TEL AVIV UNIVERSITY אוניברסיטת תל-אביב

THE IBY AND ALADAR FLEISCHMAN
FACULTY OF ENGINEERING

הפקולטה להנדסה
על שם איבי ואלדר פליישמן

Automated Knee Injury Detection and Localization using 3D MRI Images

A thesis
submitted in partial fulfilment
of the requirements for the Degree
of
Master of Science

by

Chen-Han Tsai

Tel Aviv University

August 2020

TEL AVIV UNIVERSITY אוניברסיטת תל-אביב

THE IBY AND ALADAR FLEISCHMAN
FACULTY OF ENGINEERING

הפקולטה להנדסה
על שם איבי ואלדר פליישמן

Automated Knee Injury Detection and Localization using 3D MRI Images

A thesis
submitted in partial fulfilment
of the requirements for the Degree
of
Master of Science

by

Chen-Han Tsai

This research was carried out at Tel Aviv University
in the School of Electrical Engineering
Faculty of Engineering
under the supervision of Prof. Nahum Kiryati
and Dr. Arnaldo Mayer

August 2020

This research was carried out under joint supervision of Prof. Nahum Kiryati and Dr. Arnaldo Mayer.

The main results presented in this thesis have been published in the Medical Imaging with Deep Learning (MIDL) 2020 Conference. The paper is attached to this thesis as Appendix A.

C.-H. Tsai, N. Kiryati, E. Konen, I. Eshed and A. Mayer, "Knee Injury Detection using MRI with Efficiently-Layered Network (ELNet)". Medical Imaging with Deep Learning (MIDL), Montreal, Quebec, Canada, 6-8 July 2020, Proceedings of Machine Learning Research (PMLR), Volume 121.

Additional work carried out during my M.Sc. research period, *beyond the scope of this thesis*, has been accepted to the MICCAI 2020 LABELS Workshop. The paper is attached to this thesis as Appendix B.

C.-H. Tsai, N. Kiryati, E. Konen, M. Sklair-Levy and A. Mayer, "Labeling of Multilingual Breast MRI Reports", 5th MICCAI Workshop on Large-scale Annotation of Biomedical data and Expert Label Synthesis (MICCAI LABELS), Lima, Peru, October 2020.

Abstract

Magnetic Resonance Imaging (MRI) is a widely-accepted imaging technique for knee injury analysis. Its advantage of capturing the knee structure in three dimensions makes it the ideal tool for locating potential tears in the knee. In order to alleviate the ever growing workload faced by musculoskeletal (MSK) radiologists, automated diagnostics for patients' triage are becoming a real need, reducing delays in the reading of pathological cases. In this research, we explore the development of an automated knee diagnosis algorithm suitable for triage systems.

The focus of our work is on the design of a convolutional neural network (CNN) capable of classifying variable-sized 3D MRI images. We introduce the Efficiently-Layered Network (ELNet), a novel CNN architecture optimized for knee injury detection. ELNet features a lightweight model size along with several novelties such as multi-slice normalization and BlurPool, and these improved designs allow ELNet to be easily trained from scratch.

We evaluate ELNet on the two publicly available knee MRI datasets: the MRNet dataset and the KneeMRI dataset. In both datasets, ELNet obtained favorable performance in comparison with previous state-of-the-art (SOTA) MRNet for a wide range of knee injury detection tasks. Furthermore, we compared our model's ability in locating tears in the knee to that of an MSK radiologist, and in the majority of the cases, our model's prediction closely resembles that of the radiologist.

Featuring a lightweight architecture with robust performance, ELNet may serve as a promising backbone for knee injury triage solutions. In addition to detecting knee injuries, ELNet may be extended to additional pathologies involving diagnosis with 3D medical images.

Table of Contents

Acknowledgments

First and foremost, I would like to thank Professor Nahum Kiryati and Doctor Arnaldo Mayer for allowing me to take part in such a memorable research experience. Although there were many twists and turns, highs and lows throughout this research journey, your dedication and support are what kept me motivated, and I truly appreciate your advice and encouragements throughout the past two years.

I would like to thank the Division of Diagnostics Imaging and in particular, the Computational Imaging Lab (CILAB) at the Sheba Medical Center for providing me with the resources throughout this entire research. Most of the research was performed at this lab, and I appreciate the support.

In addition, I would also like to thank Doctor Iris Eshed for providing us with clinical feedbacks throughout this research. Thank you for taking the time to ensure the quality of our research.

Lastly, I would like to thank my family for their constant support throughout all these years. Thank you for your constant encouragement, and thank you for your prayers. I miss you and I love you all very much.

List of Figures

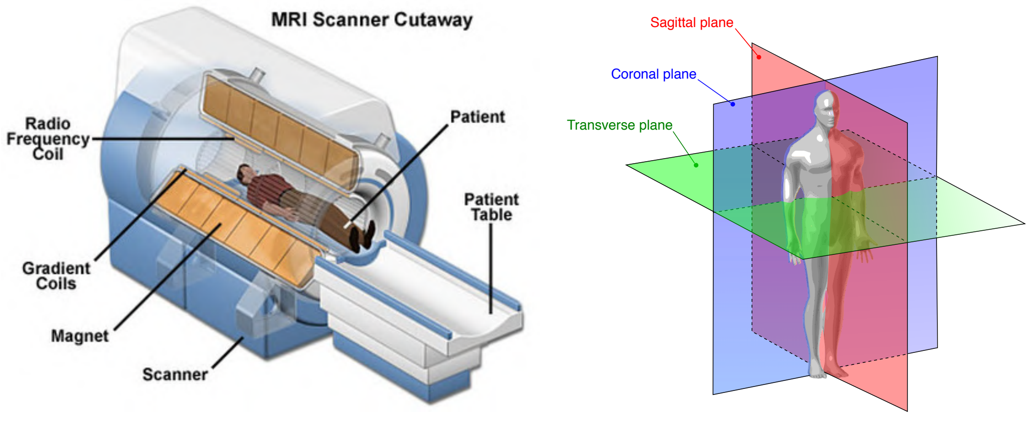
List of Tables

1 Introduction

Magnetic Resonance Imaging (MRI) is a medical imaging technique that has long been considered the most robust knee examination tool. Its widespread use is partly due to its capability to capture detailed structures in the knee while remaining a non-invasive procedure [6, 4]. Given its ability to reconstruct the knee in three dimensions, MRI has become the tool-of-choice for musculoskeletal (MSK) radiologists in an extensive range of examinations such as knee osteoarthritis and internal derangement of the knee. [8, 1]. Considering the increasing workload faced by MSK radiologists, automated tools for patients' triage are needed to reduce delays caused by the manual assessment of each case. In this research, we focus on developing an automated knee diagnosis algorithm suitable for triage systems.

