applied sciences

MDPI

Article

Automatic Meniscus Segmentation Using YOLO-Based Deep Learning Models with Ensemble Methods in Knee MRI Images

Mehmet Ali Şimşek [1,*], Ahmet Sertbaş [2], Hadi Sasani [3] and Yaşar Mahsut Dinçel [4]

[1] Department of Computer Technologies, Vocational School of Technical Sciences, Tekirdag Namik Kemal University, Tekirdag 59030, Turkey
[2] Department of Computer Engineering, Faculty of Engineering, University of Istanbul-Cerrahpasa, Istanbul 34320, Turkey; asertbas@iuc.edu.tr
[3] Department of Radiology, Faculty of Medicine, Tekirdag Namik Kemal University, Tekirdag 59030, Turkey; hsasani@nku.edu.tr
[4] Department of Orthopedics and Traumatology, Faculty of Medicine, Tekirdag Namik Kemal University, Tekirdag 59030, Turkey; ymd61@hotmail.com
* Correspondence: masimsek@nku.edu.tr

Abstract: The meniscus is a C-shaped connective tissue with a cartilage-like structure in the knee joint. This study proposes an innovative method based on You Only Look Once (YOLO) series models and ensemble methods for meniscus segmentation from knee magnetic resonance imaging (MRI) images to improve segmentation performance and evaluate generalization capability. In this study, five different segmentation models were trained, and masks were created from the YOLO series. These masks are combined with pixel-based voting, weighted multiple voting, and dynamic weighted multiple voting optimized by grid search. Tests were conducted on internal and external sets and various metrics. The dynamic weighted multiple voting method optimized with grid search performed the best on both the test set (DSC: 0.8976 ± 0.0071, PPV: 0.8561 ± 0.0121, Sensitivity: 0.9467 ± 0.0077) and the external set (DSC: 0.9004 ± 0.0064, PPV: 0.8876 ± 0.0134, Sensitivity: 0.9200 ± 0.0119). The proposed ensemble methods offer high accuracy, reliability, and generalization capability for meniscus segmentation.

Keywords: meniscus segmentation; magnetic resonance imaging; YOLO series; ensemble methods; voting methods

Academic Editor: Juan Martinez-Romo

Received: 21 January 2025
Revised: 17 February 2025
Accepted: 26 February 2025
Published: 4 March 2025

Citation: Şimşek, M.A.; Sertbaş, A.; Sasani, H.; Dinçel, Y.M. Automatic Meniscus Segmentation Using YOLO-Based Deep Learning Models with Ensemble Methods in Knee MRI Images. *Appl. Sci.* **2025**, *15*, 2752. https://doi.org/10.3390/app15052752

1. Introduction

Meniscus tissue is one of the soft tissues in the human joint located between the femur and tibia in the knee joint. The meniscus comprises C-shaped fibro-cartilaginous tissue and consists of the lateral and medial meniscus. These structures have essential functions: providing joint stability, shock absorption, joint fluid distribution, and load transfer [1,2]. Meniscal injuries usually lead to changes in joint biomechanics that affect load distribution and contact stresses [3]. Meniscal injuries are common knee injuries that can be seen in all age groups and usually occur as a result of sporting activities, aging-related degeneration, or traumatic effects.

Accurate diagnosis of meniscal injuries is essential to determine the type and extent of damage. Magnetic resonance imaging (MRI) is the gold standard for non-invasive diagnosis. MRI is widely used in the preliminary diagnosis of knee injuries due to its noninvasive properties and exceptional ability to provide clear visualization of soft tissue (high-contrast resolution) [4]. This method, which assesses the structural integrity of meniscal tissue with high precision, is often time-consuming and requires expert interpretation. In recent years,

several studies have been conducted using deep learning (DL) algorithms to detect and evaluate meniscal injuries on MRI images [5,6].

Meniscal segmentation automatically or manually determines the boundaries of meniscal tissue and separates it from surrounding tissues [7,8]. Segmentation is usually performed using pixel-based or region-based methods. With the rise of deep learning, image segmentation has been divided into three main categories: (i) semantic segmentation, (ii) instance segmentation, and (iii) panoptic segmentation. Semantic segmentation assigns each pixel in an image to a specific class, while instance segmentation separately identifies different objects belonging to the same class. Panoptic segmentation combines these approaches by labeling each pixel as an object or background [9,10].

Some DL-based studies for detecting and classifying meniscal injuries support this process by segmenting meniscal tissue [4,11]. Segmentation is not only used in injury detection, but also in treatment planning and surgical interventions. In particular, meniscal segmentation from knee MRI images plays an important role in analyzing the length, width, height, cross-sectional area, and surface area of the meniscus for meniscal allograft transplantation using a 3D reconstruction model based on the patient's normal meniscus [12].

The segmentation process can be complicated because MRI images have multiple sequences showing different tissues at various intensity and contrast levels. This causes structures with no clear boundaries, such as the meniscus, to appear in different shapes and intensities in each sequence. However, DL-based segmentation methods offer a more efficient approach to overcoming these challenges than traditional methods by providing high accuracy, speed, and automation advantages. As a result, these innovative methods provide a versatile contribution to the detection and treatment of meniscal injuries.

In recent years, studies on DL-based meniscus tissue segmentation have increased and differed in imaging methods, segmentation architectures, and targeted tasks. Although most of the studies in the literature are based on MRI images, alternative methods such as ultrasonography are also used. Especially for knee osteoarthritis (KOA), where early diagnosis is difficult and evaluation methods are limited, ultrasound stands out as a simple, economical, and non-invasive tool. Ultrasonography offers a powerful alternative to traditional methods by providing fast and effective automated segmentation in meniscal assessment. In this context, segmentation models developed with deep learning techniques achieve successful results in tasks such as the analysis of ultrasound images and automatic measurement of the meniscal area [13].

In terms of segmentation algorithms, U-Net-based architectures stand out with their simple structure and high performance, providing precise segmentation at the pixel level [12–16]. In addition, Mask R-CNN-based architectures are effective in discriminating different regions of meniscal tissue (e.g., medial and lateral) and injury types (e.g., tear and degeneration), and excel in instance segmentation tasks [11,17,18].

The objectives of the studies also determine the choice of these methods. For example, some studies in the literature focus only on segmenting meniscal tissue [12,15,18,19], while others use segmentation as a preprocessing step to detect meniscal injuries [11,14,16,17].

This study used the YOLOv8, YOLOv9, and YOLOv11 series of YOLO algorithms for meniscus segmentation. Segmentation masks were created with five models (YOLOv8x-seg, YOLOv8x-seg-p6, YOLOv9c-seg, YOLOv9e-seg, YOLOv11-seg) that can perform the segmentation task. By utilizing the different structural properties of these models, the strengths of each model are combined through ensemble methods. The masks obtained using the YOLO series are combined with innovative methods such as pixel-based multiple voting and dynamic weight optimization to improve the performance of the YOLO series on meniscus segmentation (Figure 1). Despite the significant advances in the literature, studies using ensemble methods for meniscus segmentation and injury classification are limited.

Therefore, ensemble methods aim to fill an important gap in this field. Furthermore, this is the first known study using the YOLO series for meniscus segmentation. The proposed method is validated with a test set (internal dataset) and an external set (external dataset). The contributions of this study to the literature are as follows:

- Using YOLO series models, the proposed ensemble methods (pixel-based voting, weighted multiple voting, and dynamic weighted multiple voting optimized by grid search) improved the meniscus segmentation performance.
- The proposed ensemble-based approach sheds light on the development of studies using ensemble methods as a preprocess in the detection or classification of meniscal tears.
- By improving the accuracy and reliability of the results of meniscus segmentation, the proposed method has a significant impact on the planning of surgical interventions and the improvement of the quality of life of those affected.
- The experiments on the internal and external sets demonstrated the generalization ability of the proposed method and its applicability in different clinical settings.

In the following sections of the study, the datasets, segmentation models, and ensemble methods are described in the materials and methods section. In the results section, the performance of the proposed method is evaluated using several metrics and is compared with other methods. In the discussion section, the strengths and weaknesses of the proposed method are discussed and compared with the literature. In the conclusion section, the general findings are summarized, and the potential impacts of the method in the health field are emphasized.

