big data and cognitive computing

MDPI

Article

Comparative Analysis of Backbone Networks for Deep Knee MRI Classification Models

Nataliya Shakhovska † **and Pavlo Pukach** *,†

AI Department, Lviv Polytechnic National University, 79000 Lviv, Ukraine; nataliya.b.shakhovska@lpnu.ua
* Correspondence: pavlo.p.pukach@lpnu.ua; Tel.: +38-063-014-8936
† These authors contributed equally to this work.

Abstract: This paper focuses on different types of backbone networks for machine learning architectures which perform classification of knee Magnetic Resonance Imaging (MRI) images. This paper aims to compare different types of feature extraction networks for the same classification task, in terms of accuracy and performance. Multiple variations of machine learning models were trained based on the MRNet architecture, choosing AlexNet, ResNet, VGG-11, VGG-16, and Efficientnet as the backbone. The models were evaluated on the MRNet validation dataset, computing Area Under the Receiver Operating Characteristics Curve (ROC-AUC), accuracy, f1 score, and Cohen's Kappa as evaluation metrics. The MRNet-VGG16 model variant shows the best results for Anterior Cruciate Ligament (ACL) tear detection. For general abnormality detection, MRNet-VGG16 is dominated by MRNet-Resnet in confidence between 0.5 and 0.75 and by MRNet-VGG11 for confidence more than 0.8. Due to the non-uniform nature of backbone network performance on different MRI planes, it is advisable to use an LR ensemble of: VGG16 on a coronal plane for all classification tasks; on an axial plane for abnormality and ACL tear detection; Alexnet on a sagittal plane for abnormality detection, and an axial plane for meniscal tear detection; and VGG11 on a sagittal plane for ACL tear detection. The results also indicate that the Cohen's Kappa metric is valuable in model evaluation for the MRNet dataset, as it provides deeper insights on classification decisions.

Keywords: knee MRI; computer-assisted diagnostics; MRNet; deep learning; computer vision

Citation: Shaknovska, N.; Pukach, P. Comparative Analysis of Backbone Networks for Deep Knee MRI Classification Models. *Big Data Cogn. Comput.* **2022**, *6*, 69. https://doi.org/10.3390/ bdcc6030069

Academic Editor: Moulay A. Akhloufi

Received: 24 May 2022
Accepted: 16 June 2022
Published: 21 June 2022

1. Introduction

Magnetic Resonance Imaging (MRI) is one of the most efficient techniques for knee examination developed up to this day. It has become the preferred modality for imaging the knee to show pathology and guide patient management and treatment [1].

It allows us to take sequences of images called "slices" across three different planes - axial, coronal, and sagittal (presented in Figure 1), building a complete representation of any body part. MRIs provide great insight for expert physicians, including musculoskeletal (MSK) radiologists, who conclude the diagnosis, and choose treatment methods.

As the population grows, the demand for physicians, and MSK radiologists, in particular, increases proportionally. Recent statistical studies are forecasting significant shortages of expert radiologists, along with other specialists in the field of medicine, all across the world.

In USA only, the shortage of radiologists and other physicians could surpass 35,000 by 2034, according to a recently published annual analysis of physician supply and demand by the Association of American Medical Colleges' [2].

Across all care segments, the number could climb as high as 124,000. Both aging and population growth are primary drivers of these shortages, with the 65-and-up segment projected to swell by more than 42% over the next decade.

The report did not give a specific number for radiology alone, lumping the specialty together with anesthesiology, neurology, emergency medicine, and addiction specialists.

This segment is projected to see a shortfall of between 10,300 and 35,600 physicians by 2034. All of these findings are part of the mentioned analysis, published on 11 June 2021.

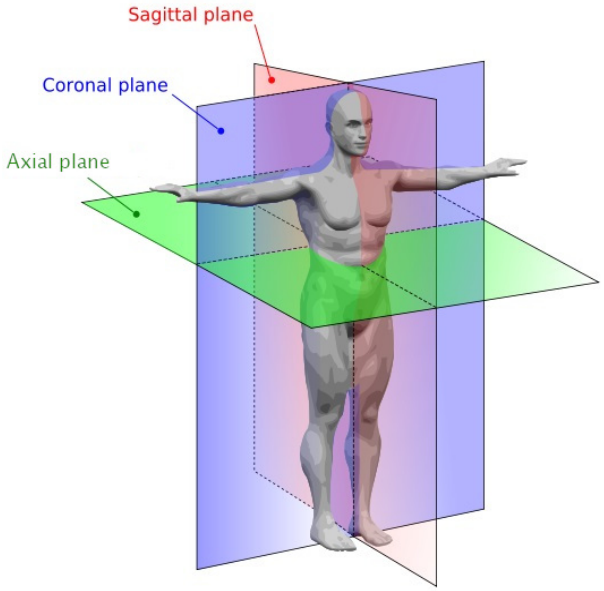


Figure 1. Planes of MRI imaging. axial (green), coronal (blue) and sagittal (red).

This brings to the point that there is a need for robust computer systems, which could aid MSK radiologists in identifying pathological conditions in patients' MRI scans. The primary application of such systems should be to lessen the ever-increasing workload of MSK radiologists.

