diagnostics

MDPI

Review

Knee Injury Detection Using Deep Learning on MRI Studies: A Systematic Review

Athanasios Siouras [1,2,*], Serafeim Moustakidis [3], Archontis Giannakidis [4], Georgios Chalatsis [5], Ioannis Liampas [6], Marianna Vlychou [7], Michael Hantes [5], Sotiris Tasoulis [1] and Dimitrios Tsaopoulos [2]

[1] Department of Computer Science and Biomedical Informatics, School of Science, University of Thessaly, 35131 Lamia, Greece; stas789@gmail.com
[2] Centre for Research and Technology Hellas, 38333 Volos, Greece; d.tsaopoulos@certh.gr
[3] AIDEAS OÜ, 10117 Tallinn, Estonia; s.moustakidis@aideas.eu
[4] School of Science and Technology, Nottingham Trent University, Nottingham NG11 8NS, UK; archontis.giannakidis@ntu.ac.uk
[5] Department of Orthopedic Surgery, Faculty of Medicine, University of Thessaly, 41500 Larissa, Greece; ghalatsis@hotmail.com (G.C.); hantesmi@otenet.gr (M.H.)
[6] Department of Neurology, School of Medicine, University Hospital of Larissa, University of Thessaly, Mezourlo Hill, 41500 Larissa, Greece; liampasioannes@gmail.com
[7] Department of Radiology, School of Health Sciences, University Hospital of Larissa, University of Thessaly, Mezourlo, 41500 Larissa, Greece; mvlychou@uth.gr
* Correspondence: asiouras@uth.gr

Abstract: The improved treatment of knee injuries critically relies on having an accurate and cost-effective detection. In recent years, deep-learning-based approaches have monopolized knee injury detection in MRI studies. The aim of this paper is to present the findings of a systematic literature review of knee (anterior cruciate ligament, meniscus, and cartilage) injury detection papers using deep learning. The systematic review was carried out following the PRISMA guidelines on several databases, including PubMed, Cochrane Library, EMBASE, and Google Scholar. Appropriate metrics were chosen to interpret the results. The prediction accuracy of the deep-learning models for the identification of knee injuries ranged from 72.5–100%. Deep learning has the potential to act at par with human-level performance in decision-making tasks related to the MRI-based diagnosis of knee injuries. The limitations of the present deep-learning approaches include data imbalance, model generalizability across different centers, verification bias, lack of related classification studies with more than two classes, and ground-truth subjectivity. There are several possible avenues of further exploration of deep learning for improving MRI-based knee injury diagnosis. Explainability and lightweightness of the deployed deep-learning systems are expected to become crucial enablers for their widespread use in clinical practice.

Keywords: ACL; deep learning; knee injury; machine learning; meniscus

check for updates

Citation: Siouras, A.; Moustakidis, S.; Giannakidis, A.; Chalatsis, G.; Liampas, I.; Vlychou, M.; Hantes, M.; Tasoulis, S.; Tsaopoulos, D. Knee Injury Detection Using Deep Learning on MRI Studies: A Systematic Review. *Diagnostics* **2022**, *12*, 537. https://doi.org/10.3390/diagnostics12020537

Academic Editor: Sameer Antani

Received: 23 December 2021
Accepted: 17 February 2022
Published: 19 February 2022

1. Introduction

1.1. Backdrop

Knee injuries account for the largest percentage of sport-related, severe injuries (i.e., injuries that cause more than 21 days of missed sport participation) [1–4]. Anterior cruciate ligament (ACL) ruptures represent more than 50% of the cases, affecting 200,000 individuals in the United States each year [1,5–7]. Knee cartilage lesions affect around 900,000 individuals in the United States every year, resulting in over 200,000 surgical procedures [5–8]. Menisci injuries are the second most common knee impairment, with an incidence of 12–14% [9] and a prevalence of 60–70 cases per 100,000 in the United Kingdom [2]. ACL injuries alone account for an expenditure of more than $7 billion in the United States [10]. Both short- and long-term pain, disability, and negatively affected, health-related quality of life have all been strongly associated with knee injuries [11–13]. In regard to young and athletic individuals, the more time they spend engaging in occupational and/or recreational activities, the higher predisposition to knee injuries they have, which, in turn, contributes to a higher likelihood of developing osteoarthritis (OA) [14]. On average, half of the individuals, that have an injury that involved ACL and/or meniscal tear develop radiographically confirmed knee OA ten to 20 years post-injury [15,16]. Another two possible consequences of knee injuries are: (i) structural muscle injuries of the lower limb [17]; and (ii) tendinopathies [18]. All the above reflect the direct and indirect (lost wages, productivity, and disability) socio-economic burden conferred on the society by knee injuries. The high prevalence of knee injuries in the general population, and the resulting socio-economic impact, have created a necessity for developing accurate and cost-effective procedures that can detect and quantify the severity of knee injuries. Early diagnosis and, consequently, treatment of ligament rupture, menisci tear, and/or cartilage lesion can prevent early onset of knee OA [1].

Arthroscopy is considered the “gold-standard” for the diagnosis of intra-articular knee pathologies, but is limited by potential complications and its invasive nature [19]. Therefore, magnetic resonance imaging (MRI) is the most widely used, non-invasive imaging technique for diagnosing knee injuries [20,21]. However, the MRI-based diagnosis of knee injuries can be a very challenging procedure, with the experience of clinicians playing a critical role in image interpretation. Human-based image interpretation pitfalls, such as subjectivity, distraction, and fatigue, as well as diagnostic uncertainties, often lead to erratic diagnoses, hindering the optimal management of knee injuries [22,23]. Moreover, clinical-diagnostic discrepancies among non-musculoskeletal radiologists and orthopedic surgeons are commonly encountered in everyday clinical practice [11].

Due to the above-listed factors, as well as the exponentially increasing number of clinical examinations, the idea of using computers for improving the challenging task of image interpretation of medical examinations has been recently adopted by the scientific community [24]. Imaging data proliferation, algorithmic advances, and recent technological advances in fast computing have already resulted in a strong push towards the utilization of artificial intelligence (AI) algorithms in medical image analysis. The term AI broadly refers to any method that enables computers to mimic human intelligence [25]. Deep learning (DL) in particular is a class of machine-learning (ML) algorithms that is currently driving the AI boom [26]. Numerous applications of DL in medical image analysis have been reported, including skin cancer classification, diabetic retinopathy detection, lung nodule detection, and mammography cancer detection, among others [27]. The aforementioned AI-empowered solutions are expected to revolutionize medical sectors by improving the accuracy and productivity of different diagnostic and therapeutic measures in clinical practice [20].

Drawing attention to the diagnosis of knee injuries, several early DL studies have exhibited better performance than traditional ML techniques, while in some cases they have proved to be even superior to radiologists [26]. However, the previously published review studies in the MRI field were either focused on other application domains (e.g., fracture detection [28]) or limited to the performance of the proposed networks without paying attention to their specifics (learning methodology, processing stages, technical limitations etc.) [29]. In the light of the advancements of AI technology and the increasing number of studies in this field, in this paper, we conducted a systematic literature review, covering all DL-oriented techniques that have been employed in the diagnostic process of knee injuries. The aim of this systematic review is to identify all recent studies that investigate the use of DL technology in the MRI-based diagnosis of the injured knee. It was decided that the primary focus would be on studies that examine at least one of the following pathologies: (i) injuries of knee ligaments; (ii) meniscus tears; and (iii) cartilage lesion.

1.2. Machine Learning in a Nutshell: Definitions and Terminology

To enhance the understanding of the readers and for the sake of completeness, this section quickly presents the relevant terminology and definitions with respect to ML and DL algorithms used in the studies involved in the present review. ML is a branch of AI that focuses on the development of algorithms that automatically learn to make accurate predictions by relying on experience (data) rather than on hard-coded instructions.

Supervised ML systems (Figure 1) operate in two phases: the learning phase (training) and the testing one. In a traditional ML pipeline, a feature extraction/selection stage (also referred to as feature engineering) is first implemented to extract or identify the most informative features [16]. These features can be extracted from the input images, employing various algorithms including grey-level co-occurrence matrix (GLCM), first- and second-order statistics, and shape/edge features, among others [30]. Next, a ML model is fit to the extracted features and the optimal model parameters are obtained. During the testing phase, the trained model is shown previously unseen samples (represented as images or features extracted from images), which are then classified. As opposed to traditional programming, where the rules are manually crafted by a programmer, a supervised ML algorithm automatically formulates rules from the data.

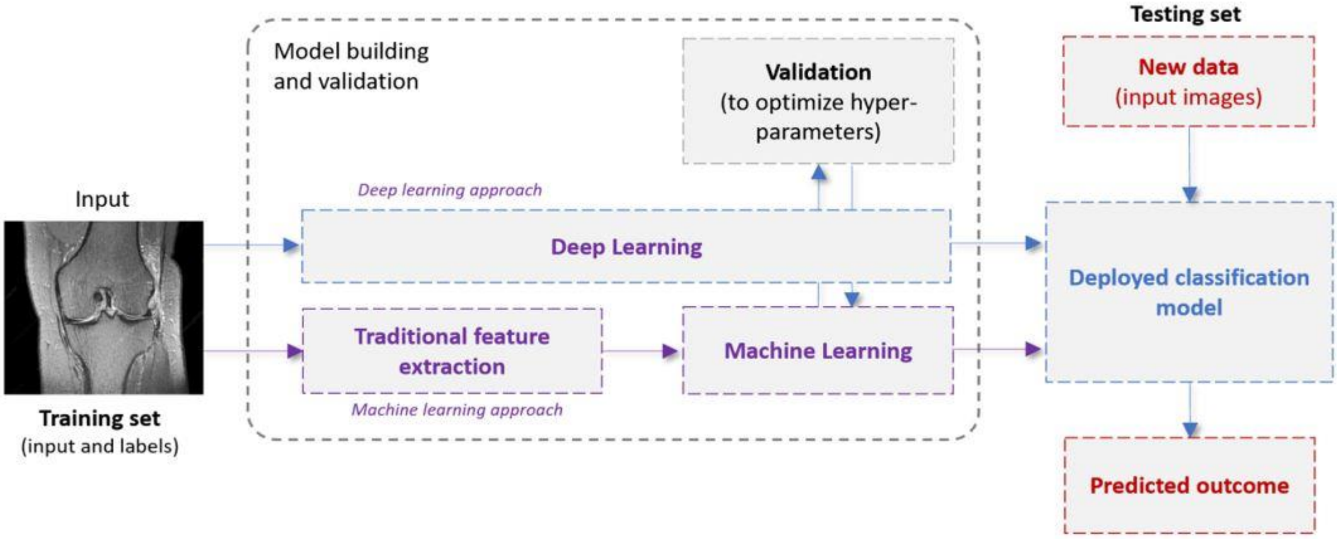

Figure 1. Examples of typical machine-learning and deep-learning pipelines.

