Economical hybrid novelty detection leveraging global aleatoric semantic uncertainty for enhanced MRI-based ACL tear diagnosis

Athanasios Siouras, Serafeim Moustakidis, George Chalatsis, Tuan Aqeel Bohoran, Michael Hantes, Marianna Vlychou, Sotiris Tasoulis, Archontis Giannakidis, Dimitrios Tsaopoulos

PII: S0895-6111(24)00101-0

DOI: https://doi.org/10.1016/j.compmedimag.2024.102424

Reference: CMIG102424

To appear in: *Computerized Medical Imaging and Graphics*

Received date: 22 April 2024
Revised date: 23 August 2024
Accepted date: 23 August 2024

Please cite this article as: Athanasios Siouras, Serafeim Moustakidis, George Chalatsis, Tuan Aqeel Bohoran, Michael Hantes, Marianna Vlychou, Sotiris Tasoulis, Archontis Giannakidis and Dimitrios Tsaopoulos, Economical hybrid novelty detection leveraging global aleatoric semantic uncertainty for enhanced MRI-based ACL tear diagnosis, *Computerized Medical Imaging and Graphics,* (2024) doi:https://doi.org/10.1016/j.compmedimag.2024.102424

Title: Economical hybrid novelty detection leveraging global aleatoric semantic uncertainty for enhanced MRI-based ACL tear diagnosis.

Authors: Athanasios Siouras[1], Serafeim Moustakidis[2], George Chalatsis[3], Tuan Aqeel Bohoran[4], Michael Hantes[3], Marianna Vlychou[5], Sotiris Tasoulis[1,*], Archontis Giannakidis[4,7,*], Dimitrios Tsaopoulos[6,*]

Affiliations:

[1]Department of Computer Science and Biomedical Informatics, School of Science, University of Thessaly, 35131 Lamia, Greece

[2]AIDEAS OÜ, 10117 Tallinn, Estonia

[3]Department of Orthopedic Surgery, Faculty of Medicine, University of Thessaly, 41500 Larissa, Greece

[4]School of Science and Technology, Nottingham Trent University, Nottingham NG11 8NS, UK

[5]Department of Radiology, School of Health Sciences, University Hospital of Larissa, University of Thessaly, Mezourlo, 41500 Larissa, Greece

[6]Centre for Research and Technology Hellas, 38333 Volos, Greece

[7]Archimedes Research Unit in Artificial Intelligence, Data Science and Algorithms, 15125, Marousi, GREECE

*Joint senior authors

Corresponding author:

Archontis Giannakidis, PhD

Assistant Professor in Data Science

School of Science and Technology, Nottingham Trent University

Tel: +44(0)115-84-83968

Email: archontis.giannakidis@ntu.ac.uk

Clifton Campus | Office: New Hall Block 245b | Nottingham | NG11 8NS | UK

Abstract

This study presents an innovative hybrid deep learning (DL) framework that reformulates the sagittal MRI-based anterior cruciate ligament (ACL) tear classification task as a novelty detection problem to tackle class imbalance. We introduce a highly discriminative novelty score, which leverages the aleatoric semantic uncertainty as this is modeled in the class scores outputted by the YOLOv5-nano object detection (OD) model. To account for tissue continuity, we propose using the global scores (probability vector) when the model is applied to the entire sagittal sequence. The second module of the proposed pipeline constitutes the MINIROCKET timeseries classification model for determining whether a knee has an ACL tear. To better evaluate the generalization capabilities of our approach, we also carry out cross-database testing involving two public databases (KneeMRI and MRNet) and a validation-only database from University General Hospital of Larissa, Greece.

Our method consistently outperformed (p-value<0.05) the state-of-the-art (SOTA) approaches on the KneeMRI dataset and achieved better accuracy and sensitivity on the MRNet dataset. It also generalized remarkably good, especially when the model had been trained on KneeMRI. The presented framework generated at least 2.1 times less carbon emissions and consumed at least 2.6 times less energy, when compared with SOTA.

The integration of aleatoric semantic uncertainty-based scores into a novelty detection framework, when combined with the use of lightweight OD and timeseries classification models, have the potential to revolutionize the MRI-based injury detection by setting a new precedent in diagnostic precision, speed and environmental sustainability. Our resource-efficient framework offers potential for widespread application.

Keywords: ACL, object detection, MRI, injury, classification, novelty detection, aleatoric uncertainty

1. Introduction

1.1. Background

The economic and societal impacts of anterior cruciate ligament (ACL) injuries are far-reaching, as they represent over 50% of all knee-related injuries and affect each year millions of individuals globally [1]. Only in the United States, the financial burden of ACL injuries exceeds $7 billion annually [2]. Young and athletic populations are particularly susceptible to these injuries, as engaging in professional or recreational activities can raise the risk of a knee injury [3]. This, in turn, contributes to a heightened likelihood of developing knee osteoarthritis [4], [5]. Therefore, prioritizing cost-effective and reliable diagnostic methods for ACL injuries is vital to alleviate the strain on healthcare systems and promote the well-being of affected individuals [2].

Although arthroscopy is considered the "gold standard" for diagnosing intra-articular knee pathology, it has limitations due to its invasive nature and potential complications [6]–[8]. Recently, magnetic resonance imaging (MRI) has emerged as the most widely adopted non-invasive imaging technique for evaluating ACL injuries [9]. However, the accurate image interpretation of knee MRI data necessitates the costly expertise of trained clinicians. In addition, given the substantial human subjectivity element, variations in ACL injuries diagnosis are not uncommon [10], [11]. Consequently, there is a pressing need for the development of MRI-based assessment methods that are both robust and user-friendly in order to enhance the diagnostic process for ACL injuries [12], [13].

In recent years, deep learning (DL) techniques have demonstrated promising capabilities in addressing the challenges associated with medical image interpretation by automatically learning complex, nonlinear patterns [14]–[16]. One notable area of computer vision that has benefitted greatly by DL is object detection (OD), which encompasses the simultaneous localization and classification of objects within images or videos. The DL-powered OD has the

potential to significantly enhance the analysis and interpretation of knee MRI datasets, and further advancements in the field of ACL tear diagnosis.