Thankfully, at the intersection of artificial intelligence (AI) and radiology fields, a variety of computer science-derived methods have been already applied to do decision support for health practitioners in diagnostic tasks. As these methods improve and show certain success in accomplishing their intended purpose, the interest in them grows proportionally from both AI scientists and radiologists, also considering the yet unlimited potential of artificial intelligence applications [3].

Within AI, deep learning (DL) algorithms can learn many different tasks that directly apply to musculoskeletal radiology, including image reconstruction, synthetic image creation, tissue segmentation, and detection and characterization of musculoskeletal diseases and conditions on radiographs, ultrasonography, CT, and MR images.

Ideally, such systems should also help the radiologists focus specifically on rare and challenging anomalies and cases, by automatically labeling common injuries with confidence rates alike those of boards of expert radiologists. In this paper, an evaluation of the state-of-art deep learning knee MRI classification models is presented.

Despite the fact that there had been no fundamental breakthroughs in the field of convolutional neural networks for computer vision since the introduction of Resnet in 2015 [4], a lot of research has been made on providing variations of the major model architectures, that either improve on feature extraction, are lighter in terms of disk size and faster to train, or both.

Amongst all the related works, there have been few attempts to train the baseline MRNet architecture on more contemporary computer vision architectures. And so far, there was no recorded attempt to bring out and document the gradual improvement of MRNet prediction accuracy, as the effort of the researchers in the computer vision field brought more robust architectures with improved accuracy.

The reference MRNet architecture and related papers depend on a pre-trained Alexnet network as a feature extraction layer. In the work of Tsai et al. (2020) [5], the authors use

their own convolutional feature extraction layer, claimed as having greater efficiency and reduced number of trainable parameters, compared to AlexNet [6].

A relatively new feature extractor utilized in MRNet was mentioned only in [7], where they have used a variation of the Resnet architecture—Resnet18.

The paper aims to check the significance and the impact of contemporary models on feature extraction on the knee MRI classification task. Such comparison is needed as guidance for building applied machine learning model architectures, focused on automatic knee injury detection in enterprise systems.

In this paper, we attempt to compare the MRNet baseline architecture performance for the knee MRI classification task, using modern computer vision feature extraction architectures. We are also planning to display the gradual prediction accuracy improvement, aligned with the evolution of these backbone networks: VGG11, VGG16 [8], Resnet [4], Efficientnet [9].

Even taking into account the fact that the Resnet architecture as a backbone was already mentioned in the work of Azcona et al. (2020) [7], we still would like to re-train this architecture as a backbone, to display the gradual improvement of validation accuracy with the usage of more recent and progressive computer vision tools, and provide transparent prediction accuracy scores, that might have been otherwise affected by the differences in model hyper-parameters or training options between our model, and the version listed in Azcona et al. (2020).

Another significant aspect of our research is the fact, that the models trained during our experiments have unified architecture, except for the feature extractor part, and they were all trained from scratch in a reproducible manner. This is the perfect environment for comparing the prediction accuracy, as non-deterministic factors like random seeds for classifier layer weight initialization, and training parameters (learning rate, weight decay, number of epochs) are unified for each model evaluation presented in this paper.

Furthermore, our model evaluation strategies utilize an additional metric, which was not yet measured in any of the related works—the Cohen's Kappa score. Here, it is considered a significant one because of the natural imbalance of the MRNet dataset.

Therefore, the paper contribution is formulated as follows:

- A comparative analysis is conducted for a set of classification models based on the original MRNet architecture, with different backbone networks as feature extractors for MRNet blocks, including Alexnet, VGG11, VGG16, Resnet, and Efficientnet. MRNet-VGG16 shows the best results for ACL tear detection. For general abnormality detection, MRNet-VGG16 is dominated by MRNet-Resnet for confidence between 0.5 and 0.75 and by MRNet-VGG11 for confidence more than 0.8.
- This research brings out the fact that the performance of the above-mentioned backbone models are different across MRI planes. To achieve the best classification accuracy, it is advisable to use a Logistic Regression (LR) ensemble of different backbone architectures: VGG16 on the coronal plane for all classification tasks; on an axial plane for abnormality and ACL tear detection; Alexnet on a sagittal plane for abnormality detection, and axial plane for meniscus tear detection; VGG11 on a sagittal plane for ACL tear detection.
- The interrater reliability approach to model evaluation, in the form of Cohen's kappa scores, calculated for each diagnosis, and each backbone type, is a valuable evaluation metric, as it can provide deeper insights into model performance than traditional ROC-AUC computation.

2. Materials and Methods

2.1. Training Datasets and State of the Art Models

One of the first attempts at utilizing artificial intelligence methods for the purpose of knee injury detection on MRI scans, which resulted in a publicly available dataset of labeled knee MRI scans gathered from a substantial amount of clinical cases was presented by [10]. The resulting dataset was named KneeMRI.

The KneeMRI dataset was gathered retrospectively from exam records made on a Siemens Avanto 1.5T MR scanner, and obtained by proton density-weighted fat suppression technique at the Clinical Hospital Centre Rijeka, Croatia, from 2006 until 2014. The dataset consists of 917 12-bit grayscale volumes of either left or right knees.

Each volume record was assigned a diagnosis concerning the condition of the anterior cruciate ligament in a double-blind fashion, i.e., each volume record was labeled according to the ligament condition: (1) healthy, (2) partially injured, or (3) completely ruptured. A wider rectangular region of interest (ROI) was manually extracted from the original volumes and is also annotated [10]. This dataset was built to provide scientists, involved with machine vision and/or machine learning, an easy way of working with the data.