DL [31] is a subfield of ML that sets an alternative architectural paradigm by shifting the process of extracting features from images to the underlying learning mechanism. The most informative features for the task at hand are extracted by the algorithm itself. The mainstream DL architecture for computer vision applications is the convolutional neural network (CNN). A CNN typically consists of multiple building blocks (layers such as convolutional, pooling, and fully connected) that automatically extract increasingly abstract spatial hierarchies of features. The CNN training is carried out via a backpropagation algorithm. The huge popularity of CNNs is attributed to certain characteristics they possess, such as weight sharing and spatial invariance.

Transfer learning is a common strategy where a network, that was pre-trained on a big dataset, is partly re-used to provide decisions on a problem with a different dataset. The main idea behind transfer learning is that generic features learned on a large dataset could be useful and applicable to other domain tasks with a potentially limited amount of accessible data. Numerous pre-trained networks are currently available, such as DenseNet [32], AlexNet [33], and VGG [34]. When employing DL with transfer learning for feature extraction, the pre-trained network is treated as an arbitrary feature extractor: the input image propagates through multiple layers until it reaches a pre-specified layer, the outputs of which are considered as the finally extracted features. Table 1 provides a brief presentation of the main ML and DL algorithms that were reported in the papers of this review.

Table 1. Brief presentation of the feature extraction techniques, as well as the ML and DL models, and the main procedures that were reported in the papers of our survey.

Category	Models	Description
Feature extraction	Histogram of oriented gradient (HOG) [35]	This is a feature descriptor used in computer vision and image processing for the purpose of object detection. The technique counts occurrences of gradient orientation in localized portions of an image.
	Generalized search tree (GIST) [30]	GIST descriptor represents holistic spatial scene properties (spatial envelope) of an image. It summarizes gradient information on different spatial scales and orientations by splitting the image into a grid of cells on several scales and convolving each cell using a Gabor filter bank from different perspectives.
	Gray-level co-occurrence matrix (GLCM) [36]	GLCM is a way of extracting second-order statistical texture features. In particular, the texture of an image is estimated by calculating how often pairs of pixels with specific values and a certain spatial relationship occur.
Traditional Machine Learning	k-nearest neighbor (K-NN) [37]	KNN algorithm is a simple, easy-to-implement supervised ML algorithm that can be used to solve both classification and regression problems. It works by (i) finding the distances between a query and all the examples in the data, (ii) selecting the K nearest neighbors of the query, and (iii) voting for the most frequent label (in the case of classification) or averaging the labels (in the case of regression).
	Support vector machines (SVMs) [38]	SVMs is a supervised method that identifies a hyperplane that best divides the data into two classes. To separate the two clouds of data points, there are many possible hyperplanes that could be chosen. The objective of the SVM algorithm is to find a slab that has the maximum thickness, i.e., the maximum distance between data points of the different classes.
	Shallow artificial neural networks (ANNs) [39]	The ANN vaguely simulates the way the human brain analyzes and processes information. They consist of sequential layers: input, hidden and output layers. The hidden layer processes and transmits the input information to the output layer.
Deep Learning	Convolutional neural networks (CNNs) [40]	This is a class of DL algorithms commonly used in computer vision and pattern recognition. CNNs are a specific type of neural networks that are generally composed of the following layers: (i) input layer, (ii) convolution layers, (iii) pooling layers and (iv) fully connected layers. The convolution layers use filters that perform convolution operations as they are scanning the input with respect to its dimensions. Pooling is a down-sampling operation, which is typically applied after a convolution layer. The fully connected layers operate on a flattened input where each input is connected to all neurons in the next layer and are usually found towards the end of CNN architectures to optimize objectives such as class scores.
	Region based convolutional neural networks (R-CNNs) [41]	The method of detecting and classifying objects in an image is known as object detection. R-CNN (regions with convolutional neural networks) is a deep learning technique that blends rectangular area proposals with convolutional neural network functionality. The R-CNN algorithm is a two-stage detection method.
	Deep residual networks [42]	A residual neural network (ResNet) is an ANN variant that uses residual mapping and shortcut connections to tackle the problem of vanishing and exploding gradients that is characteristic of deep CNNs. As a consequence of this, deep residual networks achieve better performance when compared to plain very deep networks, whereas their training is easier as well. Typical ResNet models are implemented with double- or triple-layer skips that contain nonlinearities such as rectified linear unit (ReLUs) and batch normalization in between.

Table 1. *Cont.*

Category	Models	Description
	3D-CNNs [43]	A 3D CNN is simply the 3D generalization of 2D CNNs. It takes as input a 3D volume or a sequence of 2D frames (e.g., slices in an MRI scan). Then kernels move through 3 dimensions of data producing 3D activation maps. Overall, they learn powerful representations of volumetric data.
	Computer Vision Transformers [44]	When data is modelized as a sequence of embeddings, the Transformer model is a basic yet scalable technique that can be used for any type of data. Even without typical convolutional pipelines, transformers can be utilized to provide SOTA results in Computer Vision. It is a DL network that extracts inherent properties of the interest domain via the self-attention technique.
Procedure	Training	The standard procedure involves a dataset of paired images and labels (x, y) for training and testing, an optimizer (e.g., stochastic gradient descent, Adam [45]), and a loss function to update the model parameters. The aim of the training is to find the optimal values for the network parameters so that the loss function is minimized.
	Data augmentation	Data augmentation is a strategy that artificially generates more training samples to increase the diversity of the training data. This can be done via applying affine transformations (e.g., rotation, scaling), flipping or cropping to original labeled samples.
	Dropout	Dropout is a regularization method that randomly drops some units from the neural network during training, encouraging the network to learn a sparse representation. It is used to reduce overfitting.
	Loss function	The metric to assess the discrepancy between model predictions and labels is called loss function. The gradients of the loss function are used to update the weights of the neural networks.
	Transfer learning	This aims to transfer knowledge from one task to another different but related target task. This is often achieved by reusing the weights of a pre-trained model, to initialize the weights in a new model for the target task. Transfer learning can help to decrease the training time and achieve lower generalization error.

2. Materials and Methods

2.1. Reporting

The present (systematic) review was performed in accordance with the preferred reporting items for systematic reviews and meta-analyses (PRISMA) guidelines [46]. Each step of the review process (literature search, study selection, and data extraction) was independently performed by 2 authors (A.S., S.M.). Discrepancies were resolved by a 3rd author (D.T.). The present study was not registered in a database prior to its conduction.

2.2. Literature Search

A structured literature search was conducted in the following databases: (a) MEDLINE (through PubMed), (b) CENTRAL (through Cochrane Library), and (c) EMBASE (through Elsevier). Articles cited by the retrieved papers, as well as articles citing the retrieved papers (using Google Scholar), were also identified through a supplementary manual search. Grey literature was examined based on conference abstracts, English abstracts (from articles not published in English), and the OpenGrey database. The potential eligibility of the articles was initially decided based on their title and abstract. Full texts were investigated to verify whether the initial qualifiers fulfill the inclusion criteria. The structured search strategy per database is quoted in Table S1.

2.3. Eligibility Criteria

The inclusion criteria were as follows: (i) papers were published between the 1st of January 2013 (the dawn of the DL era) and the 15th of November 2021 (date of final literature search); (ii) MRI images were used for the evaluation of knee injuries; (iii) knee-injury detection was conducted via AI-based algorithms, including both traditional ML and DL techniques; and (iv) the performance of the AI-based algorithms was compared to clinical, human-based evaluations.

Papers were excluded according to the following criteria: (i) articles published before 2013; (ii) papers investigating OA or other bone pathology not directly linked with knee injuries; (iii) studies performed in animals; (iv) non-original research articles, such as protocols, reviews, meta-analysis, etc.; (v) articles not written in English (however, English abstracts were assessed as part of the grey literature); and (vi) book chapters, editorials and commentaries.

2.4. Data Extraction

Extracted data were placed into a custom Microsoft Excel spreadsheet using a standardized table. The following information was included for each of the articles: first author, publication year, database, description of data and models, and learning algorithm, including pre-processing, size of training and test samples, validation method, and obtained results.

2.5. Statistical Analysis

Multiple evaluation criteria were employed to assess the predictive capacity of the proposed learning algorithms. The most common evaluation metric considered in this review study was accuracy, along with the receiver operator characteristic curves (ROCs) that visualize the performance of a classification model at various likelihood ratio thresholds. These curves plot two factors: true positive rate (sensitivity = TP/(TP + FN)) versus the false positive rate (specificity = FP/(FP + TN)), where TP, FN, FP, and TN denote true positives, false negatives, false positives, and true negatives, respectively. The quantitative output of this curve is the AUC, which could be interpreted as an aggregated measure of performance across all possible classification thresholds.