1.2. Clinical Background

MRI has become an invaluable noninvasive tool for assessing the integrity and healing of the ACL graft after reconstruction. Increasingly, MRI signal intensity measurements are utilized in clinical studies to evaluate injuries. Despite this, the heterogeneity in MRI acquisition and interpretation methodologies complicates the comparison of signal intensity between scans and different studies. The efficacy of using MRI in this context is quantified by a model that rates the normalcy of the ACL graft in terms of integrity, contour, direction, and thickness, with scores up to 100%. Interpretation of MR images is influenced by the radiologist's experience, study protocol parameters like signal scaling factors, voxel volume, pulse sequence weighting, and patient positioning. However, MRI is still valuable for monitoring the healing of the tendon-bone interface after ACL reconstruction, correlating with functional recovery of the knee joint, and assisting decisions regarding return to sports. Advances in artificial intelligence and deep learning further enhance the utility of MRI in this field. High-resolution imaging, such as with a 3T MR scanner, and specialized protocols like quantitative MRI with ultrashort echo time T2* and T2* mapping, offer improved evaluations of ACL injury. However, these advanced imaging protocols require longer scanning times and are not feasible for routine use in busy radiology departments. Recent studies have shown that even conventional imaging protocols can provide adequate assessments of ACL graft maturation.

1.3. Related Work

The topic of detecting ACL injuries from MRI has recently received considerable attention by the DL community [14]. Awan et al. [17] utilized class balancing and data augmentation to develop a customized 14-layer ResNet-based convolutional neural network (CNN) featuring six distinct paths. Jeon et al. [18] proposed an interpretable lightweight 3D deep-neural network model, outperforming (p-value<0.05) previous state-of-the-art (SOTA) models on the Chiba and Stanford knee datasets [19]. Astuto et al. [20] employed 3D CNNs to diagnose and grade ACL damage. Dunnhofer et al. [21] developed MRPyrNet, a novel CNN architecture which focuses on small, localized regions. By integrating MRPyrNet into existing diagnostic pipelines, the researchers significantly improved diagnostic capabilities, particularly for ACL and meniscal tears, due to the architecture's ability to exploit relevant appearance features.

All the above studies exemplify the potential of DL techniques to revolutionize the detection and assessment of ACL injuries in clinical practice. These papers have nevertheless used an ordinary classification framework for the detection of ACL injuries. Yet training a network on datasets with severely imbalanced classes poses a serious challenge to the classification-based predictive modeling, since the algorithm is more prone to errors in the minority than the majority class. This can be particularly hazardous in the clinical context where the minority class is typically the most valuable. In addition, existing studies derived their decision by relying on a single slice, which had typically been obtained in a manual fashion by the clinicians. Nonetheless, the investigated tissue realistically spans across multiple slices. Moreover, the network architecture in previous studies was designed and evaluated using MRI data from a single database. Therefore, the knowledge about the ability of these algorithms to generalize is limited. Finally, previous papers disregarded the compute demand by the DL model. Yet, this is a significant matter since these computations necessitate mind-boggling amounts of generated power for fuelling them [22].

1.4. Our contribution

In this paper, in order to enhance ACL tear diagnosis by tackling the class imbalance issue, presented in most publicly available knee MRI datasets, we propose to reformulate this task as a novelty detection problem [23]. In brief, our framework first attempts to accurately model "normality" by training an OD model (YOLOv5-nano [24]) exclusively on a dataset of healthy knees, devoid of any instance that we would like the framework to detect. To obtain a discriminative novelty score, we leverage the aleatoric semantic uncertainty as this is naturally and implicitly modeled in the metadata (class scores) outputted by the OD model. Given that the ACL tissue spans across multiple slices, we propose to make use of the global class scores obtained when the OD model is applied to the whole image sequence. Therefore, the novelty score is a vector of generated probabilities which indicate the likelihood that the object inside the bounding box belongs to the “healthy knee” class, expecting alternate (lower) values when instances diverge from the training dataset. As the final step of the proposed hybrid framework, we utilize a timeseries classification model (MINIROCKET [25]) for determining whether a knee has an ACL tear or not.

As well as representing the SOTA, both machine learning modules of our pipeline are also known to be highly resource-efficient, fostering green DL research and aligning with efforts to democratize it. Notably, such economical strategies are more relevant to being adopted by lightweight edge devices which in turn could boost their use.

The optimal visualization of ACL necessitates having a sagittal scanning plane [26]. This is also backed by a growing literature demonstrating that sagittal plane factors are responsible for ACL injury mechanisms [27]. Consequently, only data from the sagittal plane is used in this study.

Another critical contribution of this study is that we implement both single- and cross-database testing to better evaluate the generalization capabilities of our approach, involving two public databases namely MRNet [19] and KneeMRI [28], and a third validation-only dataset sourced from the Department of Orthopaedic Surgery and Musculoskeletal Trauma at the University General Hospital of Larissa (UGHL) in Greece. Our hybrid (two-phase) novel detection framework is rigorously compared to two SOTA methods [18], [21] with regard to accuracy, resource efficiency and environmental cost.

2. Materials and methods

2.1. Datasets and annotation

In this study, we utilized two public datasets (MRNet [19] and KneeMRI [28]) and one validation-only dataset sourced from UGHL in Greece.

MRNet [19] was created between 2001 and 2012 for the development and evaluation of artificial intelligence (AI) algorithms in medical imaging. It contains 1,370 knee examinations of 1,104 patients, along with labels indicating the presence or absence of abnormalities in four different categories: ACL tear, meniscus tear, abnormal cartilage, and abnormal bone marrow. A subset of this database was used, that is 266 healthy knees and 319 knees with ACL rupture. The number of slices in each 3D MRI scan ranges from 17 to 61, and the size of each MRI slice is 256×256 pixels.

The KneeMRI [28] dataset was retrospectively collected using proton density weighted fat suppression and a Siemens Avanto 1.5T MRI scanner at the Clinical Hospital Centre Rijeka, Croatia from 2006 until 2014. The dataset comprises 917 left- or right-knee 12-bit grayscale volumes. Following radiologist reports, the authors labeled each scan according to ACL disorder: non-damaged (690 scans), partially injured (172 scans), and totally ruptured (55 scans). Each MRI examination comprises 21–45 slices and the spatial resolution of each slice is 320×320 or 290×300 pixels.

At UGHL, MRI examinations of 129 ACL ruptured knees were conducted from 2018 onwards using a 3.0T MRI scanner (Signa HDxt 3.0T, GE Healthcare) with a quadrature knee coil, field of view (FOV) of 18x18 pixels, thickness of 1.0 (in millimeters), and a resolution of 254×256 pixels. The examinations were performed preoperatively on knees that were slated for surgery. The MRI acquisition protocol included the capture of sagittal and axial T1-weighted as well as T2-weighted images, supplemented by sagittal and coronal proton density weighted Turbo Spin Echo (TSE) images with fat saturation. A musculoskeletal radiologist with two decades of experience (M.V.) reviewed the MRI scans.