For this paper, the MRNet dataset was chosen as a base for training and tuning our machine learning models. It was created by the Stanford University School of Medicine, as means to facilitate the development of deep-learning models able to predict abnormalities in presented knee MRI scans.

The MRNet dataset consists of 1370 knee MRI exams performed at Stanford University Medical Center. The dataset contains 1104 (80.6%) abnormal exams, with 319 (23.3%) ACL tears and 508 (37.1%) meniscal tears; labels were obtained through manual extraction from clinical reports [11]. Examples of middle slice knee MRI images are shown in Figure 2.

The most common indications for the knee MRI examinations in the dataset included acute and chronic pain, follow-up or preoperative evaluation, and injury/trauma. Examinations were performed with GE scanners (GE Discovery, GE Healthcare, Waukesha, WI, USA) with a standard knee MRI coil and a routine non-contrast knee MRI protocol that included the following sequences: coronal T1 weighted, coronal T2 with fat saturation, sagittal proton density (PD) weighted, sagittal T2 with fat saturation, and axial PD weighted with fat saturation. A total of 775 (56.6%) examinations used a 3.0-T magnetic field; the remaining used a 1.5-T magnetic field [11].

(**a**) Axial MRI slice example (**b**) Coronal MRI slice example (**c**) Sagittal MRI slice example

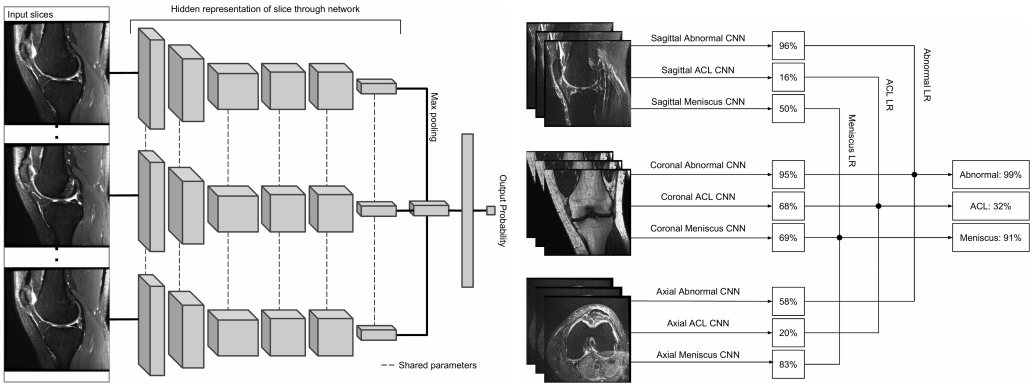
Figure 2. Input image examples of MRNet. Three images represent (**a**) axial, (**b**) coronal and (**c**) sagittal knee MRI planes.

Since the introduction and initial publication of the MRNet dataset, many attempts at building robust decision support systems for the purpose of knee MRI classification have been made. Most of them utilized the deep learning approach and convolutional neural networks.

The primary state-of-art model is the original MRNet architecture, implemented by [11]. In their work, the authors have built a composite model for knee injury detection, with a dedicated model for each MRI slice.

The overall model architecture is shown in Figure 3. A single MRnet block is a max-pooling layer, on top of a pre-trained backbone network - AlexNet [6] as a feature extractor, followed by a fully-connected layer with softmax activation for classification. Then, for each MRI plane - axial, coronal, and sagittal, a single MRnet block is trained to classify injuries and abnormalities. Lastly, logistic regression is used as an approach to combine

these different MRNet blocks into a single binary classifier for ACL tears, Meniscus tears, and general knee abnormalities.



(**a**) Single block architecture in MRNet (**b**) Overall model architecture involving Logistic Regression (LR)

Figure 3. MRNet model architecture: (**a**) single block, (**b**) overall model.

In Azcona et al. (2020) [7], the authors have shown that the three-model MRNet architecture with logistic regression has the best validation performance, compared to different variations of the same architecture, including self-trained AlexNets as a feature extraction layer, and single models, which operate on a concatenation of slices throughout all three MRI planes. Thus, in this paper, the decision is made to use the same base MRNet architecture for comparing different backbone networks for a single block.

In Tsai et al. (2020) [5], the authors present an Efficiently-Layered Network (ELNet) architecture optimized for knee diagnosis detection using MRI. The main contribution of their work is a novel feature extracting network that incorporates multi-slice normalization along with BlurPool downsampling, instead of max pooling or adaptive average pooling, which could be seen in other related works.

2.2. Overfitting Prevention

As stated in Shorten, C., Khoshgoftaar T.M. (2019) [12], it is a generally accepted notion that bigger datasets result in better Deep Learning models. However, assembling enormous datasets can be a very daunting task due to the manual effort of collecting and labeling data. Limited datasets are an especially prevalent challenge in medical image analysis.

MRNet could be considered a small dataset, so the overfitting mitigation techniques listed in the mentioned paper are very useful for this application, to help the built models generalize on previously unknown and non-trained data.

For overfitting prevention, we utilize an approach of image augmentation. This is a method of synthetically boosting the number of input images, by applying random image transformations. Each input image is processed, and a set of new "transformed" images is created from it. The process of image augmentation for input knee MRI images is shown in Figure 4. This helps to create a more vast and diverse input for the machine learning models, also helping with the overfitting issue.