2.6. Quality Assessment

Quality assessment was performed using a modified methodologic index for non-randomized studies (MINORS) [47]. A seven-item checklist was considered, including information with respect to the following items: disclosure, study aim, input feature, ground truth label determination, dataset distribution, performance metric, and explanation of the applied AI models. Data were extracted and recorded using standardized forms (Microsoft Excel spreadsheet). To resolve conflicts over article selection, quality assessment, and data extraction, both observers (A.S., S.M.) convened a consensus meeting. The items were scored with 0 (not reported), 1 (reported but inadequate), or 2 (reported and adequate). The average modified MINORS score among all studies was 9.82 ± 1.99. It should be mentioned that the range of the score per item was between 0 and 44.

As shown in Figure 2, all the reported studies (22) clearly stated the study aim, input features, and the performance achieved using appropriate metrics. A clear distribution and description of the dataset were reported in twenty studies (90.09%). Fifteen studies (68.18%) clearly described how they established the ground truth (AI's reference standards), whereas the others were subjected to AI models that were inadequately trained. The most prevalent causes of quality point loss were failures to describe ground truth assignment. Last, but not least, more than half of the studies (54.54%) failed to disclose a conflict of interest declaration.

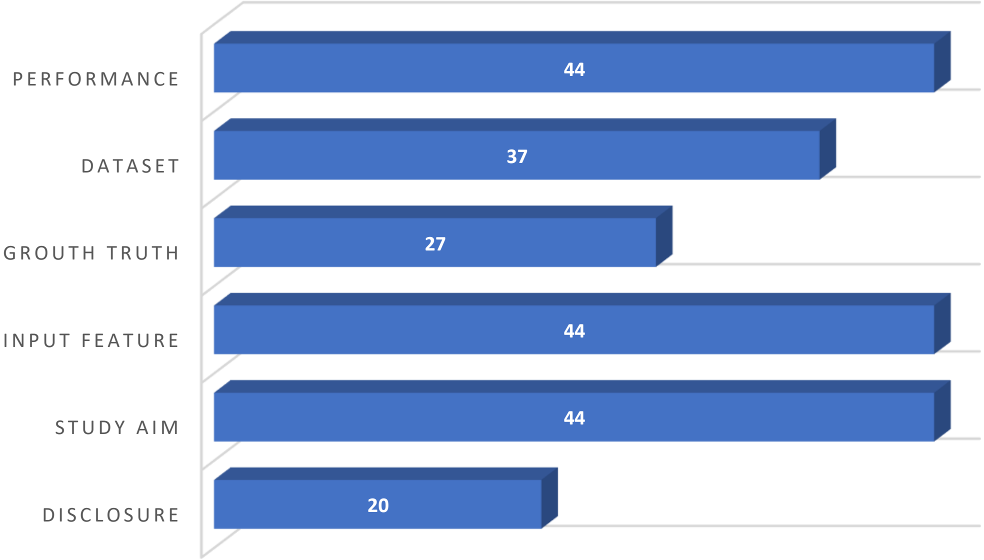

Figure 2. Quality assessment outcomes using the MINORS tool.

3. Results

In total, 407 studies were retrieved: 172 from MEDLINE (through PubMed), 170 from EMBASE (through ELSEVIER), 24 from CENTRAL, 40 from the structured search in Google Scholar, and 1 conference abstract (grey literature). Fifty-nine papers were selected after applying the proposed inclusion/exclusion criteria. Thirty-seven studies were further excluded due to irrelevant content (for example, those focusing only on segmentation or other scientific fields). Taking everything into consideration, 22 articles were finally included in the present systematic review. A flow chart of the literature search design is presented at Figure 3.

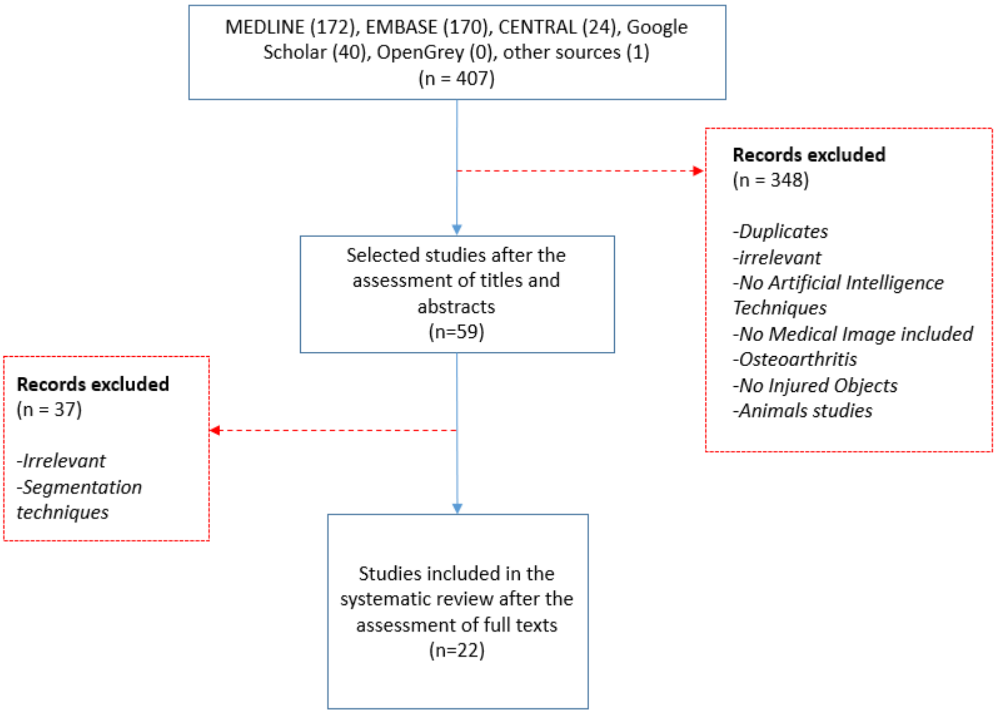
MEDLINE (172), EMBASE (170), CENTRAL (24), Google Scholar (40), OpenGrey (0), other sources (1)
(n = 407)
Records excluded
(n = 348)
-Duplicates
-irrelevant
-No Artificial Intelligence Techniques
-No Medical Image included
-Osteoarthritis
-No Injured Objects
-Animals studies
Selected studies after the assessment of titles and abstracts
(n=59)
Records excluded
(n = 37)
-Irrelevant
-Segmentation techniques
Studies included in the systematic review after the assessment of full texts
(n=22)

Figure 3. Flow chart presenting the design of the literature search.

The retrieved articles were categorized into the following application domains: (i) detection of ACL injuries alone (10 studies); (ii) detection of meniscus tears alone (7 studies); (iii) detection of cartilage lesions (1 study); and (iv) combined ACL and meniscus tears plus other knee injuries (4 studies). The main results of each study have been quoted, while the individual study validity has been determined based on its methodological strengths and weaknesses. Important methodological features of the retrieved articles have been commented.

The identified studies focusing on ACL and meniscus tears detection have been grouped into three categories: (i) those employing traditional ML pipelines; (ii) DL studies in which transfer learning is reported; and (iii) papers that propose the use of custom-made DL architectures.

3.1. ACL Injury Detection

3.1.1. Machine Learning

Štajduhar and colleagues [48] utilized two different feature extraction techniques: histogram oriented gradient (HOG) [35] and generalized search tree (GIST) [30]. These feature extraction techniques were subsequently paired with two commonly used ML models: support vector machines (SVMs) [38] and random forests [49]. They found that the best performing ML model was the one that combined HOG with linear-kernel SVM, producing an AUC of 0.89 for differentiating between ACL-injured and healthy subjects, and an area under curve (AUC) of 0.94 for detecting only completely ruptured ACL. Abdulah et al. [50] described a diagnostic system consisting of image pre-processing, feature extraction based on segment-derived spatial descriptors (perimeter, area, and shape), and, finally, classification. They compared k-nearest neighbor (K-NN) with back propagation artificial neural network (BP-ANN) for ACL tear classification. BP-ANN achieved a classification accuracy of 94.44% whereas K-NN reached an accuracy of 87.33%. Another study [51] tested an SVM algorithm on a dataset that was comprised of 100 non-injured ACLs, 100 partially-torn ACLs, and 100 completely-torn ACLs. All datasets underwent pre-processing. Features were extracted using shape descriptors, such as objects' contour circularity, aspect ratio, angle, and number of sides. It was reported that the SVM model had an accuracy of 100% for classifying ACL MRI samples as normal, partial-tear, or complete-tear. The authors also sought to compare the diagnostic capability of their AI model with that of two medical experts on a subset of 10 samples. No statistically significant differences between the AI model and the radiologists were found.

3.1.2. Deep Learning with Transfer Learning

Bien et al. [27] used transfer learning in order to train a fully automated CNN for classifying MRI series and they combined the predictions from 3 series per exam using logistic regression. The accuracy and the AUC of the model for detecting ACL tears were 86.7% and 0.965, respectively. These results were juxtaposed with the assessments by three musculoskeletal (MSK) radiologists on a testing set of 120 knee MR images. Radiologists achieved significantly higher sensitivities for tear diagnosis than the AI model (AUC: 0.91 vs. 0.76, p-value = 0.002). The accuracy achieved by the radiologists (92%) was higher than the one achieved by the AI model (86.7%). Azcona et al. [52] proposed and evaluated the performance of four architectures: (i) deep residual network with transfer learning; (ii) custom deep residual network using a fixed number of slices; (iii) multi-plane deep residual network; and (iv) multi-plane multi-objective deep residual network. They found that transfer learning combined with a carefully tuned data augmentation strategy were the crucial factors in achieving best performance. The authors modified the last layer to output a probability instead of a one-hot softmax vector for a number of classes and they also used transfer learning with pre-trained weights from ImageNet. By using the aforementioned DL architectures and data augmentation strategies for ACL detection, they achieved an AUC of 0.96 on the validation data.