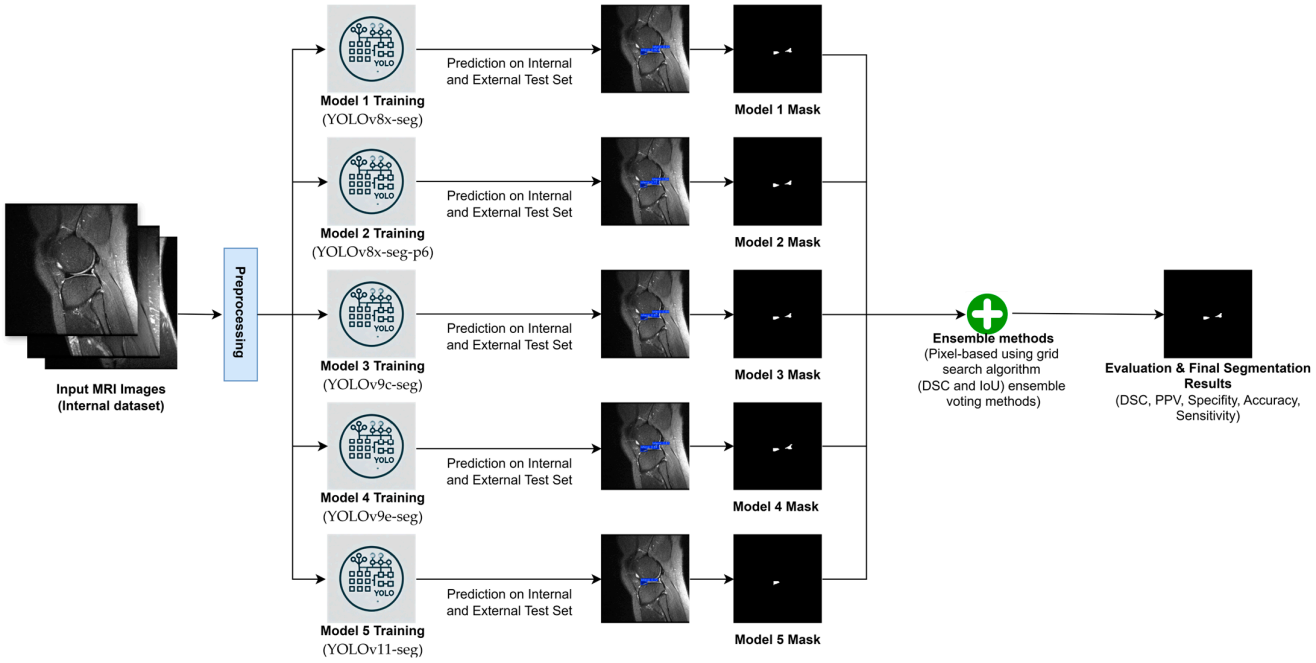


Figure 1. General workflow of the proposed segmentation approach.

2. Materials and Methods

2.1. Ethical Approval

The present study is a retrospective randomized study approved by the local ethics committee.

2.2. Source of Datasets

This study used two different datasets, an internal dataset and an external set, for meniscal segmentation. The internal dataset was obtained from the FastMRI initiative database provided by New York University (NYU) (https://fastmri.med.nyu.edu/) [20,21]. The FastMRI dataset contains 10,012 consecutive DICOM (Digital Imaging and Communications in Medicine) images from 9290 patients from clinical knee MRI examinations, providing a wide range of clinical images representing different tissue contrasts and various imaging planes. This dataset is also notable for including k-space data [21]. The external set was obtained using the MRNet dataset of 1370 knee MRI examinations performed at Stanford University Medical Center [22].

2.3. Internal Dataset

To create the internal dataset, a random selection was made from the DICOM format radiologic images of 600 patients in the FastMRI database. The selected images were limited to radiologic images of 471 patients after the application of the exclusion criteria (Figure 2). T2 and PD-weighted images were selected. The age range of the selected patients was from 2 to 85 years, with a mean age of 47.79 years; 44.58% (n = 210) were female, 42.02% were male, and 13.38% (n = 63) had no gender information. Right-knee MRI images were used in 53.08% (n = 250), and left-knee MRI images were used in 46.92% (n = 221) of the patients. While 16.35% (n = 77) of the patients had no meniscal tear, 83.65% (n = 394) had meniscal tears. In the group with meniscal tears, 53.81% (n = 212) of the tears were in the right knee, 46.19% (n = 182) in the left knee, 67.51% (n = 266) in the medial meniscus, and 32.49% (n = 128) in the lateral meniscus.

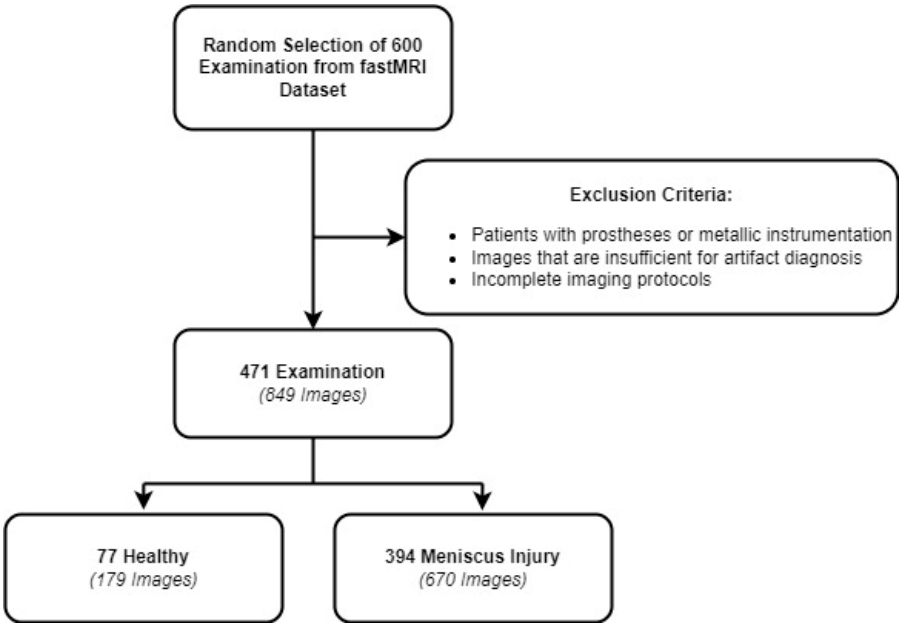


Figure 2. The process of creating the internal dataset.

In total, 849 images in portable network graphics (PNG) format were obtained from radiologic images of 471 patients. An average of 1.8 images were extracted for each patient. Of these images, 21.08% (n = 179) were healthy and 78.92% (n = 670) had meniscal damage. Of the healthy images, 51.96% (n = 93) belonged to the lateral meniscus and 48.04% (n = 86) to the medial meniscus. Of the images with meniscus damage, 32.24% (n = 216) belonged to the lateral meniscus and 67.76% (n = 454) to the medial meniscus. The medial meniscus was torn approximately three times more frequently than the lateral meniscus due to

reduced mobility caused by meniscocapsular attachment points [23]. The difference in distribution between the lateral and medial meniscus in images with meniscal tears is due to the anatomical structure and limited mobility of the medial meniscus.

The acquired images were divided into training, validation, and test sets in a ratio of 7:2:1, respectively. Stratified random sampling was used to maintain the proportions of lateral meniscus, medial meniscus, healthy, and torn meniscus in each set. This approach prevented imbalances in data distribution and ensured homogeneous class representation in each set. To show the distribution of the dataset more clearly, Table 1 shows the number of healthy and torn meniscus images in the training, validation, and test sets.

Table 1. Distribution of the internal dataset to the training, validation, and test groups.

		Train		Valid		Test	
		Image	**Labeling**	**Image**	**Labeling**	**Image**	**Labeling**
Healthy	LM	65	115	19	34	9	16
	MM	60	115	17	33	9	17
Meniscus Injury	LM	151	268	43	77	22	39
	MM	318	577	91	165	45	82

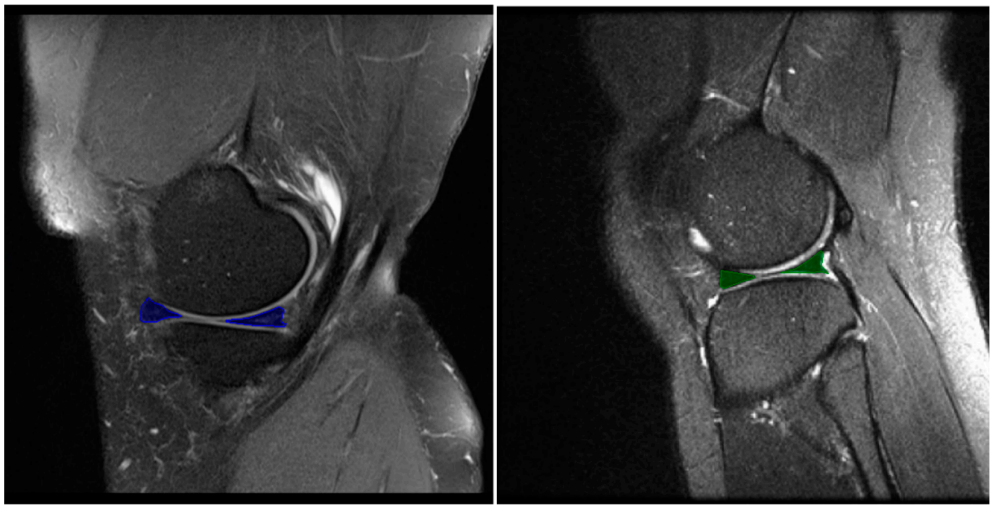
MM: medial meniscus, LM: lateral meniscus.

2.4. External Set

External validation is critical to evaluate the accuracy of a model on different datasets and to test its generalizability [24]. In this study, the MRNet dataset was used to evaluate the performance of the proposed method on an external dataset. The purpose of creating the external validation set with a similar distribution to the test set is to obtain fair and comparable results. In this context, an external validation set of 85 images was created by preserving general features such as the number of healthy and meniscal tear images in the test set and the lateral and medial meniscal image distributions. The comparisons in this study were performed between the test set of the internal dataset and the external set (external validation).

2.5. Image Selection and Labeling

In this study, image selection and labeling of the meniscal region of interest (ROI) areas were performed by a radiologist with 12 years of experience and an orthopedic and traumatologist with 12 years of experience. Considering that sagittal slices contain the most information on meniscal tears and because they better meet the clinical requirements [16,25], only sagittal slice images were used in this study. In image selection, images including both the anterior and posterior horns in the sagittal planes were preferred. Still, images showing the medial region of the meniscus were also included in the dataset. These labels produced by experienced radiologists and orthopedists were used as the gold standard (Figure 3). This study created separate ROIs for image segmentation for the anterior and posterior horn regions. A total of 1548 segmentation areas were labeled from 849 images. Of these labels, 35.70% (n = 549) belonged to the lateral meniscus (LM) and 63.30% (n = 999) to the medial meniscus (MM).



