROC-AUCs were 0.970 for normal, 0.964 for sarcopenia, 0.971 for functional sarcopenia, and 0.989 for severe sarcopenia, with an overall macro ROC-AUC of 0.973(Figure 4).

Input Variable Importance

The decrease in validation accuracy upon random permutation was greatest for mean grip (0.2254), followed by SMI (0.2137), SPPB (0.1897), sex (0.1793), and height (0.0535) (Figure 5).

Training Behavior of the Final Model

Over 300 epochs of training the final model, the training loss generally decreased and training accuracy increased (Figures 6 and 7).

At the final epoch, the training loss was 0.0286 and training accuracy was 0.9853.

However, as the training accuracy for the full dataset was calculated using data already used for training, it was not used as a measure of generalization performance. The internal generalization performance of the model was evaluated based on the five-fold OOF results presented above.

External Validation

External validation included 151 distinct participants from three institutions: 148 women and three men.

Of the 151 external participants, 136 were classified according to the existing reference label. The results were as follows: accuracy = 0.9007 (95% CI 0.8543–0.9470), balanced accuracy = 0.9230 (95% CI 0.8769–0.9623), macro F1-score = 0.8967 (95% CI 0.8420–0.9432), weighted F1-score 0.9024 (95% CI 0.8554–0.9474), and macro ROC-AUC = 0.9801 (95% CI 0.9640–0.9924) (Table 4).

Table 4. Overall external validation performance for 151 participants

Metric	**External validation**	95% bootstrap CI
Number of institutions	3	—
Number of participants	151	—
Correct classifications	136/151	—
Accuracy	0.9007	0.8543–0.9470
Balanced accuracy	0.9230	0.8769–0.9623
Macro F1-score	0.8967	0.8420–0.9432
Weighted F1-score	0.9024	0.8554–0.9474
Macro ROC-AUC	0.9801	0.9640–0.9924

In the class-specific analysis, 58 of the 69 normal participants were correctly classified (recall, 0.841). The recall was 0.947 for sarcopenia (18/19), 0.935 for functional sarcopenia (29/31), and 0.969 for severe sarcopenia (31/32). The precision and F1-score for severe sarcopenia were 1.000 and 0.984, respectively, whereas the precision for sarcopenia was 0.720 (Table 5).

Table 5. Class-specific external validation performance

Class	n	Precision	Recall	F1-score	ROC-AUC
Normal	69	0.951	0.841	0.892	0.967
Sarcopenia	19	0.720	0.947	0.818	0.974
Functional sarcopenia	31	0.853	0.935	0.892	0.980
Severe sarcopenia	32	1.000	0.969	0.984	0.999

In total, 15 misclassifications were noted. Among the normal participants, six were classified as having sarcopenia and five as having functional sarcopenia. One participant with sarcopenia and two with functional sarcopenia were classified as normal, and one participant with severe sarcopenia was classified as having sarcopenia (Figure 8).

The class-specific one-vs.-rest ROC-AUCs were 0.967 for normal, 0.974 for sarcopenia, 0.980 for functional sarcopenia, and 0.999 for severe sarcopenia, with a macro ROC-AUC of 0.9801 (Figure 9).

The point estimates and 95% bootstrap CIs for the internal OOF and external validation performance are shown in Figure 10. Higher point estimates were observed in the external data; yet, as the participant composition and class distribution differed and no formal hypothesis test was performed between the two datasets, this was not interpreted as evidence of improved performance.

Reproducibility of Reference Labels

When Rule A was applied to the model input values, the agreement with the existing labels was 93.5% (317/339) for the development data and 94.0% (142/151) for the external data. The agreement for Rule B, corresponding to the pre-revision wording presented in Section 2.1, was lower at 91.7% (311/339) and 92.1% (139/151), respectively (Table 6). Nine participants (six development and three external)

386 with low SMI and low SPPB but normal grip strength were all labeled as having sarcopenia; therefore,
387 the existing labels agreed with Rule A rather than Rule B.

388 Discordance not explained by Rule A occurred in 22 development participants and nine external
389 participants. Participants meeting all three criteria but labeled as having sarcopenia were eight and three
390 in number, respectively; those with low grip strength and low SPPB but labeled as normal were 11 and
391 three in number, respectively; and those with low SMI and low grip strength but labeled as normal were
392 three each.

393 Table 6. Concordance between recorded reference labels and guideline rules applied to model inputs

Low SMI	**Low grip**	**Low SPPB**	**Rule A class**	**Development labels**	**External labels**	**Discordant (Dev/Ext)**
–	–	–	Normal	Normal 102	Normal 25	0 / 0
–	–	Yes	Normal	Normal 19	Normal 11	0 / 0
–	Yes	–	Normal	Normal 38	Normal 16	0 / 0
–	Yes	Yes	Functional	Functional 48; Normal 11	Functional 31; Normal 3	11 / 3
Yes	–	–	Normal	Normal 20	Normal 11	0 / 0
Yes	–	Yes	Sarcopenia*	Sarcopenia 6	Sarcopenia 3	0 / 0
Yes	Yes	–	Sarcopenia	Sarcopenia 22; Normal 3	Sarcopenia 13; Normal 3	3 / 3
Yes	Yes	Yes	Severe	Severe 62; Sarcopenia 8	Severe 32; Sarcopenia 3	8 / 3
Concor			Rule A	317/339 (93.5%)	142/151 (94.0%)	22 / 9

Low SMI	Low grip	Low SPPB	Rule A class	Development labels	External labels	Discordant (Dev/Ext)
dance						
			Rule B	311/339 (91.7%)	139/151 (92.1%)	28 / 12

Low SMI: < 7.0 kg/m² (men), < 5.7 kg/m² (women); low grip: mean grip < 28 kg (men), < 18 kg (women); low SPPB: ≤ 9 points. Rule A: sarcopenia = low SMI plus low grip or low SPPB; severe = all three. Rule B (pre-revision wording of Section 2.1): sarcopenia = low SMI plus low grip.

*Classified as normal under Rule B.

Performance Comparison with Rule-Based Classification and Comparator Models

Without any training, in the development data, Rule A achieved an accuracy of 0.9351 (95% CI 0.9086–0.9587), balanced accuracy of 0.9263 (0.8885–0.9601), and macro F1-score of 0.9087 (0.8673–0.9436); in the external data, the corresponding values were 0.9404 (0.9007–0.9735), 0.9388 (0.8879–0.9803), and 0.9264 (0.8722–0.9685). These point estimates were higher than those of the MLP for the internal OOF (0.8702, 0.8188, 0.8188) and external (0.9007, 0.9230, 0.8967) data (Table 7, Figure 11).

Comparing Rule A with the original MLP in the same participants, the accuracy difference was +0.0649 (95% paired bootstrap CI 0.0324–0.1003) in the development data and +0.0397 (−0.0066–0.0861) in the external data. In the development data, Rule A alone was accurate for 29 participants and the MLP alone for seven (exact McNemar p=0.00031); in the external data, the corresponding numbers were nine and three (p=0.146).

Among the comparator models, random forest achieved an internal OOF accuracy of 0.8820 and an external accuracy of 0.9404, similar to or higher than the MLP, while multinomial logistic regression (0.7876, 0.8344) and the decision tree (0.8053, 0.8146) displayed lower values. The tendency for external point estimates to exceed internal estimates was also observed for rule-based classifications and comparator models, suggesting differences in participant composition and class distribution in the external data.

