*This manuscript is currently under peer review at Surgical and Radiologic Anatomy.*

*Under review at Surgical and Radiologic Anatomy*

*(Springer Nature)*

# Title

## Anatomical variations of the peroneus brevis tendon on axial MRI: A validated cross-sectional classification at the lateral malleolus

Rafał Zych[1], Dan Mocanu[2,3], Ymer Hagberg[2,3], Katarzyna Bokwa-Dąbrowska[2,3], Michael Huuskonen[2], Isaac Romanus[2], Dawid Dziedzic[1], Pawel Szaro[1,2,3]

1. Department of Clinical and Descriptive Anatomy, Medical University of Warsaw, Poland

2. Department of Musculoskeletal Radiology, Sahlgrenska University Hospital, Gothenburg, Sweden

3. Department of Radiology, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden.

# Title

## Anatomical variations of the peroneus brevis tendon on axial MRI: A validated cross-sectional classification at the lateral malleolus

Rafał Zych[1], Dan Mocanu[2,3], Ymer Hagberg[2,3], Katarzyna Bokwa-Dąbrowska[2,3], Michael Huuskonen[2], Isaac Romanus[2], Dawid Dziedzic[1], Pawel Szaro[1,2,3]

1. Department of Clinical and Descriptive Anatomy, Medical University of Warsaw, Poland

2. Department of Musculoskeletal Radiology, Sahlgrenska University Hospital, Gothenburg, Sweden

3. Department of Radiology, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden.

# Abstract

**Purpose**

The shape of the peroneus brevis tendon may contribute to the development of split tears, but a classification of its anatomical variants on transverse section in vivo is lacking. This study aimed to validate a proposed classification system for peroneus brevis tendon morphology on the transverse cross section on magnetic resonance imaging (MRI).

**Methods**

We analysed 130 normal peroneus brevis tendons (power >0.8, $\alpha = 0.05$, effect size = 0.31) on consecutively selected ankle MRI performed at 3T. We excluded patients with fractures, ligament injuries, or peroneal tendon pathology. Seven independent raters classified tendons

into four morphological categories: general flat, flat with a lateral bulge, flat with a medial bulge, and oval. Inter-rater reliability was assessed using Cohen's kappa, Gwet's AC1, and Fleiss' kappa. We evaluated classification robustness using F1 score, receiver operating characteristic analysis, and precision-recall metrics.

**Results**

Consensus classification showed 37.7% of tendons were general flat, 20.8% flat with a lateral bulge, 19.2% flat with a medial bulge, and 22.3% oval. Inter-rater agreement was substantial, with Fleiss' kappa of 0.69 (95% CI 0.63–0.75). Cohen's kappa ranged from 0.59 to 0.87, and Gwet's AC1 from 0.58 to 0.87, confirming strong pairwise agreement. Classification performance was high (area under the curve = 0.89), and majority voting further improved agreement (F1 score = 0.93). The most common misclassification occurred between oval and flat with a medial bulge.

**Conclusion**

This proposed classification is robust and reliable, offering a validated framework for consistent tendon evaluation in both clinical and research settings.

## Key words

Peroneus brevis tendon, Tendon morphology, Magnetic resonance imaging, Tendon classification, Inter-rater reliability, Ankle anatomy

# Introduction

The peroneus brevis tendon is morphologically variable, particularly in its distal insertion. Numerous anatomical studies - ranging from classical dissections to modern systematic analyses - have documented a broad spectrum of these insertional variants[24, 37]. MacAlister (1875) was among the first to describe additional slips to the abductor digiti minimi and peroneus longus tendons[19]. Le Double (1897) later described up to eight types of distal insertions[18], and Bhargava et al. (1961) further reported tendon slips inserting onto the fifth metatarsal head, cuboid, calcaneus, and even the distal phalanx of the fifth toe[4].

More recent cadaveric and foetal studies have attempted to formalize this variability. Olewnik et al. proposed a two-type system based on classical dissection of adult limbs: a single distal insertion (Type I) and bifurcated insertions (Type II), with accessory bands inserting dorsally or fusing with other tendons such as the peroneus tertius[25]. Zielinska et al. expanded this into a three-type classification in human foetuses, identifying a rare, trifurcated configuration (Type III), and further highlighted that these variants are present from early developmental stages[37]. These findings were recently synthesized in a comprehensive review and new classification system, which consolidates adult and foetal variations and proposes clinical implications for tendon repair, graft harvesting, and surgical planning[37]. These investigations have provided an exceptionally detailed picture of the distal anatomy of the peroneus brevis tendon. However, their focus remains on the terminal insertions—typically at or near the fifth metatarsal—rather than the proximal segment of the tendon. The cross-sectional shape of the tendon, particularly at the level of the lateral malleolus, has not been systematically studied, despite this being the region where longitudinal split tears most frequently originate.

A peroneus brevis split tear is a common but often overlooked pathology by both radiologists and clinicians at the level of the lateral malleolus [5, 15, 30, 32]. Despite the clinical significance, its diagnosis remains challenging due to the complex anatomical relationship of the peroneal tendons and the inherent limitations of the available imaging modalities [10]. The peroneus brevis and peroneus longus tendons lie adjacent to each other, tightly confined within the retromalleolar groove and stabilised posteriorly by the superior peroneal retinaculum [6, 21]. Tight anatomical relationships between the content of the superior peroneal tunnel [2] make it difficult to assess peroneus brevis pathology on ultrasound and magnetic resonance imaging (MRI). Moreover, given the lack of large-scale epidemiological studies, the exact incidence of peroneus brevis split tears remains unknown. Reports suggest that a substantial proportion of cases may go undiagnosed, further highlighting the limitations of the current diagnostic approaches[5].

Peroneus brevis split tears vary, ranging from complete tears, where the tendon is fully divided into two parts, to non-complete tears, where only a part of the cross-section area is torn[7]. Non-complete tears are particularly difficult to detect with imaging, as their radiological features may overlap with normal anatomical variations. One proposed imaging marker is an alteration in tendon shape on imaging [20, 31]. Tendon shapes described by some authors as ‘chevron’ or ‘bisected’ [20] have been suggested to be associated with split tears. Nevertheless, these descriptions remain unclear and lack precise morphological definitions that can be applied systematically in future studies. However, no studies have examined the cross sectional shape of the peroneus brevis at the level where split tears most commonly occur[16].

Despite clinical observations, there is no consensus in the literature regarding the normal anatomical variability of the peroneus brevis tendon shape on the cross section. Its shape has been hypothesised to influence the susceptibility to split tears, with some authors suggesting

that flatter peroneus brevis tendons are at greater risk due to compression by the peroneus longus against the malleolar sulcus[20]. However, there is no validated classification system for cross sectional peroneus brevis morphology on MRI.

In contrast, the morphology of the retromalleolar groove, a low-lying muscle belly, or the presence of the peroneus quartus has been studied extensively [12, 21, 22, 29, 33, 34]. Variability in its depth and contour has been linked to peroneal tendon instability, emphasising the role of anatomical predisposition in tendon pathology.

Our study addresses this gap by proposing and validating a novel classification system for the cross-sectional shape of the peroneus brevis tendon based on in vivo MRI. This region is clinically important, as the tendon lies in close proximity to both the fibula and peroneus longus tendon, and is subject to mechanical forces that may predispose it to longitudinal splitting [2, 11]. Unlike existing classifications, which emphasize distal branching patterns, our system is based on transverse MRI sections captured at the lateral malleolus—the precise site where split tears often begin.

Despite extensive anatomical research on distal insertional variants of the peroneus brevis tendon, its shape in axial cross-section at the level of the lateral malleolus remains unclassified. This lack of a standardized system represents a significant gap in musculoskeletal radiological anatomy. Without clear imaging criteria, normal anatomical variants may be mistaken for pathology, and true pathological changes may be overlooked. Given the clinical importance of accurate interpretation, any proposed classification must be validated for consistency and applicability. To our knowledge, no prior study has systematically classified the cross-sectional morphology of the peroneus brevis tendon in vivo on MRI. We therefore aimed to evaluate the reliability of our proposed classification system, identify common misclassification patterns, and explore factors influencing diagnostic accuracy.

# Methods

## Study design and observer validation framework

This study evaluated the inter-rater agreement regarding a proposed classification of the peroneus brevis tendon form on the transverse section at the level of the lateral malleolus. This study was conducted and reported in accordance with the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) [14].

## MRI examinations

The dataset was compiled retrospectively from MRI examinations performed at Sahlgrenska University Hospital in 2021–2024 with a 3T machine. Overall, 494 ankle MRI scans were re-evaluated. Of these, 360 were excluded for the following reasons: fractures involving the lateral malleolus (n = 87), lateral ligament injuries or pathological changes in the peroneus tendons or lateral malleolar pain on the referral letter (n = 147), a history of peroneus tendon pathology or postoperative changes in lateral ligaments or peroneal tendons (n = 121), infections (n = 6), and ganglion cysts impacting the peroneal tendons (n = 3). Thus, a total of 130 patients were included and analysed.

## Raters

A structured visual assessment approach was used for grading. Seven independent raters with different backgrounds (2 musculoskeletal radiologists with 7 and 10 years of experience in musculoskeletal radiology, 1 radiology resident, 1 medical doctor, 2 physiotherapists, and 1

fifth-year medicine student), visually assessed each tendon and assigned one of four predefined variations based on morphology. The classification system was developed from a preliminary study by our team and refined through pilot consensus rounds. To establish a reference standard, a consensus classification agreement among musculoskeletal radiologists.