(a) (b)

Figure 3. ROI areas for meniscal segmentation in the sagittal section. (**a**) An image with meniscal injury in the posterior and anterior horn of the medial meniscus in the test set. (**b**) Representation of healthy lateral meniscus in the external set.

2.6. Preprocessing of Images

Since the obtained MR images had different resolutions, they were preprocessed and converted to the input dimensions (640 × 640 pixels) of the YOLO algorithm. Apart from this, no preprocessing was performed on the images.

2.7. Segmentation Models

The YOLO series models offer high speed and accuracy and stand out in object detection tasks. They also provide effective results in instance segmentation tasks. These models, which are especially critical for accurately detecting anatomical structures in biomedical images, are among the fastest and most accurate approaches for real-time object segmentation thanks to their one-stage design idea and powerful feature extraction capabilities. YOLO's object segmentation models have become the focus of recent studies, offering the advantage of higher accuracy and speed than two-stage segmentation models [26,27].

This study used the Ultralytics [28] framework to train segmentation models. Five different models (YOLOv8x-seg, YOLOv8x-seg-p6, YOLOv9c-seg, YOLOv9e-seg, YOLOv11-seg) from the YOLOv8, YOLOv9, and YOLOv11 series that fulfill the segmentation tasks supported by this framework were used in our study. In the following sections of the study, these models are named Model 1, Model 2, Model 3, Model 4, and Model 5, respectively.

2.8. Ensemble Methods

These machine learning approaches aim to create more powerful and effective prediction models by combining the output of multiple individual predictors. They provide an effective approach to solving complex prediction problems, especially in biomedical fields. These methods often offer higher accuracy and reliability by combining the results of different types of base predictors [29,30]. In this study, the strengths of different YOLO models were combined to improve segmentation accuracy. Masks from five different models were used for this purpose.

2.9. Ensemble Voting Methods

Ensemble Voting Methods allow for more reliable and accurate detection by combining the predictions of different classifiers [31]. In addition to the masks generated through YOLO models, voting-based ensembles were evaluated using three different strategies:

(i) pixel-based voting, (ii) weighted multiple voting, and (iii) dynamic weighted multiple voting with grid search. These strategies aim to achieve better results by combining the diversity and strengths of different models. Empirical results and quantitative data are presented in detail in the findings section. Within the scope of multiple voting methods, prediction masks generated from the models obtained after the training of the YOLO series were created, and these masks were used as input images (I_{Mn}). These masks consist of black (0) and white (255) pixel values.

The pixel-based voting method is calculated as shown in Equation (1). In this equation, the input images for each model are (I_{M1}, I_{M2}, I_{M3}, I_{M4}, I_{M5}) and (x, y) are the pixel positions. $W(x, y)$ represents the average of the sum value for the (x, y) pixel position. The thresholding process is shown in Equation (2) and allows $O(x, y)$ for the assignment of white (255) or black (0) according to a value of 128.

$$W(x,y) = \frac{I_{M1}(x,y) + I_{M2}(x,y) + I_{M3}(x,y) + I_{M4}\,(x,y) + I_{M5}(x,y)}{5} \tag{1}$$

$$O(x,y) = \begin{cases} 255, & W(x,y) \geq 128 \\ 0, & W(x,y) < 128 \end{cases} \tag{2}$$

In the weighted multiple voting method, the three most successful models are selected from five models, and each input image is assigned a weight value between [0.1:0.9]. The weighted multiple voting method is shown in Equation (3), where w_i represents the weight coefficient of the input image. The thresholding process for the new mask image is completed as shown in Equation (2). The sample weight values of the input images are (0.5, 0.3, 0.2), respectively. The three highest-performing YOLO models were selected for training.

$$W(x,y) = w_1 I_{M1}(x,y) + w_2 I_{M2}(x,y) + w_3 I_{M3}(x,y) \tag{3}$$

Grid search and dynamic weighted multiple voting are used to determine the optimal weight combinations by the pixel-based weighting of the prediction masks generated by different YOLO segmentation models. This method goes beyond static weights and enables a systematic search for dynamic weight values to optimize the prediction performance of each model. For this purpose, the grid search algorithm is used to determine the weights for each mask dynamically. Equation (4) is the equation for calculating the dynamic weighted voting method. The thresholding process for the new mask image was performed as shown in Equation (2). With this method, the segmentation performance of each model is calculated by the dynamic weights (w_i) and by grid search, and a composite prediction mask is created. This method provides a dynamic contribution to the segmentation performance of the models by taking into account their differences, and a significant improvement is realized to increase the segmentation accuracy.

$$W(x,y) = w_1 I_{M1}(x,y) + w_2 I_{M2}(x,y) + w_3 I_{M3}(x,y) + w_4 I_{M4}(x,y) + w_5 I_{M5}(x,y) \tag{4}$$

2.10. Grid Search Algorithm

This algorithmic approach aims to determine the configuration that gives the best results by evaluating the performance of the mask weights of each model. It offers a more structured and efficient optimization process than the trial-and-error method [32]. It evaluates a finite set of user-specified values through a Cartesian product [33]. In this study, the weight coefficients were scanned in steps [0.1, 0.9], and their sum was set to 1. To select the best weight combination, dice similarity coefficient (DSC) and intersection over union (IoU) metrics were used to evaluate them, and the best weight combination was selected.

2.11. Experimental Setup and Hyperparameters

This study trains the models on the Google Colab platform using the Ultralytics framework, YOLOv8.3.39 library, torch-2.5.1+cu121, and Python-3.10.12. Google Colab is a GPU-powered notebooking environment provided by Google servers. The training processes were run on NVIDIA A100-SXM4-40GB GPU (Nvidia, Santa Clara, CA, USA, 40,514 MiB) hardware, and segmentation models were created. All the rest of the processing was performed in Jupyter Notebook (version 6.5.4) and Python 3.10.

In this study, the dataset used for the segmentation task was kept constant, and only the model's hyperparameters were tweaked. In experiments with the same dataset, the model's performance was evaluated with lower epoch numbers. During the hyperparameter optimization, the basic parameters of the model, such as learning rate, momentum, and batch size, were tested with different combinations, and this process was carried out to ensure that the model produced more efficient results in a shorter time. Table 2 shows the hyperparameters used.

Table 2. Hyperparameters of the experiments.

Hyperparameter	Description
workers	8
batch	16
device	0
epochs	100
lr0	0.005
lrf	0.1
momentum	0.9
weight_decay	0.0005
warmup_epochs	5
warmup_momentum	0.8
warmup_bias_lr	0.05
optimizer	SGD

lr0: initial learning, lrf: final learning rate.

2.12. Performance Evaluation

It is important to use different evaluation parameters to comprehensively analyze the performance of the trained models and the proposed method, and to identify their strengths and weaknesses [32,34]. Therefore, recall (R), precision (P), and F1 score metrics were used to evaluate the performance of the trained models. DSC and IoU metrics were used to select the best weight combination in the grid search algorithm. DSC, PPV (Positive Predictive Value), Specificity, accuracy, and Sensitivity metrics were used to evaluate the image segmentation performance. DSC and PPV measure how accurately the model predicts positive pixels, Sensitivity measures how well true positives are captured, and Specificity measures how effectively false positives are excluded. Accuracy generally refers to the model's correct prediction rate. Furthermore, Cohen's Kappa, Jaccard Index, and Matthews Correlation Coefficient (MCC) metrics were used to analyze the consistency between external and internal validation and to assess the balance between classes.

During this evaluation process, additional analyses were performed using an external dataset to measure the generalizability of the method and to determine how it performs in real-world scenarios. Equations (5)–(13) shows the metrics used in this study.

$$R = \frac{TP}{TP + FN} \tag{5}$$

$$P = \frac{TP}{TP + FP} \tag{6}$$

$$F1\ score = 2\frac{P\ x\ R}{P + R} \tag{7}$$

$$DSC = \frac{2\ \times\ TP}{2\ \times\ TP + FP + FN} \tag{8}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{9}$$

$$PPV = \frac{TP}{TP + FP} \tag{10}$$

$$Specificity = \frac{TN}{TN + FP} \tag{11}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{12}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{13}$$

In these equations, TP (True Positive) is the number of pixels correctly predicted as positive, TN (True Negative) is the number of pixels correctly predicted as negative, FP (False Positive) is the number of pixels incorrectly predicted as positive when they are negative, and FN (False Negative) is the number of pixels incorrectly predicted as negative when they are positive.

3. Results

Five models (YOLOv8x-seg, YOLOv8x-seg-p6, YOLOv9c-seg, YOLOv9e-seg, YOLOv9e-seg, YOLOv11-seg) were trained for the segmentation of ROIs in the meniscus region using the existing YOLO algorithms YOLOv8, YOLOv9, and YOLOv11 series. In the rest of the paper, these models will be referred to as Model 1, Model 2, Model 3, Model 4, and Model 5, respectively. To improve the prediction results obtained from these five models, ensemble-methods-based approaches were used, and pixel-based voting, weighted multiple voting, and dynamic weighted multiple voting optimized by grid search were applied. In particular, dynamic weighted multiple voting with grid search significantly improved the model performance.