1.1 Background

The use of MRI has long been a well established practice in the assessment of knee conditions. When patients undergo a knee MRI examination, they are placed on an examination table inside an MRI scanner, and a knee coil is situated around the knee (see Figure 1.1 a). A static magnet in the MRI scanner generates a strong magnetic field B_0 (ranging between 1.5-3 Teslas) that forces the protons in the body (specifically the knee) to precess along the longitudinal direction parallel (or anti-parallel) to B_0. The gradient coils then generate a secondary magnetic field B_1 (an RF pulse) to alter the precession of the protons caused by B_0. This RF pulse causes the net longitudinal magnetization to decrease (as the protons begin precessing at the same phase), and the transverse magnetization increases as a result.



(a) (b)

Figure 1.1: (a) A patient undergoing an MRI exam (b) The coronal, axial, and sagittal imaging planes. ***(Figures taken from atmph.org and med.libretexts.org)***

The RF pulse is applied for a short duration before it is turned off, and the decrease in transverse magnetization is detected as a signal by the RF coils situated on the transverse plane. The signal is then sample and stored in the K-space, and the Fourier Transform is applied to obtain the reconstructed MRI images of knee. Typical knee MRI's are captured from the sagittal, axial, and coronal image planes (see Figure 1.1 b), and such reconstruction allows the radiologist to examine different types of knee injuries from multiple perspectives.

Most knee injuries occur either in the ligaments or the tendons, and are often observed in individuals taking part in sports activities. In a study by Nicolini *et al.*, the two most common knee injuries amongst such individuals are tears in the anterior cruciate liagment (ACL) and tears in the meniscus (see Figure 1.2) [16].

The ACL is one of the ligaments connecting the femur (thighbone) and the tibia (shinbone), and most ACL tears occur during sudden movements of the knee in changing directions. The meniscus is a cartilage located between the femur and the tibia, and a common cause for meniscus tear is often related to rapid twisting motion of the knee. Other forms of knee injuries include cartilage lesions and osteoarthritis, and in this work, we aim to address the diagnosis of such knee injuries.

In a clinical setting, one of the many roles of an MSK radiologist is to evaluate

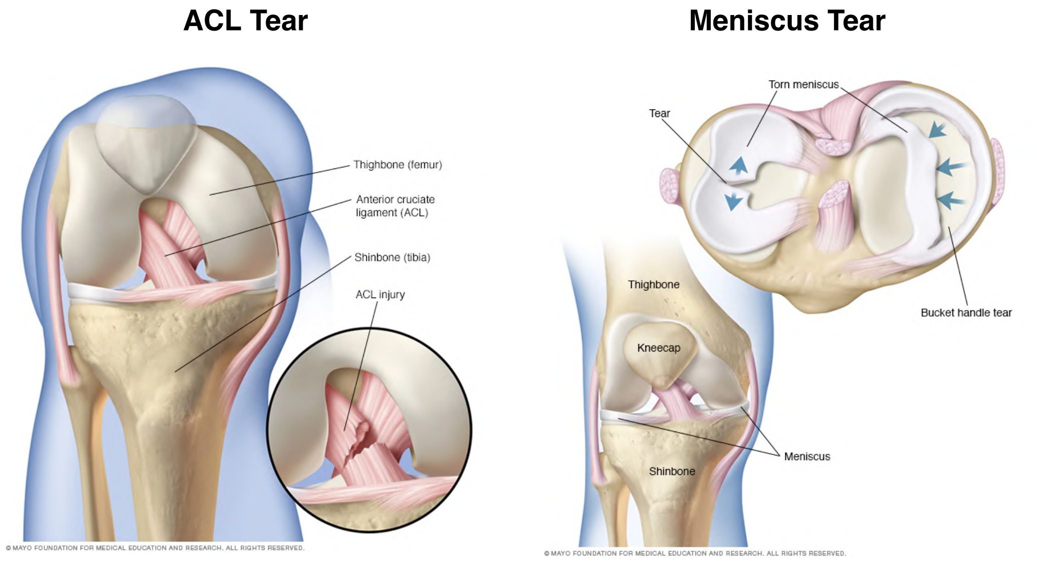

ACL Tear
Meniscus Tear
Thighbone (femur)
Anterior cruciate ligament (ACL)
Shinbone (tibia)
ACL injury
© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH. ALL RIGHTS RESERVED.
Torn meniscus
Tear
Thighbone
Kneecap
Meniscus
Shinbone
Bucket handle tear
© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH. ALL RIGHTS RESERVED.

Figure 1.2: Detailed visualization of an ACL tear and a meniscus tear. ***(Figures taken from mayoclinic.org)***

the condition of a patient's knee from the acquired MRI scan. However, one major concern that radiologists face on a daily basis is the growing number of patients awaiting evaluation each day. Since training a radiologist may take anywhere between 8-13 years, this lengthy process in addition to the increase in cases, strains the efficiency of assessing knee exams. An ordinary knee MRI may take anywhere between 5 days to 2 weeks before being examined, and this protracted delay can be significantly improved by the use of a triage system (see Figure 1.3).

The purpose of a triage system is to automatically prioritize cases based on the level of severity detected in the given acquisition. Specifically, each newly acquired knee MRI scan will be inspected by a knee diagnosis algorithm, and a flag will be raised for cases with a suspected injury. These flagged cases will then be placed at the top of the queue so that the radiologists can assess these scans first. This prioritization benefits both the patients and the radiologists, as the patients receive expedited treatment of the injured knee, and the overall efficiency of the examination workflow is improved.

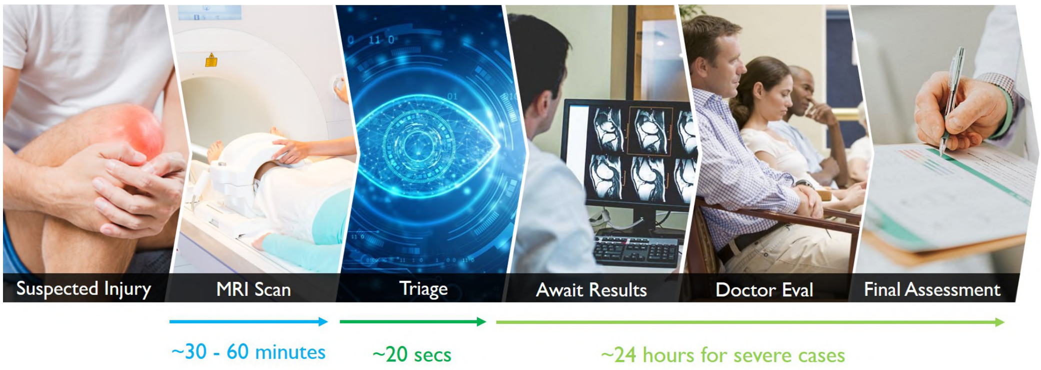


Figure 1.3: Integration of a triage system into the knee examination workflow

1.2 Prior Works

The core component of the envisioned triage solution is a fully-functional knee diagnosis algorithm. Several studies have suggested plausible techniques for this purpose.