In the two public datasets each MRI slice comes pre-annotated. Regarding the UGHL dataset, there was only a single overarching annotation for the entire MRI scan.

2.2. Data preprocessing

Image augmentation was employed to enhance the generalization capacity of our pipeline by modifying image brightness (within a range from -32% to +32%), altering hue (ranging from -39° to +39°), adjusting exposure (between -30% and +30%), incorporating blurring (up to 1.75 pixels), and adding random noise (affecting up to 7% of pixels). The concluding step prior to initiating OD training involved exporting the annotations in an appropriate format.

2.3. Hybrid network architecture

The proposed hierarchical hybrid architecture (Figure 1) is composed of two key elements.

At first, we incorporate the latest advancements in OD technology. Specifically, we apply YOLOv5-nano [24], a SOTA OD network and part of the "You Only Look Once" series, to the complete sagittal MRI scan. The reason why we opted for the nano version of YOLOv5 for our task is due to its lightweight nature. When compared with the other available versions of YOLOv5, the nano variant trades off detection accuracy for computational efficiency. In brief, the architecture of YOLOv5 is divided into three essential parts (Figure 1):

- Backbone: The core CNN that processes input images and extracts features at various levels.
- Neck: Layers that mix and combine features from the backbone, preparing them for accurate predictions.
- Head: This part uses the combined features to predict object bounding boxes and their classes.

This structure allows YOLOv5 to efficiently detect objects (by predicting bounding boxes) and classify them in a single unified framework.

In this study, solely healthy ACLs are fed into YOLOv5-nano, in line with our aim to build a novelty-detection framework. To obtain an appropriate novelty score, we propose to capitalize on the metadata outputted by the OD model. In more detail, we exploit the aleatoric semantic uncertainty which is naturally and implicitly modeled in the class probability score, every time our framework is fed with an observation that does not belong to the trained distribution. These class scores are typically estimated by employing a combination of techniques such as softmax activation function, class-specific filters in the final layer of the neural network, and class detection thresholds. Taking into account that the ACL tissue spans across multiple MRI slices, we propose to make use of the global class probabilities obtained from the whole sagittal image sequence. As a result, the novelty score is a vector of generated probabilities which indicate the likelihood that the object inside the bounding box belongs to the “healthy knee” class (rather than background), and the expectation is that these will take different (lower) values every time the input instance diverges from the training population.

In the second phase of the proposed hybrid framework, we utilize a timeseries classification model (MINIROCKET [25]) for determining whether a knee has an ACL injury or not by applying the probability vector to it. MINIROCKET (MINImally RandOm Convolutional KErnel Transform) marked a significant advancement in the field of time series classification due to

its exceptional computational efficiency and speed, particularly on larger datasets. It offers the advantage of mostly deterministic operation, with an option for full determinism. This ensures both consistent and reliable results. In a nutshell, MINIROCKET's architecture is characterized by its use of a small, fixed set of convolutional kernels, each with a length of 9 and weights restricted to -1 and 2. This simplification not only maintains a similar level of accuracy compared to more complex models but also significantly reduces computational complexity. The algorithm is tailored to the length of the input time series through its dilation process and uses alternated padding for each kernel/dilation combination, adding to the transformation's consistency. In terms of feature handling, MINIROCKET simplifies its approach by focusing primarily on the Proportion of Positive Values (PPV) pooling, resulting in a total of about 10,000 features. This feature reduction effectively balances accuracy with computational efficiency, making the algorithm versatile for various applications. MINIROCKET's speed is further enhanced through several optimization techniques. These include computing PPV for a kernel and its inverse simultaneously, reusing convolution outputs for multiple feature computations, and performing most of the computations for all kernels at once for each dilation. While it brings in several advancements, MINIROCKET maintains linear scalability in relation to the number of features and the length of input time series. Although it requires slightly more memory than traditional methods due to the temporary storage of additional vectors during transformation, this increase is generally minimal. Utilizing the features it generates, MINIROCKET is capable of training linear classifiers, such as ridge regression or logistic regression, making it highly suitable for real-time and large-scale time series classification tasks where a balance between high accuracy and low computational expense is crucial. The ridge classifier was used in this study.

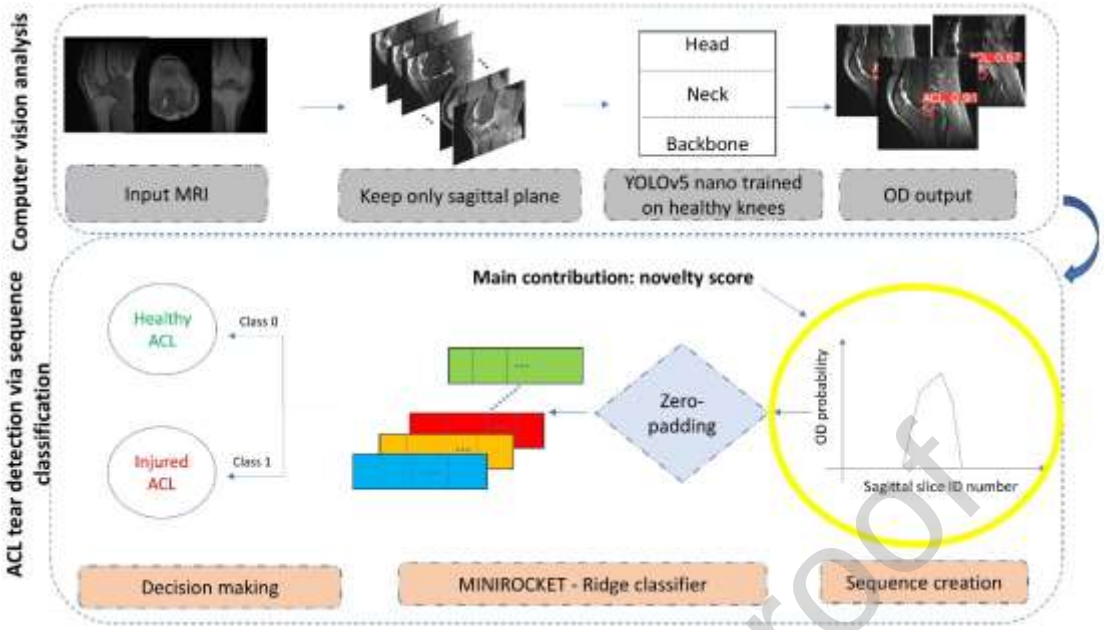


Figure 1: An illustration of the proposed hybrid DL. Highlighted is our core innovative contribution, that is a novelty score that capitalizes on the global aleatoric semantic uncertainty as this is modelled in the class scores outputted by object detection model. A novelty score that also incorporates critical clinical (ACL geometry) information and can be rapidly acquired.