Data augmentation has been shown to produce promising ways to increase the accuracy of classification tasks. Even traditional ways of image augmentation, without involving additional Generative Adversarial Networks (GAN), proved to be very effective, as described in Wang, J., Perez, L. (2017) [13].

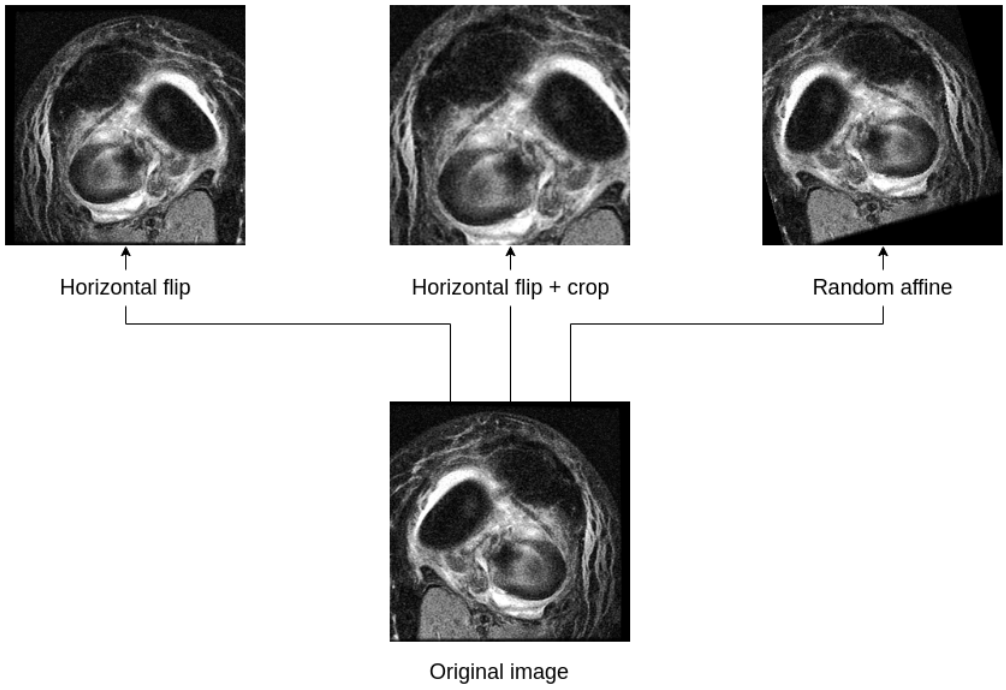


Figure 4. Random input image transforms applied to a middle axial image slice to augment the original MRNet imaging data.

2.3. Model Evaluation Strategy using Cohen's Kappa

When building a binary classification model, there are a lot of ways to interpret and evaluate the results based on the model predictions. Well-known evaluation metrics like Area Under the Receiver Operating Characteristics Curve (ROC-AUC) are based on the confusion matrix. However, not all metrics tend to perform well on imbalanced datasets.

Imbalanced data refers to those types of datasets where the target class has an uneven distribution of observations, i.e., one class label has a relatively high number of observations, whereas the other class has a low number of observations compared to the total amount.

The MRNet dataset, on which this paper is based, can be considered an imbalanced dataset. Having a total of 1370 knee MRI exams, the dataset contains as high as 1104, or 80.6% of abnormal exams. The amount of knee MRIs that are labeled as normal is only 19.4%.

Taking this into account, the evaluation results for the models described in our paper cannot rely solely on ROC-AUC or plain prediction accuracy metrics, as the imbalance, in the form of lacking normal knee examinations, is significant. Thus, for the purpose of further and more accurate evaluation of the machine learning models, we also introduce the calculation of the Cohen's Kappa score.

In his original paper, J. Cohen (1960) [14] describes his Kappa statistic as simply the proportion of chance-expected disagreements between two judges which do not occur, or alternatively, the proportion of agreement after a chance agreement is removed from consideration.

This statistic is frequently used to test interrater reliability. The importance of rater reliability lies in the fact that it represents the extent to which the data collected in the study are correct representations of the variables measured [15].

In Cohen's terms, we are expected to have two independent judges, or classifiers, with the categories under classification being independent, mutually exclusive, and exhaustive. To achieve these conditions, we interpret the first judge as our classification model for a given diagnosis - ACL tear, meniscus tear, or general abnormality. The second judge is the set of validation labels. Although in Cohen's terminology, there are no "correct" or

"incorrect" judgments to be made, in our case, the predictions made by the second judge are considered "true".

Cohen's kappa value can be calculated as follows:

$$K = \frac{p_0 - p_e}{1 - p_e} \tag{1}$$

where p_0 is the observed accuracy, or the proportion of predictions, where the judges have agreed, p_1 is the rate of predictions, where agreement is expected by chance, or the expected accuraccy.

Like many other evaluation metrics for classification models, Cohen's kappa calculation relies on the confusion matrix. In contrast to calculating overall accuracy, Cohen's kappa takes imbalance in class distribution into account for the validation dataset.