In the work by Štajduhar *et al*, a semi-supervised approach was proposed to classify whether a tear was present in the ACL of a given knee [20]. A 2D patch of the ACL image is first manually extracted, then a histogram of oriented gradients (HOG) descriptor is generated representing the patch. The HOG descriptor is then fed into a support vector machine (SVM) to classify whether the ACL is torn based on the selected patch. Although this method demonstrates strong classification performance, the manual extraction required to select the patch is a tedious process in itself, rendering this method impractical for the purpose of fully-automated diagnosis.

In a study by Liu *et al.*, a cartilage lesion detection algorithm was proposed [15]. The algorithm consists of two separate CNN networks; one in charge of segmentation, and another for classification. The purpose of the segmentation network is to label the areas in the scan resembling the cartilage. Random patches are then sampled from the predicted cartilage regions, and the classification network classifies these patches as to whether they contain lesions or not. Although this approach is fully automated, the amount of labeling required to train the segmentation network is extremely expensive and far-fetched.

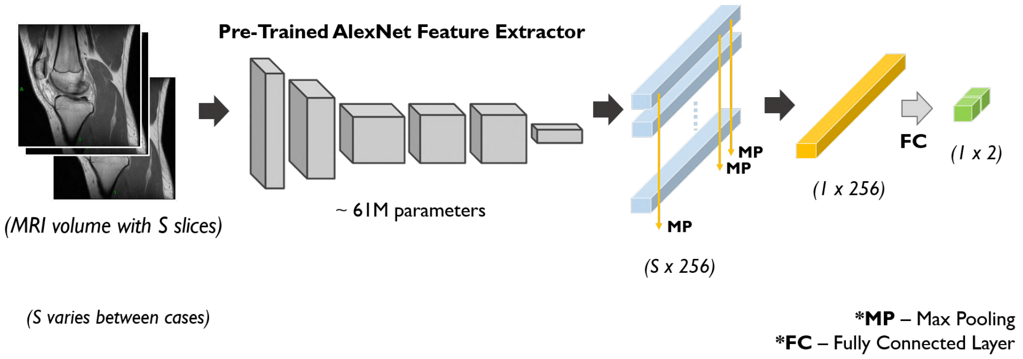

Pre-Trained AlexNet Feature Extractor
(MRI volume with S slices)
~ 61M parameters
MP
MP
MP
(S x 256)
(1 x 256)
FC
(1 x 2)
(S varies between cases)
*MP – Max Pooling
*FC – Fully Connected Layer

Figure 1.4: Design of MRNet-single architecture

The former state-of-the-art (SOTA) fully automated knee diagnosis algorithm is the MRNet proposed by Bien *et al.* [3]. The MRNet is a combination of three smaller CNN networks (referred to individually as "MRNet-single"), and each MRNet-single performs classification on scans acquired from a single plane orientation (see Figure 1.4).

The MRNet-single is designed to first extract a set of feature vectors by passing each slice of the scan through a pre-trained AlexNet [13]. These vectors are then stacked to form a 2D array, and max-pooling is applied to obtain a single vector representing the entire volume. This vector is then fed into a fully connected layer followed by softmax activation to obtain the predicted class probabilities.

The three MRNet-singles are trained separately to perform classification from scans acquired in the coronal, axial, and sagittal plane orientations. Their class predictions are then aggregated by logistic regression to compute the final class probabilities.

Although the MRNet architecture demonstrated promising performance in detecting a wide range of knee injuries, several concerns still persist. First, the AlexNet architecture used in feature extraction is an outdated design that does not contain improved design choices such as normalization layers or skip connections. The addition of such improvements will most likely yield a better optimization process as we demonstrate in our work.

The second concern relates to the model size of MRNet. In total, each MRNet contains roughly 183M trainable parameters, and this is a large model size that cannot be neglected. When saved to disk, each MRNet takes around 750 MB of storage, and in a clinical setting, such a large file size would imply longer inference time and prolonged over-the-air (OTA) update during a system upgrade.

The final concern facing the MRNet design is the reliance on three simultaneous image plane orientations for performing classification. Since most radiologists assess cases based on just a single plane orientation (occasionally two orientations are used), this prompts the question of whether there exists a better algorithm that can match the performance of MRNet while relying on images captured from a single plane orientation.

In this research, we present the Efficiently Layered Network (ELNet), a novel CNN architecture optimized for diagnosing a wide range of knee injuries using MRI. We propose several novelties (detailed in Section 2) that address the shortcomings of past approaches, and the resulting model is one that matches the previous SOTA MRNet with an almost 1000× smaller model size.

ELNet relies on images captured in just a single plane orientation, and its performance matches the MRNet without relying on a pre-trained model. In addition to detecting a knee injury, our model is also capable of locating the tear within the knee. This is extremely helpful in a triage system as the radiologists may receive both our model's prediction of an injury as well as the relevant MRI slices indicating a tear. In this research, we demonstrate ELNet's effectiveness for diagnosing some of the most common knee injuries, and we hope our work may help accelerate the process for which knee MRI examinations are carried out.

2 Methods

In this section, we present the ELNet architecture in detail. The illustration of ELNet is given in Figure 2.3 and the details are listed in Table 2.1.

2.1 *Inspiration*

Convolutional Neural Networks (CNN's) are a family of deep neural networks that have been widely adopted as the method of choice in performing a wide range of visual recognition tasks. The earliest and the most well-known CNN architecture is AlexNet, introduced by Krizhevsky *et al.* in 2012 [13]. The feature extractor in AlexNet primarily consists of 5 convolutional layers followed by the ReLU activation, and the linear classifier is a combination of three fully connected layers. Max-Pooling was also applied after several convolutional layers, and it operates by selecting the maximum value from a sliding window moving across the obtained feature representation (see Figure 2.1). With a stride of 2, the down-sampled feature representation is half the spatial height and width of that of the original representation.