R, P, and F1 score metrics were used to measure the performance of the models at the end of training. The best metric values (best) and the last epoch value (last) obtained by the models for 100 epochs in the mask generation task are given in Table 3 and Figure 4. When Table 4 and Figure 4 are carefully analyzed, Model 1 showed the highest performance in terms of best metric values with P (0.9379), R (0.9277), and F1 score (0.9328) values. Considering the last epoch values, Model 1 maintained its superiority over the other models, although it showed a slight decrease in P (0.9312), R (0.9084), and F1 score (0.9196) metrics. Figure 4 shows that during the training process, the parameter values of the models initially increase rapidly and then reach a stable point. In particular, Figure 4 shows that Model 1 reaches a stable performance in the early epoch values, while Model 5 fluctuates a lot during the training process.

Table 3. Shows the performance results of the models in the precision, recall, and F1 score metrics.

	Best			Last		
Model	**P**	**R**	**F1 Score**	**P**	**R**	**F1 Score**
Model 1	**0.9379**	**0.9277**	**0.9328**	**0.9312**	**0.9084**	**0.9196**
Model 2	0.8874	0.8574	0.8722	0.8738	0.8268	0.8496
Model 3	0.9064	0.9093	0.9078	0.9064	0.8740	0.8899
Model 4	0.9092	0.8866	0.8977	0.8714	0.8801	0.8757
Model 5	0.8790	0.8563	0.8675	0.8396	0.8192	0.8293

P: precision, R: recall. Values in bold represent the best results obtained for each metric.

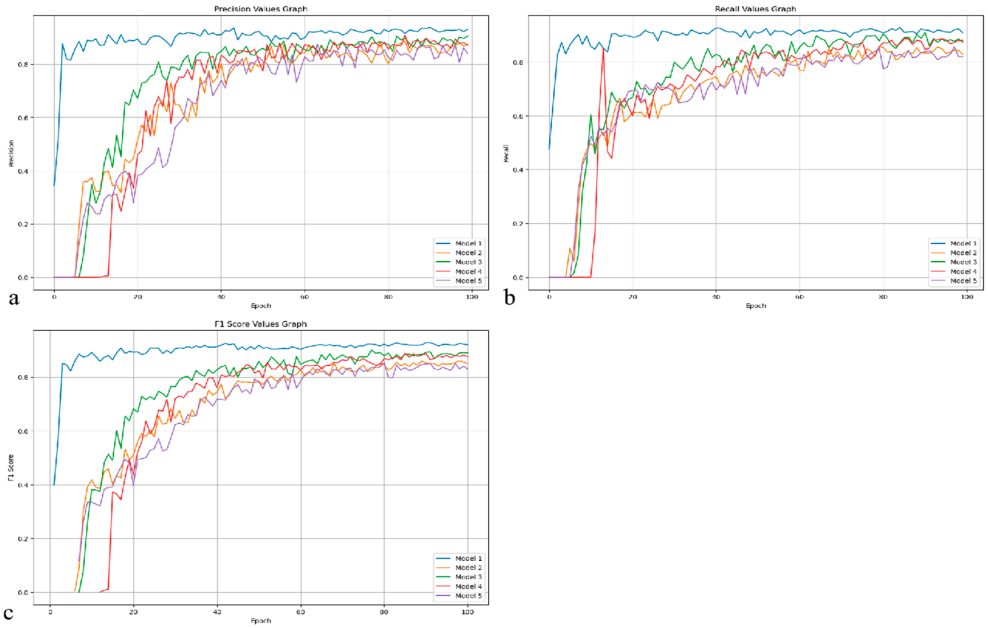


Figure 4. Curves showing precision, recall, and F1 score values of the models over 100 epochs. (**a**) P value curve. (**b**) R value curve. (**c**) F1 score value curve.

The models were trained on the internal dataset and evaluated with the test and external sets. Figure 5 shows the results of the qualitative images obtained. Table 4 shows the number of images from the test set and the external set for which the models failed to generate ROI areas during the evaluation. Model 1 successfully predicted ROI areas for all test and external set images. Model 2, on the other hand, fell behind the other models as the model with the highest number of unsegmented ROI areas.

Table 4. Number of unsegmented images in test and validation sets.

	Test Set	External Set
Model 1	0	0
Model 2	6	8
Model 3	0	1
Model 4	1	1
Model 5	3	4

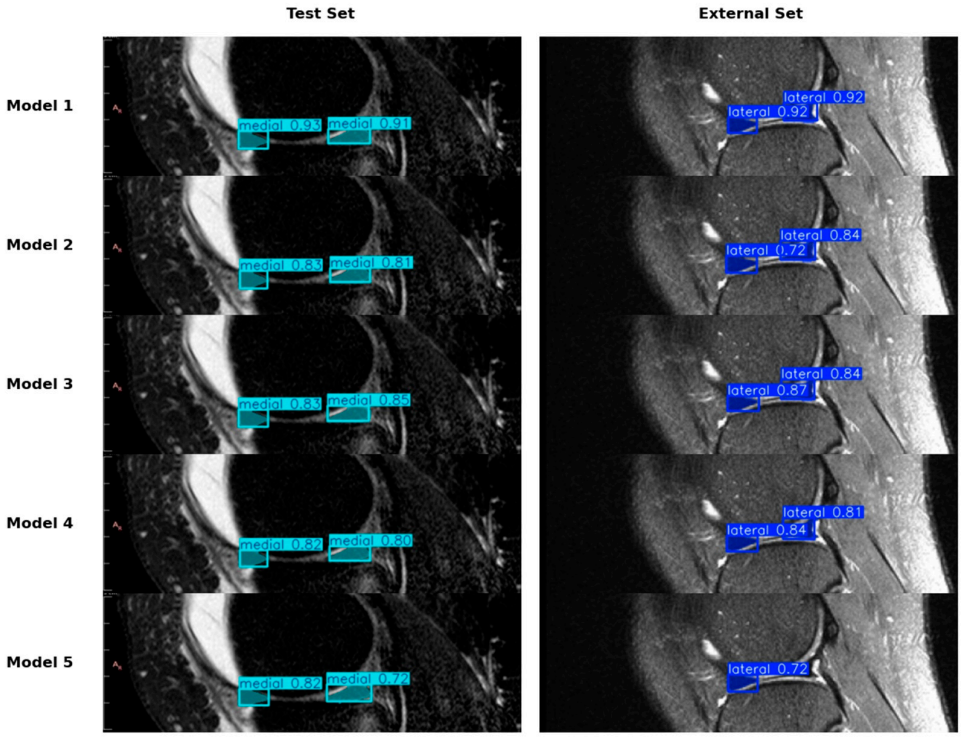


Figure 5. Qualitative visual analysis results were obtained by the models in the test set and external set. The image in the test set is the medial meniscus with meniscal damage, the image in the external set is the lateral meniscus with healthy meniscus.

The trained models were evaluated in the test set and external set with a confidence threshold of 0.5, and ROI masks of the meniscus region predicted by the model were created from these images. The confidence threshold was set to 0.5 to ensure a balanced performance between P and R values. The resulting masks were used for qualitative visual analysis and evaluated in the proposed method's pipeline. Figure 6 shows examples of the masks produced due to this process. The original images, ground truth, and the prediction masks of Model 1 in the test set are also shown in Figure 6.

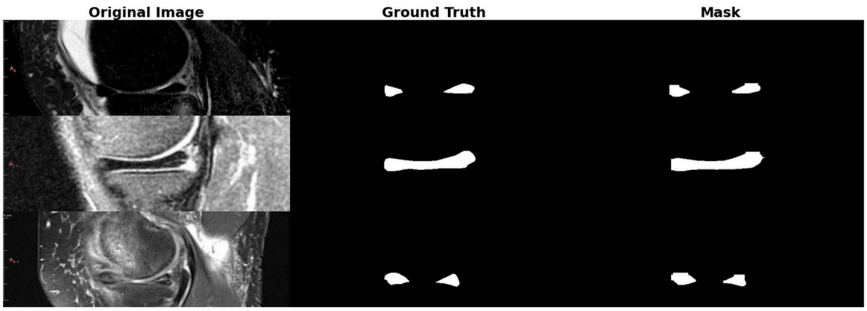


Figure 6. Masks of the meniscus region predicted by Model 1.

The masks created using YOLO models were developed with Ensemble Voting Methods and evaluated both in the test and external sets. Several methods were tested in the study, including pixel-based voting (Method 1), weighted ensemble voting (Method 2), and dynamic weighted ensemble voting optimized by grid search (Method 3). The masks for these methods are generated by the multiple voting methods described in the Materials and Methods section. The quantitative results of the models and the proposed method on the test set and the external set are presented in Tables 5–7. Performance metrics are reported with 95% confidence intervals, and confidence intervals are denoted by the ± symbol to

represent the variance around the mean values of the metrics. Table 5 summarizes the performance evaluations of the models applied to the test and external set and the masks obtained from multiple voting methods.

Table 5. Qualitative evaluation results of the masks created using YOLO models and multiple voting methods on the test set and external set.