Having a confusion matrix for a single classification task (Abnormal, ACL tear, Meniscus tear) as in Table 1, where TP = True Positives, FP = False Positives, FN = False Negatives, TN = True Negatives, and taking N as a total number of observations, we can calculate required values for the Kappa statistic in our case with the following formulas:

$$p_0 = \frac{TP + TN}{N} \tag{2}$$

$$p_e = \frac{R_{cond} + R_{norm}}{N} \tag{3}$$

$$R_{cond} = \frac{(TP + FP) \cdot (TP + FN)}{N} \tag{4}$$

$$R_{norm} = \frac{(TN + FP) \cdot (TN + FN)}{N} \tag{5}$$

The observed accuracy p_0 is simply the number of instances that were classified correctly throughout the confusion matrix. That represents the level of agreement between the true data and the predicted data. R_{cond} and R_{norm} can be described as rates for a given label, compared to the total number of observations. This is the part that defines the imbalance of the validation dataset.

Table 1. Confusion matrix for a MRNet-based classification model.

	Predicted Pathologic Conditions	**Predicted Normal Cases**
True pathologic conditions	*TP*	*FP*
True normal cases	*FN*	*TN*

Although the interpretation of the kappa score may be more difficult than for traditional evaluation metrics, as it is ranged in the interval $[-1, 1]$, we believe it has the potential to unveil better classification models by taking out the random guessing factor.

2.4. Resulting Architecture and Training Process

In this work, a few classification models are built and trained on the original MRNet architecture, with the usage of more contemporary backbone networks as feature extractors for a single MRNet block, including VGG11, VGG16, Resnet, and Efficientnet. The original AlexNet version of the MRNet architecture is trained as well to compare the newer model evaluation results.

Our evaluation strategy is to measure classification performance metrics, such as ROC-AUC, plain accuracy score, and F1 score. Additionally, as described in Section 2.3, the Cohen's Kappa score is evaluated for each model and each diagnosis type.

Each trained model block corresponds to the original MRNet architecture, defined by [11]. The block operates on a single MRI plane and learns to extract relevant features from this plane, in the process of training.

Throughout all of the models, presented in this paper, the only varying layers are the feature extraction layer, and different classifier input dimensions, since not all of the backbones extract a feature vector of the same dimensions from the input image sequence. The resulting architecture of model blocks used in this paper is presented in Figure 5.

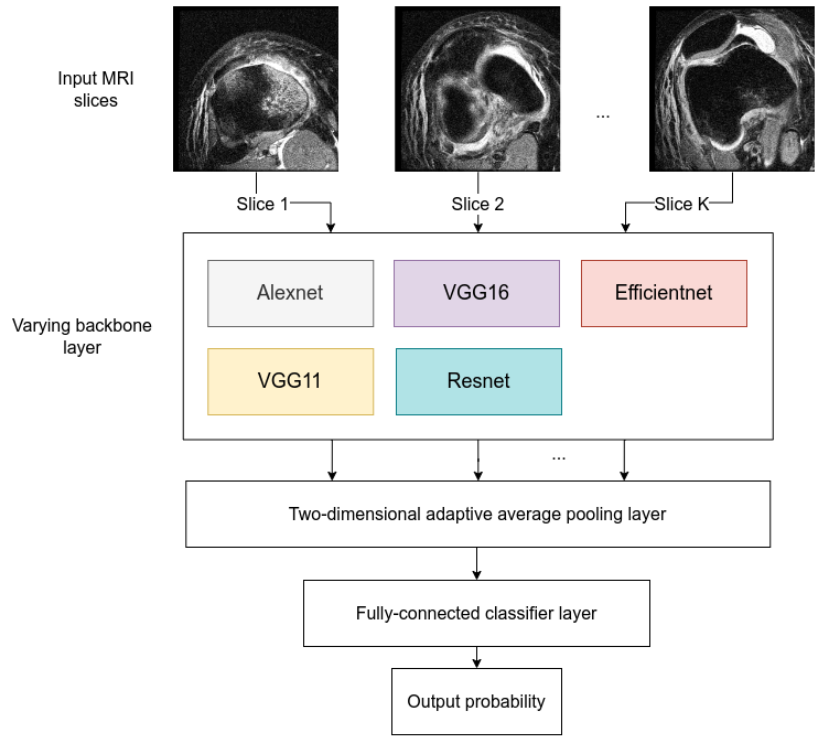


Figure 5. Adapted architecture of a single MRNet block. The input is processed by varying CV architectures (backbones), and adaptive average pooling is applied to the backbone output for each MRI slice. Afterwards, a fully-connected classifier layer returns the probability for a particular diagnosis.

Each model contains three training blocks, which were trained exclusively on axial, coronal, and sagittal MRI slices respectively. The Logistic regression (LR) approach is used to combine the outputs of three separate per-plane MRNet blocks. This provides the final layer of aggregated predictions as a result of consensus between decisions made by each of the per-plane blocks. The resulting macro-architecture involving LR models is the same as in Figure 2. Only individual block changes were applied in this paper.

Input image transformations were applied to prevent overfitting, exactly as described in Section 2.2.

We started by training each of the baseline MRNet blocks, on each of the MRI planes. The training process was conducted on the Google Cloud Platform, specifically using the Vertex AI service. A total of 45 models were trained in the cloud (3 training jobs for each diagnosis type, multiplied by 3 MRI planes, multiplied by the number of backbones(5)). Each training job was accelerated using 2×NVIDIA Tesla P100 GPU instances.