2.4. Implementation and training

To better assess the generalization capabilities of our framework, we performed a total of six experiments, also allowing a cross-database testing. The two public databases (MRNet and KneeMRI) were utilized for training and evaluation purposes, with the third database (UGHL) being employed for external validation. The model evaluation protocols that we employed for the public databases adhere to the existing literature to ensure rigorous benchmarking and facilitate reliable comparisons across studies. In particular, all papers that discuss MRNet have undergone testing on its validation set. Likewise, every publication referring to KneeMRI has adopted a 5-fold cross-validation technique.

Therefore, the six experiments we carried out were:

- Exp1: Training on MRNet (Train) and testing on MRNet (Validation)
- Exp2: Training on MRNet (Train) and testing on the entire KneeMRI dataset
- Exp3: Training on MRNet (Train) and testing on the entire UGHL database
- Exp4: Training on KneeMRI and testing on KneeMRI (cross-validated)
- Exp5: Training on KneeMRI and testing on the entire MRNet.
- Exp6: Training on KneeMRI and testing on the entire UGHL.

We trained the OD model using exclusively healthy knees. Transfer learning was employed to address the limited dataset size. Zero padding was performed to the class probability vector outputted by YOLOv5-nano to create uniform data structures, essential for the legit application of our subsequent Ridge Classifier.

The pseudocode outlining the training and inference protocolss for the proposed hybrid architecture is provided below.

Table 1: Pseudocode of the hybrid architecture training and inference protocols.

Training protocol

DATASET: D (X)

D represents the dataset MRNet or KneeMRI.

Splitting the dataset:

Partition D into training and validation sets using a split function S:

$$(X_{train}, X_{val}) = S(D)$$

where

$$|X_{train}| = 0.8\ |D|,\ |X_{val}| = 0.2\ |D|$$

Data preprocessing or modifications Keeping Only Sagittal Plane Data:

Extract the sagittal plane data from X_{tr-} using function $F_{S.}$ Mathematically, this can be represented as:

$$X_{trs} = F_S(X_{tr});\ F_S: X_{train} \rightarrow X_{trs}$$

- Extract healthy data from X_{trs} using a function F_H:

$$X_{trhs} = F_H(X_{trs});\ F_H: X_{trs} \rightarrow X_{trhs}$$

Training phase for the first model:

Train YOLOv5-nano model 1 (M_1) on X_{trhs}

Data Preparation for the second model:

$Predictions_{train}$ (P_{tr}) = PREDICT using M_1 on X_{trs}

P'_{tr} = EXTRACT probabilities from P_{tr} per slice and zero padding.

Training phase for the second model: Train MINIROCKET (Ridge Classifier) model 2 (M_2) on P'_{tr}
Inference protocol
Initialize new MRI (Y) Keep sagittal slices from Y → Y_s FOR each slice i in Y_s: APPLY M_1 to slide i EXTRACT probability P_i from the output of M_1 PERFORM zero padding on P_i to match Model M_2 input dimensions → Vector (V) FOR V: APPLY Model M_2 EXTRACT final decision L_i RETURN L_i as final decisions for Y

We trained (Table 2) the YOLOv5-nano model for 150 epochs with images resized to 640x640 pixels. For the RidgeClassifier, the default hyperparameters were employed: In this configuration, the alpha parameter, which controls the strength of regularization, was set to 1.0. The maximum number of iterations for the algorithm to converge, was not limited (max_iter= None). The parameter indicating whether to calculate the intercept for this model was enabled (fit_intercept=True). The tolerance for algorithm stopping was set to 0.0001. The solver, which is the algorithm used for optimization was set to automatically choose the best method ('auto'). Finally, the random_state parameter, used to seed the random number

generator for reproducibility was not set (None), allowing the algorithm to use a random seed. Our framework was implemented[1] using Python version 3.10 on a system with operating system Ubuntu Linux 22.04 and an NVIDIA RTX A6000 (48GB) graphics card.

Table 2: Model structured details for YOLOv5-nano and RidgeClassifier.

Model Component	YOLOv5-nano	RidgeClassifier
Parameters	1.9 million	
Layer Settings	CSP-Darknet53 backbone, SPPF, PANet	N/A
Activation Functions	SiLU (Swish) for hidden layers, Sigmoid for output layer	N/A
Optimization Methods	SGD, Adam	Linear Regression with L2 Regularization
Training Epochs	150	N/A
Image Size	640x640 pixels	N/A
Alpha Parameter	N/A	1.0
Max Iterations	N/A	None
Fit Intercept	N/A	True
Tolerance	N/A	0.0001
Solver	N/A	Auto

[1] https://github.com/ThanosUTH/ACL-Tear-Diagnosis-

Random State	N/A	None
System Specs	Python 3.10, Ubuntu Linux 22.04, NVIDIA RTX A6000 (48GB)	

2.5. Model evaluation.

The ACL tear classification task involved evaluation metrics such as Receiver Operating Characteristic (ROC) curve and the respective Area Under the Curve score (AUC), accuracy, specificity, and sensitivity. Additionally, we employed the carbon dioxide equivalent emissions (CO_2eq) (g) produced and energy (kWh) spent during training, equivalent distance traveled by car (km), training time (h:min:s), and average inference time (ms) in order to provide a comprehensive assessment of the system's resource efficiency and generated carbon emissions. All evaluation metrics of the proposed methodology were juxtaposed with those yielded by two SOTA methods [18], [21] for comparison purposes.

2.6. Statistical analysis

We conducted a statistical analysis to compare the performance of our proposed method against two state-of-the-art approaches: the SOTA methods [18], [21]. We used two datasets for this evaluation. For the first dataset, we used the MRNet dataset with its predefined training and testing splits. Our proposed method, along with the methods from Dunnhofer et al. and Jeon et al., was trained on the training set and evaluated on the testing set. We calculated performance metrics such as Accuracy, Sensitivity, Recall, Specificity, AUC. To assess the statistical significance of the differences in performance, we performed permutation tests, where the training labels were shuffled 1000 times, and p-values were obtained by comparing the actual accuracy with the distribution of accuracies from the shuffled data.