	DSC	PPV	Specificity	Accuracy	Sensitivity
Test Set (Internal)					
Model 1	0.8720 (±0.0088)	0.8069 (±0.0145)	0.9991 (±0.0001)	0.9989 (±0.0001)	**0.9538 (±0.0089)**
Model 2	0.7650 (±0.0513)	0.7524 (±0.0473)	0.9992 (±0.0001)	0.9983 (±0.0004)	0.8083 (±0.0605)
Model 3	0.8453 (±0.0189)	0.8090 (±0.0150)	0.9992 (±0.0000)	0.9987 (±0.0002)	0.9068 (±0.0294)
Model 4	0.8369 (±0.0268)	0.8028 (±0.0239)	0.9992 (±0.0001)	0.9987 (±0.0002)	0.8910 (±0.0347)
Model 5	0.7860 (±0.0397)	0.7701 (±0.0356)	0.9992 (±0.0001)	0.9984 (±0.0002)	0.8354 (±0.0518)
Method 1	0.8583 (±0.0181)	0.8286 (±0.0151)	0.9993 (±0.0000)	0.9988 (±0.0002)	0.9068 (±0.0276)
Method 2	0.8743 (±0.0153)	0.8525 (±0.0136)	**0.9994 (±0.0001)**	0.9990 (±0.0001)	0.9092 (±0.0223)
Method 3	**0.8976 (±0.0071)**	**0.8561 (±0.0121)**	**0.9994 (±0.0000)**	**0.9992 (±0.0000)**	0.9467 (±0.0077)
External set					
Model 1	0.8852 (±0.0072)	0.8643 (±0.0148)	0.9993 (±0.0001)	0.9988 (±0.0001)	0.9159 (±0.0143)
Model 2	0.7642 (±0.0570)	0.7714 (±0.0562)	0.9994 (±0.0000)	0.9981 (±0.0004)	0.7812 (±0.0634)
Model 3	0.8585 (±0.0243)	0.8406 (±0.0255)	0.9992 (±0.0001)	0.9986 (±0.0002)	0.8916 (±0.0299)
Model 4	0.8260 (±0.0317)	0.8450 (±0.0269)	0.9993 (±0.0001)	0.9984 (±0.0002)	0.8418 (±0.0430)
Model 5	0.7800 (±0.0448)	0.8098 (±0.0427)	0.9993 (±0.0001)	0.9981 (±0.0003)	0.7842 (±0.0544)
Method 1	0.8574 (±0.0254)	0.8576 (±0.0257)	0.9994 (±0.0000)	0.9986 (±0.0002)	0.8753 (±0.0331)
Method 2	0.8852 (±0.0082)	0.8843 (±0.0145)	0.9994 (±0.0001)	0.9988 (±0.0002)	0.8966 (±0.0174)
Method 3	**0.9004 (±0.0064)**	**0.8876 (±0.0134)**	**0.9995 (±0.0000)**	**0.9990 (±0.0001)**	**0.9200 (±0.0119)**

The dark color indicates the best performance of the comparison methods.

Table 6. Qualitative evaluation results for the MM and LM of the masks created using YOLO models and multiple voting methods on the test and external sets.

	MM		LM	
	DSC	Sensitivity	DSC	Sensitivity
Test Set (Internal)				
Model 1	0.8838 (±0.0105)	0.9382 (±0.0127)	0.8748 (±0.0151)	0.9425 (±0.0172)
Model 2	0.8039 (±0.0459)	0.8436 (±0.0613)	0.6974 (±0.1163)	0.7468 (±0.1297)
Model 3	0.8477 (±0.0247)	0.9022 (±0.0386)	0.8412 (±0.0306)	0.9148 (±0.0472)
Model 4	0.8319 (±0.0403)	0.8733 (±0.0503)	0.8456 (±0.0253)	0.9219 (±0.0378)
Model 5	0.8054 (±0.0422)	0.8396 (±0.0581)	0.7522 (±0.0827)	0.8281 (±0.1043)
Method 1	0.8682 (±0.0186)	0.9137 (±0.0307)	0.8410 (±0.0387)	0.8947 (±0.0562)
Method 2	0.8717 (±0.0229)	0.8979 (±0.0336)	0.8790 (±0.0143)	0.9288 (±0.0192)
Method 3	**0.9006 (±0.0084)**	**0.9435 (±0.0104)**	**0.8923 (±0.0134)**	**0.9522 (±0.0111)**

Table 6. *Cont.*

	MM		LM	
	DSC	**Sensitivity**	**DSC**	**Sensitivity**
		External set		
Model 1	0.8891 (±0.0084)	0.8956 (±0.0189)	0.8782 (±0.0139)	0.9511 (±0.0158)
Model 2	0.7787 (±0.0673)	0.7644 (±0.0741)	0.7390 (±0.1084)	0.8105 (±0.1218)
Model 3	0.8535 (±0.0375)	0.8548 (±0.0443)	0.8673 (±0.0155)	0.9555 (±0.0108)
Model 4	0.8117 (±0.0477)	0.7888 (±0.0604)	0.8509 (±0.0270)	0.9341 (±0.0385)
Model 5	0.8175 (±0.0336)	0.7880 (±0.0513)	0.7147 (±0.1080)	0.7775 (±0.1246)
Method 1	0.8538 (±0.0382)	0.8401 (±0.0460)	0.8636 (±0.0236)	0.9366 (±0.0363)
Method 2	0.8852 (±0.0106)	0.8695 (±0.0231)	0.8854 (±0.0133)	0.9439 (±0.0160)
Method 3	**0.9055 (±0.0069)**	**0.9008 (±0.0151)**	**0.8916 (±0.0128)**	**0.9534 (±0.0129)**

The dark color indicates the best performance of the comparison methods. MM: medial meniscus, LM: lateral meniscus.

Table 7. Qualitative evaluation results of the masks created using YOLO models and multiple voting methods on the test set and external set for healthy meniscus and torn meniscus.

	Healthy		Tear	
	DSC	**Sensitivity**	**DSC**	**Sensitivity**
		Test Set (Internal)		
Model 1	0.8713 (±0.0203)	0.9541 (±0.0143)	0.8830 (±0.0095)	0.9359 (±0.0121)
Model 2	0.7836 (±0.1006)	0.8449 (±0.1197)	0.7601 (±0.0603)	0.7985 (±0.0710)
Model 3	0.8210 (±0.0504)	0.9022 (±0.0815)	0.8519 (±0.0202)	0.9080 (±0.0316)
Model 4	0.8354 (±0.0508)	0.9052 (±0.0761)	0.8373 (±0.0318)	0.8872 (±0.0399)
Model 5	0.8257 (±0.0254)	0.9341 (±0.0516)	0.7753 (±0.0500)	0.8089 (±0.0632)
Method 1	0.8616 (±0.0168)	0.9439 (±0.0298)	0.8574 (±0.0228)	0.8968 (±0.0341)
Method 2	0.8502 (±0.0534)	0.9001 (±0.0764)	0.8808 (±0.0138)	0.9116 (±0.0213)
Method 3	**0.8859 (±0.0178)**	**0.9556 (±0.0143)**	**0.9008 (±0.0077)**	**0.9443 (±0.0090)**
		External set		
Model 1	0.8705 (±0.0202)	**0.9591 (±0.0120)**	0.8891 (±0.0075)	0.9042 (±0.0170)
Model 2	0.7382 (±0.1379)	0.8113 (±0.1573)	0.7712 (±0.0640)	0.7731 (±0.0707)
Model 3	0.8541 (±0.0209)	0.9567 (±0.0124)	0.8597 (±0.0305)	0.8741 (±0.0369)
Model 4	0.8527 (±0.0296)	0.9574 (±0.0169)	0.8188 (±0.0396)	0.8108 (±0.0521)
Model 5	0.7695 (±0.1066)	0.8493 (±0.1286)	0.7829 (±0.0505)	0.7667 (±0.0608)
Method 1	0.8642 (±0.0224)	0.9575 (±0.0144)	0.8556 (±0.0319)	0.8532 (±0.0405)
Method 2	0.8741 (±0.0180)	0.9490 (±0.0150)	0.8882 (±0.0093)	0.8826 (±0.0205)
Method 3	**0.8829 (±0.0194)**	0.9565 (±0.0100)	**0.9051 (±0.0061)**	**0.9102 (±0.0140)**

The darker color indicates the best performance of the comparison methods.

As seen in Table 5, Method 3 shows successful results both on the test set and external set, in short, on all metric values. These results show that Method 3 succeeded in both segmentation accuracy and detecting positive samples correctly. In the external set, Method 3 achieved the highest value of 0.9004 (±0.0064) for the DSC metric, followed by 0.8976 (±0.0071) in the test set. Regarding PPV, Method 3 also stood out with values of 0.8561 (±0.0121) and 0.8876 (±0.0134) in both test and external sets, respectively. This shows that Method 3 performs strongly in maintaining the accuracy of optimistic predictions. Similar success was observed regarding Specificity and accuracy values, with very high values of

0.9995 (±0.0000) in both datasets. The results of these metrics show that the false positive rate is extremely low, and the overall accuracy of the model is high. In the Sensitivity metric, Method 3 outperformed the other methods with values of 0.9467 (±0.0077) on the test set and 0.9200 (±0.0119) on the external set. These results show that Method 3 can identify positive examples with high accuracy. It is seen that the methods built with multiple voting methods generally give better results than the models trained with YOLO. In model training, Model 1, which has the best performance, performs better than the other models.

The performance analyses of the methods and models were not limited to general evaluations on the test set and external set; detailed analyses specific to the lateral and medial meniscus regions and healthy and torn meniscus conditions were also performed. In these analyses, DSC and Sensitivity metrics were used to emphasize the overall performance and Sensitivity of the model. These methods' regional and case-based success were compared based on these two metrics. The performance results for the LM and MM ROI areas are presented in Table 6, while the results for the healthy and torn meniscus cases are presented in Table 7.