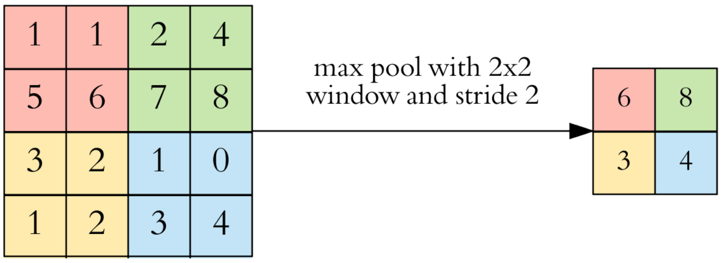


Figure 2.1: Example of Max-Pool Operation ***(Figure taken from computersciencewiki.org)***

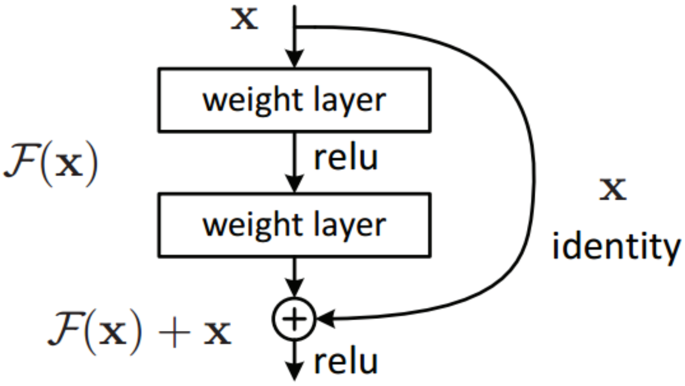


Figure 2.2: Skip Connection in ResNet ***(Figure taken from [9])***

The VGG architecture, proposed by Simonyan *et al.* in 2014, improved on the design of AlexNet by building deeper layers into the feature extractor [18]. In 2015, the ResNet architecture proposed by He *et al.* [9] enhanced the VGG architecture by the addition of skip connections (see Figure 2.2) and batch normalization [11]. Since then, skip connections have become a standard in CNN design given its robustness to vanishing gradients and a smoother loss landscape during optimization [14]. Similarly, batch normalization mitigates the effect of covariate shift which also reduces the issue of gradient saturation during training.

The concept of a *block* was also introduced in the ResNet feature extractor. It is defined as a series of

$$[\text{2D Convolution} \rightarrow \text{Batch Normalization} \rightarrow \text{ReLU activation}]$$

and a skip connection is added between the input and output of each *block*. These *blocks* serve as the core components of the ResNet, and in our model, we define a *Block* module in a similar manner.

2.2 Block Modules

The backbone of ELNet's design centers around *Block* modules (see Figure 2.4). Inspired by ResNet [9], we define a *Block* as a sequence of:

$$[\text{2D Convolution} \rightarrow \text{Multi-slice Normalization} \rightarrow \text{ReLU activation}]$$

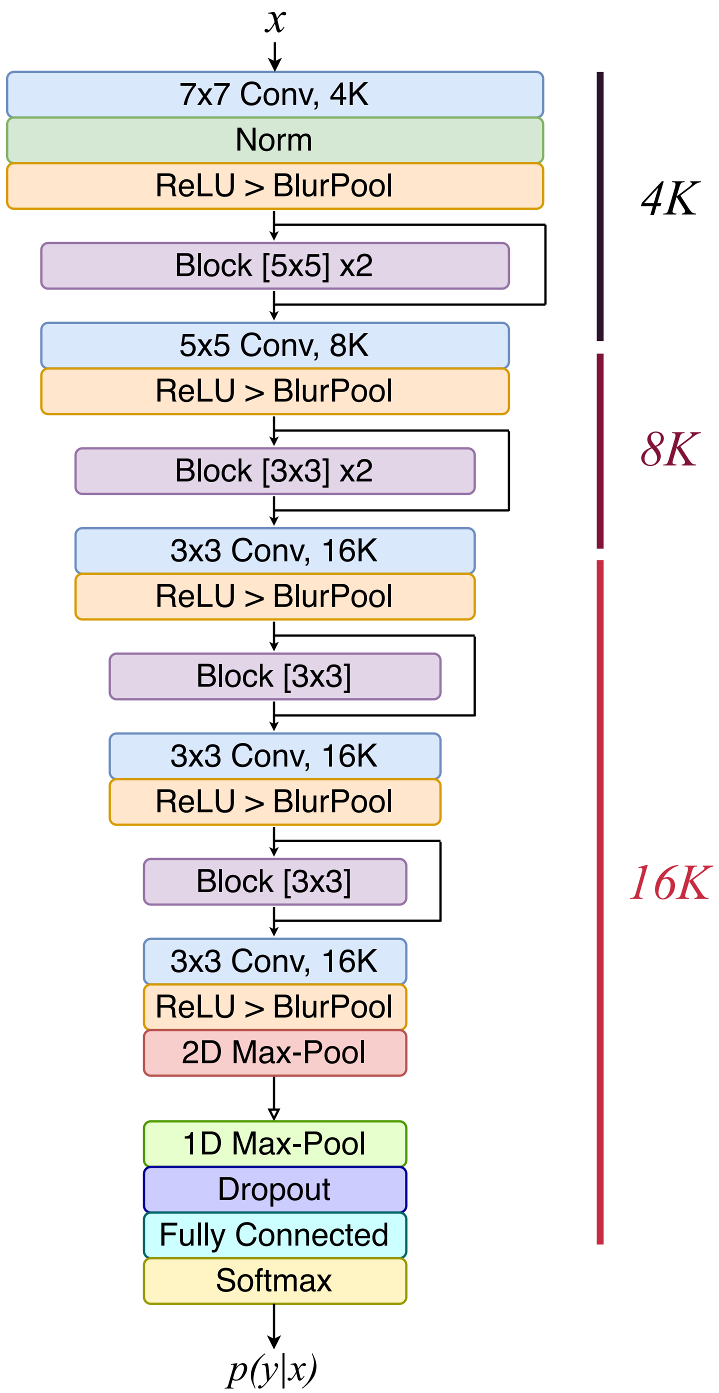


Figure 2.3: The ELNet Architecture ('K' is a hyperparameter that determines the channel dimensions, typically 'K=4')

Output Size	Layer Operation	Trainable Parameters
$S \times 4K \times 128 \times 128$	7×7 Conv, $4K$	$196K$
	Normalization	$4K$
$S \times 4K \times 62 \times 62$	ReLU → BlurPool	
	Block $[5 \times 5] \times 2$	$800K^2 + 16K$
$S \times 8K \times 62 \times 62$	5×5 Conv, $8K$	$800K^2$
$S \times 8K \times 29 \times 29$	ReLU → BlurPool	
	Block $[3 \times 3] \times 2$	$1152K^2 + 32K$
$S \times 16K \times 29 \times 29$	3×3 Conv, $16K$	$1152K^2$
$S \times 16K \times 13 \times 13$	ReLU → BlurPool	
	Block $[3 \times 3]$	$2304K^2 + 32K$
	3×3 Conv, $16K$	$2304K^2$
$S \times 16K \times 5 \times 5$	ReLU → BlurPool	
	Block $[3 \times 3]$	$2304K^2 + 32K$
	3×3 Conv, $16K$	$2304K^2$
$S \times 16K$	ReLU → BlurPool → 2D Max-Pool	
$16K$	1D Max-Pool → Dropout	
2	Fully Connected → Softmax	$32K + 2$
Total Trainable Parameters		$13120K^2 + 348K + 2$