For the second dataset, we employed 5-fold cross-validation to evaluate the models. Each method, including our proposed one, was trained and tested across five different splits of the data, with each subset serving as a test set once. We averaged the accuracies from these folds and used paired t-tests to compare the cross-validated accuracies of our proposed method against the methods by Dunnhofer et al. and Jeon et al. This approach allowed us to rigorously determine if the performance improvements of our proposed method were statistically significant and not due to random chance.

3. Results

3.1. Accuracy evaluation

Table 3 outlines the comparison results between the proposed framework and the SOTA methods in terms of ACL tear classification accuracy in single-database experiments (Exp1, Exp4). Our method consistently outperformed (p-value<0.05) SOTA approaches on the KneeMRI dataset and achieved better accuracy and sensitivity on the MRNet dataset. Figure 2 shows the ROC curves and the corresponding AUC scores of our classification models for the same experiments. As shown, the achieved AUC scores are outstanding. Figure 3 illustrates the bounding boxes and the respective class scores outputted by the OD component of our framework in Exp1 and Exp4, when this is fed with representative ACL sagittal slices of both healthy and injured knees. The overall probability vector obtained when applying the whole sagittal image sequence for representative cases is shown in Figure 4. It is obvious that the class scores are different (lower) every time our OD model is fed with a scan that diverges from the training dataset. Apparently, probability scores exceeding 0.80 are suggestive of an intact ACL.

Table 3: Comparison between the proposed and other SOTA approaches in terms of ACL tear classification performance in single database experiments. Boldface indicates best performance.

Approach	Experiment	Database	Accuracy	AUC	Sensitivity	Specificity	P-value
Dunnhofer et al. (2022) [21]	Exp1	MRNet	0.886	**0.976**	0.815	**0.944**	**0.01***
	Exp4	KneeMRI	0.834	0.914	0.806	0.843	**0.00***
Jeon et al. (2021) [18]	Exp1	MRNet	0.911	0.963	0.944	0.901	**0.00***
	Exp4	KneeMRI	0.826	0.851	0.801	0.874	**0.00***
Proposed	Exp1	MRNet	**0.949**	0.960	**0.963**	0.920	-
	Exp4	KneeMRI	**0.960**	**0.960**	**0.917**	**0.968**	-

***p-value<0.05**

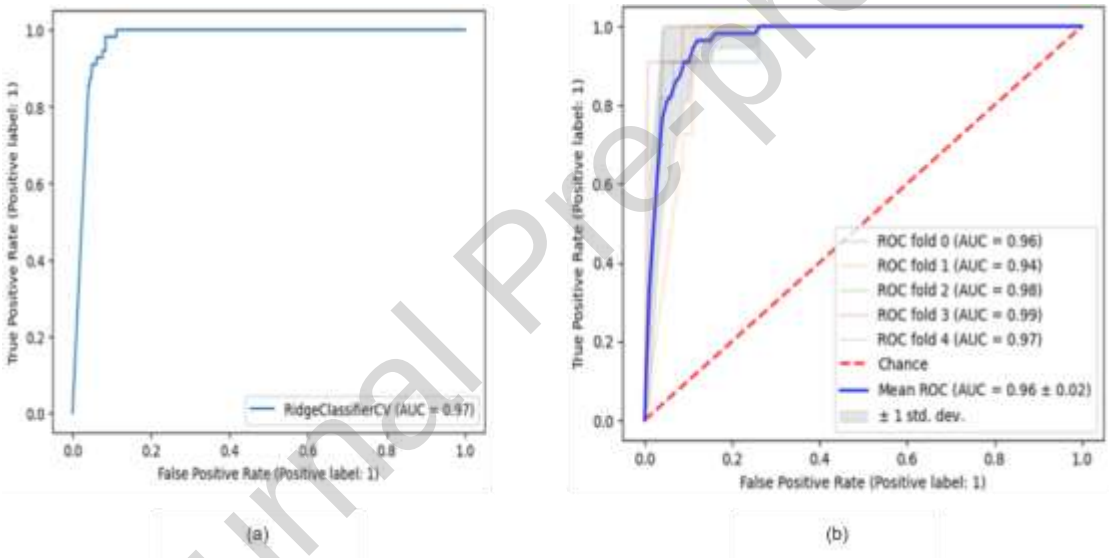


Figure 2: ROC curves and corresponding AUC scores illustrating the classification performance of the proposed models in Exp1 (a) and Exp4 (b).

Figure 3: Indicative results (i.e. bounding boxes and class scores) of the OD component of our models in Exp1 (a), Exp4 (b), when these are fed with typical ACL sagittal slices of both healthy and injured knees.

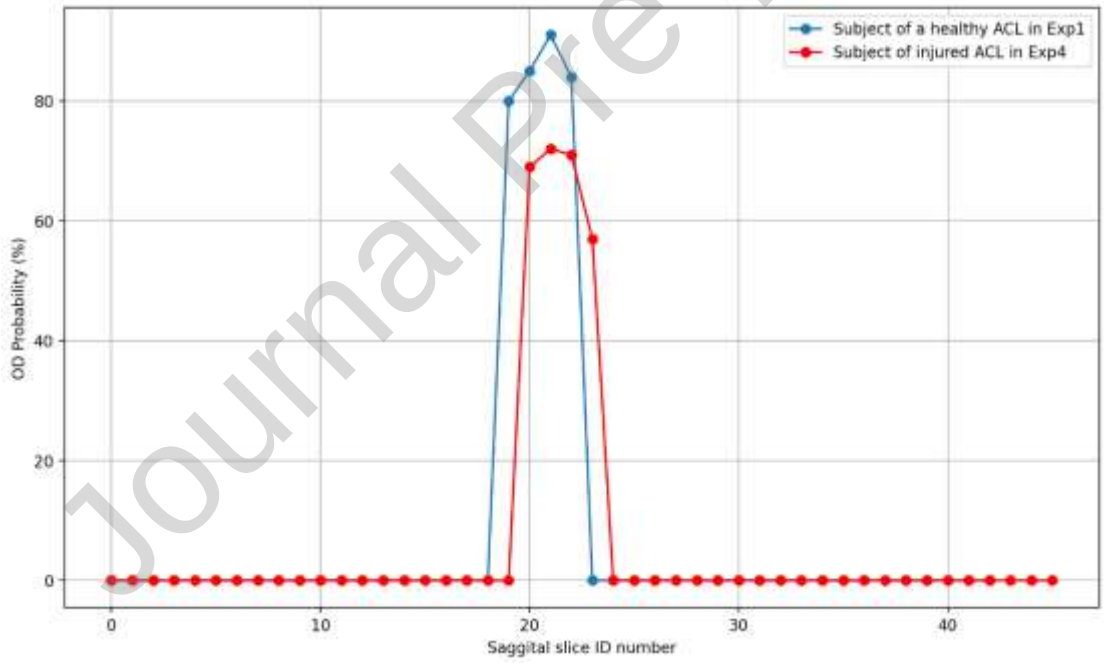


Figure 4: Indicative probability vectors obtained by the OD component of our framework in Exp1 and Exp4, when the whole sagittal plane image sequence is applied to it.