Table 7. Performance of guideline rules, the MLP, and comparator models

Method	Int. Acc	Int. BAcc	Int. Macro F1	Ext. Acc	Ext. BAcc	Ext. Macro F1
Rule A (no training)	0.935	0.926	0.909	0.940	0.939	0.926
Rule B (no training)	0.917	0.885	0.876	0.921	0.899	0.896
MLP, original OOF/H5	0.870	0.819	0.819	0.901	0.923	0.897
Random forest	0.882	0.829	0.827	0.940	0.950	0.928
Decision tree (depth 4)	0.805	0.806	0.757	0.815	0.857	0.809
Multinomial logistic regression	0.788	0.755	0.716	0.834	0.870	0.827

Int., internal (out-of-fold for trained models; whole development cohort for rules, which require no training); Ext., external; Acc, accuracy; BAcc, balanced accuracy. The MLP results were derived from the original saved OOF probabilities and frozen H5 model outputs.

Characteristics of Misclassified Cases

The original MLP misclassified 44 development and 15 external participants. Of the 44 development misclassifications, 32 (72.7%) were in the near cut-off group, whose misclassification rate (19.8%, 32/162) was higher than that of the non-near cut-off group (6.8%, 12/177). In the external data, all 15 misclassifications occurred in the near cut-off group (101 participants), with misclassification rates of 14.9% (15/101) in the near cut-off group and 0% (0/50) in the non-near cut-off group. Of the misclassifications, 15 in the development data and six in the external data had existing labels that did not agree with Rule A. Restricting to participants whose label agreed with Rule A, MLP accuracy was 0.9085 (288/317) in the development data and 0.9366 (133/142) in the external data.

Sex-Specific Subgroup Performance

In the 125 development cohort men, the accuracy, balanced accuracy, and macro F1-score of the original MLP were 0.840, 0.782, and 0.782; in the 214 development cohort women, they were 0.888, 0.822, and 0.822, respectively. The recall for sarcopenia in women was 0.579 (11/19), which was the lowest among the four classes. In the 148 external cohort women, the corresponding values were 0.905, 0.930, and 0.902, respectively, and two of the three external cohort men were correctly classified. Rule A accuracy was 0.912 for development cohort men, 0.949 for development cohort women, and 0.939 for external cohort women (Table 8).

Table 8. Sex-specific performance of the original MLP and Rule A

Subgroup	**n**	**MLP Acc**	**MLP BAcc**	**MLP Macro F1**	**Rule A Acc**	**Rule A BAcc**
Development, men	125	0.840	0.782	0.782	0.912	0.906
Development, women	214	0.888	0.822	0.822	0.949	0.944
External, women	148	0.905	0.930	0.902	0.939	0.939
External, men	3	2/3 correct	–	–	–	–

External cohort men (n=3) were too few for subgroup metrics.

Probability Calibration

For the original MLP, the top-class probability exceeded 0.99 for 79.4% (269/339) of development and 78.8% (119/151) of external participants. The mean top-class probabilities were 0.973 and 0.969, and the accuracies were 0.870 and 0.901, respectively. Accuracy in the 0.90–<0.999 probability range was 71.8% (51/71) in development data and 85.3% (29/34) in external data. ECE was 0.110 and 0.081, respectively, and class-specific calibration slope ranges were 0.24–0.32 (development) and 0.28–0.61

(external), both below 1 (Table 9, Figure 12). Multiclass Brier scores were 0.230 and 0.173, respectively, which were lower than the corresponding prevalence-only reference values (0.611 and 0.688), and log losses were 0.694 and 0.486, respectively.

Table 9. Calibration of original MLP class probabilities

Measure	**Internal OOF (N=339)**	**External (N=151)**
Multiclass Brier score (prevalence-only reference)	0.230 (0.611)	0.173 (0.688)
Log loss	0.694	0.486
Expected calibration error (top class, 10 bins)	0.110	0.081
One-vs.-rest calibration slope, range across classes	0.24–0.32	0.28–0.61
Mean top-class probability/accuracy	0.973 / 0.870	0.969 / 0.901
Predictions with top-class probability > 0.99	79.4%	78.8%
Accuracy when 0.90 ≤ top-class probability < 0.999	71.8% (51/71)	85.3% (29/34)
Accuracy when top-class probability ≥ 0.999	96.6% (228/236)	96.1% (99/103)

Discussion

In this study, we developed a four-class classification model using five clinical variables and evaluated classification performance using internal OOF data based on 151 external participants. The external accuracy was 0.9007, and macro ROC-AUC was 0.9801.

In internal OOF analysis, the recall was 92.7% for the normal class but only 72.2% and 77.1% for sarcopenia and functional sarcopenia, respectively. This outcome shows that the overall accuracy alone cannot fully describe class-specific classification patterns. A sample size imbalance may have contributed to this difference; however, the current analysis could not establish the cause.

Permutation importance displayed the largest performance decrease when the mean grip and SMI were permuted. This model dependence may be linked to the clinical importance of these variables [1–3] but does not signify independent diagnostic value or a causal effect [19]. Specifically, when the reference label is generated from the cut-off values of the same input variables, high agreement may simply reflect the reproduction of the existing classification rule. Therefore, the label generation method and determination procedure should be clearly reported, and any additional clinical value of the model is difficult to claim without a comparison with the existing rule-based classification. As the input information and diagnostic targets also differ from those in prior machine learning studies, performance values alone cannot establish superiority over those studies [11,12].

The subsequent analyses confirmed this hypothesis. The reference labels aligned better with a definition combining low muscle mass with low strength or low physical performance, as in AWGS 2019 (Rule A), than with the pre-revision definition presented in Section 2.1 [2]; classification obtained by applying this rule directly to the model inputs, without any training, yielded higher accuracy, balanced accuracy, and macro F1-score values than the MLP in both the internal and external data. Most MLP misclassifications occurred near the cut-off values. Therefore, the model in its current form is best interpreted as approximating a determination rule that follows from the input variables, and additional diagnostic performance cannot reasonably be claimed for a classification fully determined by applying the cut-off values. To demonstrate the contribution of the model, an independent reference standard unrelated to the input variables (e.g., expert panel adjudication or outcome data) should be used, or the model should be demonstrated to provide a clearly defined and validated function the rule cannot (e.g., risk stratification of borderline cases). Reports have indicated that the superiority of machine learning

in clinical prediction has not been confirmed in comparisons with low methodological bias [23], highlighting the need for comparison against simple rules.

The 31 label discordances not explained by Rule A may have arisen from differences in measurement timing, determination process, or application of the cut-off values. As such cases could act as label errors in both model training and evaluation, the basis for these determinations should be assessed against raw data, with the labels corrected as required and analysis repeated.

The external validation participants comprised 148 women and three men, with a substantial sex imbalance. Therefore, the data were insufficient to assess generalization performance in men. Further validation in an external population with more balanced sex and class distributions is required.

Additional study limitations include the development sample size and class imbalance [24,25]; external validation ideally requires at least approximately 100 events [26], whereas the external data in this study comprised 19–69 participants per class. Bootstrap CIs reflect the sampling variability of the fixed and original model outputs and do not include retraining variability. In this analysis, retraining variability across different random seeds was not evaluated. Clinical utility (e.g., decision curve analysis) was not evaluated; therefore, the clinical benefit cannot be judged from ROC-AUC or accuracy alone [18,27]. The effects on the diagnostic time, inter-clinician agreement, and patient outcomes were also not evaluated. Future studies should follow the TRIPOD+AI reporting principles to clearly report participant selection and diagnostic definitions and conduct further validation [28].

Even in the original model, the mean top-class probability exceeded observed accuracy, and class-specific calibration slopes (0.24–0.32 in development data, 0.28–0.61 in external data) demonstrated overly extreme probabilities [29,30]. Misclassification occurred even at top-class probabilities of 0.999 or higher, in eight of 236 development participants and four of 103 external participants. Using these probabilities for clinical decisions would require fitting a calibration method (e.g., temperature scaling) in a separate validation procedure using the development data, with re-evaluation of the calibration using independent external data.