To assess inter-rater reliability, each rater independently classified all cases in a blinded manner. Intra-rater reliability was assessed after a washout period. Because assessing intra-rater variability using a subset of data is a valid approach that has been used in previous studies [1], a subset comprising 30% of the subjects was randomly selected by an individual not involved in the study. The same raters re-evaluated these cases after a 3-week washout period.

## Data acquisition and preprocessing

### Sampling

Convenience sampling was used to select all available MRI scans of the ankle that met the inclusion criteria. A power analysis was performed prior to the study to determine the sample size. With power > 0.8, alpha = 0.05, an effect size of 0.31 based on our preliminary study and a 10% buffer, 130 patients were required.

### Preprocessing

Analysis of the peroneus brevis tendon shape was based on proton density–weighted axial images. A senior musculoskeletal radiologist selected a representative image at the level of the lateral malleolus, inferior to the syndesmosis, at the level of the ankle joint space. This level was chosen based on our preliminary study, which demonstrated that the tendon shape remains consistent within this region. The selected images were exported, and an electronic

form was created, including the images and four classification options (general flat, flat with a medial bulge, flat with a lateral bulge, and oval tendon), which was then distributed to the raters. Each rater received detailed instructions including definitions of tendon forms and examples (not including the evaluated cases) regarding tendon forms and assessed tendons using a standardised protocol for the study.

## Proposed classification

The proposed classification (nominal variable) is based on the visual assessment of the ratio of thickness (the shortest diameter) to the width of the tendon (the longest diameter). The four types are defined below.

### General flat tendon

The thickness of the tendon (defined as the anteroposterior [AP] dimension) is relatively consistent across its length. The width of the tendon is significantly greater than its thickness. The tendon in this form can be either straight, curved, or folded.

### General flat with a lateral bulge

The tendon is flattened, but the lateral edge bulges, increasing the thickness in the lateral part.

### General flat with a medial bulge

The tendon is flattened, but the medial edge bulges, resulting in an increase in thickness in the medial part.

### Oval tendon

The peroneus brevis tendon takes the shape of an oval, resembling the cross-section of the peroneus longus tendon. The thickness is slightly smaller than the width.

## Gold standard

The final result was a consensus between two musculoskeletal radiologists, which was considered to be the gold standard.

## Statistical analysis

R version 4.4.3 (R Core Team, Vienna, Austria) and RStudio version 2024.12.1+563 (2024.12.1+563) were used for statistical analysis and data visualisation. The ggplot2 package was used to visualise the data. Inter-rater agreement across the seven raters was assessed based on Fleiss' kappa with a bootstrapped 95% confidence interval. For pairwise agreement, Cohen's kappa and Gwet's AC1 were calculated between all rater pairs. These statistics were also used to assess intra-rater reliability after a 3-week washout period. Gwet's AC1 was included as a robust alternative to Cohen's kappa, particularly in light of potential prevalence and marginal distribution imbalances.

The F1 score, defined as the harmonic mean of precision and recall, was calculated for each rater to assess the performance of the proposed classification. This metric reflects the balance between correctly identifying tendon classifications (precision) and detecting all cases of a given class (recall). The F2 score ($\beta = 2$) was calculated to account for cases where recall was prioritised. In addition, a majority vote approach was applied, where the most frequently assigned tendon form among the seven raters was selected as the predicted class. Then, the

majority vote classifications were compared to the gold standard (defined as a consensus between two musculoskeletal radiologists) to calculate the overall classification performance metrics.

Receiver operating characteristic (ROC) curve analysis was conducted by using a cumulative one-versus-rest strategy, comparing each tendon form against all others. Macro-averaged values for precision, recall, the F1 score, and the area under the curve (ROC-AUC) are reported to ensure balanced evaluation across all tendon categories.

To identify misclassification patterns, confusion matrices were generated for each rater and for the majority vote. These matrices were used to visualise common errors and to highlight areas of disagreement or uncertainty between raters.

## Ethical considerations

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The Swedish Ethical Authority approved this study and waived informed consent (number 2024-07283-02).

# Results

## Study cohort

The study cohort consisted of 130 patients with a mean age of 39.7 years and a standard deviation of 13.4 years (range: 18–64 years). There were 73 males (56.2%) and 57 females

(43.8%). The left side was examined in 69 patients (53.1%), while the right was examined in 61 patients (46.9%).

## Consensus-based tendon classification

The distribution of tendon forms based on the consensus between raters was as follows: 37.7% of cases (n = 49) were classified as general flat, 20.8% (n = 27) as flat with a lateral bulge, 19.2% (n=25) as flat with a medial bulge, and 22.3% (n = 29) as oval tendon. Fig. 1 shows an example of each type.

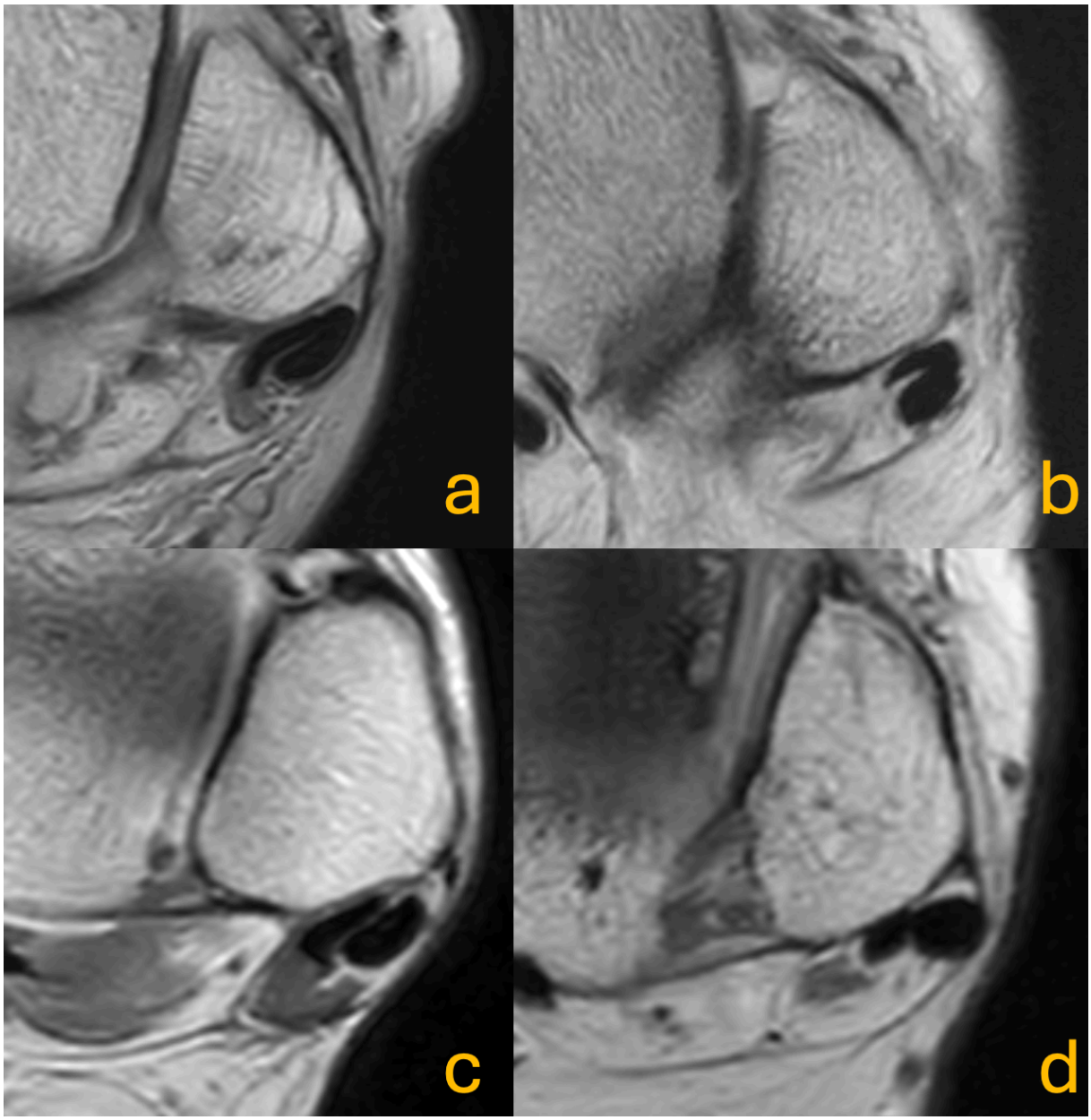


**Fig. 1** Axial section at the level of the lateral malleolus and the ankle joint space (magnetic resonance images with proton-density weighting, left sides). The four morphological types of the peroneus brevis tendon are: **A** general flat, **B** flat with a lateral bulge, **C** flat with a medial bulge, and **D** oval tendon

## Inter-rater agreement

## Pairwise agreement between raters

The alluvial plot (Fig. 2) illustrates the distribution of the four peroneus brevis tendon forms provided by the seven raters, alongside the consensus ratings. The plot visually represents the agreement and discrepancies between the individual raters' assessments, with colour-coded flows linking the raters' choices to the final consensus.