Table 6 shows that Method 3 is the most successful MM and LM segmentation method. For MM, 0.9006 (±0.0084) DSC and 0.9435 (±0.0104) Sensitivity values were obtained in the test set, and 0.9055 (±0.0069) DSC and 0.9008 (±0.0151) Sensitivity values were obtained in the external set. For LM, it showed superior performance with 0.8923 (±0.0134) DSC and 0.9522 (±0.0111) Sensitivity in the test set and 0.8916 (±0.0128) DSC and 0.9534 (±0.0129) Sensitivity in the external set.

Table 7 shows the performance of the masks for healthy and torn meniscus segmentation. In particular, Method 3 achieves the best healthy and torn meniscus segmentation results in both datasets. Model 1 stands out as the second most successful method for healthy meniscus after Method 3. In the external set, Model 1 stood out with DSC values of 0.8705 (±0.0202) and Sensitivity values of 0.9591 (±0.0120). However, Method 3 gave more consistent results with lower uncertainty than Model 1. Method 2 ranked second in the segmentation of torn meniscus. It performed close to Method 3 with DSC values of 0.8882 (±0.0093) and Sensitivity values of 0.8826 (±0.0205) in the external set.

Considering the quantitative results obtained by the proposed methods on the test and external sets, Method 3 performed the best. Cohen's Kappa, Jaccard Index, and Matthews Correlation Coefficient (MCC) metrics were used to evaluate the overall consistency of the method and its balance between classes. These metrics aim to demonstrate the method's success beyond chance-based classification and its Sensitivity to class imbalances. In the test set, these metrics were calculated as 0.8972 (±0.0071), 0.8158 (±0.0117), and 0.8991 (±0.0068), respectively. In the external set, 0.8999 (±0.0065), 0.8999 (±0.0065), and 0.9016 (±0.0061) values were obtained, respectively.

The DSC and Sensitivity differences between the test set and the external set of Method 3 were compared to evaluate the model's overall segmentation success and generalization ability. According to the *t*-test analysis, the *p*-value for DSC was calculated as 0.364. Since this value was more significant than 0.05, it was determined that there was no statistically significant difference between the DSC metrics in the test set and the external set. On the other hand, the *p*-value for Sensitivity is calculated as 0.0000087, which is much smaller than 0.05, so there is a statistically significant difference between the Sensitivity metrics between the two datasets.

4. Discussion

This study aims to extract meniscal ROIs from knee MRI images and improve segmentation results using ensemble-based approaches. We trained segmentation models based

on state-of-the-art YOLO versions. We optimized the results using innovative ensemble methods such as pixel-based voting, weighted multiple voting, and dynamic weighted multiple voting with grid search. Meniscal volume is known to be less than 0.1% of the entire knee joint MRI scan [35] and, in the 2D images used in this study, meniscal ROI areas represent less than 1.5% (Figure 7). The manual segmentation of such small tissues has limited applicability due to its high risk of error and time-consuming nature. Therefore, deep learning-based approaches offer practical solutions to overcome these challenges. Furthermore, this is the first known study in which segmentation models based on the YOLO series are integrated with innovative ensemble methods (e.g., grid-search-based dynamic weighting) to improve performance.

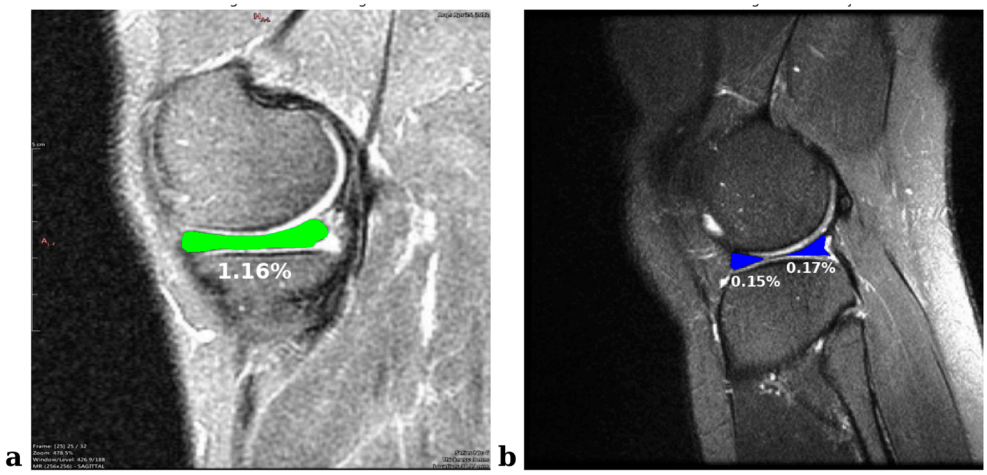


Figure 7. An example of knee MRI images that represent the sizes of meniscal ROI areas. (**a**) Medial meniscus with meniscal tear in the internal set. (**b**) Healthy lateral meniscus in the external set.

The YOLO series has an innovative architecture capable of fast and highly accurate object detection and stands out for its speed and accuracy advantages in biomedical imaging and segmentation tasks [36]. Ensemble methods combine the strengths of these models to improve their results further. Pixel-based voting, weighted multiple voting, and dynamic weighted voting with grid search combine the outputs from different models to provide more accurate and consistent results. These methods are powerful solutions for challenging segmentation tasks, especially in complex biomedical images. Incomplete segmentation results are reduced due to inhomogeneous intensity levels in the detection of meniscus ROI areas on MRI images. The ground truth comparison of the proposed method (Method 3) and the mask obtained using Model 1, which shows the best result in training, is given in Figure 8. As can be seen in Figure 8, one of the main reasons for incomplete segmentation is the small size of the meniscus ROI areas and the inhomogeneous intensity levels in the MRI images. Meniscal tissue shows similar intensities to other soft tissues on MRI images, resulting in limited contrast, especially in the posterior horn regions of the lateral and medial meniscus. This may lead to incomplete or incorrect boundaries in the segmentation masks. In addition, artifacts and low signal intensity in some regions of the images also contribute to incomplete segmentation. In Figure 8, when the masks produced by Model 1 are compared with ground truth, such missing boundaries are more evident. However, the proposed Method 3 overcomes some of these shortcomings through dynamic weighting and obtains results that are more consistent with the ground truth. This shows the success of Method 3, especially in balancing intensity differences and reducing incomplete segmentation.

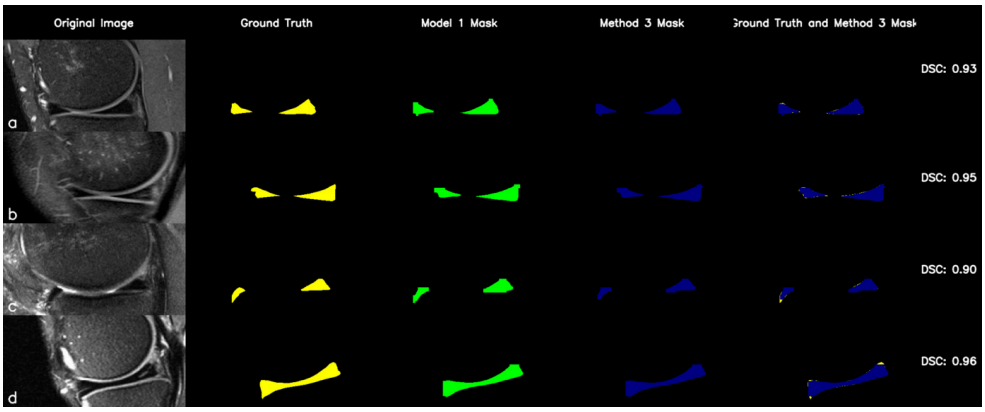


Figure 8. Comparison of segmentation masks obtained by the proposed method (Method 3) and Model 1 with ground truth. (**a**,**b**) Internal test images are healthy and torn meniscus samples, respectively. (**c**,**d**) External test images are healthy and torn meniscus samples, respectively.

Tables 3 and 4 show that Model 1 performs the best among the trained models. This can be explained by the contribution of the increase in the number of parameters in the YOLO series to the performance improvement [37]. The parameter numbers (Parameters/M) of the models are ~71.70 M (Model 1), ~5.09 M (Model 2), ~27.62 M (Model 3), ~59.68 M (Model 4), and ~2.83 M (Model 5). Although Model 1 outperformed the other models with its high number of parameters, qualitative evaluations also revealed cases where other models incorrectly estimated ROI areas. The main goal of ensemble methods is to improve segmentation accuracy by combining the strengths of each YOLO model and minimizing the inconsistencies between the outputs obtained from different models. In this study, the combination of masks from five different YOLO models demonstrates the contribution of this diversity to segmentation performance.

The quantitative results of the models and the proposed method on the test set and the external set are presented in Tables 5–7. In general, the findings in Table 5 show that Method 3 performs better and is more balanced than the other methods in terms of all metrics. In particular, the high performance of Method 3 on the external set highlights the model's ability to cope with data diversity and its generalization success. The consistency provided by multiple voting methods provides a generalizable and reliable solution for clinical applications. For example, Method 3 showed consistent results in metrics such as DSC (0.8976 ($\pm$0.0071), 0.9004 ($\pm$0.0064)) and Specificity (0.9467 ($\pm$0.0077), 0.9200 ($\pm$0.0119)) in the test set and external set, indicating that this method can work with similar accuracy on different patient groups.

Table 6 gives quantitative evaluation results according to menu types. Although the images containing LM are fewer than MM in the datasets, DSC and Sensitivity in the test set and external set obtained very close results. For example, the DSC for MM in the test set was 0.9006 ($\pm$0.0084), while in the external set, it was 0.9055 ($\pm$0.0069). This shows that the generalization ability is strong across different meniscus types.