Conclusions

This study developed and externally validated an MLP-based model for four-class sarcopenia classification. Although the model demonstrated high classification accuracy, it did not outperform direct application of the guideline-based classification rules. These findings suggest that machine learning models trained on labels defined by the same input variables may primarily reproduce existing diagnostic rules rather than provide additional clinical value. Further validation using independent reference standards and more diverse populations is necessary to establish the potential clinical utility of the model.

List of Abbreviations

AWGS: Asian Working Group for Sarcopenia
CIs: confidence intervals
ECE: expected calibration error
EWGSOP2: European Working Group on Sarcopenia in Older People 2
FNIH: Foundation for the National Institutes of Health
IRB: institutional review board
KWGS: Korean Working Group on Sarcopenia
MLP: multilayer perceptron
OOF: out-of-fold
ReLU: rectified linear unit
ROC-AUC: receiver operating characteristic-area under the curve
SD: standard deviation
SMD: standardized mean difference
SMI: skeletal muscle index
SPPB: Short Physical Performance Battery

Declarations

Ethics approval and consent to participate: This study was reviewed and approved by the Institutional Review Board (IRB) of Wonkwang University Hospital (IRB no. WKUH 2023-08-031;

approval date, September 8, 2023) and conducted in accordance with the Declaration of Helsinki. All participants provided written informed consent prior to enrollment.

Consent for publication: Not applicable.

Availability of data and materials: The datasets analyzed in this study are not publicly available due to institutional and IRB restrictions on patient-level clinical data, but are available from the corresponding author on reasonable request and with permission of the Wonkwang University Hospital IRB.

Competing interests: The authors declare that they have no competing interests.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (RS2021NR60106).

Authors' contributions: S.-H.N. and D.-W.L developed and validated the machine learning model. J.H.K. collected the data and established the clinical classification criteria. D.-W.L. developed and managed the data collection environment. S.-K.C. and C.-W.J. designed and supervised the study and drafted the manuscript. All authors read, critically revised, and approved the final manuscript.

Acknowledgments:

Not applicable

References

1. Cruz-Jentoft AJ, Bahat G, Bauer J, Boirie Y, Bruyère O, Cederholm T, et al. Sarcopenia: revised European consensus on definition and diagnosis. Age Ageing. 2019;48:16–31.

2. Chen LK, Woo J, Assantachai P, Auyeung TW, Chou MY, Iijima K, et al. Asian Working Group for Sarcopenia: 2019 consensus update on sarcopenia diagnosis and treatment. J Am Med Dir Assoc. 2020;21:300–7.e2.

3. Baek JY, Jung HW, Kim KM, Kim M, Park CY, Lee KP, et al. Korean Working Group on Sarcopenia Guideline: Expert consensus on sarcopenia screening and diagnosis by the Korean Society of Sarcopenia, the Korean Society for Bone and Mineral Research, and the Korean Geriatrics Society. Ann Geriatr Med Res. 2023;27:9–21.

4. Cruz-Jentoft AJ, Sayer AA. Sarcopenia. Lancet. 2019;393:2636–46.

5. Petermann-Rocha F, Balntzi V, Gray SR, Lara J, Ho FK, Pell JP, et al. Global prevalence of sarcopenia and severe sarcopenia: a systematic review and meta-analysis. J Cachexia Sarcopenia Muscle. 2022;13:86–99.

6. Cruz-Jentoft AJ, Baeyens JP, Bauer JM, Boirie Y, Cederholm T, Landi F, et al. Sarcopenia: European consensus on definition and diagnosis: report of the European Working Group on Sarcopenia in Older People. Age Ageing. 2010;39:412–23.

7. Studenski SA, Peters KW, Alley DE, Cawthon PM, McLean RR, Harris TB, et al. The FNIH sarcopenia project: rationale, study description, conference recommendations, and final estimates. J Gerontol A Biol Sci Med Sci. 2014;69:547–58.

8. Guralnik JM, Simonsick EM, Ferrucci L, Glynn RJ, Berkman LF, Blazer DG, et al. A short physical performance battery assessing lower extremity function: association with self-reported disability and prediction of mortality and nursing home admission. J Gerontol. 1994;49:M85–94.

9. Pavasini R, Guralnik J, Brown JC, di Bari M, Cesari M, Landi F, et al. Short Physical Performance Battery and all-cause mortality: systematic review and meta-analysis. BMC Med. 2016;14:215.

10. Bohannon RW. Grip strength: an indispensable biomarker for older adults. Clin Interv Aging. 2019;14:1681–91.

11. Zhang H, Yin M, Liu Q, Ding F, Hou L, Deng Y, et al. Machine and deep learning-based clinical characteristics and laboratory markers for the prediction of sarcopenia. Chin Med J (Engl). 2023;136:967–73.

12. Luo X, Ding H, Broyles A, Warden SJ, Moorthi RN, Imel EA. Using machine learning to detect sarcopenia from electronic health records. Digit Health. 2023;9:20552076231197098.

13. Bae JH, Seo JW, Kim DY. Deep-learning model for predicting physical fitness in possible sarcopenia: analysis of the Korean physical fitness award from 2010 to 2023. Front Public Health. 2023;11:1241388.

14. Seok M, Kim W. Sarcopenia prediction for elderly people using machine learning: a case study on physical activity. Healthcare (Basel). 2023;11:1334.

15. Lee J, Yoon Y, Kim J, Kim YH. Metaheuristic-based feature selection methods for diagnosing sarcopenia with machine learning algorithms. Biomimetics (Basel). 2024;9:179.

16. Zupo R, Moroni A, Castellana F, Gasparri C, Catino F, Lampignano L, et al. A machine-learning approach to target clinical and biological features associated with sarcopenia: findings from northern and southern Italian aging populations. Metabolites. 2023;13:565.

17. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. J Big Data. 2019;6:27.

18. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21:128–38.

19. Fisher A, Rudin C, Dominici F. All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. J Mach Learn Res. 2019;20:1–81.

20. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics. 2010;26:1340–7.

21. Breiman L. Random forests. Mach Learn. 2001;45:5–32.

600 22. Dietterich TG. Approximate statistical tests for comparing supervised classification learning
601 algorithms. Neural Comput. 1998;10:1895–923.

602 23. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic
603 review shows no performance benefit of machine learning over logistic regression for clinical
604 prediction models. J Clin Epidemiol. 2019;110:12–22.

605 24. Varoquaux G. Cross-validation failure: small sample sizes lead to large error bars. Neuroimage.
606 2018;180:68–77.

607 25. Riley RD, Ensor J, Snell KIE, Harrell FE Jr, Martin GP, Reitsma JB, et al. Calculating the sample
608 size required for developing a clinical prediction model. BMJ. 2020;368:m441.

609 26. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of
610 a multivariable prognostic model: a resampling study. Stat Med. 2016;35:214–26.

611 27. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models.
612 Med Decis Making. 2006;26:565–74.

613 28. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI
614 statement: updated guidance for reporting clinical prediction models that use regression or
615 machine learning methods. BMJ. 2024;385:e078378.

616 29. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the
617 Achilles heel of predictive analytics. BMC Med. 2019;17:230.

618 30. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. Proc Mach
619 Learn Res. 2017;70:1321–30.

620

621 Table 1. Model parameters

Parameter	**Setting**
Model	Multilayer perceptron
Input variables	Sex, height, SMI, mean grip, SPPB
Architecture	5–36–28–9–4
Architecture selection	Pre-specified; no automated hyperparameter search
Hidden activation	ReLU
Output activation	Softmax
Loss function	Sparse categorical cross-entropy
Optimizer	NAdam
Learning rate	0.001
Epochs	300 (pre-specified)
Batch size	10
Early stopping	Not used
Class weighting	Balanced
Dropout	None
Random seed	42

Parameter	Setting
Internal validation	Stratified five-fold CV
Permutation repeats	10 per feature per fold

622

623

624 **Figure Captions**

625 Figure 1. Study design and model evaluation.

626 The development cohort comprised 339 participants. Internal evaluation was based on stratified five-
627 fold cross-validation with fold-specific preprocessing. The final model and scaler were fitted on the
628 full development cohort and frozen before external evaluation. The external dataset comprised 151
629 participants across three institutions, with one observation per participant.