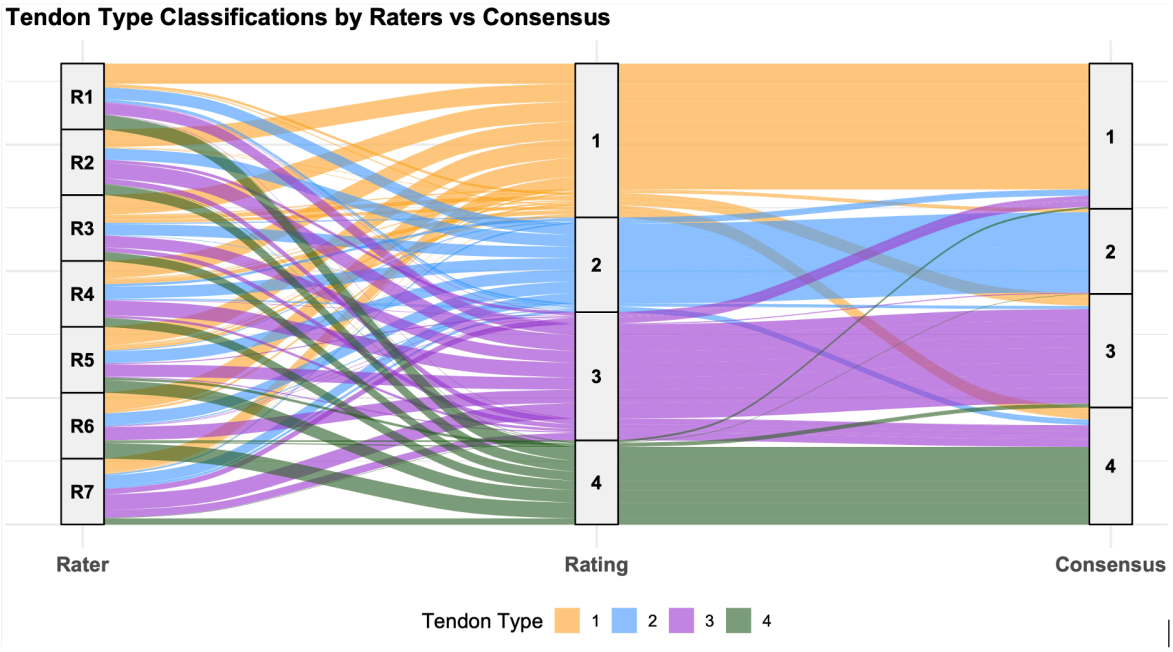


**Fig. 2** Alluvial plot showing how the raters rated the tendon forms – 1, general flat; 2, flat with a lateral bulge; 3, flat with a medial bulge; and 4, oval tendon – relative to the consensus

Cohen's kappa – which ranged from 0.589 to 0.865 – and Gwet's AC1 – which ranged from 0.580 to 0.865 – indicated substantial to almost perfect agreement according to Landis and Koch [17] (Table 1 and Figs. 3–5).

**Table 1** Agreement statistics for each rater compared with the consensus

| **Rater** | **Cohen's kappa (k)** | **95% confidence interval** | **Gwet's AC1** | **95% confidence interval** |
|---|---|---|---|---|
| Rater 1 | 0.823 | 0.78–0.86 | 0.823 | 0.79–0.86 |
| Rater 2 | 0.793 | 0.74–0.83 | 0.791 | 0.75–0.83 |
| Rater 3 | 0.707 | 0.65–0.76 | 0.705 | 0.66–0.75 |
| Rater 4 | 0.741 | 0.69–0.79 | 0.739 | 0.70–0.78 |
| Rater 5 | 0.781 | 0.73–0.82 | 0.781 | 0.74–0.82 |
| Rater 6 | 0.865 | 0.83–0.89 | 0.865 | 0.83–0.89 |
| Rater 7 | 0.589 | 0.52–0.65 | 0.580 | 0.51–0.64 |