Each of the blocks was trained for 10 epochs, with model evaluation on the validation dataset after the end of each epoch. At the end of each epoch, a snapshot of the block was taken and uploaded to Google Cloud Storage. After the training process, only the blocks with the best evaluation scores were selected.

The resulting 45 blocks were then gathered from Cloud Storage and grouped by backbone type. A logistic regression model was then trained over the per-plane models, to weigh each one's classification decisions in correspondence to the true validation label data.

LR training was done on a local machine, using NVIDIA GeForce GTX 1080 Ti to speed up the process.

Afterward, the combined model predictions were gathered on the validation dataset. Corresponding model output was saved to CSV files, to be used in model evaluation. The evaluation metrics and plots were built on top of this prediction data.

3. Results

3.1. Overall Model Accuracy

The first obvious and significant evaluation metric for this classification task is the overall model accuracy. Table 2 shows the accuracy comparison between the models for each diagnosis, as well as the average accuracy across the three diagnoses.

Since the classification model returns probabilities instead of actual positive or negative labels, the model predictions were transformed with a probability threshold of 0.5. All probabilities greater than or equal to 0.5 were considered a "positive" label.

Table 2. Accuracy comparison per model, per diagnosis (threshold = 0.5). Maximum values are highlighted in bold.

Diagnosis	Alexnet	VGG11	VGG16	Resnet	Efficientnet
abnormal	0.825	**0.858**	0.842	**0.858**	0.850
acl	0.683	0.792	**0.808**	0.583	0.550
meniscus	0.70	0.733	**0.750**	0.608	0.583
average	0.738	0.752	**0.799**	0.683	0.661

3.2. F1-Score Evaluation

Another significant evaluation metric is the F1 score because it fits well with the actual imbalance of the dataset. It balances precision and recall for the positive class, which is dominant in this dataset.

Table 3 shows F1-score, calculated for each trained model and for each diagnosis. Output predictions were transformed in the same way as for overall model accuracy, using a threshold value of 0.5.

Table 3. F1-Score evaluation per model, per diagnosis (threshold = 0.5). Maximum values are highlighted in bold.

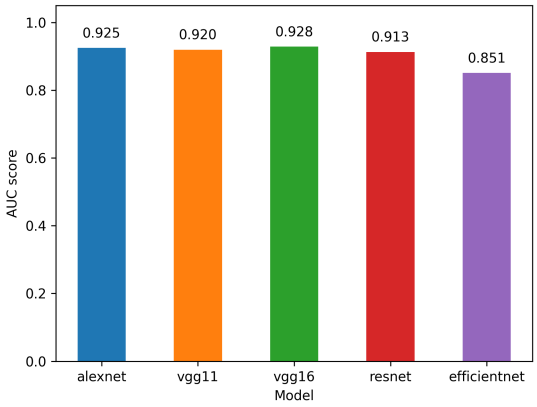
Diagnosis	Alexnet	VGG11	VGG16	Resnet	Efficientnet
abnormal	0.900	**0.917**	0.909	**0.917**	0.912
acl	0.486	0.713	**0.753**	0.167	0.000
meniscus	0.673	0.628	**0.717**	0.299	0.324
average	0.686	0.752	**0.793**	0.461	0.412

3.3. ROC-AUC Scores

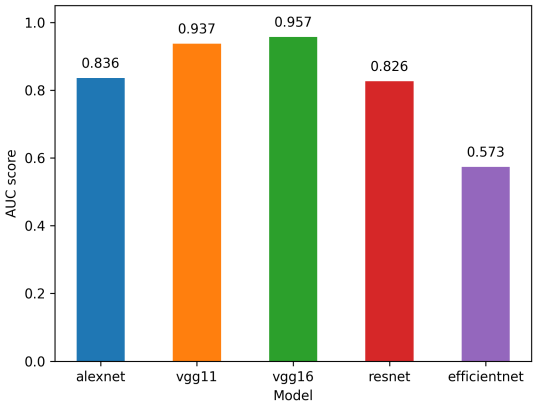
Next, the receiver operating characteristics are displayed. Figure 6 shows the ROC-AUC scores for each model. Each panel corresponds to a certain diagnosis type.

3.4. Cohen's Kappa Scores

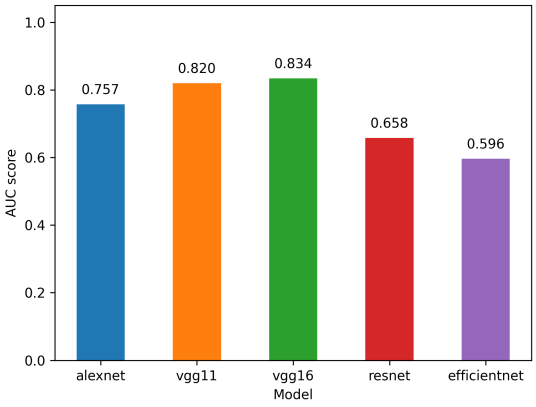
We believe that the most significant evaluation metric for this research is the Cohen's Kappa score. Due to the imbalanced nature of the MRNet dataset, it is likely that putting the model evaluation in an interrater reliability perspective could bring more useful information than traditional evaluation scores.



(**a**) General abnormality classification ROC-AUC stats



(**b**) ACL tear classification ROC-AUC stats



(**c**) Meniscus tear classification ROC-AUC stats

Figure 6. ROC-AUC stats for each model and for each diagnosis type: (**a**) general abnormality classification, (**b**) ACL tear classification, (**c**) meniscus tear classification.