3.2. Resource efficiency evaluation

Table 4 lists the generated CO_2eq emissions, the power spent, and the equivalent distance a car could cover during the final training of 150 epochs for the proposed and SOTA methods. Also given are the training time and the average inference time. The proposed framework was found to be at least 2.1 times more eco-friendly and consume at least 2.6 times less fuel in comparison to the SOTA methods. Moreover, in terms of computational time, 1.5 times shorter training phase was achieved, while inference speed was at least 1.4 times higher.

Table 4: The resource efficiency comparison of the proposed method with the SOTA approaches. Evaluation was based on carbon emissions, energy usage, and equivalent car travel distance during the final training. Additionally, total training and average inference times were recorded. All experiments were carried out utilizing an NVIDIA RTX A6000 (48GB) GPU workstation. Boldface indicates best performance.

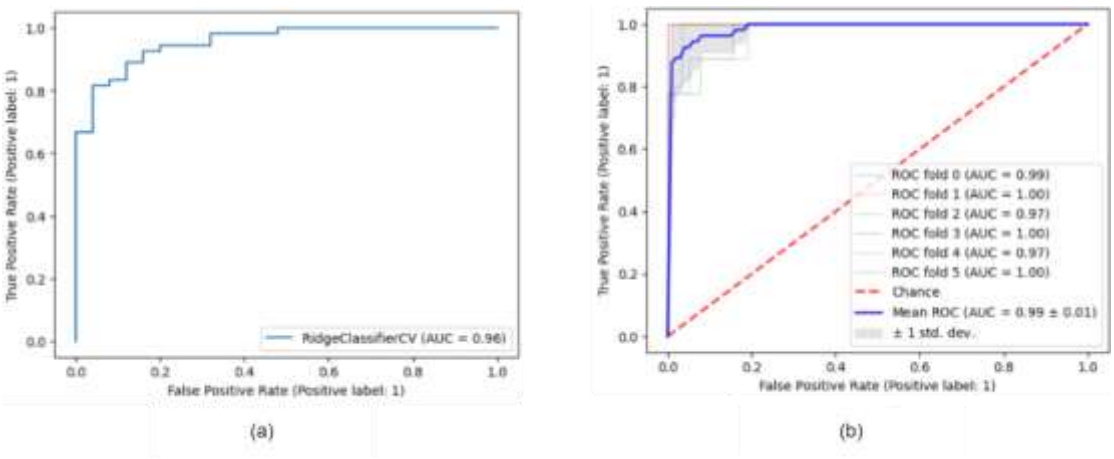
Model	CO_2eq (g)	Energy (kWh)	Equivalent distance traveled by car (km)	Training time (h:min:s)	Average inference time (ms)
Jeon et al. (2021) [18]	51.921046	0.43277567	0.431238	0:55:00	2
Dunnhofer et al. (2022) [21]	347.505372	0.908115	2.886257	1:35:37	6
Proposed	**20.264802**	**0.208915**	**0.168312**	**0:36:06**	**1.4**

3.3. Generalization capabilities evaluation

Table 5 outlines the results of the proposed framework in terms of ACL tear classification accuracy in cross-database settings (Exp2, Exp3, Exp5, Exp6). Figure 5 shows the ROC curves and the corresponding AUC scores of our classification models for cross-database experiments with testing databases that involved more than one class (Exp2, Exp5). Figure 6 illustrates the bounding boxes and the respective class scores outputted by the OD component of our framework in Exp2, Exp3, Exp5, and Exp6. It can be seen that the proposed framework exhibited excellent generalization capabilities, holding performance across different databases, especially when it was trained on the KneeMRI dataset.

Table 5: ACL tear classification evaluation of the proposed framework in cross-database experiments. *AUC, Sensitivity and Specificity are absent in Exp3 and Exp6 because UGHL involves only samples from one class (ACL ruptured knees).*

Experiment	Training Set	Testing Set	Accuracy	AUC	Sensitivity	Specificity
Exp2	MRNet	KneeMRI	88.99	97.00	98.10	88.12
Exp3	MRNet	UGHL	83.67	-	-	-
Exp5	KneeMRI	MRNet	91.14	96.00	96.30	80.00
Exp6	KneeMRI	UGHL	94.00	-	-	-



(a) (b)

Figure 5: ROC curves and corresponding AUC scores illustrating the classification performance of the proposed models in Exp2 (a) and Exp5 (b).

Figure 6: A visual presentation of results in cross-database experiments, showcasing sagittal planes, boundary boxes, and class scores of both healthy and injured knees. (a) Exp2, (b) Exp5, (c) Exp3 and (d) Exp6. Note: The absence of a boundary box in a graphic indicates a healthy knee class score less than 0.4, signaling potential injury. There were no healthy knees in the testing set of Exp3 and Exp6.

4. Discussion

This research successfully implemented an innovative hybrid DL framework that revised the sagittal MRI-based ACL tear classification task as a novelty detection problem to address the class imbalance challenge presented in public datasets. To obtain a fit-for-purpose novelty score, we capitalized on the aleatoric semantic uncertainty as this is modelled in the class scores outputted by the YOLOv5-nano OD model. Taking into consideration that the investigated tissue typically spans across multiple MRI slices, we proposed the utilization of

the global scores (probability vector) obtained when the OD model is applied to the whole sagittal plane image sequence. The second part of the proposed hybrid framework comprised the MINIROCKET timeseries classification model to determine whether a knee has an ACL tear or not. To investigate our approach's ability to generalize, we performed both single- and cross-database testing involving two public databases (KneeMRI and MRNet) and a validation-only database from UGHL in Greece.

Our method was invariably superior to the SOTA approaches [18,21] on the KneeMRI dataset and achieved better accuracy and sensitivity on the MRNet dataset. The cross-database experiments we carried out verified the robustness and excellent generalization capabilities of our method, especially when the model had been trained on KneeMRI. In addition, the presented pipeline generated at least 2.1 times less carbon emissions and consumed at least 2.6 times less energy, when compared with the SOTA approaches. It is worth noting here that the numbers reported in Table 4 pertain to the final training only. However, a typical DL development process requires hundreds of training runs due to experimenting with multiple models and hyperparameter tuning. Therefore, this aspect is crucial and the concerns over the environmental footprint of DL research and its impact on accelerating global climate change have been growing [22]. Our results go along with efforts to propose "greener" DL approaches and democratize DL research [16].