Table 7 shows the relationship between healthy meniscal tissue and torn meniscal tissue. The detection of ROI areas of torn meniscus was more successful than healthy meniscus in both the test and external sets. For example, the DSC and Sensitivity for a torn meniscus in the test set were 0.9008 ($\pm$0.0077) and 0.9443 ($\pm$0.0090), respectively, while these values were 0.9051 ($\pm$0.0061) and 0.9102 ($\pm$0.0140) in the external set. One of the main reasons for this difference is that there are four times more torn meniscus images than healthy meniscus images in the datasets. This imbalance of data diversity is considered an essential factor affecting segmentation accuracy.

Although Method 3 achieved good results using ensemble methods, the high parameter count of Model 1 provided an advantage in terms of segmentation accuracy and offered the potential to reduce the need for manual segmentation significantly. However, the DSC and Sensitivity differences between the test set of Method 3 and the external set were compared to evaluate the model's overall segmentation success and generalization ability. According to the *t*-test analysis, the *p*-value for DSC is 0.364, which is greater than 0.05, indicating that the DSC differences between the test set and the external set are not statistically significant. On the other hand, the *p*-value for Sensitivity was calculated as 0.0000087, and since this value is less than 0.05, the difference between the two datasets is statistically significant. These results demonstrate the generalization success of Method 3 and its potential to work reliably on different datasets. This suggests that, while the method maintains overall segmentation accuracy, its ability to recognize all positive regions may vary depending on differences between datasets.

DL-based methods commonly used in the literature for meniscus segmentation include U-Net [12–16] and Mask R-CNN [11,17,18]. Although the first version of YOLO was released in 2016 [38], it gained the segmentation task later in the YOLO series. After U-Net was introduced as a convolutional network for biomedical image segmentation [39], U-Net-based architectures were developed for meniscus segmentation. Some of these studies have focused on the preprocessing of meniscal tear detection [11,14,16,17], while other studies have focused only on the success of segmentation [12,15,18,19]. Although sagittal [11,15–18,40] slices have generally been used for meniscal segmentation, there have also been studies in coronal slices [12] or both slices [35]. More research has been completed on the segmental section because it contains more information about the meniscus and better defines the clinical pattern [16,25]. Jeon et al. [12] proposed a U-Net-based method for meniscal segmentation and calculated DSC for MM and LM to be 85.18% and 84.33%, respectively [12]. In another U-Net-based study for the automatic segmentation of knee cartilage and meniscal structures, DSC and overall performance for MM and LM were calculated as 0.87%, 0.89%, and 0.88%, respectively [15]. In another U-Net-based study, DSC was reported in the range of 094–0.93 [16]. In the proposal of Li et al. [11] for a fully automatic 3D DCNN for the detection and segmentation of meniscal clefts in knee MR images, DSC and Sensitivity were 0.924 and 0.95, respectively. The overall performance evaluations of the proposed method (Method 3) were 0.8976 ($\pm$0.0071) and 0.9004 ($\pm$0.0064) for DSC in the test and external set, respectively, indicating that our study contributes to the literature.

While the related studies are based on a single model for meniscus segmentation, this study uses ensemble methods and operates according to the masks obtained from five different models. In this way, the performance of YOLO models is improved. In this study, beyond proposing a new convolutional neural network model, a new model is presented. YOLO models offer real-time performance and provide a significant advantage in overcoming the difficulties of manual segmentation. Furthermore, the accuracy of YOLO is improved by using a dynamic weighted multiple voting method with grid search, which is unique in the literature.

The internal dataset was divided into train, valid, and test sets using stratified random sampling to maintain class balance, make model performances consistent, and make quantitative and qualitative observations more consistent. One of the limitations of this study is that the total number of healthy and torn meniscus images in the dataset is different. Model 1 showed a superior performance due to its high number of parameters and complexity. However, this also reveals that the model needs more computational power. Other models are not as powerful as Model 1. The existence of another model close

to the performance of Model 1 will improve the performance of ensemble methods and contribute to the performance of the proposed method (Method 3).

The ability of the proposed model to automate the segmentation of the meniscus could significantly reduce the time required for manual labeling by radiologists and orthopedic surgeons. This efficiency gain could improve workflows in the clinical setting. In addition, the dynamically weighted ensemble approach increases the reliability of the model by reducing inter-observer variability, making it a promising tool for both diagnostic and surgical planning applications.

4.1. Pros of the Study, Advantages, and Future Work

This study will be extended with a new YOLO-based architecture, providing a pre-treatment for meniscal degeneration detection. Furthermore, the method's representativeness will increase if more extensive and balanced datasets are used. In addition, the integration of alternative imaging modalities, such as ultrasound, may increase this study's applicability to different imaging protocols and open a new research area in this field.

4.2. Limitations of the Study

This study used only the YOLO series, but no comparison was made with different segmentation models (e.g., U-Net, Mask R-CNN). This shortcoming limits the comparative power of the study between methods. Although performed by experts in the field, the manual process of meniscal ROI selection may lead to some labeling errors. Although the study results were tested by external validation, the diversity of these datasets is limited across different clinical protocols or devices. The model's performance on datasets from different MRI devices or imaging protocols has not yet been evaluated. In addition, evaluations on a single slice (sagittal) are another study limitation.

Although the proposed method shows strong segmentation performance, some limitations should be noted. The model performs exceptionally well in cases where meniscal tears have clear boundaries. However, for meniscal tears with irregular shapes or low contrast in MRI scans, the segmentation accuracy decreases slightly. This is probably due to the fact that, in these cases, it is difficult to distinguish meniscal tissue from the surrounding structures.

5. Conclusions

This study proposes an innovative method for meniscus segmentation using approaches based on YOLO series models and ensemble methods. Pixel-based voting, weighted multiple voting, and dynamic weighted multiple voting optimized by grid search are used for ensemble methods, and their performances are compared. Combining the strengths of these methods with the strengths of YOLO models contributes to the performance improvement in meniscus ROI area detection. According to the DCS, PPV, and Sensitivity metrics, the values of the proposed method in the test set were calculated as 0.8976 ($\pm$0.0071), 0.8561 ($\pm$0.0121), and 0.9467 ($\pm$0.0077), respectively. The same metrics were calculated as 0.9004 ($\pm$0.0064), 0.8876 ($\pm$0.0134), and 0.9200 ($\pm$0.0119) in the external set. In particular, the dynamic weighted multiple voting methods optimized with grid search showed superior performance in both datasets.

This study's findings show that the proposed method for meniscal segmentation offers high reliability and applies to clinical averages. Addressing the limitations mentioned in the discussion section and increasing the diversity of the dataset will make it possible to test the method in a wider range of clinical applications in the future.

This study makes an important contribution to the literature as it is the first known study to use the YOLO series for meniscal segmentation and optimizes its performance with ensemble methods.

Author Contributions: Conceptualization, M.A.Ş. and A.S.; Data curation, H.S. and Y.M.D.; Formal analysis, M.A.Ş.; Funding acquisition, M.A.Ş.; Investigation, M.A.Ş.; Methodology, A.S.; Project administration, A.S.; Resources, M.A.Ş.; Software, M.A.Ş.; Supervision, A.S., H.S., and Y.M.D.; Validation, A.S., H.S., and Y.M.D.; Visualization, M.A.Ş.; Writing—original draft, M.A.Ş.; Writing—review and editing, A.S., H.S. and Y.M.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The current study is a retrospective randomized study and was approved by the local ethics committee (Research Protocol Number: 2024.02.01.02, Date: 04 April 2024).

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: This manuscript was prepared using part of Mehmet Ali Şimşek's PhD thesis, conducted at İstanbul University-Cerrahpaşa.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pawar, R.S.; Yadav, S.K.; Kalyanasundaram, D. Evaluation of the stresses on the knee meniscus tissue under various loading conditions and correlation with resulting meniscal tears observed clinically: A finite element study. *J. Braz. Soc. Mech. Sci. Eng.* **2024**, *46*, 304. [CrossRef]
2. Bandyopadhyay, A.; Ghibhela, B.; Mandal, B.B. Current advances in engineering meniscal tissues: Insights into 3D printing, injectable hydrogels and physical stimulation based strategies. *Biofabrication* **2024**, *16*, 022006. [CrossRef]
3. Skarpas, G.A.; Maniatis, K.; Barmpounakis, N.; Kakavas, G. Meniscal Repair with ArthroZheal® an Autologous Bioactive Fibrin Scaffold. A New Technique and Treatment Option. *Surg. Technol. Online* **2024**, *44*, 327–332. [CrossRef]
4. Wang, Y.; Ying, M.; Yang, Y.; Chen, Y.; Wang, H.; Tsai, T.; Liu, X. Multitask learning for automatic detection of meniscal injury on 3D knee MRI. *J. Orthop. Res.* **2024**, *43*, 703–713. [CrossRef] [PubMed]
5. Shaik, G.M.V.; Zhou, B.; Liu, Q. A Comparison Study in Detecting Knee Osteoarthritis Severity with Deep Leaning. In Proceedings of the 2024 20th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Guangzhou, China, 27–29 July 2024; pp. 1–6. [CrossRef]
6. Roblot, V.; Giret, Y.; Antoun, M.B.; Morillot, C.; Chassin, X.; Cotten, A.; Zerbib, J.; Fournier, L. Artificial Intelligence to Diagnose Meniscus Tears on MRI. *Diagn. Interv. Imaging* **2019**, *100*, 243–249. [CrossRef]
7. Saygili, A.; Albayrak, S. Meniscus segmentation and tear detection in the knee MR images by fuzzy c-means method. In Proceedings of the 2017 25th Signal Processing and Communications Applications Conference (SIU), Antalya, Turkey, 15–18 May 2017; pp. 1–4. [CrossRef]
8. Saygili, A.; Albayrak, S. Knee Meniscus Segmentation and Tear Detection from MRI: A Review. *Curr. Med. Imaging Former. Curr. Med. Imaging Rev.* **2020**, *16*, 2–15. [CrossRef]
9. Hao, S.; Zhou, Y.; Guo, Y. A Brief Survey on Semantic Segmentation with Deep Learning. *Neurocomputing* **2020**, *406*, 302–321. [CrossRef]
10. Hafiz, A.M.; Bhat, G.M. A survey on instance segmentation: State of the art. *Int. J. Multimed. Inf. Retr.* **2020**, *9*, 171–189. [CrossRef]
11. Li, Y.-Z.; Wang, Y.; Fang, K.-B.; Zheng, H.-Z.; Lai, Q.-Q.; Xia, Y.-F.; Chen, J.-Y.; Dai, Z.-S. Automated meniscus segmentation and tear detection of knee MRI with a 3D mask-RCNN. *Eur. J. Med. Res.* **2022**, *27*, 247. [CrossRef] [PubMed]
12. Jeon, U.; Kim, H.; Hong, H.; Wang, J. Automatic Meniscus Segmentation Using Adversarial Learning-Based Segmentation Network with Object-Aware Map in Knee MR Images. *Diagnostics* **2021**, *11*, 1612. [CrossRef]
13. Zheng, J.; Tian, M.; Zhou, M.; Cai, J.; Liu, C.; Lin, T.; Si, H. Radiological segmentation of knee meniscus ultrasound images based on boundary constraints and multi-scale fusion network. *J. Radiat. Res. Appl. Sci.* **2024**, *17*, 101037. [CrossRef]