630 Figure 2. Class distribution in the development cohort (N = 339).

631 Labels indicate participant counts and percentages.

632 Figure 3. Internal five-fold out-of-fold confusion matrix (N = 339).

633 Cells show counts and row percentages. Color indicates the proportion within each observed class;
634 rows sum to 100%.

635 Figure 4. One-vs.-rest ROC curves calculated from saved internal out-of-fold probabilities for 339
636 participants.

637 Macro AUC is the unweighted mean of the four class-specific AUCs. Dashed diagonal: chance
638 discrimination.

639 Figure 5. Permutation feature importance.

Points represent the mean decrease in validation accuracy; error bars show the standard deviation across five fold-level means. Each fold-level mean was calculated from 10 permutations per feature. Values are model dependence measures, not causal effects.

Figure 6. Final model training loss over 300 epochs, redrawn from the saved epoch-level history.

The final model was fitted for all 339 development participants. These are training set values and do not confirm generalization performance. No smoothing was applied.

Figure 7. Final model training accuracy over 300 epochs, redrawn from the saved epoch-level history. The final model was fitted for all 339 development participants. These are training set values and do not confirm generalization performance. No smoothing was applied.

Figure 8. External confusion matrix for all 151 participants, with one observation per participant. Cells show counts and row percentages. Color indicates the proportion within each observed class.

Figure 9. External one-vs.-rest ROC curves calculated from frozen H5 model probabilities for all 151 participants.

There was one observation per participant. Macro AUC is the unweighted mean of class-specific AUCs.

Figure 10. Internal out-of-fold performance (339 participants) and external performance (151 participants).

Error bars show 95% percentile bootstrap confidence intervals from 5,000 participant-level resamples per dataset. Models were not refitted during resampling. The score axis is restricted; no formal between-cohort hypothesis test was performed.

Figure 11. Guideline rules applied directly to the model inputs versus trained models.

MLP values are those reported in Sections 3.2 and 3.5. Error bars show 95% percentile bootstrap confidence intervals from 5,000 participant-level resamples.

Figure 12. Mean top-class probability and observed accuracy by probability bin for the original MLP. Internal results are based on saved OOF probabilities (N=339); external results are based on frozen H5 model outputs (N=151). Bins have width 0.1; only occupied bins are shown, and the final bin includes 1. Numbers above bars indicate participants per bin.

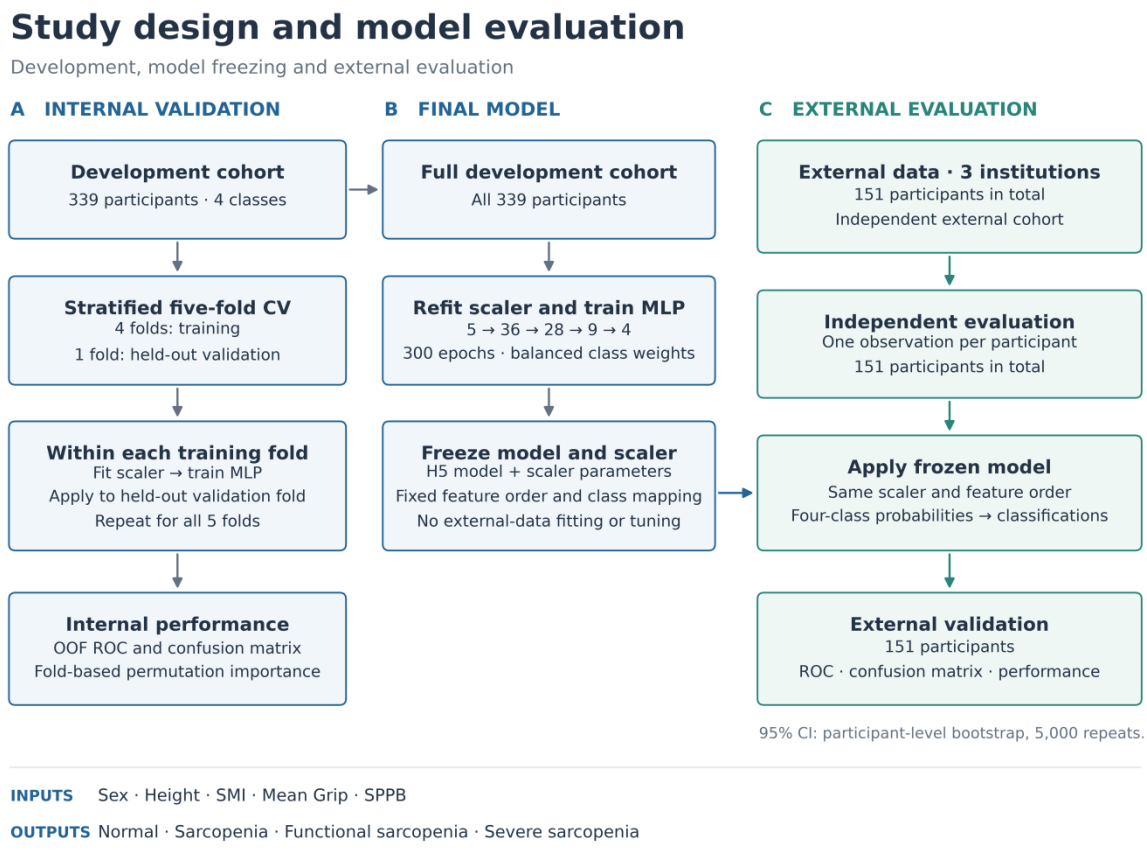

Study design and model evaluation
Development, model freezing and external evaluation
A INTERNAL VALIDATION
B FINAL MODEL
C EXTERNAL EVALUATION
Development cohort
339 participants · 4 classes
Full development cohort
All 339 participants
External data · 3 institutions
151 participants in total
Independent external cohort
Stratified five-fold CV
4 folds: training
1 fold: held-out validation
Refit scaler and train MLP
5 → 36 → 28 → 9 → 4
300 epochs · balanced class weights
Independent evaluation
One observation per participant
151 participants in total
Within each training fold
Fit scaler → train MLP
Apply to held-out validation fold
Repeat for all 5 folds
Freeze model and scaler
H5 model + scaler parameters
Fixed feature order and class mapping
No external-data fitting or tuning
Apply frozen model
Same scaler and feature order
Four-class probabilities → classifications
Internal performance
OOF ROC and confusion matrix
Fold-based permutation importance
External validation
151 participants
ROC · confusion matrix · performance
95% CI: participant-level bootstrap, 5,000 repeats.
INPUTS Sex · Height · SMI · Mean Grip · SPPB
OUTPUTS Normal · Sarcopenia · Functional sarcopenia · Severe sarcopenia