3.1.3. Custom-Made Deep-Learning Networks

Another study [8] evaluated three customized CNN models with variations in the input fields of view (i.e., full slice, cropped slice, and dynamic patch-based sampling) as well as in dimensionality (single slice, three slices, or five slices) for the detection of complete ACL tears. The importance of limiting the input field-of-view to the intercondylar region for high algorithm performance was demonstrated. The incremental value of contextual information of adjacent image slices in improving network classification accuracy was also exhibited. The model that utilized dynamic sampling had an accuracy of 96.7% and an AUC of 0.97. Liu et al. [53] trained multiple CNNs and applied them to a test set comprised of 50 MR images of ACL tears with normal thickness and 50 MR images with intact ACLs. The best model they came up with for detecting the presence or absence of a full thickness ACL tear produced an AUC of 0.98 (95% CI: 0.93–1.00, p-value < 0.001). However, there was no statistically significant difference in diagnostic performance between the AI model (AUC: 0.98, 95% CI: 0.93- 1.00) and the clinical radiologist performance: Radiologist AUC: 0.98 (95% CI: 0.95–1.00); Fellow AUC: 0.98 (95% CI: 0.95- 1.00); Resident 1 AUC: 0.93 (95% CI: 0.88–0.98); Resident 2 0.97 (95% CI: 0.94–1.00); Resident 3 0.98 (95% CI: 0.95–1.00).

Namiri et al. [54] employed two CNN types for classification of ACL injuries: the first one involved three-dimensional (3D) kernels, whereas the second one made use of two-dimensional (2D) filters. The overall accuracies using the 3D CNN and the 2D CNN were 89% (225 of 254) and 92% (233 of 254), respectively (p-value= 0.27), whereas both CNNs had a weighted Cohen k of 0.83. The 2D CNN and 3D CNN performed similarly in classifying intact ACLs (2D CNN: sensitivity of 93% and specificity of 90%; 3D CNN: sensitivity of 89% and specificity of 88%). The classification of full tears by both networks was also comparable (2D CNN: sensitivity of 82% and specificity of 94%; 3D CNN: sensitivity of 76% and specificity of 100%). The 2D CNN classified all reconstructed ACLs correctly. A separate study [6] proposed to perform CNN-based classification by relying on the architecture of 3D DenseNet [32]. They compared this DL approach with two other variants, namely VGG16 [34] and ResNet [42]. The accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of the proposed customized architecture were calculated respectively. The average AUCs were 0.95, 0.86, and 0.96 for ResNet, VGG16, and their proposed network, respectively. The diagnostic accuracies achieved by the proposed model, the residents, and the senior radiologists were 95.7%, 81.4%, and 89.9%, respectively.

Germann et al. [24] trained a deep convolutional neural network (DCNN) on 512 MR images of ACL tears from different patients (ACL tears were present in 45.7% and absent in 54.3% of the subjects). The network had a pre-processing step that involved the selection, rescaling, and cropping of coronal and sagittal-fluid-sensitive views. Next, the coronal and sagittal MRI scans were processed independently in parallel and were then concatenated before being processed by one dense layer fat-suppressed MRI scan. Finally, a soft-max layer extracted the confidence level for the ACL tear. Three fellowship-trained full-time academic MSK radiologists independently evaluated the MRI examinations for full-thickness ACL tears. ACL tears were present in 45.7% and absent in 54.3% of the subjects. The DCNN had a sensitivity of 96.1%, which was not significantly different from that of the readers (97.5–97.9%; all p-values $\geq$ 0.118). However, the sensitivity of the DCNN (93.1%) was significantly lower than that of the readers (99.6–100%, all p-values < 0.001), and a similar trend was observed in the AUC values (DCCN: 0.94, readers: 0.99–0.99, all $p < 0.001$). Finally, a related study [55] used a customized 14-layer ResNet-based CNN with six different directions by using class balancing and data augmentation. The proposed ResNet-14 achieved AUC values of 0.98, 0.97, and 0.99 for detecting a healthy tear, partial tear, and fully ruptured tear, respectively. Jeon et al. [56] proposed a 3D deep-neural-network model for diagnosing ACL tears from a knee MRI test that is both interpretable and lightweight. They used squeeze modules and fewer convolutional filters to represent the homogeneity of the features, as well as attention modules and Gaussian positional encoding to strengthen the searching of local features. Their model outperformed the prior SOTA on the Chiba and Stanford knee datasets, achieving average ROC and AUC values of 0.983 and 0.980, respectively. Recently, Dai et al. introduced TransMed [57] as a multi-modal medical picture categorization system. It combines the benefits of CNN and transformer to efficiently extract low-level characteristics from pictures and construct long-range relationships between modalities. The accuracy and the AUC of the model for detecting ACL tears were 94.9% and 0.98, respectively. These results were of higher accuracy than the MRNet technique. Astuto et al. [58] made use of 3D CNNs, which were designed to identify and grade ACL injuries in MRI investigations. The reported binary lesion sensitivity for ACL tissue is 88%. The specificity of the results is 89%. The AUC is 0.90.

3.2. Meniscus Tear

3.2.1. Machine Learning

Fu et al. [59] compared the performance of two SVM models in detecting meniscus tears. One model was trained on selected MR features (from a pool of 180 spatial and textural features using GLCM), while the other model implemented the SVM model without any feature selection. The SVM model without feature selection produced an AUC of 0.73, while their model with feature selection yielded an AUC value of 0.91. Zarandi et al. [60] performed MR image segmentation, followed by the application of a perceptron neural network (PNN) for classifying meniscal tears. The model accomplished a 90% classification accuracy (meniscus tear versus no meniscus tear) on a testing dataset of 50 MRI studies. Precision (%) was also reported for five different settings of meniscus tear, including: (1) medial anterior horn and posterior horn normal (88.82%); (2) lateral anterior horn and posterior horn normal (92.13%); (3) medial anterior horn normal and posterior horn torn (84.24%); (4) lateral anterior horn normal and posterior horn torn (91.96%); and (5) lateral anterior horn torn and posterior horn normal (87.64%).

3.2.2. Deep Learning with Transfer Learning

Another group [27] utilized MRNet as the primary building block of their prediction system, that is CNN mapping a 3D MRI series to a probability. The input to MRNet had dimensions: s × 3 × 256 × 256, where s was the number of images in the MRI series (3 is the number of color channels). In diagnosing a meniscus tear, this group reported an accuracy of 72.5% (95% CI: 0.639–0.797) and an AUC of 0.85 (95% CI: 0.78–0.91). Furthermore, they compared the performance of the proposed model with unassisted MSK radiologists for detecting a meniscus tear (intact, degenerative changes without tear, or postsurgical changes without tear). When compared to the MSK radiologists in the study, the AI model had a statistically significant lower specificity (AUC: 0.88, 95% CI: 0.85–0.91 versus AUC: 0.741, 95% CI: 0.62–0.84; *p*-value = 0.003) and accuracy (0.85, 95% CI: 0.82–0.87 versus 0.725, 95% CI: 0.64–0.80, *p*-value = 0.015). The sensitivity was also shown to be lower for the AI model (0.82, 95% CI: 0.78–0.85) compared to MSK radiologists (0.71, 95% CI: 0.59–0.81; *p*-value = 0.504), although this was not statistically significant. Azcona and colleagues [52] leveraged the baseline MRNet architecture and replaced the AlexNet feature extractor with more modern residual architectures, such as Resnet18, Resnet50, and Resnet152. They applied a series of transformations including horizontal flips and photometric augmentations (with respect to random contrast, gamma, and brightness). They reported an AUC performance of 0.91 on the validation data by using ResNet18.

3.2.3. Custom-Made Deep-Learning Networks

Couteaux et al. [61] used a region-based convolutional neural network (R-CNN) model for tear detection and localization (anterior or posterior). The anterior meniscus was classified as torn when at least one network had detected a torn anterior meniscus and the posterior meniscus was classified as torn when the strict majority of the networks had detected a torn posterior meniscus. A weighted AUC score of 0.91 was achieved by the proposed network on a test set of 700 MRIs. Another paper [62] also used an R-CNN trained on a dataset of 700 MRI images to perform three tasks, namely the detection of meniscus tear presence, position, and orientation. Their AI model produced an AUC of 0.94 on the task of detecting the presence of a meniscal tear, 0.92 for detecting the position of the two meniscal horns, and 0.83 for detecting the orientation of the tear. The overall combined AUC was 0.90.

Another group [63] created a DL model that combined meniscus segmentation and a 3D CNN for accomplishing both the detection and severity staging of meniscus lesions. The segmentation task for both cartilage and the meniscus was implemented using 2D U-Net [64]. The model was first built to recognize the presence of a lesion (including intrasubstance abnormalities), and, subsequently, to quantify the lesion severity. This model produced a lesion detection AUC performance of 0.89 on the test dataset and accuracies of 80.74%, 78.02%, and 75.00% for determining severe, mild-moderate, and no lesions, respectively. Comparisons were made between the model and experts. The authors also sought to determine the inter-rater variability between three MSK radiologists (expert 1: >20 years of experience, expert 2: 10 years of experience, and expert 3: <1 year of experience) for assessing meniscus lesion severity on selected cases. They restored an average agreement among the three experts of 86.27% for no meniscus lesions, 66.48% for mild-moderate lesions, and 74.66% for severe lesions, while the best model obtained accuracies of 80.74% for no meniscus lesions, 78.02% for mild-moderate lesions, and 75.00% for severe lesions.