In this paper, we innovatively framed the MRI-based ACL tear diagnosis as a novelty detection problem by leveraging the global aleatoric semantic uncertainty obtained by the YOLOv5-nano OD model. However other approaches do exist in the novelty detection field. For instance, a recent study [29] explored novelty detection by generating synthetic near-distribution anomalous data to bridge the gap in detecting subtle differences between normal and anomalous samples, an approach that demonstrates the flexibility of novelty detection techniques in handling near-distribution challenges. Another research [30] employed

diffusion models for novelty detection, focusing on mitigating background bias in out-of-distribution samples through a method named Projection Regret (PR), which illustrates the potential of generative models in enhancing detection sensitivity. The varied applications of novelty detection across different tasks highlight the breadth of the evolving landscape of this field and the adaptability of these techniques to specific challenges, such as class imbalance and subtle anomaly distinction.

The integration of aleatoric semantic uncertainty-based scores into a novelty detection framework is a breakthrough contribution of this study. The above attribute, when combined with the use of lightweight OD and timeseries classification models, have the potential to revolutionize injury detection from MRI data by setting a new precedent in diagnostic precision, speed and environmental sustainability.

5. Limitations

One notable limitation of our study pertains to the fact that some MRI images within the datasets, especially those taken before 2005, lack the quality synonymous with modern magnetic resonance imagers. This discrepancy in image quality could impact the performance of the algorithm. In fact, this is possibly the reason why the model that had been trained on KneeMRI dataset generalized better (than the model that had been trained on MRNet dataset) on the UGHL dataset. As such, a worthwhile avenue for future exploration would be to conduct studies using contemporary MRI scanners, which offer superior image quality. Additionally, a challenge we encountered was the presence of MRI scans with a limited number of sagittal slices. This limitation is attributed to the large intervals at which the imager captures the data (or else, the low out-of-plane spatial resolution), necessitating the use of padding in our approach. In subsequent studies, this factor should be considered, ensuring that data is collected at optimal intervals to maximize the number of image slices and reduce the need for such adjustments.

6. Conclusions

In conclusion, this study introduced a hybrid novelty detection DL pipeline for ACL tear detection from MRI sagittal plane slices. The proposed framework achieved superior accuracy, consumed notably less fuel, and generated strikingly less carbon emissions, when compared against two SOTA approaches. The generalization capabilities of our pipeline were verified by running cross-database experiments. The integration of aleatoric uncertainty-based scores that are rapidly acquired into novelty detection DL frameworks holds promise as it not only sets a new standard in the MRI-based diagnostic speed and accuracy, but it also aligns with the increasing need for environmentally sustainable and democratic computational practices. The presented resource-efficient pipeline offers potential for widespread clinical application. The research paves the way for future advancements in the DL-powered MRI-based injury detection, emphasizing environmental sustainability alongside technological innovation.

Declaration of Interest Statement

The authors have disclosed that there are no financial conflicts of interest or personal connections that might seem to affect the research presented in this document.

Acknowledgments

This work has been Co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH—CREATE—INNOVATE (project code:T1EDK-04234). Tuan Aqeel Bohoran is funded by the European Union's Horizon 2020 reasearch and innovation programme under the Marie Sklodowska-Curie grant agreement No 801604.

References

[1] T. L. Sanders *et al.*, "Incidence of anterior cruciate ligament tears and reconstruction: A 21-year population-based study," *Am. J. Sports Med.*, vol. 44, no. 6, pp. 1502–1507, 2016, doi: 10.1177/0363546516629944.

[2] R. C. Mather *et al.*, "Societal and economic impact of anterior cruciate ligament tears," *J. Bone Jt. Surg.*, vol. 95, no. 19, pp. 1751–1759, 2013, doi: 10.2106/JBJS.L.01705.

[3] D. D. Anderson *et al.*, "Post-traumatic osteoarthritis: Improved understanding and opportunities for early intervention," *J. Orthop. Res.*, vol. 29, no. 6, pp. 802–809, 2011, doi: 10.1002/jor.21359.

[4] I. Papathanasiou *et al.*, "Molecular changes indicative of cartilage degeneration and osteoarthritis development in patients with anterior cruciate ligament injury Pathophysiology of musculoskeletal disorders," *BMC Musculoskelet. Disord.*, vol. 17, no. 1, pp. 1–10, 2016, doi: 10.1186/s12891-016-0871-8.

[5] L. S. Lohmander, P. M. Englund, L. L. Dahl, and E. M. Roos, "The long-term consequence of anterior cruciate ligament and meniscus injuries: Osteoarthritis," *Am. J. Sports Med.*, vol. 35, no. 10, pp. 1756–1769, 2007, doi: 10.1177/0363546507307396.

[6] N. Phelan, P. Rowland, R. Galvin, and J. M. O'Byrne, "A systematic review and meta-analysis of the diagnostic accuracy of MRI for suspected ACL and meniscal tears of the knee," *Knee Surgery, Sport. Traumatol. Arthrosc.*, vol. 24, no. 5, pp. 1525–1539, 2016, doi: 10.1007/s00167-015-3861-8.

[7] R. Crawford, G. Walley, S. Bridgman, and N. Maffulli, "Magnetic resonance imaging versus arthroscopy in the diagnosis of knee pathology, concentrating on meniscal lesions and ACL tears: A systematic review," *Br. Med. Bull.*, vol. 84, no. 1, pp. 5–23, 2007, doi: 10.1093/bmb/ldm022.

[8] I. Hetsroni, S. Lyman, H. Do, G. Mann, and R. G. Marx, "Symptomatic pulmonary embolism after outpatient arthroscopic procedures of the knee: The incidence and risk factors in 418 323 arthroscopies," *J. Bone Jt. Surg. - Ser. B*, vol. 93 B, no. 1, pp. 47–51, 2011, doi: 10.1302/0301-620X.93B1.25498.

[9] H. O. Alanazi, A. H. Abdullah, and K. N. Qureshi, "A Critical Review for Developing Accurate and Dynamic Predictive Models Using Machine Learning Methods in Medicine and Health Care," *J. Med. Syst.*, vol. 41, no. 4, 2017, doi: 10.1007/s10916-017-0715-6.