Table 2.1: Details of the ELNet Architecture ('S' refers to the number of slices of an MRI scan)

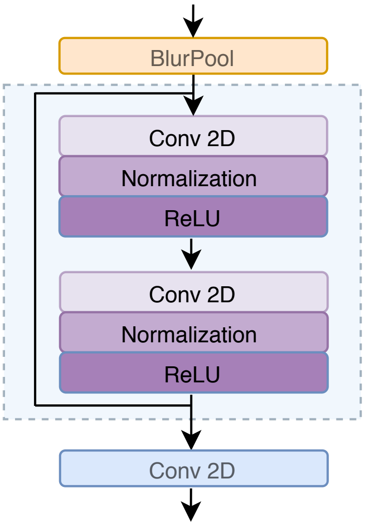


Figure 2.4: Block Module with two Repeats

Blocks are designed to allow for non-linearities in the network, and they may be repeated while ensuring equal input and output dimensions. A skip connection is added between the input and output, allowing better optimization of the network. The first two *Blocks* are repeated twice with $4K$ and $8K$ channels, and the remaining *Blocks* are fixed with $16K$ channels.

Each *Block* is followed by another 2D Convolution and ReLU activation, and they serve to increase channel dimension. The spatial height and width are reduced using a BlurPool layer. Eventually, in the final layer of the feature extractor, 2D max-pooling is applied to obtain a $16K$-dimensional feature vector for each MRI slice. Max-pooling is consecutively applied to obtain a single $16K$-dimension feature vector that combines feature information across slices. Dropout is performed before feeding into a fully-connected layer with two output logits, and the final probability $p(y|x)$ is computed by softmax [7].

In the following two subsections, we detail two innovative features of ELNet: the use of multi-slice normalization, and BlurPool.

2.3 Multi-Slice Normalization

We propose two possible variants of multi-slice normalization: a first one based on *layer normalization* [2], and a second one based on *contrast normalization* [22] (see Figure 2.5). Let's assume a feature representation $x^{(i)} \in \mathbb{R}^{S \times C \times H \times W}$ from some layer i in the network (usually a 2D-convolution), where S is the number of slices in the MRI sequence, C is the number of channels in the representation, and H, W are the spatial height and width of the representation. The network applies a normalization on x (omitting i for simplicity) by computing the appropriate mean and variance.

In the *layer normalization* variant, the mean μ_s and variance σ_s^2 are computed from x for each slice s ($1 \leq s \leq S$). In *contrast normalization*, the mean μ_{sc} and variance σ_{sc}^2 are computed for each slice s and also for each channel c ($1 \leq c \leq C$). Using the computed mean and variance, x is standardized into $\hat{x}$. An affine transform is applied to $\hat{x}$ to obtain the normalized output y. The normalization process is expressed by Equation 2.1 for *layer normalization* and Equation 2.2 for *contrast normalization* respectively:

$$\hat{x}_s = \frac{x_s - \mu_s}{\sqrt{\sigma_s^2 + \varepsilon}} \rightarrow y_s = \gamma \hat{x}_s + \beta \qquad \forall s \in \{1, \dots, S\} \tag{2.1}$$

$$\hat{x}_{sc} = \frac{x_{sc} - \mu_{sc}}{\sqrt{\sigma_{sc}^2 + \varepsilon}} \rightarrow y_{sc} = \gamma \hat{x}_{sc} + \beta \qquad \forall s \in \{1, \dots, S\},\ \forall c \in \{1, \dots, C\} \tag{2.2}$$

Parameters γ, β (C dimensional vectors) are learned independently for each normalization layer. Typically, $\gamma, \beta, \varepsilon$ are initialized to $\mathbf{1}$, $\mathbf{0}$, and 1e-8 respectively.

In comparison to batch normalization [11], multi-slice normalization offers the advantage of normalizing the feature representation of each slice independently. This is important in the case of 3D MRI images as information discrepancy among slice images may occur (e.g. only 1-2 MRI slices illustrate a tear, whereas the remaining slices do not highlight the tear). Multi-slice normalization handles this by normalizing the feature representations of each layer independently, and this offers ELNet the advantage of learning meaningful representations from the few