Fritz et al. [15] proposed that deep CNN-based meniscus tear detection be performed in a fully automated manner with a similar specificity, but a lower sensitivity, in comparison with the MSK radiologists. The AUC of the deep CNN employed was 0.88, 0.78, and 0.96 for the detection of medial, lateral, and overall meniscus tear, respectively. The sensitivity, specificity, and accuracy for medial meniscus tear detection were 93%, 91%, and 92%, respectively, for reader 1; 96%, 86%, and 92%, respectively, for reader 2; and 84%, 88%, and 86%, respectively, for the DCNN. The sensitivity, specificity, and accuracy for lateral meniscus tear detection were 71%, 95%, and 89%, respectively, for reader 1; 67%, 99%, and 91%, respectively, for reader 2; and 58%, 92%, and 84%, respectively, for the DCNN. The sensitivity for medial meniscus tears was significantly different between reader 2 and the DCNN (p-value = 0.039), but no significant differences were witnessed in all other comparisons (all p-value $\geq$ 0.092). Rizk et al. [65] used a 3D CNN architecture that incorporated meniscal localization and lesion classification. They achieved AUC values of 0.93 and 0.84 for medial and lateral meniscal tear detection, respectively, and 0.91 and 0.95 for medial and lateral meniscal tear migration detection, respectively. The combined medial and lateral meniscal tear detection models were externally validated and yielded an AUC of 0.83 without additional training and 0.89 after fine-tuning. Moreover, Dai et al. utilized TransMed [57], achieving accuracy and AUC values of 94.9% and 0.98, respectively, for detecting meniscus tears, thus improving over the MRNet technique. 3D CNNs were built by Astuto et al. [58] to identify and grade meniscus tear in MRI examinations. The reported binary lesion sensitivity and specificity values were 85% for both., whereas the AUC was 0.93. Lastly, Dai et al. used TransMed to also identify meniscus tears in the MRNet dataset. The group reported an AUC of 0.95 and an accuracy of 85.3%.

3.3. Cartilage Lesion and Other Abnormalities

Liu et al. [66] developed a fully automated DL-based cartilage lesion detection system by combining CNN-based semantic segmentation and disease classification. Segmentation was implemented via the use of a VGG-16-based encoder network consisting of a combination of 2D convolution layers, rectified-linear activations, batch normalization layers, and max-pooling layers to achieve image feature extraction and data compression at the same time. The classification CNN in the proposed pipeline was also based on the 2D VGG16. Their pipeline achieved an AUC in the range of 0.91–0.92, indicating high overall diagnostic accuracy for detecting cartilage lesions. In addition, there was good intra-observer agreement between two individual evaluations, with a k-statistic of 0.76. As previously indicated, Astuto et al. [58] also used 3D CNNs to detect cartilage lesions. The sensitivity and specificity of binary lesions were found to be 85% and 89%, respectively, whereas the AUC was 0.93. Finally, three of the reported papers [27,52,57] attempted to detect other knee abnormalities, such as osteoarthritis, effusion, iliotibial band syndrome, posterior cruciate ligament tear, fracture, contusion, plica, and medial collateral ligament sprain. MRNet [27], ResNet18 [67], and TransMed networks were employed to implement the classification tasks, achieving AUC values of 0.94, 0.94, and 0.976, respectively.

4. Discussion and Conclusions

The present systematic review (Table 2) outlined the recent application of traditional ML and DL models to the diagnosis of the most common knee injuries using MRI as the main data source. The results of the present study can be summarized as follows. Figure 4 shows an increasing trend in adopting ML-based studies in this application area, with most of the papers being published from 2017 onwards (whilst the first ML-based paper on the field was published in 2013). Medical imaging, and specifically MRI, has to be seen as one of the most instructive assets in the field of knee injury diagnosis. The proliferation of MRI data has facilitated the effective training of ML and DL networks towards the development of: (i) novel methodologies that could enhance the medical experts' domain knowledge and understanding of MRI; and (ii) new, data-driven tools that could enable a more reliable, fast, and fully automated detection of knee injuries. The main characteristics of the proposed MRI-based learning algorithms and pipelines were identified along with the data sources investigated. The following paragraphs present our findings with respect to the choice of CNN networks and the associated results in comparison with clinical assessments carried out by experts.

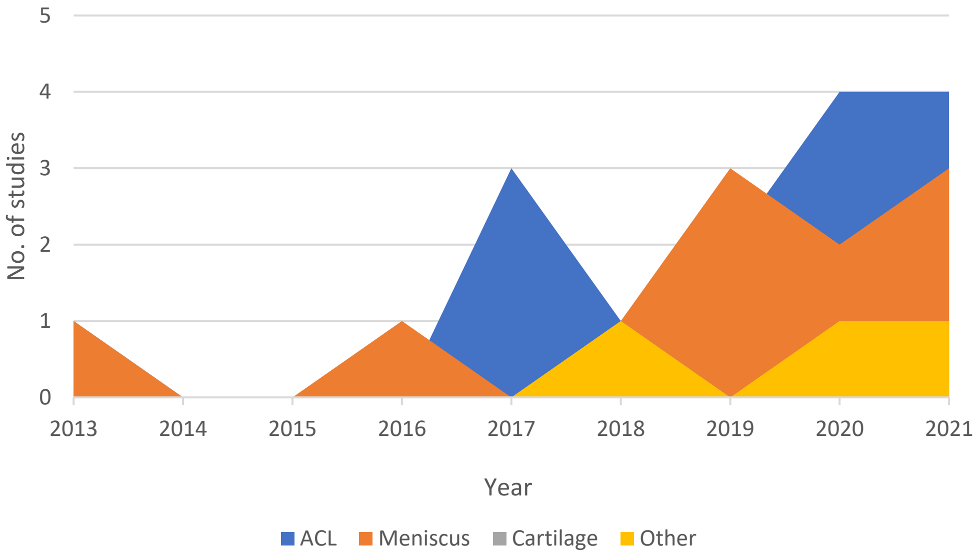

Figure 4. Temporal evolution chart depicting the number of ML papers per category published each year since 2013.

Table 2. Results of studies.

No.	Author	Year	AI Model Used	Pretrained CNN	MRI (T)	Localization Technique	Validation	Performance (Accuracy/AUC)	Application Domain
1	Awan et al. [55]	2021	CNN	ResNet-14	1.5 T	They applied normal approach to localize based upon region of interest (ROI)	5-fold cross-validation	92%/(healthy tear = 0.98, partial tear = 0.97 and fully ruptured tear = 0.99)	ACL tear
2	Jeon et al. [56]	2021	3D CNN	VGGNet, AlexNet, and SqueezeNet	3 T & 1.5 T	Custom localization technique	5-fold cross-validation	N/A/0.983 and 0.980 on the Chiba and Stanford knee datasets, respectively	ACL tear
3	Rizk et al. [65]	2021	3D CNN	CNN-based localization model	1 T (54%)–1.5 T (9.7%)–3 T (36.3%)	Custom localization technique	ten-fold cross validation	Meidal = N/A/0.93, Lateral = N/A/0.84	Meniscus tear
4	Dai et al. [57]	2021	TransMed	N/A	3 T & 1.5 T	N/A	120 exams	ACL tear = 94.9%/0.98, Abnormality = 91.8%/0.976, Meniscus tear = 85.3%/0.95	ACL tear—Meniscus tear—Abnormalities
5	Astuto et al. [58]	2021	3D CNN	N/A	3 T	V-Net	Hold out (15% of sample)	N/A/from 0.83 to 0.93	ACL tear—Meniscus tear—Cartilage Lession
6	Fritz et al. [15]	2020	DCNN	N/A	1.5 T (64%)–3 T (36%)	To visually localize the tear, the software computes the class activation map (CAM) of the last convolution layer in the CNN and maps it to an axial knee image	Hold out (10% of sample)	Medial = (86%/0.88), Lateral = (84%/0.78), Overall = (N/A/0.96)	Meniscus tear
7	Namiri et al. [54]	2020	CNN	N/A	3 T	three-dimensional V-Net	Hold out (10% of sample)	3D-model = (89%/sensitivity of 89% and specificity of 88%), 2D-model = (92%/sensitivity of 93% and specificity of 90%)	ACL tear

Table 2. *Cont.*

No.	Author	Year	AI Model Used	Pretrained CNN	MRI (T)	Localization Technique	Validation	Performance (Accuracy/AUC)	Application Domain
8	Zhang et al. [6]	2020	CNN	3D DenseNet, VGG16, ResNet	1.5 T (74%)–3 T (26%)	-	Hold out (20% of sample)	Custom = (95.7%/0.96), ResNet = (NA/0.95), VGG16 = (NA/0.86)	ACL tear
9	Germann et al. [24]	2020	DCNN	N/A	1.5 T–3 T	They cropped manually	Out of the 5802 MRI studies, 4802 were used for training, 500 for validation, and 500 for initial testing	N/A/0.94	ACL tear
10	Azcona et al. [52]	2020	CNN	MRNet, ResNet18, Resnet50 and ResNet152, ImageNet	3 T (56.6%)–1.5 T (43.4%)	-	N/A	NA/0.96–N/A/0.91–N/A/0.94	ACL tear—Meniscus tear—Abnormalities
11	Chang et al. [8]	2019	CNN	ResNet	1.5 T–3 T	The object localization CNN was implemented as a fully convolutional network based on U-net architecture	5-fold-cross-validation	96.7%/0.97	ACL tear
12	Liu et al. [53]	2019	CNN	LeNet-5, DenseNet, VGG16, AlexNet	N/A	They used object detection technique YOLO	50 subjects test set (14% of the sample)	N/A/0.98	ACL tear
13	Couteaux et al. [61]	2019	CNN	ResNet-101, ConvNet, R-CNN	N/A	To localize both menisci and identify tears in each meniscus, they used the Mask R-CNN framework	54 cases and the model with the highest validation accuracy was selected	N/A/0.90	Meniscus tear
14	Pedoia et al. [63]	2019	2D U-Net, CNN	N/A	3 T	-	Hold out (20% of sample)	Sensitivity of 89.81% and specificity of 81.98%	Meniscus tear
15	Roblot et al. [62]	2019	CNN	AlexNet, MRNet	N/A	They used object detection technique Fast RCNN & Faster RCNN	The algorithm was thus used on a test dataset composed of 700 images for external validation	72.5%/0.85	Meniscus tear