Figure 1

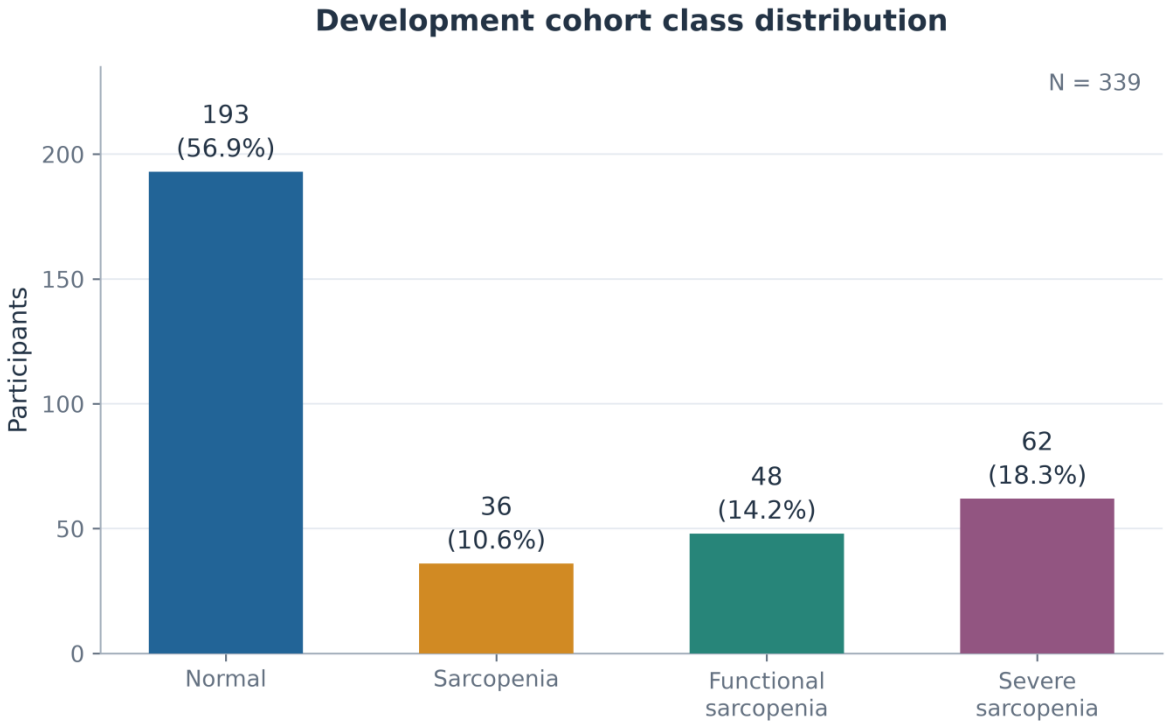

Development cohort class distribution
N = 339
Participants
0
50
100
150
200
193
(56.9%)
36
(10.6%)
48
(14.2%)
62
(18.3%)
Normal
Sarcopenia
Functional sarcopenia
Severe sarcopenia

Figure 2

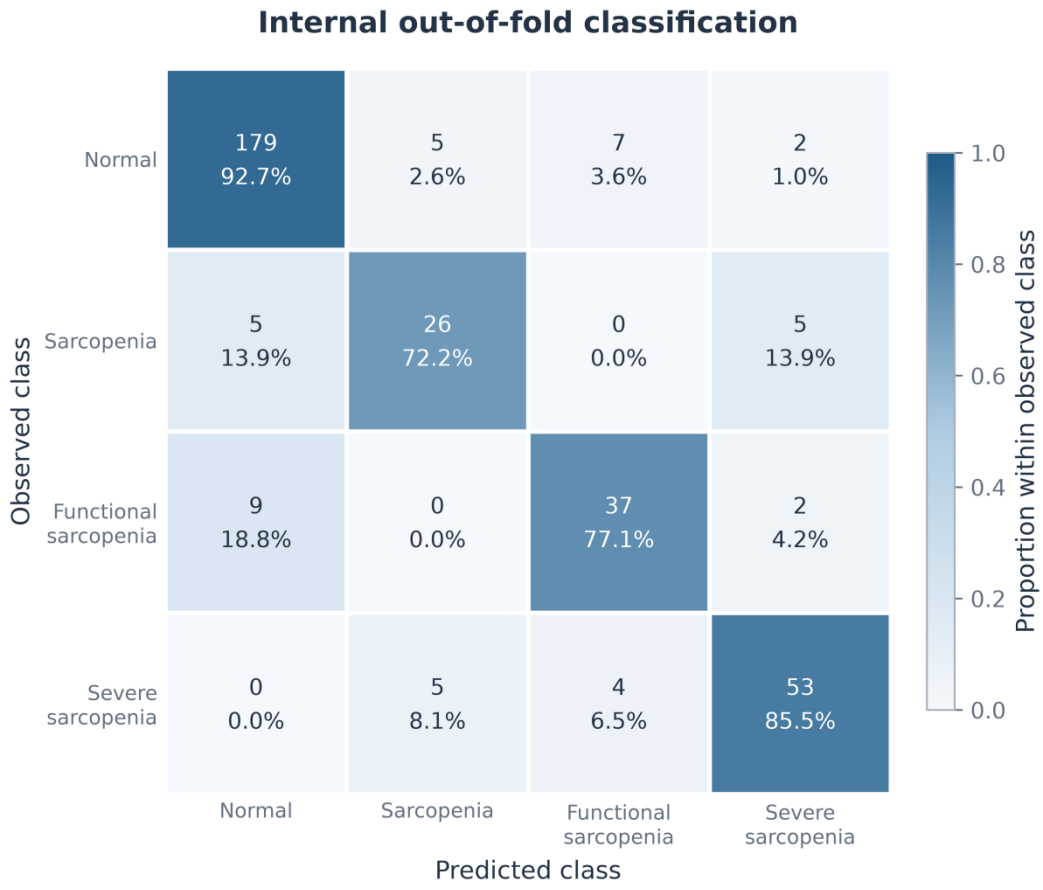

Internal out-of-fold classification
Normal
Sarcopenia
Functional sarcopenia
Severe sarcopenia
Observed class
179
92.7%
5
2.6%
7
3.6%
2
1.0%
5
13.9%
26
72.2%
0
0.0%
5
13.9%
9
18.8%
0
0.0%
37
77.1%
2
4.2%
0
0.0%
5
8.1%
4
6.5%
53
85.5%
Normal
Sarcopenia
Functional sarcopenia
Severe sarcopenia
Predicted class
1.0
0.8
0.6
0.4
0.2
0.0
Proportion within observed class

Figure 3

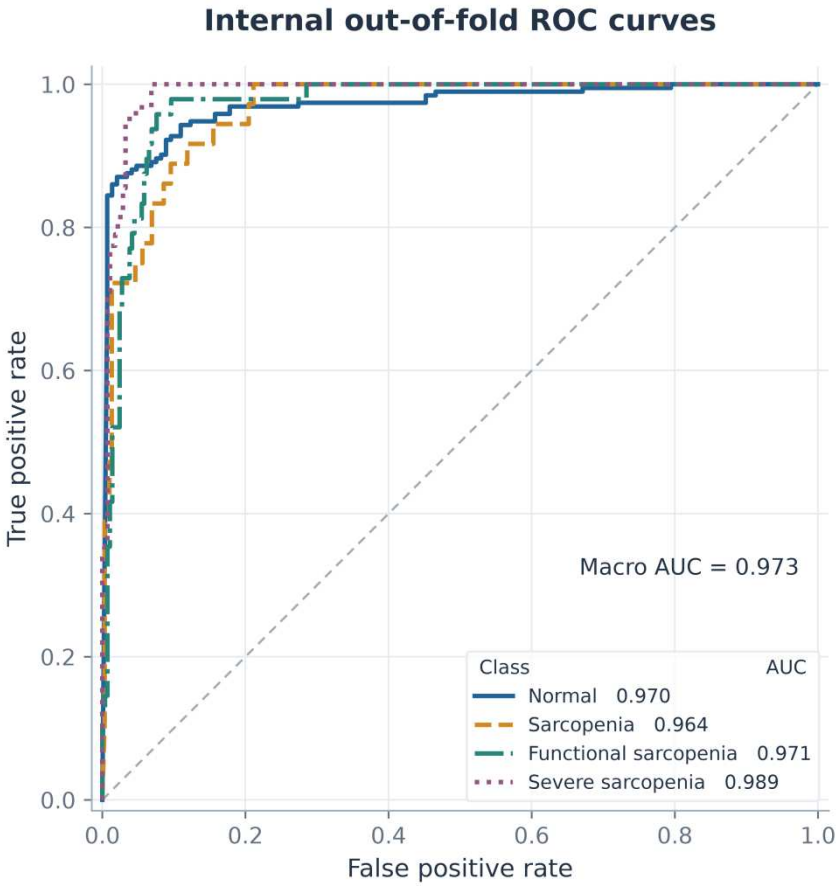

Internal out-of-fold ROC curves
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Macro AUC = 0.973
Class
AUC
Normal 0.970
Sarcopenia 0.964
Functional sarcopenia 0.971
Severe sarcopenia 0.989