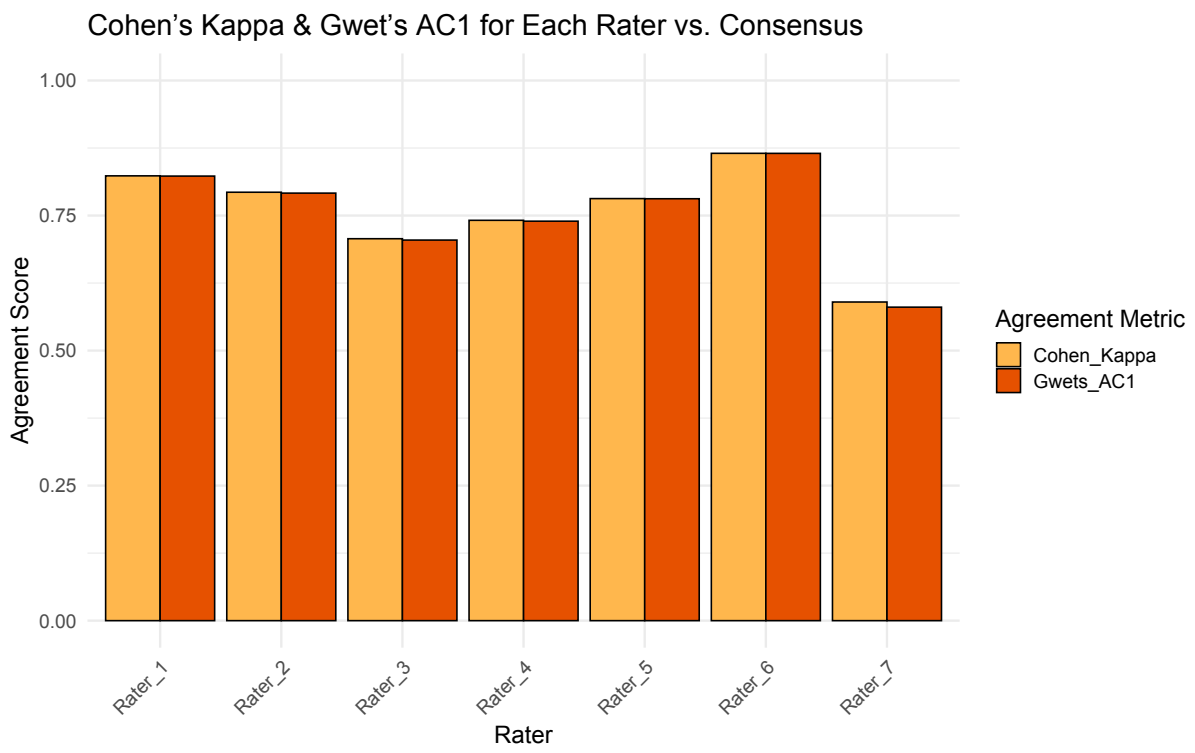


**Fig. 3** Comparison of Cohen's kappa and Gwet's AC1 for each rater versus the consensus

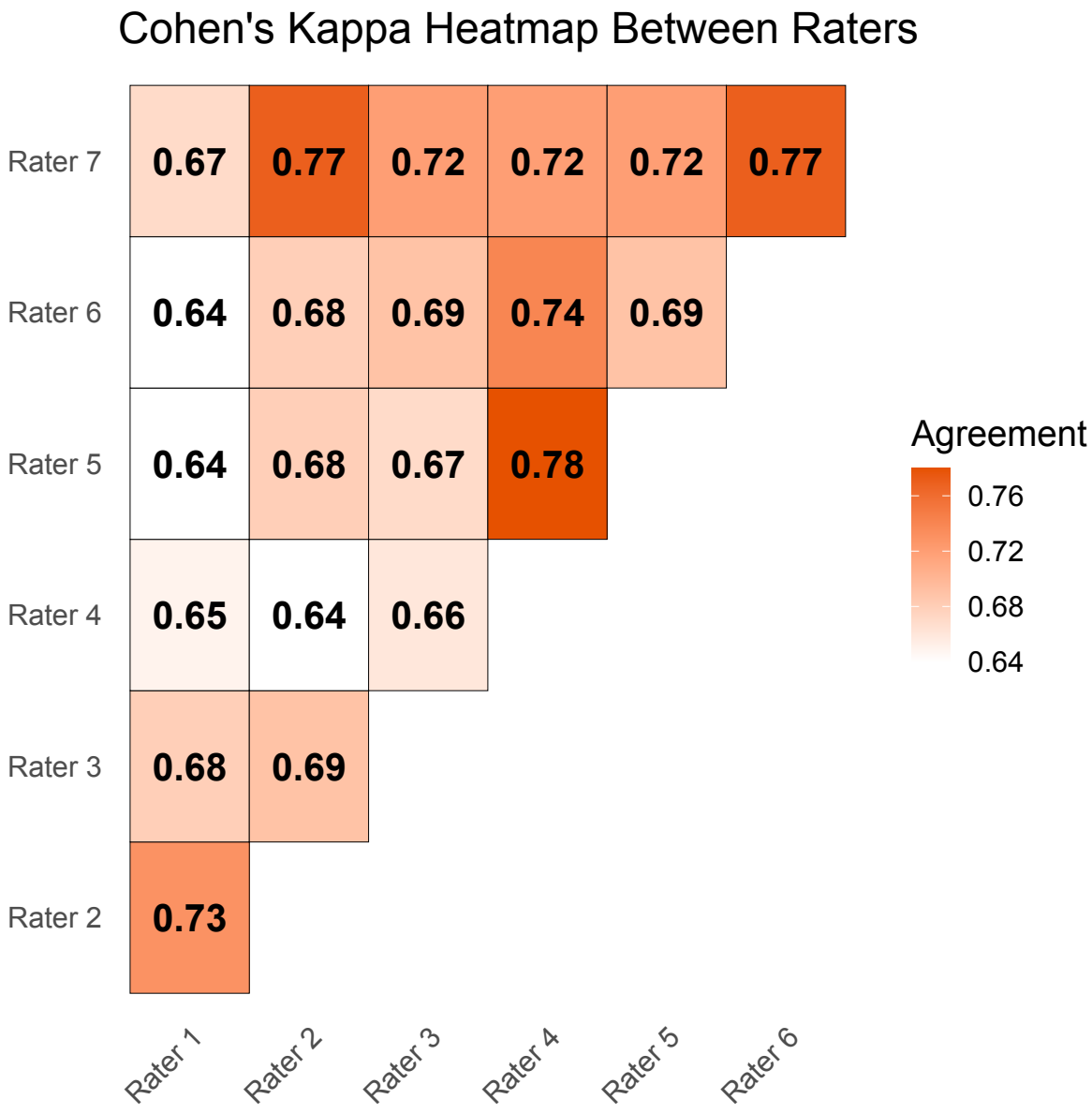


**Fig. 4** Heatmap visualisation of pairwise Cohen's kappa between raters. Darker colours indicate stronger agreement

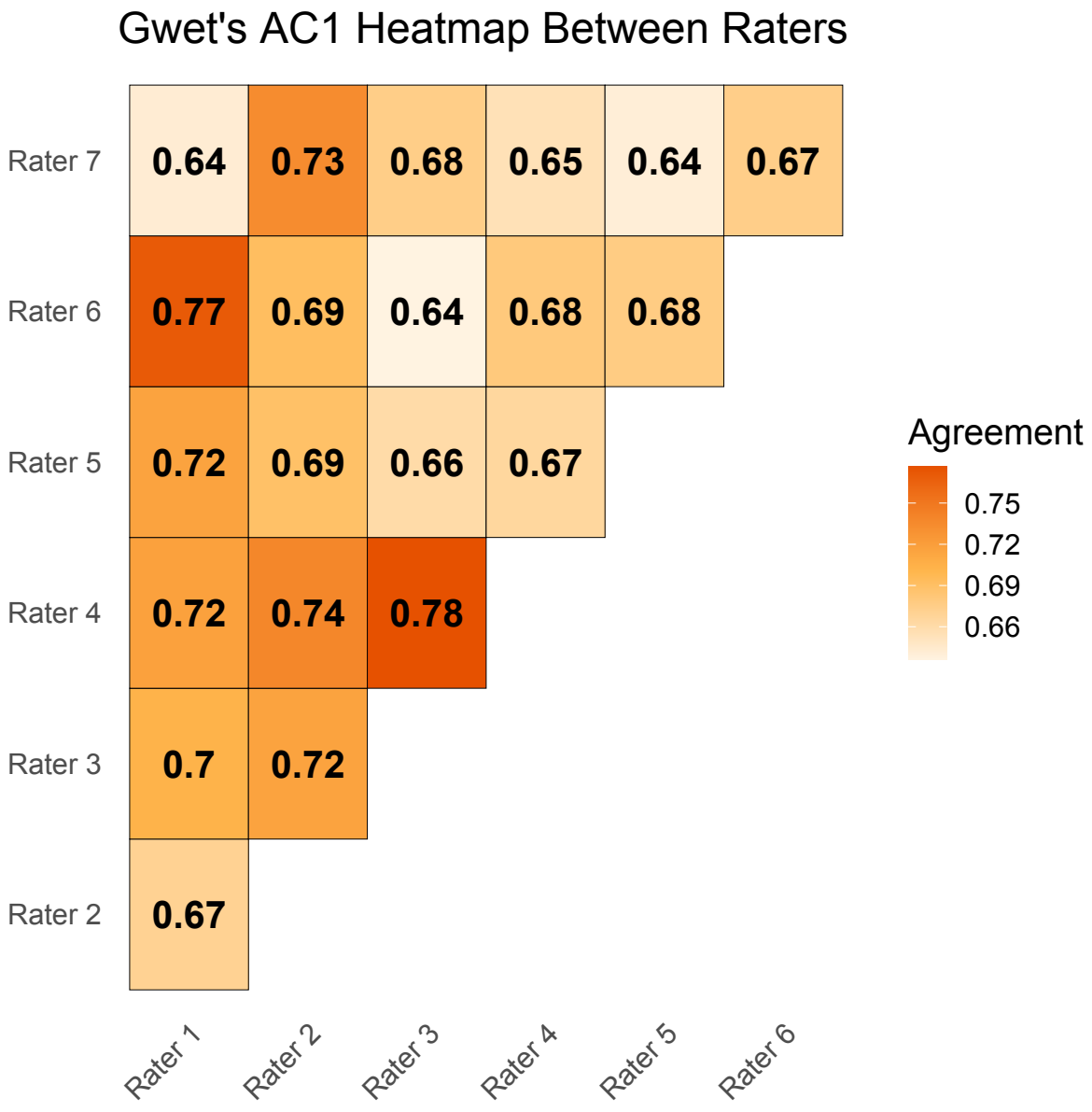


**Fig. 5** Heatmap visualisation of pairwise Gwet's AC1 between raters. Darker colours indicate stronger agreement

## Intra-rater reliability

After a 3-week washout period, Cohen's kappa ranged from 0.57 to 0.93, indicating moderate to almost perfect intra-rater agreement. Gwet's AC1 was generally higher, ranging from 0.70 to 0.93. Most raters demonstrated substantial to almost perfect reproducibility according to Landis and Koch [17].

## Overall agreement across raters

Fleiss' kappa across all raters was 0.691 (95% confidence interval 0.629–0.752), indicating substantial agreement according to Landis and Koch [17]. To assess the stability of this result, we performed bootstrapping, resampling the data 1000 times. The 95% confidence interval was 0.629–0.752 (Fig. 6). This narrow confidence interval suggests that the level of agreement between raters is consistent across different resamples. The substantial agreement found, with the lower bound of the confidence interval still above 0.6, indicates reliable rater consistency in classifying the tendon forms.

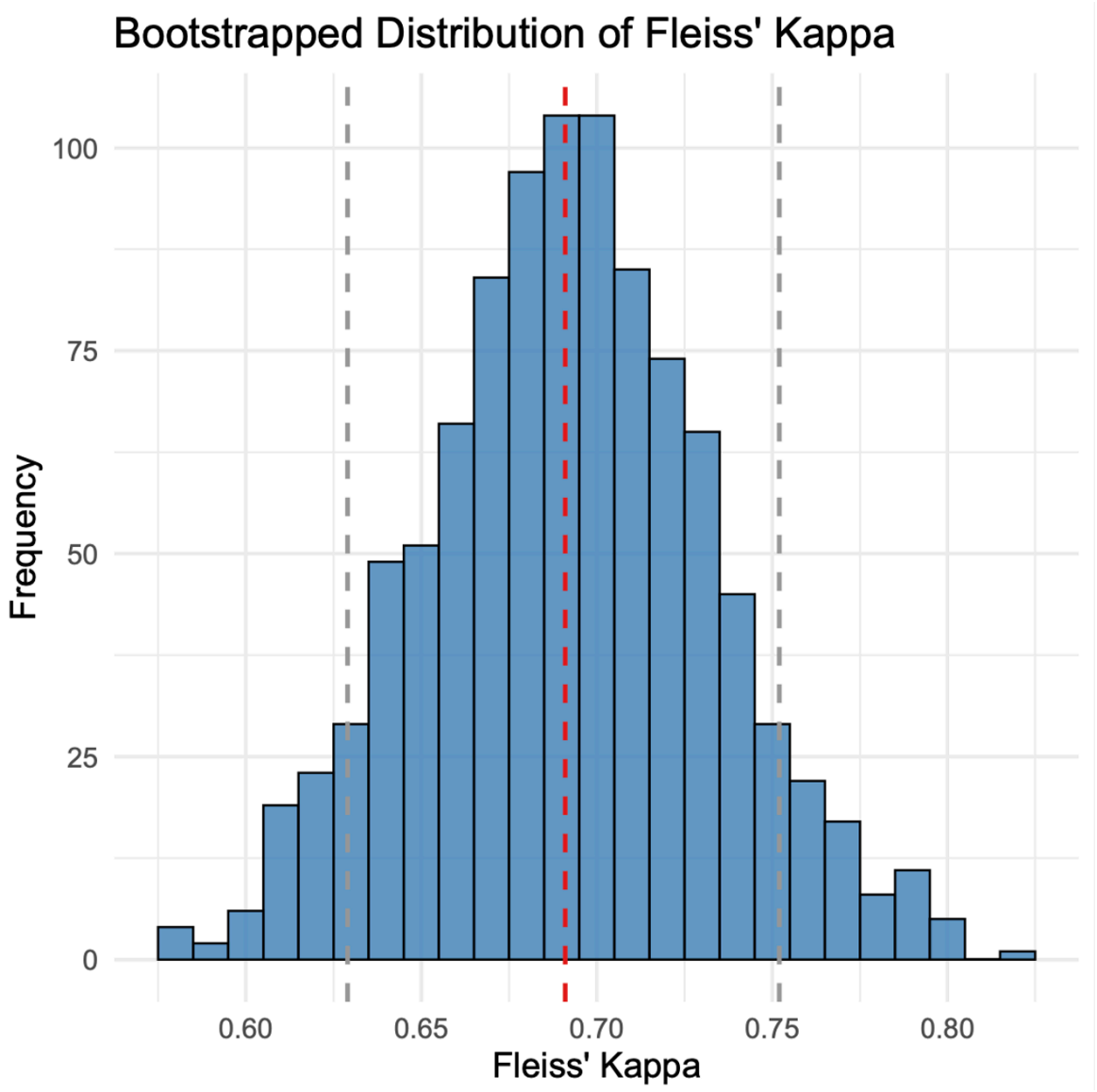


**Fig. 6** Bootstrapped distribution of Fleiss' kappa computed from 1,000 resamples. The vertical dashed line represents the mean Fleiss' kappa (0.691), and the grey dashed lines indicate the 95% confidence interval (0.629–0.752)

## Classification performance metrics

The classification performance of individual raters varied, as reflected in the F1 score, harmonic mean, and F2 score. While most raters demonstrated substantial agreement with the

consensus classification, we observed differences in balance precision and recall that affected the overall F1 scores (Table 2). Raters 1, 2, and 5 had consistently high performance, with an F1 score > 0.84, indicating strong agreement with the consensus. Rater 6 exhibited the highest classification accuracy, achieving an F1 score of 0.90, suggesting reliable differentiation between tendon forms. Conversely, Raters 3 and 7 had a lower F1 score (0.78 and 0.68, respectively), indicating greater variability in classification, likely due to challenges in distinguishing between certain tendon forms.