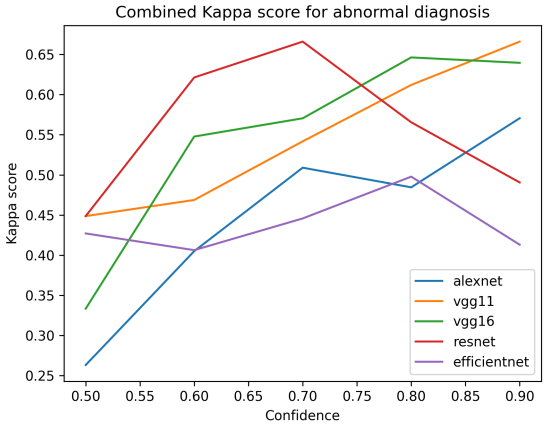
Yet again, to convert model predictions from probabilities to actual labels, different thresholds have been used, ranging from 0.5 to 0.9. This could better illustrate the overall behavior of certain models, not fixating on a single value for the threshold.

Figure 7 displays the graph of Cohen's Kappa values for certain thresholds, or confidence values.

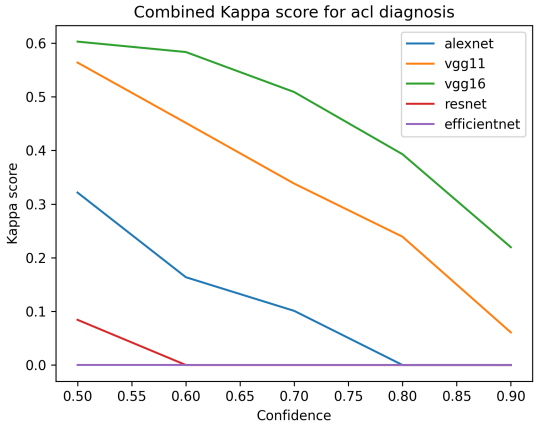
3.5. Single Block Performance for Each Backbone Type

One more very significant aspect of this model evaluation is the per-plane performance of each backbone listed in this paper. These results are displayed in Table 4, and they show the raw ROC-AUC classification score of each MRNet block, trained on specific backbones. The Logistic Regression layer is not involved in this evaluation, meaning this table shows the sole performance of a given backbone network on the three MRI slices for the three classification tasks.

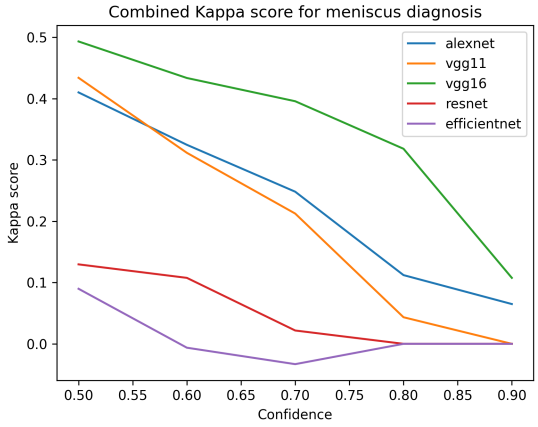
The highest evaluation score per plane and diagnosis type is highlighted in respective plane color in the table.



(**a**) General abnormality classification Cohen's Kappa scores



(**b**) ACL tear classification Cohen's Kappa scores



(**c**) Meniscus tear classification Cohen's Kappa scores

Figure 7. Cohen's Kappa scores per model and for each diagnosis type: (**a**) general abnormality classification, (**b**) ACL tear classification, (**c**) meniscus tear classification.

Table 4. ROC-AUC evaluation of each backbone per MRI plane. Highest evaluation scores per plane and for given diagnosis are highlighted with respective plane color.

Backbone	Plane	General Abnormality	ACL	Meniscus Tear
Alexnet	axial	0.845	0.738	**0.813**
	coronal	0.788	0.616	0.633
	sagittal	**0.935**	0.834	0.707
VGG11	axial	0.892	0.758	0.811
	coronal	0.732	0.926	0.762
	sagittal	0.917	**0.903**	0.753
VGG16	axial	**0.919**	**0.829**	0.759
	coronal	**0.840**	**0.950**	**0.808**
	sagittal	0.909	0.883	**0.801**
Resnet	axial	0.848	0.763	0.618
	coronal	0.477	0.462	0.659
	sagittal	0.912	0.700	0.632
Efficientnet	axial	0.845	0.586	0.642
	coronal	0.667	0.593	0.535
	sagittal	0.723	0.631	0.591

4. Discussion

As seen from the model evaluations, there is a steady overall model performance increase from using Alexnet to VGG11 and VGG16 backbones for classification. This result implies that the sole replacement of the Alexnet backbone with VGG counterparts is already beneficial in terms of classification performance.

Another fact that stands out is the rather underperformance of the newest CV architectures. Both ResNet and Efficientnet are the most contemporary models from the whole set on which the evaluations were conducted. They tend to outperform both Alexnet and VGG models in the results of the ImageNet competition 2021 [16]. However, for the purpose of knee MRI classification they did not seem to utilize their full potential. This is likely related to image pre-processing steps, and perhaps, this issue could be fixed by using more robust image augmentation methods.