Figure 4

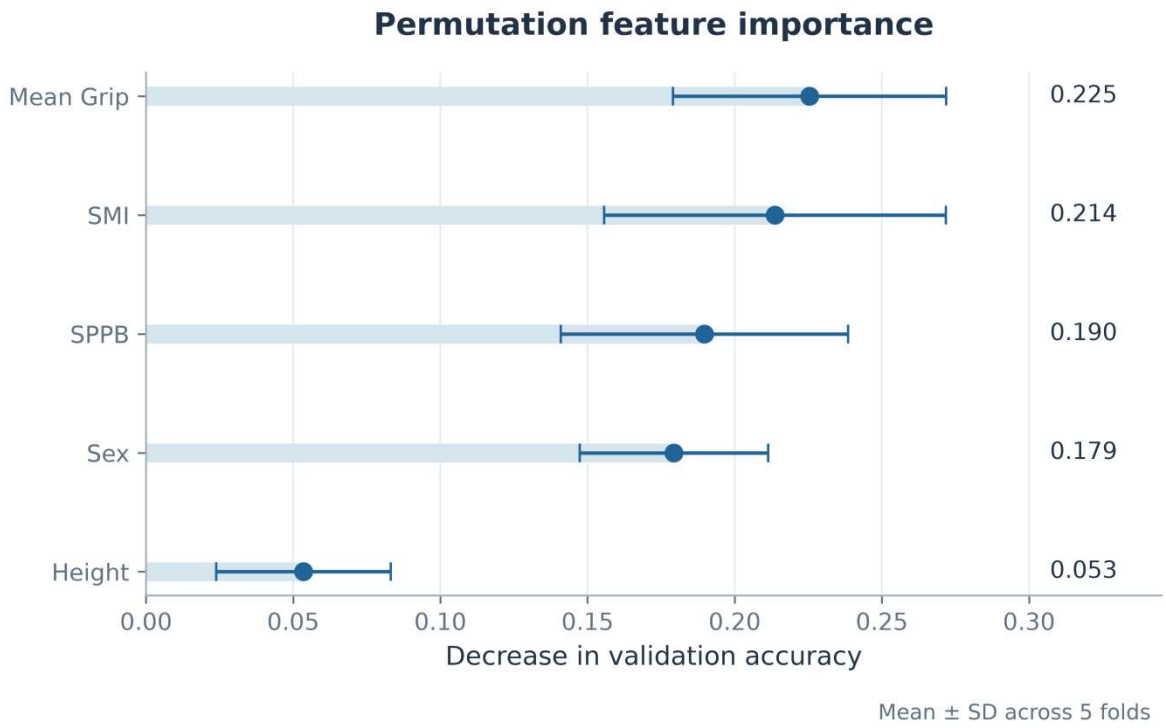

Permutation feature importance
Mean Grip
SMI
SPPB
Sex
Height
0.225
0.214
0.190
0.179
0.053
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Decrease in validation accuracy
Mean ± SD across 5 folds

Figure 5

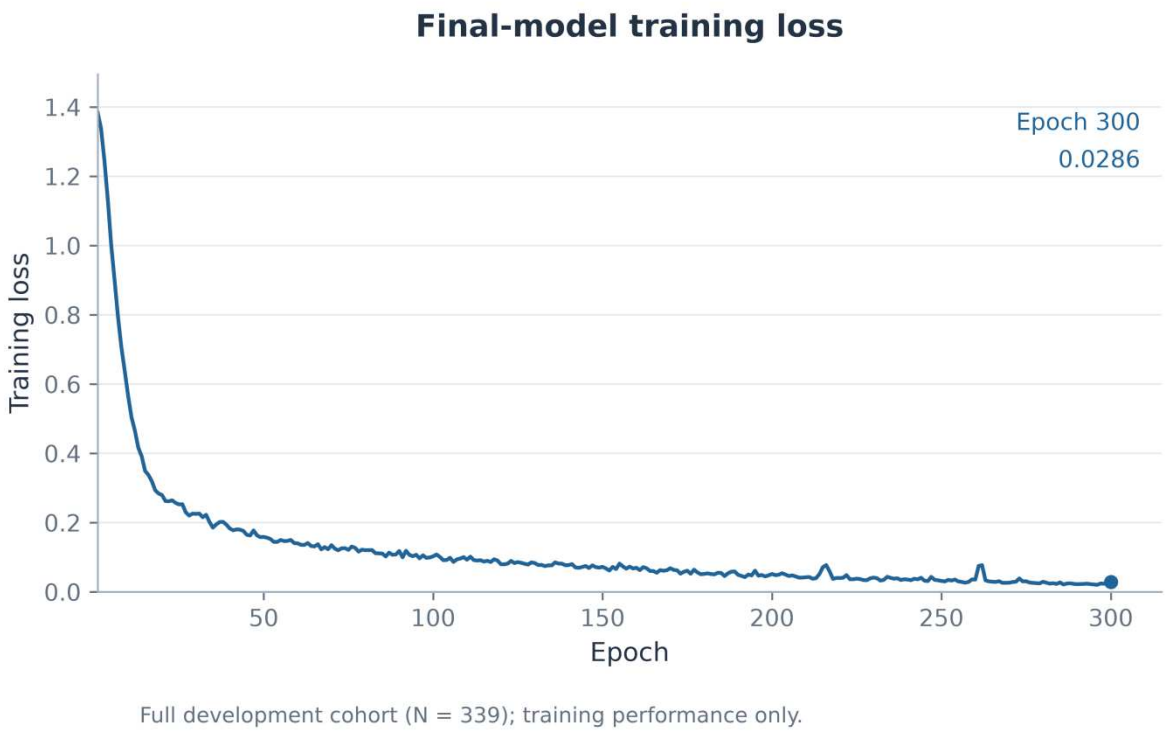

Final-model training loss
Epoch 300
0.0286
Training loss
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
50
100
150
200
250
300
Epoch
Full development cohort (N = 339); training performance only.