Table 2. *Cont.*

No.	Author	Year	AI Model Used	Pretrained CNN	MRI (T)	Localization Technique	Validation	Performance (Accuracy/AUC)	Application Domain
16	Nicholas Bien et al. [27]	2018	CNN	AlexNET, MRNet	3 T (56.6%)–1.5 T (43.4%)	-	120 exams	86.7%/0.97–72.5%/0.85–N/A/0.94	ACL tear—Meniscus tear—Abnormalities
17	Liu et al. [66]	2018	CNN	VGG16	3 T	-	fellowship trained musculoskeletal radiologist (R.K., with 15 years of clinical experience)	N/A/0.92	Cartilage lesion
18	Stajduhar et al. [48]	2017	HOG + linSVM, HOG + RF, GIST + rbfSVM, GIST + RF	N/A	1.5 T	Manual extraction of a rectangular ROI	10-fold cross validation	(Injury detection problem, complete rupture) = (N/A/0.89, N/A/0.94), (N/A/0.88, N/A/0.94), (N/A/0.889, N/A/0.91), (N/A/0.88, N/A/0.90) respectively with the models	ACL tear
19	Mazlan et al. [51]	2017	SVM	N/A	N/A	They use cropping technique	Hold out (10% of sample)	100%/N/A	ACL tear
20	Zarandi et al. [60]	2016	IT2FCM, PNN	N/A	N/A	-	Hold out (20% of sample)	0 and 1 mode: 90%/N/A Binary mode: 78%/N/A	Meniscus tear
21	Fu et al. [59]	2013	SVM	N/A	N/A	Active Contours without Edges method. This method combines Active Contours with Level Sets and is called ACLS	5-Fold cross validation	SVM model: N/A/0.73 SFFS + SVM: N/A/0.91	Meniscus tear
22	Abdullah et al. [50]	2013	BP ANN, K-NN	N/A	N/A	-	5-fold and 6-fold	BP ANN: 94.44%/N/A k-NN: 87.83%/N/A	ACL tear

Although there is no clear acceptance of a "gold-standard" methodological pipeline for diagnosing knee abnormalities using MRI data, it was observed that a number of processing steps were commonly employed in the majority of the reported studies. Figure 5 visualizes a DL pipeline that was adopted by most of the papers, including a pre-processing step, localization (optionally) by identifying regions of interest, and, finally, a CNN-based classification step. Data augmentation was employed by a significant number of papers in the detection of ACL injuries [6,27,52,54–58], in papers where meniscus injuries were investigated [27,52,57,58,62,63], and, finally, in studies focusing on cartilage lesion abnormalities [27,52]. In particular, the available MRI images were modified (via a number of image transformations such as random rotations, shifting, flipping, and the addition of noise) to expand the training dataset, and thus help to improve the performance and ability of the employed DL models to generalize. Localization was employed in papers from all three subcategories: (i) ACL studies [6,8,24,48,53–56,58]; (ii) meniscus injuries detection studies [15,58–63,65]; and (iii) for diagnosing lesion abnormalities [66]. Segmentation or objection detection algorithms were applied in the aforementioned studies to extract areas of interest, enabling the application of CNN-based models on focused and more relevant parts of the initially available images. Given that the region of interest (ROI) may appear in slightly different positions within an image and may have different aspect ratios or sizes, identifying ROIs with an automatic manner has been proven to be a crucial processing step.

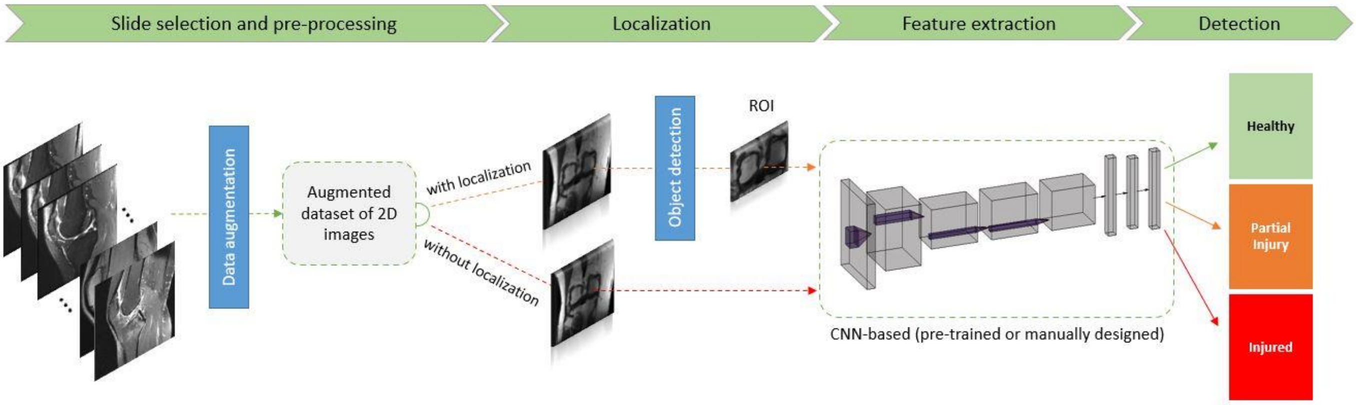

Figure 5. A typical DL pipeline for ACL detection.

CNN-inspired networks were identified as the dominant approach in the task of extracting informative features from either ROIs or entire MRIs and finally classifying them as normal (healthy) or abnormal (indicating either partial or complete tears). Transfer learning was preferred in most of the cases, allowing the training of big and powerful deep architectures, even if the amount of available data was limited. As networks require a lot of information to be trained from scratch, this technique essentially 'steals' knowledge from already pre-trained large networks. Specifically, ResNet variants were used in five papers [6,8,52,55,61] of this review, whereas VGG [34], AlexNet [33], and MRNet [27] were used three times [6,27,52,53,62,66]. Other pre-trained networks that were used at least once in this survey are: DenseNet [32], Le-Net [68], ImageNet [33], and R-CNN [41]. In five [48,50,51,59,60] out of the 22 studies of the present survey, more traditional ML pipelines were applied, including a separate feature engineering step (where features were manually extracted from images). SVM classification was the preferred classifier in most of the cases.

Despite the excellent capability of CNNs to come up with valuable image representations, these models lack the capacity for capturing long-range relationships. To deal with this limitation, recent research studies [44,69] have proposed employing Transformer-based architectures for various image recognition tasks. The Transformer [70] is a neural network architecture that relies on global self-attention mechanisms, and it was initially designed for sequence-to-sequence prediction. Papers that used this architectural paradigm have indeed achieved state-of-the-art results [71,72] in many natural language processing (NLP) tasks. Dai et al. [57] were the first to employ a Transformer-based architecture for the MRI-based knee injury detection task. In particular, their hybrid (Transformer and CNN) model was used to extract features that pick up the long-range dependencies between MRI and other modalities.

The present review demonstrated that the prediction accuracy of the DL models for the ACL and meniscus tears detection ranged from 72.5% to 100%. However, certain limitations have been identified in all studies that are included in this literature review. The lack of multi-center data has been recognized as a limitation in three papers [15,27,53], leading to the development of biased DL-detection systems that have only been tested on knee MRIs carried out at a single institution. The results of these studies relate to knee examinations using specific MR acquisition protocols for knee joint assessment. In general, classification models, trained using data acquired by a specific MRI acquisition protocol, are unsuccessful or underperform when applied to data that was obtained differently. One way to tackle this lack of ability to generalize is by using DL models that learn MRI acquisition-invariant contrast-agnostic representations [73,74]. The effect of data imbalance has been also highlighted in some cases [6,54,55] where the sample of patients was not properly balanced among all gradings, leading the algorithms to pay more attention to the majority class (typically the class of healthy subjects). Applying down-sampling in the majority class has been proved to be an unreliable approach, which led to a biased result in the case of the fully ruptured classes [55]. Verification bias was also identified [24,53], mainly because subjects involved in the studies underwent arthroscopic knee surgery, leading to increased sensitivity and decreased specificity for both the detection system and the clinical radiologists. Moreover, it should be stressed that the grades used for the training of the detection algorithms are typically dependent on subjective assessment by a limited number of radiologists (one in some cases [54]). In most of the studies, only two categories (normal versus tears) were discriminated and the need of considering additional categories was highlighted [6] to allow more detailed classifications to happen. Overall, it was stressed that the diagnostic performance of the combined use of a clinical radiologist and machine interpretation of the MRI examinations has not been evaluated [53].

The current study is a systematic review that followed the PRISMA guidelines, but did not include a more formal quantitative meta-analysis due to the observed heterogeneity of the identified studies. Moreover, diagnostic arthroscopy was not used as the gold-standard reference to identify ACL or meniscus injuries in the majority of the studies, which may restrict the clinical applicability of the findings.

Future studies should try to train and test the accuracy of AI prediction models for the detection of ACL and meniscal lesion based on the arthroscopic images, and compare the outcome with that of direct, non-arthroscopic assessments. Arthroscopy is a surrogate "gold-standard" for the validation of non-invasive assessments, such as MRI, as it provides highly magnified and direct views of articular cartilage with non-destructive interactive assessments of its structure and functional properties.

Radiological imaging data of the knee continues to grow at a disproportionate rate, vastly outnumbering the trained MSK radiologists. The workload has also increased dramatically, leading to inevitable errors in the decision-making process. Despite the identified limitations, AI systems have the potential to relieve physician burnout, utilize clinicians in fields at which they have not been specialized (MSK MRI), and reduce the cost of knee injury diagnosis for the public health system. In addition to flagging abnormal cases, if an AI algorithm could rapidly identify negative exams (increased sensitivity and negative predictive value), then, a substantial amount of time and other resources could be made free. Such a concept would be really useful in countries without easy access to medical expertise.

Advances in medical imaging, in terms of quality, sensitivity, and resolution, have enabled the discrimination between the smallest differences in the various knee tissue densities. These differences sometimes are difficult to recognize, even by a trained, specialized eye. Expert's diagnostic capacity used to be superior, but now we see this has been balanced out. As it was recently reported [24], deep CNN performance has reached performance levels akin to fellowship-trained, full-time, academic MSK radiologists in several tasks, including detection and segmentation. Despite this, AI can provide several new tools to the field of radiology imaging of the knee and medicine in general. The major hope for automated intelligent systems in the knee injury diagnosis is to increase accuracy, efficiency, and productivity in order to streamline patient care and outcomes. The newest, high-performance DL models should surpass the performance of traditional systems, meet the requirements for clinical utility, and become more user-friendly for the MSK clinician. Furthermore, there is the possibility of better training for young MSK radiologists with the help of AI.