To further assess classification reliability, we applied a majority vote approach, using the most frequently assigned class among all raters as the final prediction (Table 2). This method significantly improved classification performance, achieving an F1 score of 0.93, a harmonic mean of 0.935, and an F2 score of 0.929, showing the value of combining multiple assessments to mitigate individual rater variability (Fig. 7).

**Table 2** Per-rater and overall performance

| | **F1 score** | **Harmonic mean** | **F2 score** |
|---|---|---|---|
| **Rater 1** | 0.87 | 0.87 | 0.87 |
| **Rater 2** | 0.85 | 0.86 | 0.84 |
| **Rater 3** | 0.78 | 0.80 | 0.78 |
| **Rater 4** | 0.80 | 0.82 | 0.80 |
| **Rater 5** | 0.84 | 0.84 | 0.84 |
| **Rater 6** | 0.90 | 0.90 | 0.90 |
| **Rater 7** | 0.68 | 0.73 | 0.68 |

| | | | |
|---|---|---|---|
| **Overall (majority vote)** | 0.93 | 0.93 | 0.93 |

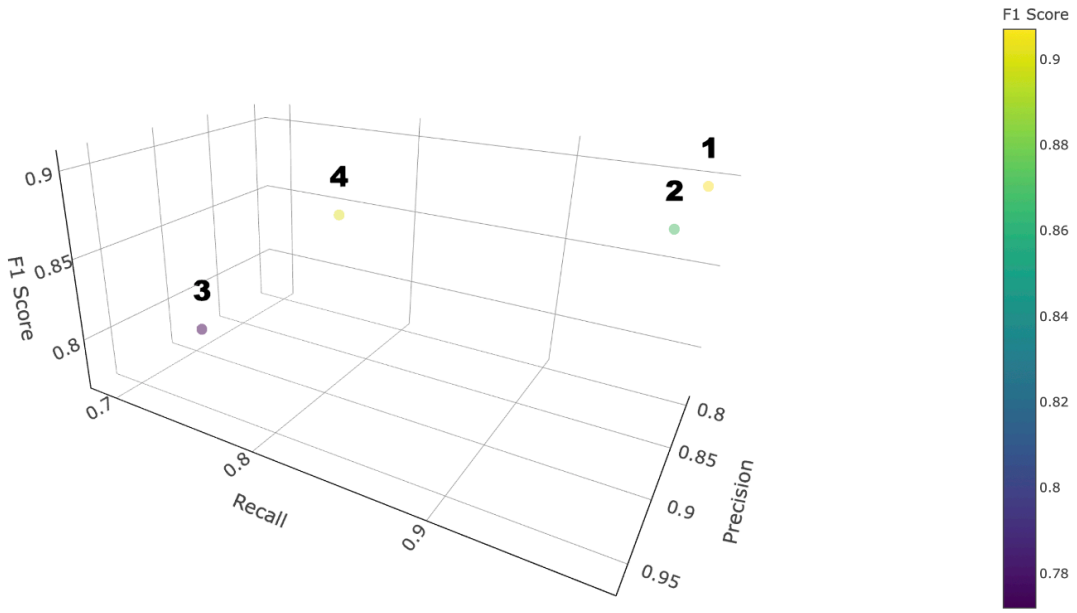


**Fig. 7** A three-dimensional (3D) scatterplot of precision, recall, and the F1 score (the colour scale) across the different tendon forms.

## Classification performance

The ROC curves demonstrated robust classification performance across all tendon forms, with the general flat tendon achieving the steepest curve, indicating clear separability from other classes (Fig. 8). A consistently high area under the curve (ROC-AUC = 0.89) confirms that the classification effectively differentiates between tendon forms. It suggests that the classification system provides reliable assessment. The high sensitivity and specificity observed in the ROC analysis highlight the model's robustness, with a steep slope at a lower

false positive rate, indicating minimal misclassification. The cumulative one-versus-rest approach, well suited for these data, ensured that we evaluated classification performance in a clinically meaningful way rather than treating each category as an independent class.

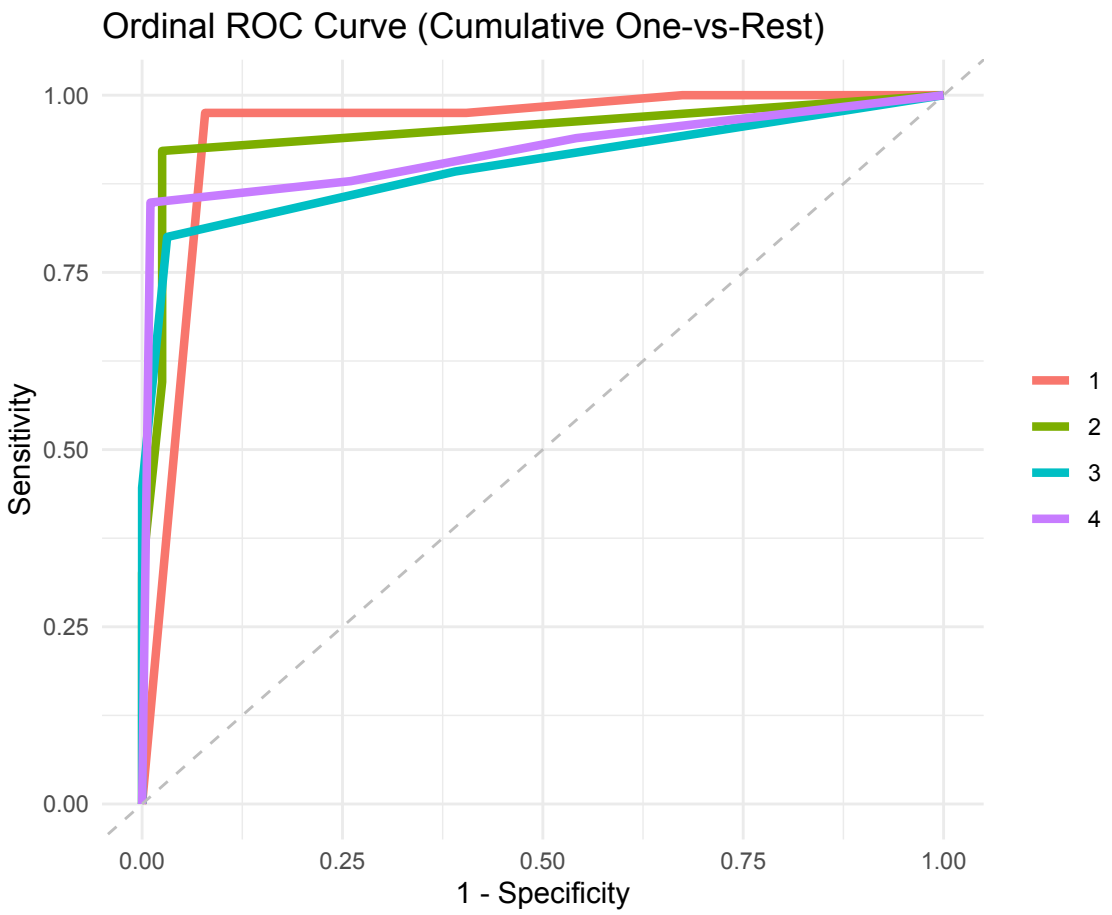


**Fig. 8** Receiver operating characteristic (ROC) curve for the tendon classification system, using a cumulative one-versus-rest approach. Tendon forms: 1, general flat; 2, flat with a lateral bulge; 3, flat with a medial bulge; and 4, oval tendon. Sensitivity (the true positive rate) is plotted against 1 – specificity (the false positive rate). The curve demonstrates the ability of the classification model to distinguish between the four tendon forms

The strong classification performance is reflected by an accuracy of 0.83, with a macro-averaged F1 score of 0.83, indicating well-balanced performance across the tendon categories.

The null accuracy of 0.31 confirms that the classification significantly outperforms random chance (Table 3).

**Table 3** Overall classification performance metrics

| | Value |
|---|---|
| Accuracy | 0.83 |
| Null accuracy | 0.31 |
| Receiver operating characteristic area under the curve | 0.89 |
| Precision (macro-average) | 0.84 |
| Recall (macro-average) | 0.83 |
| F1 score (macro-average) | 0.83 |

## Precision-recall analysis

We used a precision-recall curve to evaluate classification performance; it demonstrates the trade-off between recall and precision. The analysis revealed differences in precision, recall, and the F1 score across the tendon forms. We observed higher F1 scores (Fig. 9) for tendon forms with more distinct shapes, whereas lower F1 scores reflected greater uncertainty in classification. The high area under the precision-recall curve (0.83) indicates good performance in correctly identifying tendon forms while reducing false positives.