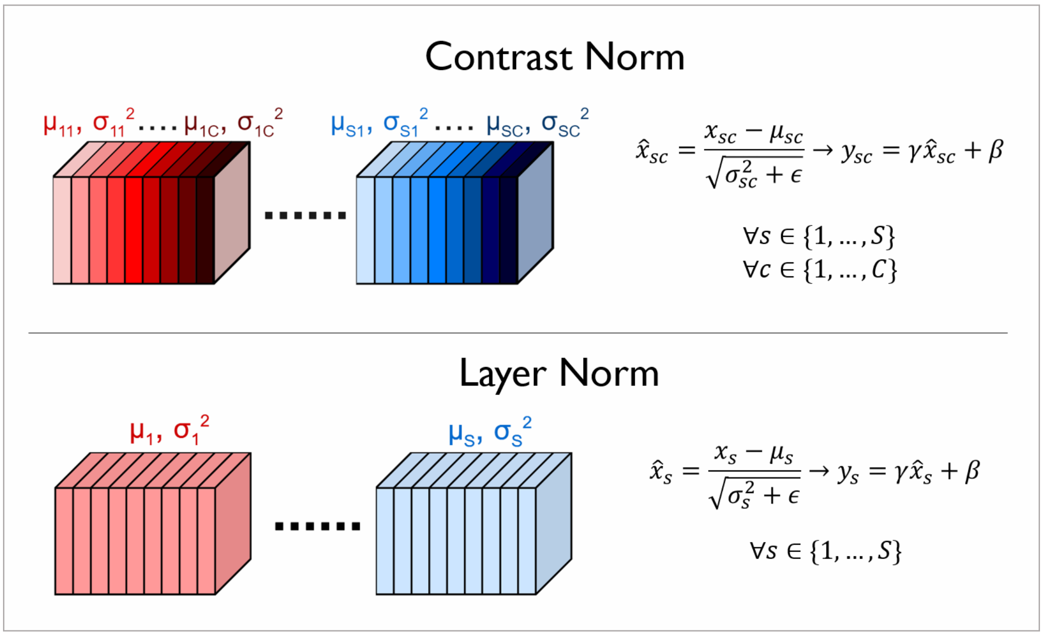

Contrast Norm
$\mu_{11}, \sigma_{11}^2 \dots \mu_{1C}, \sigma_{1C}^2$
$\mu_{S1}, \sigma_{S1}^2 \dots \mu_{SC}, \sigma_{SC}^2$
$\hat{x}_{sc} = \frac{x_{sc} - \mu_{sc}}{\sqrt{\sigma_{sc}^2 + \epsilon}} \rightarrow y_{sc} = \gamma \hat{x}_{sc} + \beta$
$\forall s \in \{1, \dots, S\}$
$\forall c \in \{1, \dots, C\}$
Layer Norm
μ_1, σ_1^2
μ_S, σ_S^2
$\hat{x}_s = \frac{x_s - \mu_s}{\sqrt{\sigma_s^2 + \epsilon}} \rightarrow y_s = \gamma \hat{x}_s + \beta$
$\forall s \in \{1, \dots, S\}$

Figure 2.5: Multi-Slice Normalization

slices that illustrate an injury. In Section 4.2, we compare the performance of ELNet's trained using multi-slice normalization to the performance of ELNet's trained with batch normalization, and an explanation is provided for why batch normalization is not the ideal normalization operation when processing 3D MRI images.

2.4 BlurPool

In the work of Zhang [23], a BlurPool operation was proposed to mitigate the shift-variance phenomenon observed in modern CNN architectures where max-pooling is often utilized. As demonstrated in the paper, the prediction probabilities of trained CNN's can fluctuate significantly when small shifts in the same input image are introduced. The author attributed this undesired phenomenon as a consequence of Max-Pooling, a down-sampling operation that has been widely adopted in the design of most CNN architectures.

The issue with Max-Pooling is that it ignores the problem of aliasing when down-sampling the input representation. From a signal processing perspective, an

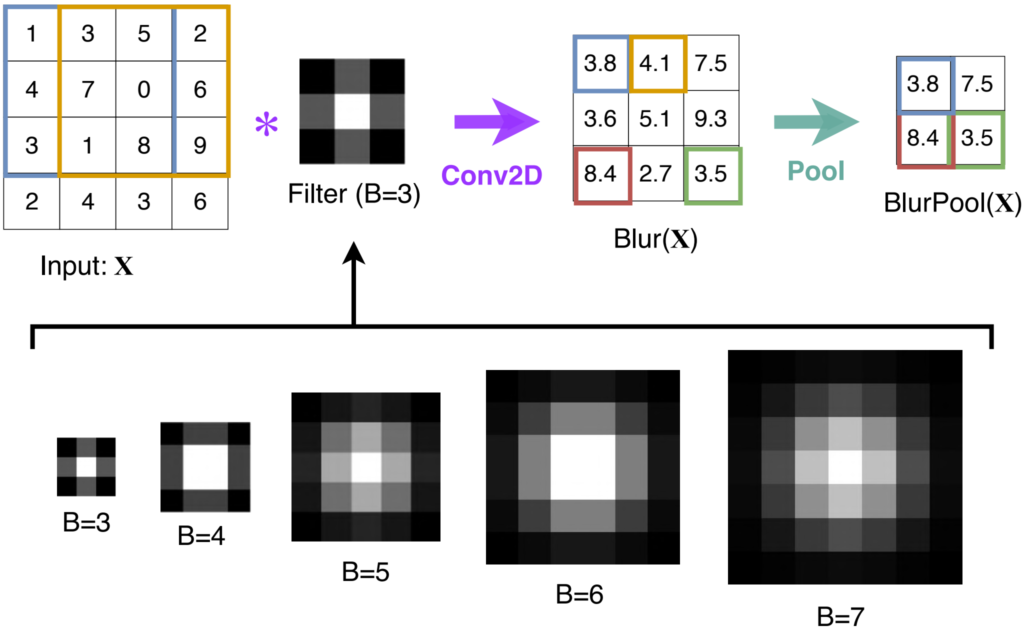


Figure 2.6: Example of a BlurPool Operation and BlurPool Kernels

anti-aliasing filter (i.e. a low-pass filter) should first be applied on the input signal to avoid aliasing during down-sampling. However, the Max-Pooling operation does not take aliasing into consideration, and the BlurPool operation proposed by *Zhang* alleviates the issue of aliasing.

BlurPool functions by first applying an anti-aliasing filter (binomial filter with kernel size B and stride 1) to the input representation, then strided pooling is applied to obtain the pooled feature map (see Figure 2.6). The resulting representation is therefore a pooled version of the blurred input representation. A more detailed analysis is available in the paper [23].

2.5 Training Pipeline

As suggested by Nyúl and Udupa [17], we perform histogram-based intensity standardization according to the training set statistics, thus enabling similar-valued pixels to be associated with the relevant tissue type. In addition, we perform randomized data augmentations to each series which includes translation, horizontal flip, scaling, and minor rotations up to ±10 degrees around the center of the volume. For volumes captured in the axial and coronal orientations, we apply an additional random rotation of a multiple of 90 degrees to the volume. Finally, all the images are resized to 256×256 before entering the network.

Aside from data augmentation, we implement oversampling to compensate for dataset imbalance. For each pathology, we select the minority class samples (allowing repeats) from our training set and apply augmentations on them until the number of minority class samples (along with their augmented copies) equals the number of majority samples.

We train ELNet using standard cross-entropy loss [7]. Optimization can be done using a simple grid-search over relevant hyperparameters such as learning rate, choice of multi-layer normalization, BlurPool kernel sizes, dropout rate, etc.

3 Experiments

In this section, we present the datasets that were used in our experiments. The metrics for evaluation are defined, and we specify the ELNet configurations in performing the knee injury detection objectives associated with the datasets. To better understand how multi-slice normalization and BlurPool contribute to improving ELNet's performance, we carried out several studies with different configurations. Finally, we created a test to evaluate ELNet's ability in locating a potential tear from a given knee MRI, and we compared ELNet's performance to that of an MSK radiologist.

3.1 Datasets

MRNet Dataset. The MRNet Dataset contains 1,370 knee MRI examinations that were carried out at the Stanford University Medical Center. Each case was labeled according to the presence/absence of an anterior cruciate ligament (ACL) tear, a meniscus tear, or other signs of abnormalities in the corresponding knee. Each exam was randomly assigned either to the training, validation, or test set [3]. It should be noted that each exam may contain multiple labels (e.g. an exam labeled positive for abnormality and ACL tear is also labeled for other forms of abnormality in addition to the ACL tear).