MRI data of the knee, complemented by massive amounts of associated, multi-dimensional data such as omics and electronic health records, are only expected to grow. To fully exploit the full potential of this wealth of data, new paradigms should arise involving processes and workflows suitable for multi-institutional collaboration. Moreover, addressing the need for trustworthy detection systems of knee injuries, a medical diagnosis algorithm should meet a number of requirements (e.g., transparency, interpretability, explainability, and ease of use) in order to gain trust from clinicians. AI explainability and lightweight deep learning are key enablers for the wide use of such systems in the everyday clinical practice. Exploiting the intersection and merits of traditional ML and DL methods, AI analytics are expected to revolutionize knee medical informatics, enabling informed and accurate diagnoses needed by precision medicine.

Notwithstanding the huge potential of AI to improve the medical domain, the DL-based methods have yet to achieve significant deployment in clinical environments. This mainly ensues as a result of: (i) the intrinsic black-box nature of the DL algorithms; and (ii) the high computational cost. Explainable AI aims at building trust in the AI algorithms by providing medical experts with a diagnostic rationale behind the AI decision processes. The goal of the lightweight DL field is to develop models that have shallower architecture and are also faster and more data-efficient, while retaining the high-performance standards. Jeon et al. [56] were the first to get to grips with the clinical deployment of the MRI-based knee injury diagnosis. To this end, they proposed to use post-inference visualisation tools (such as CAM and Grad-CAM), and they also incorporated attention modules, Gaussian positional encoding, squeeze modules, and fewer convolutional filters.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/diagnostics12020537/s1.

Author Contributions: A.S., D.T., S.T., M.H. and G.C. designed the study. A.S. collected the data. D.T., A.G. and M.V. acted as reviewers to resolve issues regarding the selection of articles for data synthesis. I.L. checked that the methodology was conducted with PRISMA. A.S. and S.M. prepared all the figures and tables. A.S. wrote the first manuscript draft with significant support from S.M. A.G. reviewed and updated the manuscript. G.C. provided input to the introduction and the discussion. All authors reviewed the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been Co-financed by: (a) 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) and (b) the Department of Mathematics at Nottingham Trent University, Nottingham, UK.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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

References

1. Musahl, V.; Karlsson, J. Anterior cruciate ligament tear. *N. Engl. J. Med.* **2019**, *380*, 2341–2348. [CrossRef] [PubMed]
2. Ahmed, I.; Bowes, M.; Hutchinson, C.E.; Parsons, N.; Staniszewska, S.; Price, A.J.; Metcalfe, A. Protocol: Meniscal tear outcome Study (METRO Study): A study protocol for a multicentre prospective cohort study exploring the factors which affect outcomes in patients with a meniscal tear. *BMJ Open* **2020**, *10*, e038681. [CrossRef] [PubMed]
3. Darrow, C.J.; Collins, C.L.; Yard, E.E.; Comstock, R.D. Epidemiology of severe injuries among United States high school athletes: 2005–2007. *Am. J. Sports Med.* **2009**, *37*, 1798–1805. [CrossRef] [PubMed]
4. Gage, B.E.; McIlvain, N.M.; Collins, C.L.; Fields, S.K.; Dawn Comstock, R. Epidemiology of 6.6 million knee injuries presenting to United States emergency departments from 1999 through 2008. *Acad. Emerg. Med.* **2012**, *19*, 378–385. [CrossRef]
5. Merkely, G.; Ackermann, J.; Lattermann, C. Articular cartilage defects: Incidence, diagnosis, and natural history. *Oper. Tech. Sports Med.* **2018**, *26*, 156–161. [CrossRef]
6. Zhang, L.; Li, M.; Zhou, Y.; Lu, G.; Zhou, Q. Deep Learning Approach for Anterior Cruciate Ligament Lesion Detection: Evaluation of Diagnostic Performance Using Arthroscopy as the Reference Standard. *J. Magn. Reson. Imaging* **2020**, *52*, 1745–1752. [CrossRef]
7. Kaeding, C.C.; Léger-St-Jean, B.; Magnussen, R.A. Epidemiology and diagnosis of anterior cruciate ligament injuries. *Clin. Sports Med.* **2017**, *36*, 1–8. [CrossRef]
8. Chang, P.D.; Wong, T.T.; Rasiej, M.J. Deep Learning for Detection of Complete Anterior Cruciate Ligament Tear. *J. Digit. Imaging* **2019**, *32*, 980–986. [CrossRef]
9. Logerstedt, D.S.; Snyder-Mackler, L.; Ritter, R.C.; Axe, M.J.; Godges, J.; Altman, R.D.; Briggs, M.; Chu, C.; Delitto, A.; Ferland, A. Knee pain and mobility impairments: Meniscal and articular cartilage lesions: Clinical practice guidelines linked to the international classification of functioning, disability, and health from the orthopaedic section of the American Physical Therapy Association. *J. Orthop. Sports Phys. Ther.* **2010**, *40*, A1–A35.
10. Mather, R.C., III; Koenig, L.; Kocher, M.S.; Dall, T.M.; Gallo, P.; Scott, D.J.; Bach Jr, B.R.; Spindler, K.P.; Group, M.K. Societal and economic impact of anterior cruciate ligament tears. *J. Bone Jt. Surg. Am. Vol.* **2013**, *95*, 1751. [CrossRef]
11. Cameron, K.L.; Thompson, B.S.; Peck, K.Y.; Owens, B.D.; Marshall, S.W.; Svoboda, S.J. Normative values for the KOOS and WOMAC in a young athletic population: History of knee ligament injury is associated with lower scores. *Am. J. Sports Med.* **2013**, *41*, 582–589. [CrossRef] [PubMed]
12. Huffman, G.R.; Park, J.; Roser-Jones, C.; Sennett, B.J.; Yagnik, G.; Webner, D. Normative SF-36 values in competing NCAA intercollegiate athletes differ from values in the general population. *JBJS* **2008**, *90*, 471–476. [CrossRef] [PubMed]
13. Lam, K.C.; Thomas, S.S.; Valier, A.R.S.; McLeod, T.C.V.; Bay, R.C. Previous knee injury and health-related quality of life in collegiate athletes. *J. Athl. Train.* **2017**, *52*, 534–540. [CrossRef] [PubMed]
14. Pouly, M.; Koller, T.; Gottfrois, P.; Lionetti, S. Artificial intelligence in image analysis-fundamentals and new developments. *Der Hautarzt Z. Fur Dermatol. Venerol. Und Verwandte Geb.* **2020**, *71*, 660–668. [CrossRef] [PubMed]
15. Fritz, B.; Marbach, G.; Civardi, F.; Fucentese, S.F.; Pfirrmann, C.W. Deep convolutional neural network-based detection of meniscus tears: Comparison with radiologists and surgery as standard of reference. *Skelet. Radiol.* **2020**, *49*, 1207–1217. [CrossRef]
16. Garwood, E.R.; Tai, R.; Joshi, G. The Use of Artificial Intelligence in the Evaluation of Knee Pathology. In *Seminars in Musculoskeletal Radiology*; Thieme Medical Publishers: New York, NY, USA, 2020; pp. 21–29.
17. Palermi, S.; Massa, B.; Vecchiato, M.; Mazza, F.; De Blasiis, P.; Romano, A.M.; Di Salvatore, M.G.; Della Valle, E.; Tarantino, D.; Ruosi, C.J.J.O.F.M.; et al. Indirect Structural Muscle Injuries of Lower Limb: Rehabilitation and Therapeutic Exercise. *J. Funct. Morphol. Kinesiol.* **2021**, *6*, 75. [CrossRef]
18. Sirico, F.; Palermi, S.; Massa, B.; Corrado, B. Tendinopathies of the hip and pelvis in athletes: A narrative review. *J. Hum. Sports Exerc.* **2020**, *15*, S748–S762.
19. Hetsroni, I.; Lyman, S.; Do, H.; Mann, G.; Marx, R. Symptomatic pulmonary embolism after outpatient arthroscopic procedures of the knee: The incidence and risk factors in 418 323 arthroscopies. *J. Bone Jt. Surg. Br. Vol.* **2011**, *93*, 47–51. [CrossRef]
20. Alanazi, H.O.; Abdullah, A.H.; Qureshi, K.N. A critical review for developing accurate and dynamic predictive models using machine learning methods in medicine and health care. *J. Med. Syst.* **2017**, *41*, 69. [CrossRef]
21. Prickett, W.D.; Ward, S.I.; Matava, M.J. Magnetic resonance imaging of the knee. *Sports Med.* **2001**, *31*, 997–1019. [CrossRef]
22. Krampla, W.; Roesel, M.; Svoboda, K.; Nachbagauer, A.; Gschwantler, M.; Hruby, W. 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.* **2009**, *19*, 1519–1528. [CrossRef] [PubMed]
23. Mohankumar, R.; White, L.M.; Naraghi, A. Pitfalls and pearls in MRI of the knee. *Am. J. Roentgenol.* **2014**, *203*, 516–530. [CrossRef] [PubMed]
24. Germann, C.; Marbach, G.; Civardi, F.; Fucentese, S.F.; Fritz, J.; Sutter, R.; Pfirrmann, C.W.; Fritz, B. 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. *Investig. Radiol.* **2020**, *55*, 499. [CrossRef] [PubMed]
25. Gyftopoulos, S.; Lin, D.; Knoll, F.; Doshi, A.M.; Rodrigues, T.C.; Recht, M.P. Artificial intelligence in musculoskeletal imaging: Current status and future directions. *Am. J. Roentgenol.* **2019**, *213*, 506–513. [CrossRef] [PubMed]
26. Shen, D.; Wu, G.; Suk, H.-I. Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* **2017**, *19*, 221–248. [CrossRef]