One more significant result is the non-uniform nature of backbone network performance on different MRI planes. Instead of sticking to a single backbone network type, it is advisable to use an LR ensemble of different ones: VGG16 on a coronal plane for all classification tasks; on an axial plane for abnormality and ACL tear detection; Alexnet on a sagittal plane for abnormality detection, and axial plane for meniscal tear detection; VGG11 on a sagittal plane for ACL tear detection.

The results also show that model evaluation using the interrater reliability approach, by measuring the Cohen's Kappa metric, could show different results than traditional model evaluation techniques. For instance, in abnormality detection, using ROC-AUC, the Alexnet model has almost the same value as VGG16, but the Cohen's Kappa graph shows it mostly as way below every model except Efficientnet. Also, these metrics help to bring out the true issues of Efficientnet in this application - for meniscus classification, the Efficientnet model has gone negative in threshold value 0.7, which can be interpreted as the model making wrong decisions, unaffected by randomness.

The possibilities of future research might include three main directions. Firstly, there is the possibility to investigate more complex image pre-processing steps. It is promising to investigate the application of the BlurPool downsampling technique listed in Tsai et al. (2020), to modern backbone networks, and use this for reducing the number of training parameters, without greatly reducing the overall model performance.

The second direction is to use optimized versions of more advanced computer vision models for feature extraction. There are still models like Inception-v4, and Inception-Resnet, which could be used as backbones. However, they are considerably heavier in terms of training resources, and thus, we were not yet able to apply these newest models to this classification task.

Lastly, but most importantly, one may experiment with using per-plane MRNet blocks with different backbones, and combine them using Logistic Regression. In our opinion, one of the most significant results obtained in this paper is the fact that not all computer vision models behave uniformly on different MRI planes. Thus, it is advisable to use the best possible backbone architecture for each MRI plane, instead of sticking to a single architecture for all planes. Choosing this option should increase the overall performance of the LR-combined classification model.

Author Contributions: Conceptualization, N.S.; methodology, N.S. and P.P.; software, P.P.; validation, N.S. and P.P.; formal analysis, N.S.; investigation, P.P.; resources, P.P.; data curation, N.S.; writing—original draft preparation, P.P.; writing—review and editing, N.S.; visualization, P.P.; supervision, N.S.; project administration, N.S. and P.P.; funding acquisition, N.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Research Foundation of Ukraine # 2021.01/0103.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of Lviv Polytechnic National University (protocol code 149-1-10, approved on 19 March 2020).

Informed Consent Statement: Patient consent for this research was waived due to the usage of a public anonymized open-source dataset. We hereby state that no patient information can be identified from the data used in this research.

Data Availability Statement: The results obtained in this paper rely solely on the publically available MRNet dataset, published by Stanford ML Group—https://stanfordmlgroup.github.io/competitions/mrnet/ (accessd on 17 June 2022).

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

References

1. Nacey, N.C.; Geeslin, M.G.; Miller, G.W.; Pierce, J.L. Magnetic resonance imaging of the knee: An overview and update of conventional and state of the art imaging. *J. Magn. Reson. Imaging* **2017**, *45*, 1257–1275. [CrossRef]
2. IHS Markit Ltd (Prepared for the AAMC). The Complexities of Physician Supply and Demand: Projections from 2019 to 2034 AAMC, Washington, DC, USA, June 2021. Available online: https://www.aamc.org/media/54681/download (accessed on 17 May 2022).
3. Gore, J.C. Artificial intelligence in medical imaging. *J. Magn. Reson. Imaging* **2020**, *68*, A1–A4. [CrossRef] [PubMed]
4. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
5. Tsai, C.; Kiryati, N.; Konen, E.; Eshed, I.; Mayer, A. Knee Injury Detection using MRI with Efficiently-Layered Network (ELNet). In Proceedings of the Third Conference on Medical Imaging with Deep Learning, Montreal, QC, Canada, 6–8 July 2020; Volume 121, pp. 784–794.
6. Krizhevsky, A.; Sutskever, I.; Hinton, E.G. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems 25 (NIPS 2012), Lake Tahoe, NV, USA, 3–8 December 2012. Available online: https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html (accessed on 17 June 2022).
7. 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), Valencia, Spain, 19–22 October 2020.
8. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
9. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv* **2019**, arXiv:1905.11946.
10. Štajduhar, I.; Mamula, M.; Miletić, D.; Unal, G. Semi-automated detection of anterior cruciate ligament injury from MRI. *Comput. Methods Programs Biomed.* **2017**, *140*, 151–164. [CrossRef] [PubMed]
11. Bien, N.; Rajpurkar, P.; Ball, R.L.; Irvin, J.; Park, A.; Jones, E.; Bereket, M.; Patel, B.N.; Yeom, K.W.; Shpanskaya, K.; et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLoS Med.* **2018**, *15*, e1002699. [CrossRef] [PubMed]
12. Shorten, C.; Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 1–48. [CrossRef]
13. Wang, J.; Perez, L. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. *arXiv* **2017**, arXiv:1712.04621.
14. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
15. McHugh, M.L. Interrater reliability: The kappa statistic. *Biochem. Med.* **2012**, *22*, 276–282. [CrossRef]
16. ImageNet Competition Leaderboard. 2021. Available online: https://paperswithcode.com/sota/image-classification-on-imagenet (accessed on 20 May 2022).