The dataset provided includes, for each case, corresponding axial, coronal and sagittal MRI acquisitions (see Figure 3.1). As reported by Bien *et al.*, a sagittal T2-weighted series, a coronal T1 weighted series, and an axial proton density weighted series were selected for this dataset. Each image is of size 256×256 and the number of slices ranges between 17- 61 (mean 31 and standard deviation 7.97). The MRNet Dataset is currently the largest public labeled knee

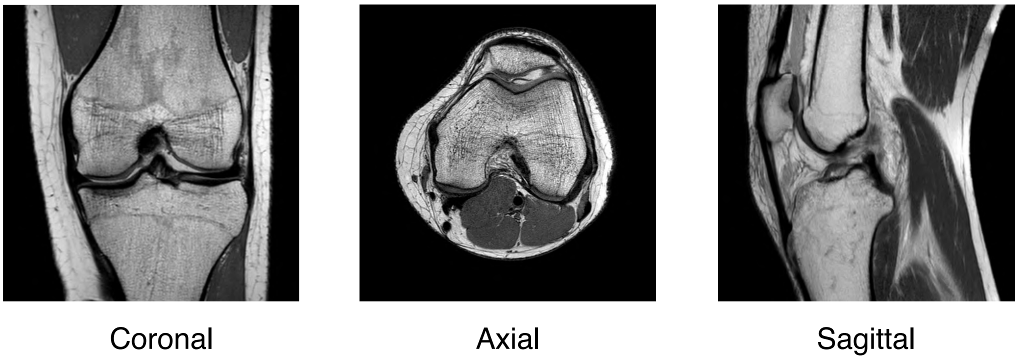


Figure 3.1: Example of knee MRI slices captured from three imaging plane orientations

MRI dataset.

KneeMRI. The KneeMRI dataset collected at the Clinical Hospital Centre Rijeka, Croatia by Štajduhar *et al.* consists of 917 exams labeled with ACL conditions in the corresponding knee. For each exam, the ligament condition was classified as either healthy (690 exams, 75.2%), partially injured (172 exams, 18.8%), or completely ruptured (55 exams, 6%). Each assessment corresponds to a T1-weighted sagittal MRI series, containing 320×320 or 290×300 images. The number of images in each series ranges between 21-45 (mean 31 and standard deviation 2.27). The dataset was divided into 10 strata with similar distributions, and we perform stratified sampling for evaluation.

3.2 Metrics

ELNet performs diagnosis on a given knee MRI by classifying whether it contains an injury (positive class) or not (negative class). The metrics we use to evaluate our model's performance are defined as follows:

$$\textbf{Accuracy (ACC)} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\textbf{Sensitivity} = \frac{TP}{TP+FN}$$

$$\textbf{Specificity} = \frac{TN}{TN+FP}$$

$$\textbf{F1} = \frac{TP}{TP+\frac{1}{2}(FP+FN)}$$

Here, TP, TN, FP and FN stand for True-Positive, True-Negative, False-Positive and False-Negative respectively.

The **Receiver Operating Characteristic (ROC)** is graphical illustration that plots a classifier's true-positive rate (TPR) against its false-positive rate (FPR) by varying the threshold of the classifier.

ROC Area Under the Curve (ROC-AUC) denotes the area under the ROC of a given classifier and is typically a value between 0.5-1.0. An ROC-AUC of 0.5 indicates that the model is performing random classification, whereas an ROC-AUC of 1.0 indicates perfect classification by the model.

$$\textbf{Matthew's Correlation Coefficient (MCC)} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

In the case of binary classification, the **MCC** is a more reliable metric that takes into account all four metrics (TP, TN, FP, FN) from the confusion matrix. Unlike previous metrics, the MCC value is independent of the dataset distribution it is computed on, and is generally a challenging success metric. The MCC ranges between -1 and 1. An MCC of 0 indicates random classification, whereas an MCC of 1 indicates perfect classification [5].

3.3 Setup for Knee Injury Detection Experiments

MRNet Dataset. In the MRNet dataset, we were provided with three imaging orientations per examination. For the three pathologies, we trained three separate ELNet's with $K = 4$, and the network weights were initialized uniformly by choosing the best random seed between 0-4 [10]. Based on experiments, we selected coronal images for detecting meniscus tears, and axial images for detecting ACL tears and abnormalities. Contrast normalization yielded the best results for detecting meniscus tears, and layer normalization for detecting ACL tears and abnormalities (notice the correspondence between the selected multi-slice normalization and imaging plane orientation). Each model was trained using Adam [12], with a learning rate between 1e-5 and 3e-5 for 200 epochs (batch size of 1), taking roughly 1.5 hours.

KneeMRI Dataset With the KneeMRI dataset, we perform 5-fold cross validation using eight out of the ten strata, and validation using the remaining two. Similar to the MRNet Dataset, we train an ELNet with K=2 using SGD+Momentum [21] for 200 epochs (batch size of 1) and the training time is roughly an hour for each fold.

By choosing K=2, and K=4 for the ELNet architectures, our trained model involves 53,178, and 211,314 trainable parameters respectively. In relation to AlexNet ($\sim$61M trainable parameters), ELNet (with K=4) contains 288× fewer parameters than AlexNet. In comparison with MRNet, ELNet with K=4 contains 866× less parameters, and ELNet with K=2 contains 1147× less parameters. Each trained model was saved using the standard PyTorch format. Model sizes are 850kB and 435kB for K=4 and K=2 respectively. Our experiments were perfomed on an NVIDIA GTX 1070 8GB GPU.

3.4 Ablation Setup

To better understand how our proposed novelties compare against typical CNN design choices, we setup an ablation test to analyze the effects of each individual ELNet module. Specifically, we compare ELNet's classification performance when multi-slice normalization and BlurPool are replaced with batch normalization [11] and max-pooling. The objectives are the three pathologies presented in the MRNet dataset. For each pathology, we train four ELNets, each with a different normalization and pooling configuration. We perform the same grid-search over relevant hyperparameters for each normalization and pooling configuration (for each pathology), and the best performing model is recorded.

3.5 Clinical Comparison Setup

To verify that ELNet is indeed performing diagnosis based on features resembling a tear in a given MRI scan, we setup an experiment to compare ELNet's prediction for where the tear resides to that of a radiologist. We randomly selected one of the five cross validation splits from the KneeMRI dataset, and evaluated the trained ELNet on the validation set of that split (so as to ensure that the model has never seen the validation set before). Samples from the validation set were randomly selected from both classes, and this resulted in 9 cases containing ACL tear and 7 cases without ACL tear. A board-certified MSK radiologist with 17 years of experience was asked to identify the most informative slice (slice containing the most area for which a tear resides) in a given MRI series. Furthermore, the radiologist had to identify the region containing the tear within the most informative slice that corresponded to an ACL injury.