27. Bien, N.; Rajpurkar, P.; Ball, R.L.; Irvin, J.; Park, A.; Jones, E.; Bereket, M.; Patel, B.N.; Yeom, K.W.; Shpanskaya, K. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLoS Med.* **2018**, *15*, e1002699. [CrossRef]
28. Langerhuizen, D.W.; Janssen, S.J.; Mallee, W.H.; Van Den Bekerom, M.P.; Ring, D.; Kerkhoffs, G.M.; Jaarsma, R.L.; Doornberg, J.N. What are the applications and limitations of artificial intelligence for fracture detection and classification in orthopaedic trauma imaging? A systematic review. *Clin. Orthop. Relat. Res.* **2019**, *477*, 2482. [CrossRef]
29. Kunze, K.N.; Rossi, D.M.; White, G.M.; Karhade, A.V.; Deng, J.; Williams, B.T.; Chahla, J. Diagnostic Performance of Artificial Intelligence for Detection of Anterior Cruciate Ligament and Meniscus Tears: A Systematic Review. *Arthrosc. J. Arthrosc. Relat. Surg.* **2020**, *37*, 771–781. [CrossRef]
30. Hellerstein, J.M.; Naughton, J.F.; Pfeffer, A. Generalized Search Trees for Database Systems. In Proceedings of the 21st VLDB Conference, Zurich, Switzerland, 11–15 September 1995.
31. Liu, F. Susan: Segment unannotated image structure using adversarial network. *Magn. Reson. Med.* **2019**, *81*, 3330–3345. [CrossRef]
32. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
33. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Processing Syst.* **2012**, *25*, 1097–1105. [CrossRef]
34. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
35. Freeman, W.T.; Roth, M. Orientation histograms for hand gesture recognition. In Proceedings of the International Workshop on Automatic Face and Gesture Recognition, Zurich, Switzerland, 26–28 June 1995; pp. 296–301.
36. Sebastian, V.B.; Unnikrishnan, A.; Balakrishnan, K. Gray level co-occurrence matrices: Generalisation and some new features. *arXiv* **2012**, arXiv:1205.4831.
37. Peterson, L.E. K-nearest neighbor. *Scholarpedia* **2009**, *4*, 1883. [CrossRef]
38. Hearst, M.A.; Dumais, S.T.; Osuna, E.; Platt, J.; Scholkopf, B. Support vector machines. *IEEE Intell. Syst. Appl.* **1998**, *13*, 18–28. [CrossRef]
39. Hassoun, M.H. *Fundamentals of Artificial Neural Networks*; MIT Press: Cambridge, MA, USA, 1995.
40. O'Shea, K.; Nash, R. An introduction to convolutional neural networks. *arXiv* **2015**, arXiv:1511.08458.
41. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
42. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
43. Kamnitsas, K.; Ledig, C.; Newcombe, V.F.; Simpson, J.P.; Kane, A.D.; Menon, D.K.; Rueckert, D.; Glocker, B. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med. Image Anal.* **2017**, *36*, 61–78. [CrossRef]
44. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.J.A.P.A. An image is worth 16×16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
45. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
46. Moher, D.; Liberati, A.; Tetzlaff, J.; Altman, D.G.; Altman, D.; Antes, G.; Atkins, D.; Barbour, V.; Barrowman, N.; Berlin, J.A. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement (Chinese edition). *J. Chin. Integr. Med.* **2009**, *7*, 889–896. [CrossRef]
47. Slim, K.; Nini, E.; Forestier, D.; Kwiatkowski, F.; Panis, Y.; Chipponi, J.J.A.J.O.S. Methodological index for non-randomized studies (MINORS): Development and validation of a new instrument. *ANZ J. Surg.* **2003**, *73*, 712–716. [CrossRef]
48. Štajduhar, I.; Mamula, M.; Miletić, D.; Ünal, G. Semi-automated detection of anterior cruciate ligament injury from MRI. *Comput. Methods Programs Biomed.* **2017**, *140*, 151–164. [CrossRef] [PubMed]
49. Pal, M. Random forest classifier for remote sensing classification. *Int. J. Remote Sens.* **2005**, *26*, 217–222. [CrossRef]
50. Abdullah, A.A.; Azz-Zahra Md Som, N.S.F. Design of an Intelligent Diagnostic System for Detection of Knee Injuries. *Appl. Mech. Mater.* **2013**, *399*, 219–224. [CrossRef]
51. Mazlan, S.S.; Ayob, M.; Bakti, Z.K. Anterior cruciate ligament (ACL) injury classification system using support vector machine (SVM). In Proceedings of the 2017 International Conference on Engineering Technology and Technopreneurship (ICE2T), Kuala Lumpur, Malaysia, 18–20 September 2017; pp. 1–5.
52. Azcona, D.; McGuinness, K.; Smeaton, A.F. A Comparative Study of Existing and New Deep Learning Methods for Detecting Knee Injuries using the MRNet Dataset. In Proceedings of the 2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA), Kuala Lumpur, Malaysia, 18–20 September 2020; pp. 149–155.
53. Liu, F.; Guan, B.; Zhou, Z.; Samsonov, A.; Rosas, H.; Lian, K.; Sharma, R.; Kanarek, A.; Kim, J.; Guermazi, A. Fully automated diagnosis of anterior cruciate ligament tears on knee MR images by using deep learning. *Radiol. Artif. Intell.* **2019**, *1*, 180091. [CrossRef]
54. Namiri, N.K.; Flament, I.; Astuto, B.; Shah, R.; Tibrewala, R.; Caliva, F.; Link, T.M.; Pedoia, V.; Majumdar, S. Deep Learning for Hierarchical Severity Staging of Anterior Cruciate Ligament Injuries from MRI. *Radiol. Artif. Intell.* **2020**, *2*, e190207. [CrossRef]

55. Awan, M.J.; Rahim, M.S.M.; Salim, N.; Mohammed, M.A.; Garcia-Zapirain, B.; Abdulkareem, K.H. Efficient detection of knee anterior cruciate ligament from magnetic resonance imaging using deep learning approach. *Diagnostics* **2021**, *11*, 105. [CrossRef]
56. Jeon, Y.S.; Yoshino, K.; Hagiwara, S.; Watanabe, A.; Quek, S.T.; Yoshioka, H.; Feng, M.J.I.J.O.B.; Informatics, H. Interpretable and lightweight 3-D deep learning model for automated ACL diagnosis. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 2388–2397. [CrossRef]
57. Dai, Y.; Gao, Y.; Liu, F.J.D. Transmed: Transformers advance multi-modal medical image classification. *Diagnostics* **2021**, *11*, 1384. [CrossRef]
58. Astuto, B.; Flament, I.K.; Namiri, N.; Shah, R.; Bharadwaj, U.; M. Link, T.; D. Bucknor, M.; Pedoia, V.; Majumdar, S.J.R.A.I. Automatic Deep Learning–assisted Detection and Grading of Abnormalities in Knee MRI Studies. *Radiol. Artif. Intell.* **2021**, *3*, e200165. [CrossRef]
59. Fu, J.-C.; Lin, C.-C.; Wang, C.-N.; Ou, Y.-K. Computer-aided diagnosis for knee meniscus tears in magnetic resonance imaging. *J. Ind. Prod. Eng.* **2013**, *30*, 67–77. [CrossRef]
60. Zarandi, M.F.; Khadangi, A.; Karimi, F.; Turksen, I. A computer-aided type-II fuzzy image processing for diagnosis of meniscus tear. *J. Digit. Imaging* **2016**, *29*, 677–695. [CrossRef] [PubMed]
61. Couteaux, V.; Si-Mohamed, S.; Nempont, O.; Lefevre, T.; Popoff, A.; Pizaine, G.; Villain, N.; Bloch, I.; Cotten, A.; Boussel, L. Automatic knee meniscus tear detection and orientation classification with Mask-RCNN. *Diagn. Interv. Imaging* **2019**, *100*, 235–242. [CrossRef] [PubMed]
62. 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] [PubMed]
63. Pedoia, V.; Norman, B.; Mehany, S.N.; Bucknor, M.D.; Link, T.M.; Majumdar, S. 3D convolutional neural networks for detection and severity staging of meniscus and PFJ cartilage morphological degenerative changes in osteoarthritis and anterior cruciate ligament subjects. *J. Magn. Reson. Imaging* **2019**, *49*, 400–410. [CrossRef]
64. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
65. Rizk, B.; Brat, H.; Zille, P.; Guillin, R.; Pouchy, C.; Adam, C.; Ardon, R.; d'Assignies, G.J.P.M. Meniscal lesion detection and characterization in adult knee MRI: A deep learning model approach with external validation. *Phys. Med.* **2021**, *83*, 64–71. [CrossRef]
66. Liu, F.; Zhou, Z.; Samsonov, A.; Blankenbaker, D.; Larison, W.; Kanarek, A.; Lian, K.; Kambhampati, S.; Kijowski, R. Deep learning approach for evaluating knee MR images: Achieving high diagnostic performance for cartilage lesion detection. *Radiology* **2018**, *289*, 160–169. [CrossRef]
67. Targ, S.; Almeida, D.; Lyman, K. Resnet in resnet: Generalizing residual architectures. *arXiv* **2016**, arXiv:1603.08029.
68. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
69. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Online, 22–25 November 2021; pp. 6881–6890.
70. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
71. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K.J.A.P.A. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
72. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.J.A.P.A. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
73. Kouw, W.M.; Loog, M.; Bartels, L.W.; Mendrik, A.M.J.A.P.A. MR acquisition-invariant representation learning. *arXiv* **2017**, arXiv:1709.07944.
74. Billot, B.; Greve, D.; Van Leemput, K.; Fischl, B.; Iglesias, J.E.; Dalca, A.V.J.A.P.A. A learning strategy for contrast-agnostic MRI segmentation. *arXiv* **2020**, arXiv:2003.01995.