Figure 6

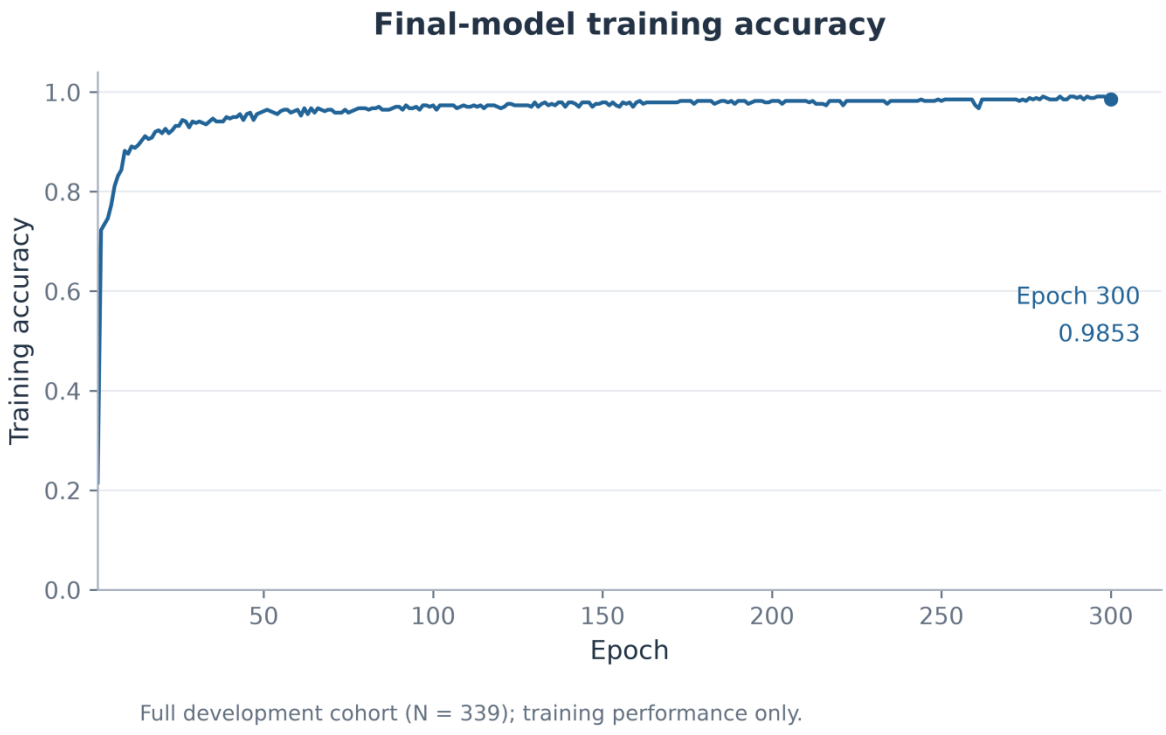

Final-model training accuracy
Training accuracy
1.0
0.8
0.6
0.4
0.2
0.0
50
100
150
200
250
300
Epoch
Epoch 300
0.9853
Full development cohort (N = 339); training performance only.

Figure 7

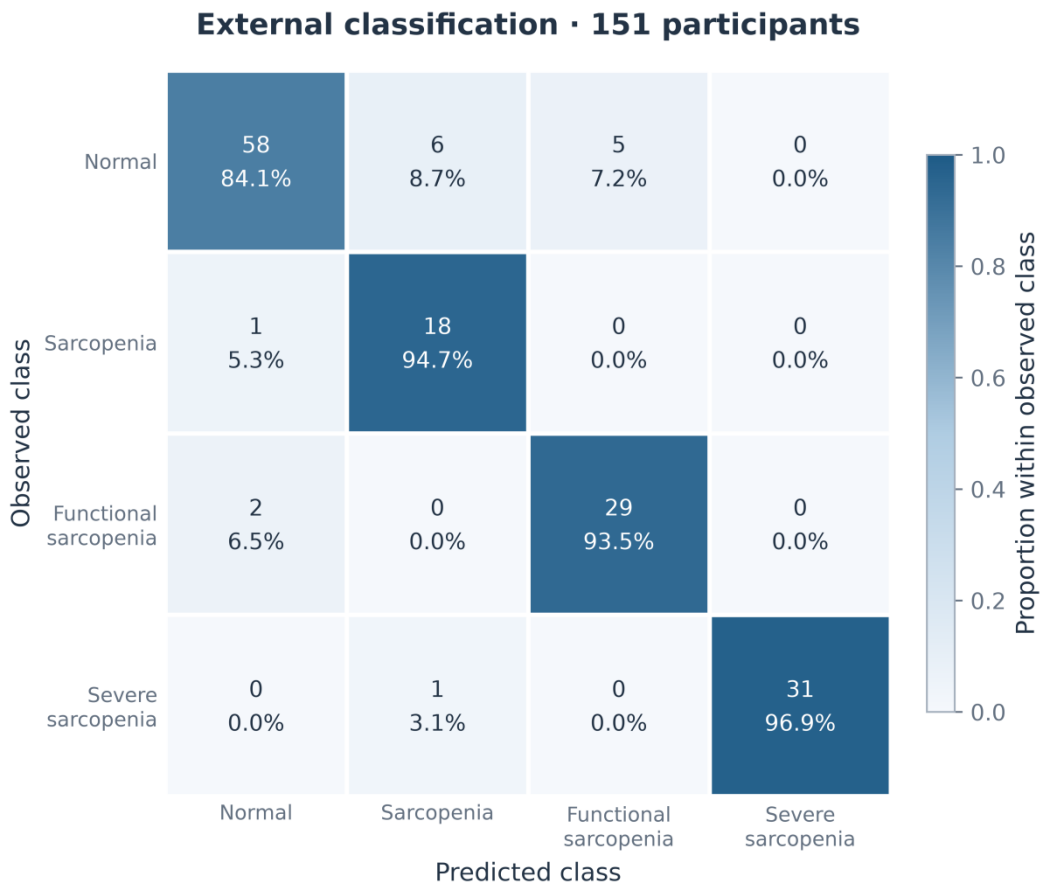

External classification · 151 participants
Normal
Sarcopenia
Functional sarcopenia
Severe sarcopenia
Observed class
58
84.1%
6
8.7%
5
7.2%
0
0.0%
1
5.3%
18
94.7%
0
0.0%
0
0.0%
2
6.5%
0
0.0%
29
93.5%
0
0.0%
0
0.0%
1
3.1%
0
0.0%
31
96.9%
Normal
Sarcopenia
Functional sarcopenia
Severe sarcopenia
Predicted class
1.0
0.8
0.6
0.4
0.2
0.0
Proportion within observed class

Figure 8

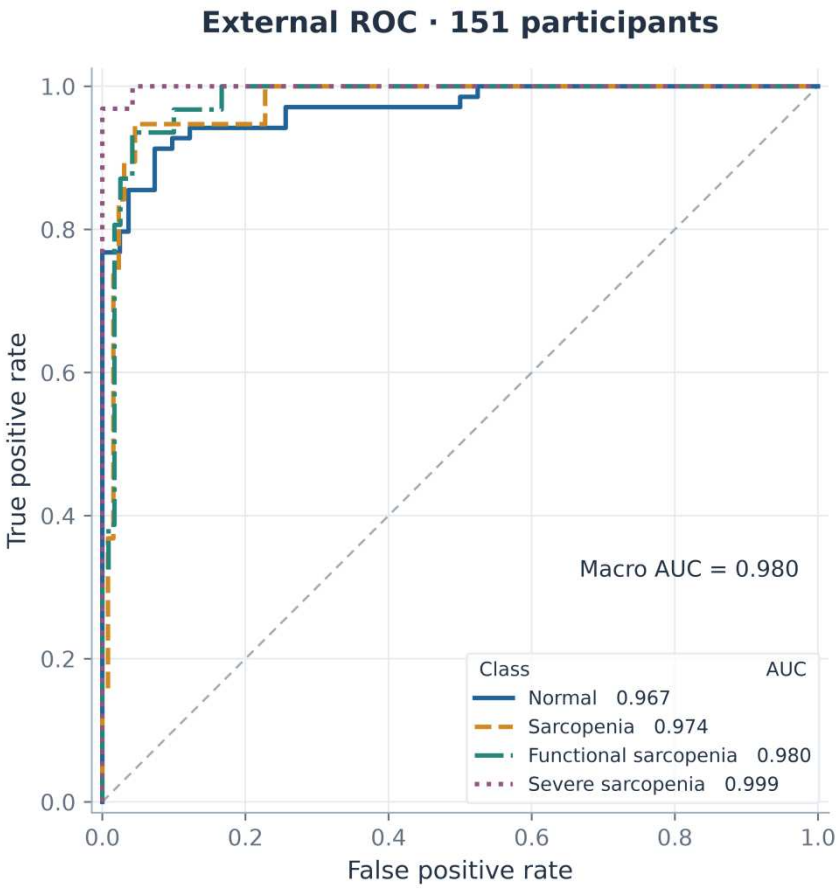

External ROC · 151 participants
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Macro AUC = 0.980
Class
AUC
Normal 0.967
Sarcopenia 0.974
Functional sarcopenia 0.980
Severe sarcopenia 0.999