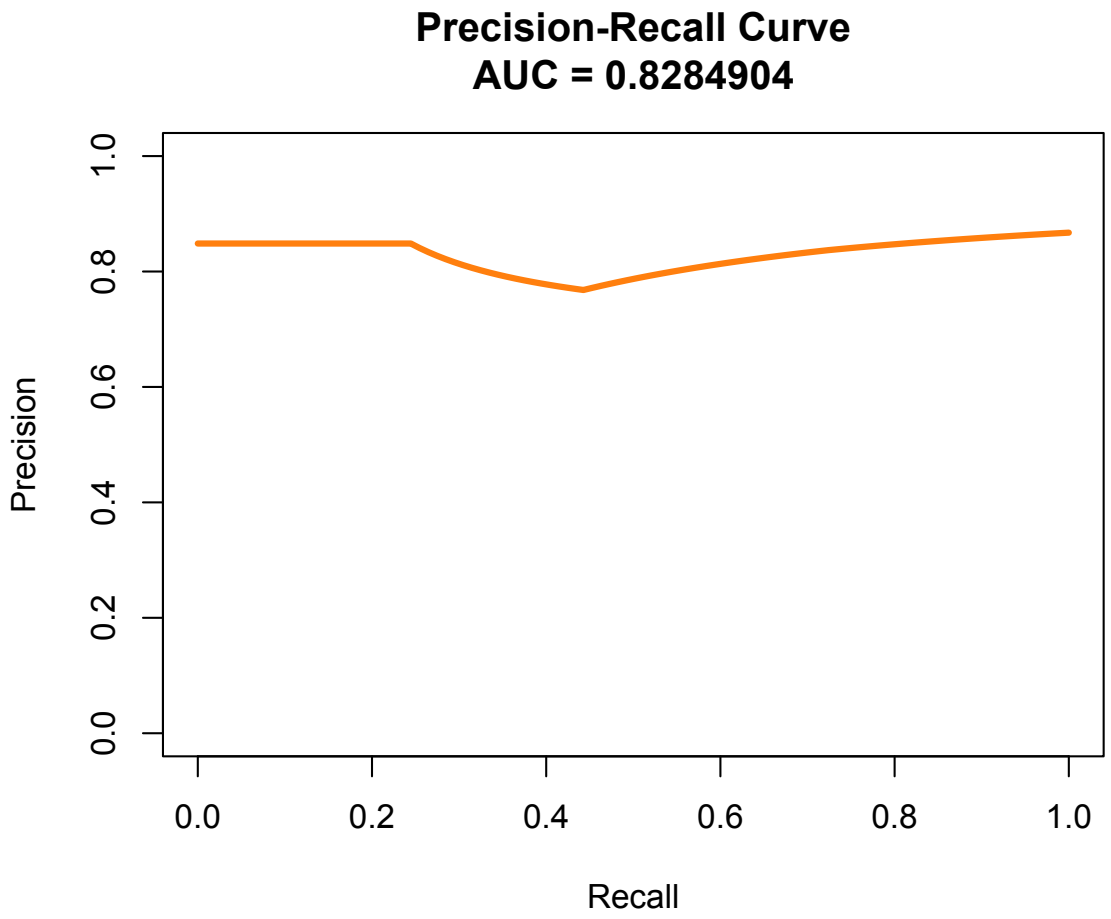


**Fig. 9** The average precision-recall curve demonstrating the relationship between recall and precision for tendon classification for all tendon forms (average curve). The curve shows the trade-off between recall and precision at different classification thresholds, with higher values indicating better performance in distinguishing tendon forms. AUC, area under the curve

## Error analysis and misclassification trends

We generated a confusion matrix for individual raters (Fig. 10) to complement Fleiss' kappa and to provide a more detailed analysis of the misclassifications. The most common error – misclassifying the oval tendon as general flat with a medial bulge (4.76%) – occurred almost twice as much as the other errors (Fig. 11). The confusion matrix shows the frequency with

which raters classified tendons as another type, indicating both areas of agreement and misclassification. The diagonal of the matrix, representing the correctly classified tendon forms, was consistently high, suggesting that raters generally agreed with the gold standard (the consensus between two musculoskeletal radiologists).

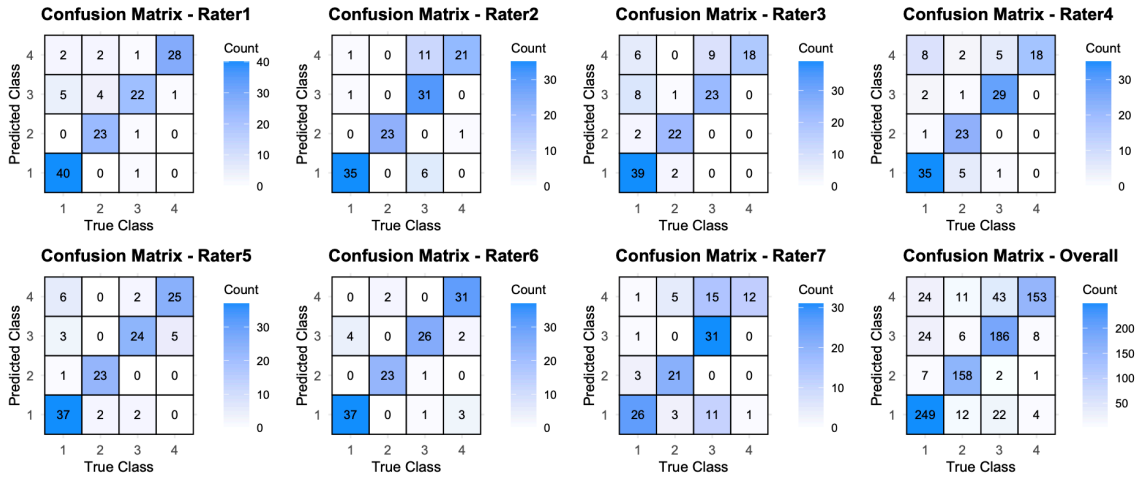


**Fig. 10** Confusion matrix heatmap illustrating classification discrepancies between each rater, the overall evaluation (majority vote), and the consensus classification. Tendon forms: 1, general flat; 2, flat with a lateral bulge; 3, flat with a medial bulge; and 4, oval tendon. Darker colours indicate higher agreement with the consensus, while lighter shades highlight misclassifications

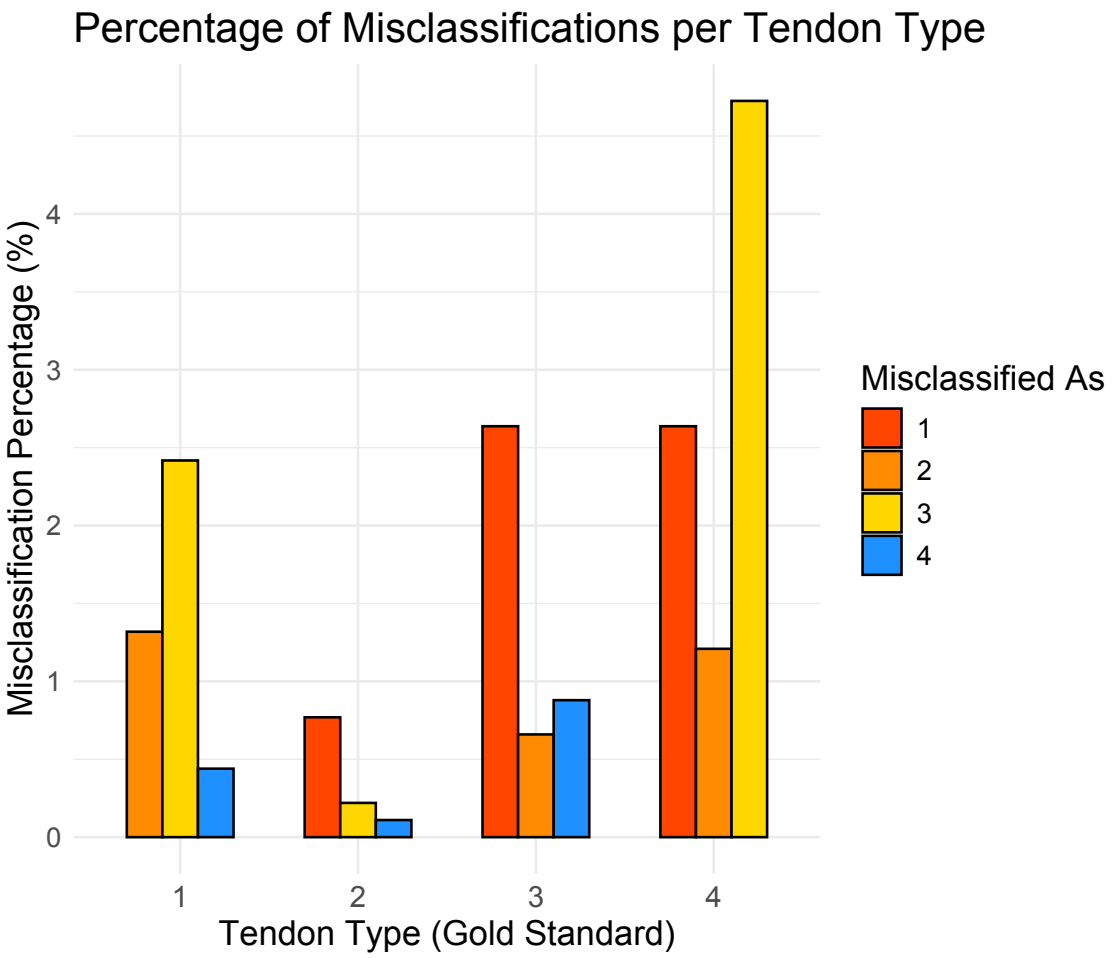


**Fig. 11** Bar chart showing the percentage (%) of times each tendon form was classified incorrectly by the raters. Tendon forms: 1, general flat; 2, flat with a lateral bulge; 3, flat with a medial bulge; and 4, oval tendon