14. Fang, Y.; Liu, L.; Yang, Q.; Hao, S.; Luo, Z. A new method for early diagnosis and treatment of meniscus injury of knee joint in student physical fitness tests based on deep learning method. *BioImpacts* **2025**, *15*, 30419. [CrossRef]
15. Gaj, S.; Yang, M.; Nakamura, K.; Li, X. Automated cartilage and meniscus segmentation of knee MRI with conditional generative adversarial networks. *Magn. Reson. Med.* **2020**, *84*, 437–449. [CrossRef]
16. Jiang, K.; Xie, Y.; Zhang, X.; Zhang, X.; Zhou, B.; Li, M.; Chen, Y.; Hu, J.; Zhang, Z.; Chen, S.; et al. Fully and Weakly Supervised Deep Learning for Meniscal Injury Classification, and Location Based on MRI. *J. Imaging Inform. Med.* **2024**, *38*, 191–202. [CrossRef] [PubMed]
17. Li, J.; Qian, K.; Liu, J.; Huang, Z.; Zhang, Y.; Zhao, G.; Wang, H.; Li, M.; Liang, X.; Zhou, F.; et al. Identification and diagnosis of meniscus tear by magnetic resonance imaging using a deep learning model. *J. Orthop. Transl.* **2022**, *34*, 91–101. [CrossRef] [PubMed]
18. Ölmez, E.; Akdoğan, V.; Korkmaz, M.; Er, O. Automatic Segmentation of Meniscus in Multispectral MRI Using Regions with Convolutional Neural Network (R-CNN). *J. Digit. Imaging* **2020**, *33*, 916–929. [CrossRef] [PubMed]
19. Byra, M.; Wu, M.; Zhang, X.; Jang, H.; Ma, Y.; Chang, E.Y.; Shah, S.; Du, J. Knee menisci segmentation and relaxometry of 3D ultrashort echo time cones MR imaging using attention U-Net with transfer learning. *Magn. Reson. Med.* **2020**, *83*, 1109–1122. [CrossRef] [PubMed]
20. Zbontar, J.; Knoll, F.; Sriram, A.; Murrell, T.; Huang, Z.; Muckley, M.J.; Defazio, A.; Stern, R.; Johnson, P.; Bruno, M.; et al. fastMRI: An Open Dataset and Benchmarks for Accelerated MRI. *arXiv* **2018**, arXiv:1811.08839.
21. Knoll, F.; Zbontar, J.; Sriram, A.; Muckley, M.J.; Bruno, M.; Defazio, A.; Parente, M.; Geras, K.J.; Katsnelson, J.; Chandarana, H.; et al. fastMRI: A Publicly Available Raw k-Space and DICOM Dataset of Knee Images for Accelerated MR Image Reconstruction Using Machine Learning. *Radiol. Artif. Intell.* **2020**, *2*, e190007. [CrossRef]
22. Bien, N.; Rajpurkar, P.; Ball, R.L.; Irvin, J.; Park, A.; Jones, E.; Bereket, M.; Patel, B.N.; Yeom, K.W.; Shpanskaya, K.; et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLoS Med.* **2018**, *15*, e1002699. [CrossRef]
23. Adams, B.G.; Houston, M.N.; Cameron, K.L. The Epidemiology of Meniscus Injury. *Sports Med. Arthrosc. Rev.* **2021**, *29*, e24–e33. [CrossRef] [PubMed]
24. Bleeker, S.E.; Moll, H.A.; Steyerberg, E.W.; Donders, A.R.T.; Derksen-Lubsen, G.; Grobbee, D.; Moons, K.G.M. External validation is necessary in prediction research: A clinical example. *J. Clin. Epidemiol.* **2003**, *56*, 826–832. [CrossRef]
25. Ma, Y.; Qin, Y.; Liang, C.; Li, X.; Li, M.; Wang, R.; Yu, J.; Xu, X.; Lv, S.; Luo, H.; et al. Visual Cascaded-Progressive Convolutional Neural Network (C-PCNN) for Diagnosis of Meniscus Injury. *Diagnostics* **2023**, *13*, 2049. [CrossRef] [PubMed]
26. Huang, S.; Sirejiding, S.; Lu, Y.; Ding, Y.; Liu, L.; Zhou, H.; Lu, H. YOLO-Med: Multi-Task Interaction Network for Biomedical Images. In Proceedings of the ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; pp. 2175–2179. [CrossRef]
27. Kang, M.; Ting, C.-M.; Ting, F.F.; Phan, R.C.-W. ASF-YOLO: A novel YOLO model with attentional scale sequence fusion for cell instance segmentation. *Image Vis. Comput.* **2024**, *147*, 105057. [CrossRef]
28. Glenn, J.; Ayush, C.; Jing, Q. YOLO by Ultralytics. Available online: https://github.com/ultralytics/ultralytics (accessed on 20 December 2024).
29. Whalen, S.; Pandey, O.P.; Pandey, G. Predicting protein function and other biomedical characteristics with heterogeneous ensembles. *Methods* **2016**, *93*, 92–102. [CrossRef] [PubMed]
30. Fodeh, S.J.; Brandt, C.; Luong, T.B.; Haddad, A.; Schultz, M.; Murphy, T.; Krauthammer, M. Complementary ensemble clustering of biomedical data. *J. Biomed. Inform.* **2013**, *46*, 436–443. [CrossRef]
31. Shandhoosh, V.; Venkatesh S, N.; Chakrapani, G.; Sugumaran, V.; Ramteke, S.M.; Marian, M. Intelligent fault diagnosis for tribo-mechanical systems by machine learning: Multi-feature extraction and ensemble voting methods. *Knowl. Based Syst.* **2024**, *305*, 112694. [CrossRef]
32. Malahina, E.A.U.; Iriane, G.R.; Belutowe, Y.S.; Katemba, P.; Asmara, J. A Grid-search Method Approach for Hyperparameter Evaluation and Optimization on Teachable Machine Accuracy: A Case Study of Sample Size Variation. *J. Appl. Data Sci.* **2024**, *5*, 1008–1025. [CrossRef]
33. Belete, D.M.; Huchaiah, M.D. Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. *Int. J. Comput. Appl.* **2022**, *44*, 875–886. [CrossRef]
34. Xie, Y.; Zhu, C.; Zhou, W.; Li, Z.; Liu, X.; Tu, M. Evaluation of machine learning methods for formation lithology identification: A comparison of tuning processes and model performances. *J. Pet. Sci. Eng.* **2018**, *160*, 182–193. [CrossRef]
35. Luo, A.; Gou, S.; Tong, N.; Liu, B.; Jiao, L.; Xu, H.; Wang, Y.; Ding, T. Visual interpretable MRI fine grading of meniscus injury for intelligent assisted diagnosis and treatment. *NPJ Digit. Med.* **2024**, *7*, 97. [CrossRef]
36. Ragab, M.G.; Abdulkadir, S.J.; Muneer, A.; Alqushaibi, A.; Sumiea, E.H.; Qureshi, R.; Al-Selwi, S.M.; Alhussian, H. A Comprehensive Systematic Review of YOLO for Medical Object Detection (2018 to 2023). *IEEE Access* **2024**, *12*, 57815–57836. [CrossRef]

37. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J. YOLOv10: Real-Time End-to-End Object Detection. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 107984–108011.
38. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
39. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Munich, Germany, 5–9 October 2015; pp. 234–241. [CrossRef]
40. Botnari, A.; Kadar, M.; Patrascu, J.M. Considerations on Image Preprocessing Techniques Required by Deep Learning Models. The Case of the Knee MRIs. *Maedica-J. Clin. Med.* **2024**, *19*, 526–535.