Figure 9

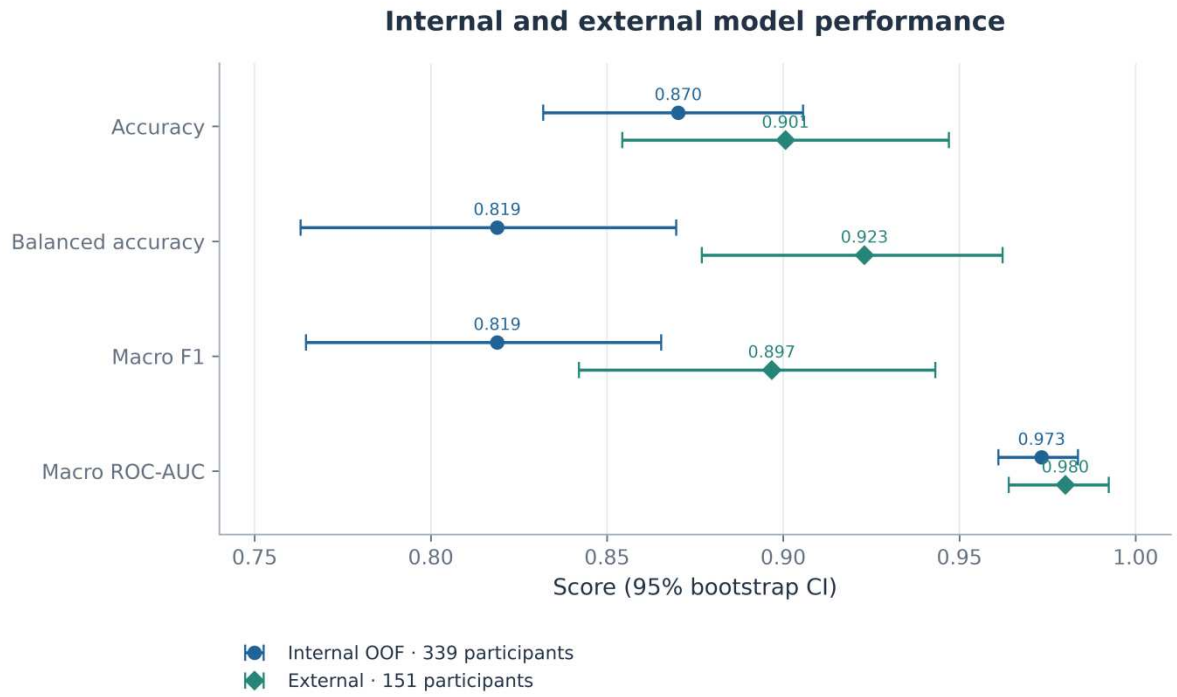

Internal and external model performance
Accuracy
Balanced accuracy
Macro F1
Macro ROC-AUC
0.870
0.901
0.819
0.923
0.819
0.897
0.973
0.980
0.75
0.80
0.85
0.90
0.95
1.00
Score (95% bootstrap CI)
Internal OOF · 339 participants
External · 151 participants

Figure 10

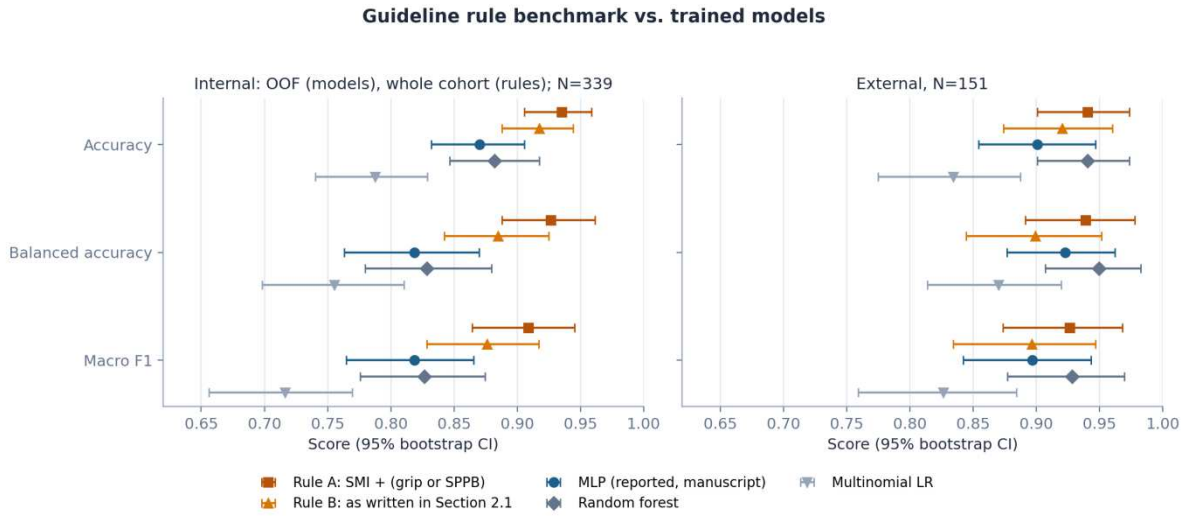

Guideline rule benchmark vs. trained models
Internal: OOF (models), whole cohort (rules); N=339
External, N=151
Accuracy
Balanced accuracy
Macro F1
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00
Score (95% bootstrap CI)
Rule A: SMI + (grip or SPPB)
Rule B: as written in Section 2.1
MLP (reported, manuscript)
Random forest
Multinomial LR

Figure 11

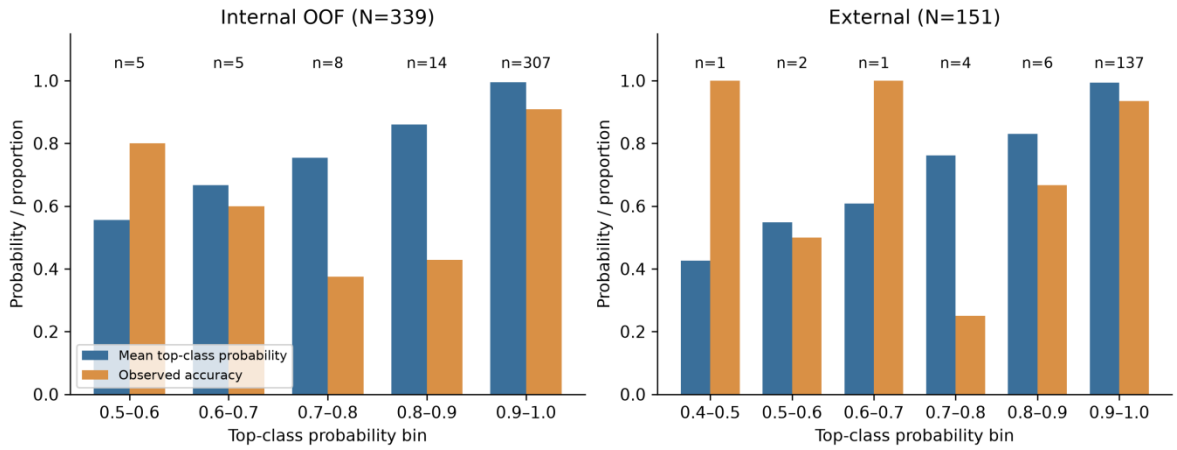

Internal OOF (N=339)
External (N=151)
Probability / proportion
Top-class probability bin
Mean top-class probability
Observed accuracy
n=5
n=5
n=8
n=14
n=307
n=1
n=2
n=1
n=4
n=6
n=137
0.0
0.2
0.4
0.6
0.8
1.0
0.4–0.5
0.5–0.6
0.6–0.7
0.7–0.8
0.8–0.9
0.9–1.0

Figure 12