# Discussion

The anatomical variability of the distal insertion of the peroneus brevis tendon has been thoroughly investigated, and several classification systems have been proposed based on cadaveric studies[24, 37]. However, no validated system currently exists for assessing peroneus brevis tendon morphology in vivo on transverse MRI at the lateral malleolus - the site where split tears most often occur. In this study, we presented a clear and reproducible

classification system with four forms based on peroneus brevis tendon morphology, assessed on transverse proton-density MRI scans. We validated this system based on 130 patients and 7 independent raters. The system showed good discriminatory ability and was applied consistently by the raters from various clinical backgrounds, reflecting the reality of multidisciplinary image interpretation. This classification holds clinical and research relevance, as morphological variations in the peroneus brevis tendon may be related to a higher risk of developing split tears; additional studies are needed to quantify this relationship. By offering a practical and reliable framework, the proposed system supports consistent assessment by both radiologists and non-radiologists.

Because there are no clear methodological recommendations on how to validate tendon classification systems, we conducted several analyses to assess the robustness of our proposed system, beginning with an evaluation of inter-rater reliability. The substantial agreement based on Fleiss' kappa suggests that, overall, the raters were able to reliably classify the tendon forms according to the consensus. The narrow confidence interval in our study further supports the robustness of our result, as it demonstrates that the kappa value is stable across multiple re-samplings of the data, with minimal uncertainty. Compared with Cohen's kappa, Gwet's AC1 accounts for prevalence bias and shows slightly higher agreement scores across raters [9]. The stability of these values supports the robustness of the classification system. The high intra-rater agreement confirms the reproducibility of tendon classification across repeated evaluations. The consistently higher Gwet's AC1 values may highlight the limitations of Cohen's Kappa in cases with uneven category distributions [35]. These findings suggest that individual raters maintained consistency in their assessments, reinforcing the reliability of the classification system. Minor discrepancies may be related to borderline tendon forms or visual assessment, warranting future metric standardisation efforts to further optimise classification consistency.

We included raters from different professional backgrounds to reflect the complexity of diagnosing peroneal tendons. All raters received detailed instructions outlining the study protocol, definitions of tendon types, and illustrative examples, which likely contributed to the high level of agreement. In our view, studies of this kind should involve specialists from various fields to ensure clinical relevance. Although our rater group did not include any orthopaedic surgeons, it did include two physiotherapists, offering a balanced mix of raters. Had we included only experienced radiologists, inter-rater agreement might have been higher [3]; however, the findings would have been less generalisable to everyday clinical practice, where patients are managed by multidisciplinary teams and magnetic resonance imaging is interpreted by different healthcare professionals.

Some previous studies have referred to the different forms of the peroneus brevis tendon on transverse cross section, but no validated classification exists. Authors have mentioned that a chevron or bisected shape of the peroneus brevis tendon [20] may represent risk factors to split rupture [27]. Overly complex descriptions of tendon shape can lead to misunderstandings. Therefore, it was necessary to develop a relatively simple classification that still captures the complexity of anatomical variation. To the best of our knowledge, this study represents the first robust validation of peroneus brevis tendon classification at the level of the lateral malleolus. We choose this level based on clinical observation and the fact that this is the most common location of a peroneus brevis split rupture [5, 23].

The anatomical variability of the distal portion of the peroneus brevis tendon has been extensively described in both classical and modern anatomical literature[4, 18, 19, 24, 26, 37]. Variants include additional slips inserting into adjacent muscles, bones, or soft tissue structures, as well as bifurcated or trifurcated tendons and fusions with surrounding musculature. Several classification systems have been proposed based on cadaveric dissections, primarily focusing on the diversity of insertion sites onto the fifth metatarsal and

beyond. These include historical reports of accessory bands to the abductor digiti minimi and peroneus longus [19],up to eight insertion types [18], and more recent cadaveric [24] and foetal [26] classification frameworks. While these studies have contributed substantially to our understanding of distal part, they do not address the cross-sectional shape of the tendon at the level of the lateral malleolus. Our study complements this body of knowledge by proposing a validated, in vivo MRI-based classification that captures morphological variability at the transverse cross section at the lateral malleolus.

Considering previous studies highlighting the complex anatomical variants of the distal peroneus brevis and its bony insertions, it was essential to evaluate inter-rater agreement among investigators from different clinical backgrounds. While the overall agreement was substantial [17], we observed variability among the raters, reflecting the challenges of distinguishing subtle anatomical differences. The misclassification patterns seen in the confusion matrix shows the limitations of the visual assessment [28]. The confusion matrix analysis revealed that the oval tendon and the general flat tendon with a medial bulge were frequently misclassified, indicating some similarity in visual classification. This suggests that subtle convexities or asymmetries in the tendon outline challenge rater perception. Such misclassifications suggest that there is similarity and overlap of anatomical features between tendon forms, which may impact how these tendon variants are classified. This issue contributed to lower inter-rater agreement. The least frequent misclassification occurred between the general flat tendon with a medial bulge and the general flat tendon with a lateral bulge. This may be explained by the fact that distinguishing the medial from the lateral side is generally easier than differentiating between a uniformly flat tendon and one with a subtle thickening on one end. It is well known that simpler classifications tend to achieve higher agreement between raters [8].

Variations in agreement reflect the complexity of anatomical variations and the individual characteristics of tendon morphology. Despite variations, our analysis demonstrated that the classification remains reliable even when radiologists represent a minority within the group of raters. This intriguing observation highlights that anatomical variability and very subtle differences between tendon forms may contribute to lower inter-rater agreement. Multicentre studies with larger and more diverse samples would help confirm whether these misclassification patterns are generalisable across different patient populations. Nevertheless, we determined the sample size prior to commencing the study to ensure we had adequate statistical power and that our results are valid.

Our raters did not perform tendon measurements, as we deliberately chose a purely visual assessment method to reflect real-world clinical scenarios where decisions must be time effective. We hypothesise that in practice, measurements would be performed, and most raters would likely rely on visual evaluation. It is possible that using measurements could have further improved agreement. We think that in the future, the use of artificial intelligence may be used to classify tendon types objectively. However, at present, no such solution exists.

The bootstrapping analysis also highlighted the stability of Fleiss' kappa, confirming the robustness of our classification system [36]. The narrow confidence intervals and similar Cohen's kappa and Gwet's AC1 indicate that the observed level of agreement is unlikely to be due to sampling variability and can be generalised beyond our study cohort. The usage of Cohen's kappa and Gwet's AC1 is an advantage of this study. While Cohen's kappa is widely used in agreement radiological studies, it is known to be sensitive to prevalence imbalances, resulting in the kappa paradox [9]. Gwet's AC1 is a prevalence- and bias-adjusted kappa statistic [13]. Similar results produced similar results, suggesting that classification reliability is not significantly affected by imbalances in tendon form distribution. The strong correlation between these two agreement measures confirms the consistency and robustness of the

proposed classification system, supporting its potential for additional studies and clinical implementation.

The ROC curve analysis demonstrated strong classification performance, with the general flat form achieving the steepest curve, suggesting clear separability from the three other forms. This is likely due to the fact that the uniform thickness of the general flat tendon distinguishes it from the other tendon forms, which exhibit greater anatomical variation in contour. The high area under the curve across all tendon types indicates that the classification system effectively differentiates between morphological variants, reinforcing its clinical applicability. The overlap in the ROC curves suggests some degree of uncertainty in classification between similar tendon forms, which we believe could be further refined with measurements or machine learning models. Future studies should investigate the performance of this classification system across other imaging modalities, including ultrasound, and in diverse international settings.

Despite the strengths of this study, several limitations should be acknowledged. Our classification is based solely on transverse section on magnetic resonance imaging without surgical correlation. Our classification is visual, and we hypothesise that integrating quantitative morphometric analysis could further refine the classification system and improve reproducibility. Our material is based on evaluation of the magnetic resonance examinations performed on clinical indications. Additional prospective studies are warranted for external validation and continued development of the classification system in a more quantitative direction.

# Conclusion

Our classification system for the peroneus brevis tendon proved to be a robust and reliable tool, demonstrating substantial inter-rater agreement and excellent overall performance. Misclassifications were rare and mostly limited to tendon types with closely related anatomical features. Its simplicity, practicality, and reproducibility make it well-suited for both clinical and research use. The system's strength lies in its ability to capture complex anatomical variation in a clear and accessible format. With appropriate rater training, it offers consistent and accurate tendon evaluation, supporting confident diagnosis and informed treatment planning across a range of clinical settings.

# Declarations

**Declaration of Generative AI and AI-assisted technologies in the writing process**

Nothing to disclose

**Ethics approval and consent to participate**

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The Swedish Ethical Authority approved this study and waived informed consent (number 2024-07283-02).