[10] W. Krampla, M. Roesel, K. Svoboda, A. Nachbagauer, M. Gschwantler, and W. Hruby, "MRI of the knee: How do field strength and radiologist's experience influence diagnostic accuracy and interobserver correlation in assessing chondral and meniscal lesions and the integrity of the anterior cruciate ligament?," *Eur. Radiol.*, vol. 19, no. 6, pp. 1519–1528, 2009, doi: 10.1007/s00330-009-1298-5.

[11] R. Mohankumar, L. M. White, and A. Naraghi, "Pitfalls and pearls in MRI of the knee," *Am. J. Roentgenol.*, vol. 203, no. 3, pp. 516–530, 2014, doi: 10.2214/AJR.14.12969.

[12] C. Germann *et al.*, "Deep Convolutional Neural Network-Based Diagnosis of Anterior Cruciate Ligament Tears: Performance Comparison of Homogenous Versus

Heterogeneous Knee MRI Cohorts with Different Pulse Sequence Protocols and 1.5-T and 3-T Magnetic Field Strengths," *Invest. Radiol.*, vol. 55, no. 8, pp. 499–506, 2020, doi: 10.1097/RLI.0000000000000664.

[13] C. Kokkotis *et al.*, "Identifying Gait-Related Functional Outcomes in Post-Knee Surgery Patients Using Machine Learning: A Systematic Review," *Int. J. Environ. Res. Public Health*, vol. 20, no. 1, 2023, doi: 10.3390/ijerph20010448.

[14] A. Siouras *et al.*, "Knee Injury Detection Using Deep Learning on MRI Studies: A Systematic Review," *Diagnostics*, vol. 12, no. 2, pp. 1–21, 2022, doi: 10.3390/diagnostics12020537.

[15] A. Giannakidis *et al.*, "Fast Fully Automatic Segmentation of the Severely Abnormal Human Right Ventricle from Cardiovascular Magnetic Resonance Images Using a Multi-Scale 3D Convolutional Neural Network," *Proc. - 12th Int. Conf. Signal Image Technol. Internet-Based Syst. SITIS 2016*, pp. 42–46, 2017, doi: 10.1109/SITIS.2016.16.

[16] T. A. Bohoran *et al.*, "Resource efficient aortic distensibility calculation by end to end spatiotemporal learning of aortic lumen from multicentre multivendor multidisease CMR images," *Sci. Rep.*, vol. 13, no. 1, pp. 1–15, 2023, doi: 10.1038/s41598-023-48986-6.

[17] M. J. Awan, M. S. M. Rahim, N. Salim, M. A. Mohammed, B. Garcia-Zapirain, and K. H. Abdulkareem, "Efficient Detection of Knee Anterior Cruciate Ligament from Magnetic Resonance Imaging Using Deep Learning Approach," *Diagnostics*, vol. 11, no. 1, 2021, doi: 10.3390/diagnostics11010105.

[18] Y. S. Jeon *et al.*, "Interpretable and Lightweight 3-D Deep Learning Model for Automated ACL Diagnosis," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 7, pp. 2388–2397, 2021, doi: 10.1109/JBHI.2021.3081355.

[19] N. Bien *et al.*, "Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet," *PLoS Med.*, vol. 15, no. 11, pp. 1–19, 2018, doi: 10.1371/journal.pmed.1002699.

[20] B. Astuto *et al.*, "Automatic deep learning–assisted detection and grading of abnormalities in knee MRI studies," *Radiol. Artif. Intell.*, vol. 3, no. 3, 2021, doi: 10.1148/ryai.2021200165.

[21] M. Dunnhofer, N. Martinel, and C. Micheloni, "Deep convolutional feature details for better knee disorder diagnoses in magnetic resonance images," *Comput. Med. Imaging Graph.*, vol. 102, no. November, p. 102142, 2022, doi: 10.1016/j.compmedimag.2022.102142.

[22] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," *Commun. ACM*, vol. 63, no. 12, pp. 54–63, 2020, doi: 10.1145/3381831.

[23] M. A. F. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal Process.*, vol. 99, pp. 215–249, 2014, doi: 10.1016/j.sigpro.2013.12.026.

[24] G. Jocher, "Yolov5." https://github.com/ultralytics/yolov5.

[25] A. Dempster, D. F. Schmidt, and G. I. Webb, "MiniRocket: A Very Fast (Almost) Deterministic Transform for Time Series Classification," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 248–257, 2021, doi: 10.1145/3447548.3467231.

[26] J. K. Lee, L. Yao, C. T. Phelps, C. R. Wirth, J. Czajka, and J. Lozman, “Anterior cruciate ligament tears: MR imaging compared with arthroscopy and clinical tests.,” *Radiol.*, vol. 166, no. 3, pp. 861–4, Mar. 1988, doi: 10.1148/radiology.166.3.3340785.

[27] M. Leppänen *et al.*, “Sagittal Plane Hip, Knee, and Ankle Biomechanics and the Risk of Anterior Cruciate Ligament Injury: A Prospective Study,” *Orthop. J. Sport. Med.*, vol. 5, no. 12, p. 232596711774548, Dec. 2017, doi: 10.1177/2325967117745487.

[28] I. Štajduhar, M. Mamula, D. Miletić, and G. Ünal, “Semi-automated detection of anterior cruciate ligament injury from MRI,” *Comput. Methods Programs Biomed.*, vol. 140, pp. 151–164, 2017, doi: 10.1016/j.cmpb.2016.12.006.

[29] H. Mirzaei *et al.*, “Fake It Till You Make It: Towards Accurate Near-Distribution Novelty Detection,” 2022, [Online]. Available: http://arxiv.org/abs/2205.14297.

[30] S. Choi, H. Lee, H. Lee, and M. Lee, “Projection Regret: Reducing Background Bias for Novelty Detection via Diffusion Models,” no. NeurIPS, pp. 1–16, 2023, [Online]. Available: http://arxiv.org/abs/2312.02615.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Highlights

- The formulation of the ACL tear diagnosis task as a novelty detection problem for the first time in order to address class imbalance.

- The integration of rapidly-obtained aleatoric semantic uncertainty-based scores into the novelty detection framework.
- The inclusion of global tissue information into the decision-making process.
- The implementation of the proposed framework in single-database experiments achieved superior performance in detecting ACL tears, while at the same time consumed notably less fuel and generated strikingly less carbon emissions during training and inference, when compared with two state-of-the-art methods.
- Cross-database experiments verified the robustness and excellent generalization capabilities of the proposed framework.