**Consent for publication**

Not applicable. The manuscript does not contain the data of individuals in any form.

**Availability of data and material**

Yes

**Competing interests**

The authors declare that they have no competing interests.

**Funding**

This work was supported by Stiftelsen Tornspiran (934: 2023-12-01) and by the Swedish state under the agreement between the Swedish Government and the county councils, the ALF-agreement (ALFGBG-1006741).

**Authors' contributions**

Conceptualization: PS; Methodology: PS, KBD, RZ; Software: PS, RZ; Data Collection: All authors; Data Curation: RZ, PS; Writing – Original Draft: RZ, PS; Writing – Review and Editing: All authors. All authors contributed to previous manuscript versions, and all have read and approved the final manuscript.

**Acknowledgements**

None

# References

1. Armstrong ME, Tornblad MK, Jones KS (2021) The accuracy of interrater reliability estimates found using a subset of the total data sample: A bootstrap analysis. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 64:1377-1382. https://doi.org/10.1177/1071181320641329.

2. Athavale SA, Swathi, Vangara SV (2011) Anatomy of the superior peroneal tunnel. J Bone Joint Surg Am 93:564-571. https://doi.org/10.2106/JBJS.J.00662.

3. Bai X, Sun SM, Xu W, et al (2020) MRI-based Bosniak Classification of Cystic Renal Masses, Version 2019: Interobserver Agreement, Impact of Readers' Experience, and Diagnostic Performance. Radiology 297:597-605. https://doi.org/10.1148/radiol.2020200478.

4. Bhargava KN, Sanyal KP, Bhargava SN (1961) Lateral musculature of the leg as seen in hundred Indian Cadavers. Indian Journal of Medical Sciences 15:181-185.

5. Bokwa-Dabrowska K, Mocanu D, Alexiev A, et al (2024) Peroneus brevis split rupture is underreported on magnetic resonance imaging of the ankle in patients with chronic lateral ankle pain. Eur J Radiol Open 13:100591. https://doi.org/10.1016/j.ejro.2024.100591.

6. Bokwa-Dabrowska K, Zych R, Mocanu D, et al (2025) Peroneus brevis split tear - A challenging diagnosis: A pictorial review of magnetic resonance and ultrasound imaging. Part 1. Anatomical basis and clinical insights. Eur J Radiol Open 14:100633. https://doi.org/10.1016/j.ejro.2024.100633.

7. Bokwa-Dąbrowska K, Mocanu D, Romanus I, et al (2025) Peroneus brevis split tear - A challenging diagnosis: A pictorial review of magnetic resonance and ultrasound imaging - Part 2: Imaging with magnetic resonance and ultrasound. Eur J Radiol Open 14:100627. https://doi.org/10.1016/j.ejro.2024.100627.

8. Chhabra A, Duarte Silva F, Mogharrabi B, et al (2024) MRI-based Neuropathy Score Reporting And Data System (NS-RADS): multi-institutional wider-experience usability study

of peripheral neuropathy conditions among 32 radiology readers. Eur Radiol https://doi.org/10.1007/s00330-023-10517-2.

9. Cicchetti DV, Feinstein AR (1990) High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 43:551-558. https://doi.org/10.1016/0895-4356(90)90159-m.

10. Davda K, Malhotra K, O'Donnell P, et al (2017) Peroneal tendon disorders. EFORT Open Rev 2:281-292. https://doi.org/10.1302/2058-5241.2.160047.

11. Ersoz E, Tokgoz N, Kaptan AY, et al (2019) Anatomical variations related to pathological conditions of the peroneal tendon: evaluation of ankle MRI with a 3D SPACE sequence in symptomatic patients. Skeletal Radiol 48:1221-1231. https://doi.org/10.1007/s00256-019-3151-5.

12. Galli MM, Protzman NM, Mandelker EM, et al (2015) An examination of anatomic variants and incidental peroneal tendon pathologic features: a comprehensive MRI review of asymptomatic lateral ankles. J Foot Ankle Surg 54:164-172. https://doi.org/10.1053/j.jfas.2014.11.005.

13. Gwet KL (2008) Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol 61:29-48. https://doi.org/10.1348/000711006x126600.

14. Kottner J, Audige L, Brorson S, et al (2011) Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol 64:96-106. https://doi.org/10.1016/j.jclinepi.2010.03.002.

15. Kuwada GT (2008) Surgical correlation of preoperative MRI findings of trauma to tendons and ligaments of the foot and ankle. J Am Podiatr Med Assoc 98:370-373. https://doi.org/10.7547/0980370.

16. Lambert HW, Leg Muscles, Bergman's Comprehensive Encyclopedia of Human Anatomic Variation2016, pp. 421-437.

17. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159-174.

18. Le Double AF, Traité des variations du système musculaire de l'homme et de leur signification au point de vue de l'anthropologie zoologique, Schleicher frères, Paris, 1897.

19. Macalister A (1875) Additional Observations on Muscular Anomalies in Human Anatomy. (Third Series) With a Catalogue of the Principal Muscular Variations Hitherto Published. The Transactions of the Royal Irish Academy 25:118-124.

20. Major NM, Helms CA, Fritz RC, et al (2000) The MR imaging appearance of longitudinal split tears of the peroneus brevis tendon. Foot Ankle Int 21:514-519. https://doi.org/10.1177/107110070002100612.

21. Matcuk GR, Jr., Patel DB, Cen S, et al (2019) Fibular groove morphology and measurements on MRI: correlation with fibularis tendon abnormalities. Surg Radiol Anat 41:75-85. https://doi.org/10.1007/s00276-018-2134-x.

22. Mirmiran R, Squire C, Wassell D (2015) Prevalence and Role of a Low-Lying Peroneus Brevis Muscle Belly in Patients With Peroneal Tendon Pathologic Features: A Potential Source of Tendon Subluxation. J Foot Ankle Surg 54:872-875. https://doi.org/10.1053/j.jfas.2015.02.012.

23. Miura K, Ishibashi Y, Tsuda E, et al (2004) Split lesions of the peroneus brevis tendon in the Japanese population: an anatomic and histologic study of 112 cadaveric ankles. J Orthop Sci 9:291-295. https://doi.org/10.1007/s00776-004-0784-5.

24. Olewnik L, Podgorski M, Ruzik K, et al (2020) New classification of the distal attachment of the fibularis brevis - Anatomical variations and potential clinical implications. Foot Ankle Surg 26:308-313. https://doi.org/10.1016/j.fas.2019.04.002.

25. Olewnik Ł, Podgórski M, Ruzik K, et al (2020) New classification of the distal attachment of the fibularis brevis - Anatomical variations and potential clinical implications. Foot Ankle Surg 26:308-313. https://doi.org/10.1016/j.fas.2019.04.002.

26. Olewnik Ł, Zielinska N, Ruzik K, et al (2024) Morphological variability of the fibularis brevis tendon in human fetuses. Ann Anat 251:152154. https://doi.org/10.1016/j.aanat.2023.152154.

27. Park HJ, Cha SD, Kim HS, et al (2010) Reliability of MRI findings of peroneal tendinopathy in patients with lateral chronic ankle instability. Clin Orthop Surg 2:237-243. https://doi.org/10.4055/cios.2010.2.4.237.

28. Shen H, Jin H, Cabrera ÁA, et al (2020) Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm Performance. Proceedings of the ACM on Human-Computer Interaction 4:1-22. https://doi.org/10.1145/3415224.

29. Sobel M, Geppert MJ, Warren RF (1993) Chronic ankle instability as a cause of peroneal tendon injury. Clin Orthop Relat Res 187-191.

30. Taljanovic MS, Alcala JN, Gimber LH, et al (2015) High-resolution US and MR imaging of peroneal tendon injuries. Radiographics 35:179-199. https://doi.org/10.1148/rg.351130062.

31. Taparia A, Kumar S, Saran S (2022) Lateral Retromalleolar Swelling and Pain - Peroneus Brevis Tendon Tear (Boomerang Sign). J Med Ultrasound 30:237-238. https://doi.org/10.4103/JMU.JMU_31_21.

32. van den Bekerom MP, Kerkhoffs GM, McCollum GA, et al (2013) Management of acute lateral ankle ligament injury in the athlete. Knee Surg Sports Traumatol Arthrosc 21:1390-1395. https://doi.org/10.1007/s00167-012-2252-7.

33. van Dijk PA, Miller D, Calder J, et al (2018) The ESSKA-AFAS international consensus statement on peroneal tendon pathologies. Knee Surg Sports Traumatol Arthrosc 26:3096-3107. https://doi.org/10.1007/s00167-018-4971-x.

34. Wang XT, Rosenberg ZS, Mechlin MB, et al (2005) Normal variants and diseases of the peroneal tendons and superior peroneal retinaculum: MR imaging features. Radiographics 25:587-602. https://doi.org/10.1148/rg.253045123.

35. Wongpakaran N, Wongpakaran T, Wedding D, et al (2013) A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. BMC Med Res Methodol 13:61. https://doi.org/10.1186/1471-2288-13-61.

36. Zapf A, Castell S, Morawietz L, et al (2016) Measuring inter-rater reliability for nominal data - which coefficients and confidence intervals are appropriate? BMC Med Res Methodol 16:93. https://doi.org/10.1186/s12874-016-0200-9.

37. Zielinska N, Tubbs RS, Gonera B, et al (2024) The tendon of the fibularis brevis muscle - Systematic overview and new classification system. Ann Anat 253:152208. https://doi.org/10.1016/j.aanat.2023.152208.