An identical task was performed on the trained ELNet from the same validation set, and the Full-Gradient representation [19] was computed. FullGrad generates a heat-map that corresponds to parts of the input that most influence the output prediction. Conceptually, the generated heat-map should be "hotter" in areas indicating an injury and "cold" elsewhere. The generated heatmap was compared to the radiologist's determination.

4 Results

In this section, we present the experimental results obtained using the experimental setups described in Chapter 3. Comparison is made between ELNet and previous approaches.

4.1 Knee Injury Detection Results

MRNet Dataset. We evaluate ELNet's performance using the validation set provided by the MRNet dataset (since the test set is not publicly available), and we compare it with the MRNet model proposed and trained by Bien *et al.* Although they evaluated their models primarily using the ROC-AUC, we perform a more thorough analysis by considering additional metrics that are just as significant, such as Sensitivity and the Matthew Correlation Coefficient (MCC). The evaluation results are presented in Table 4.1 and the ROC for each detection objective (pathology type) is plotted in Figures 4.1, 4.2, and 4.3.

In order to evaluate whether the differences between ELNet and MRNet's performance on the MRNet are statistically significant, we performed a McNemar Test between the trained ELNet and MRNet on the three classification pathologies. We obtained a p-value of 0.009, 0.387, 0.99 for the classification of meniscus tear, general abnormalities, and ACL tear respectively. Thus, we may reject the null hypothesis that the two models' performances are equal for detecting meniscus tears, and we may not reject the null hypothesis in the case of detecting general abnormalities and ACL tears. When compared with the MRNet, ELNet's performance is on par with MRNet in detecting general abnormalities and ACL tear, and exceeds MRNet's performance in detecting meniscus tear.

Architecture	Pathology	Accuracy	Sensitivity	Specificity	ROC-AUC	MCC
MRNet	Meniscus Tear	0.735	0.827	0.662	0.826	0.489
	ACL Tear	0.9	0.907	0.894	0.956	0.769
	Abnormality	0.883	0.947	0.64	0.936	0.628
ELNet	Meniscus Tear	0.88	0.86	0.89	**0.904**	**0.745**
	ACL Tear	0.904	0.923	0.891	**0.960**	**0.815**
	Abnormality	0.917	0.968	0.72	**0.941**	**0.736**

Table 4.1: Evaluation Statistics between MRNet and ELNet on the MR-Net validation set (Threshold = 0.5)

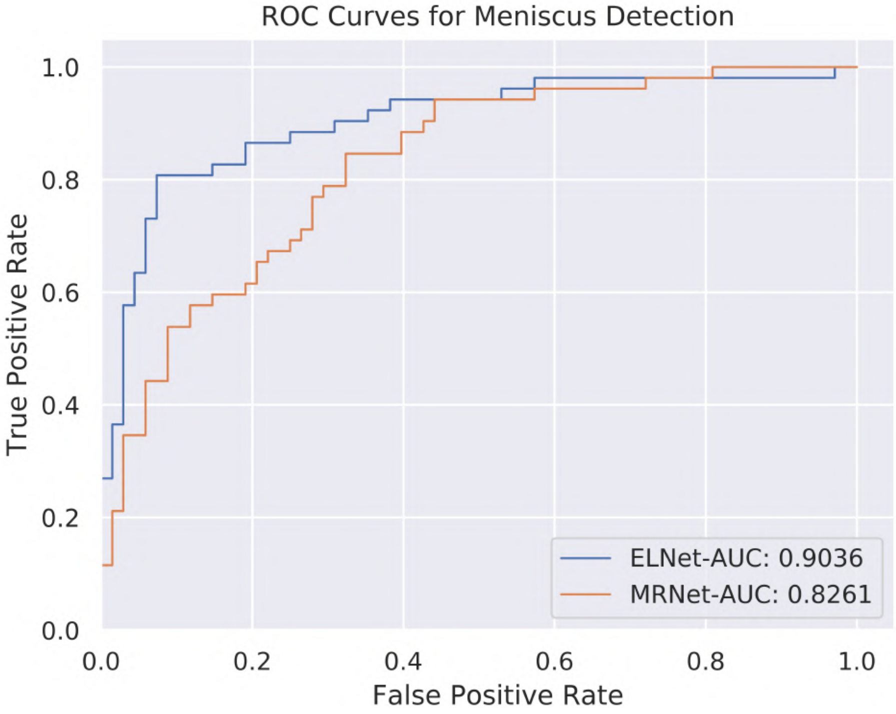


Figure 4.1: ROC for Meniscus Tear detection

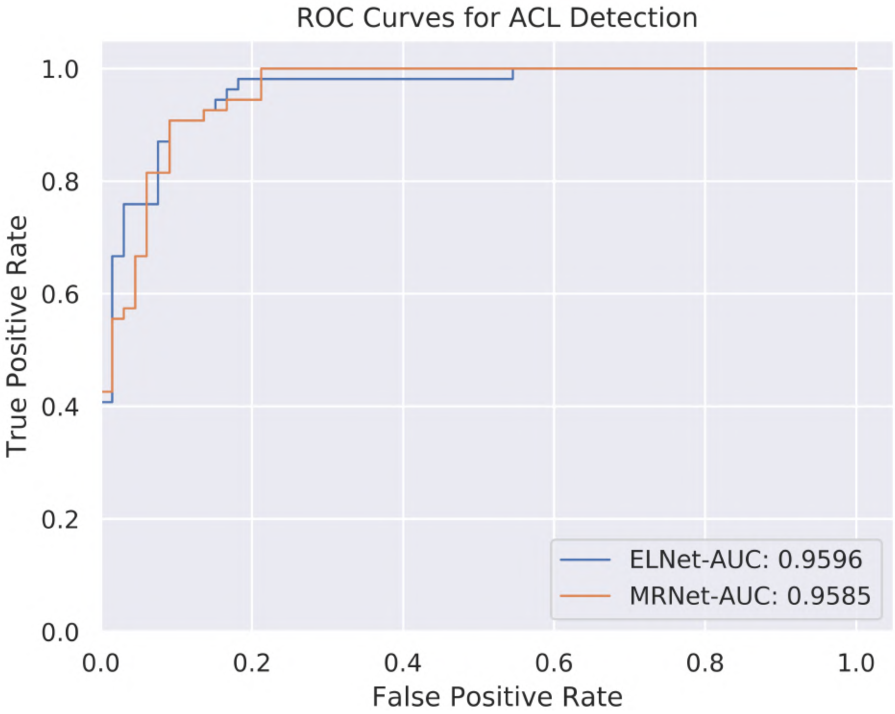


Figure 4.2: ROC for ACL Tear detection

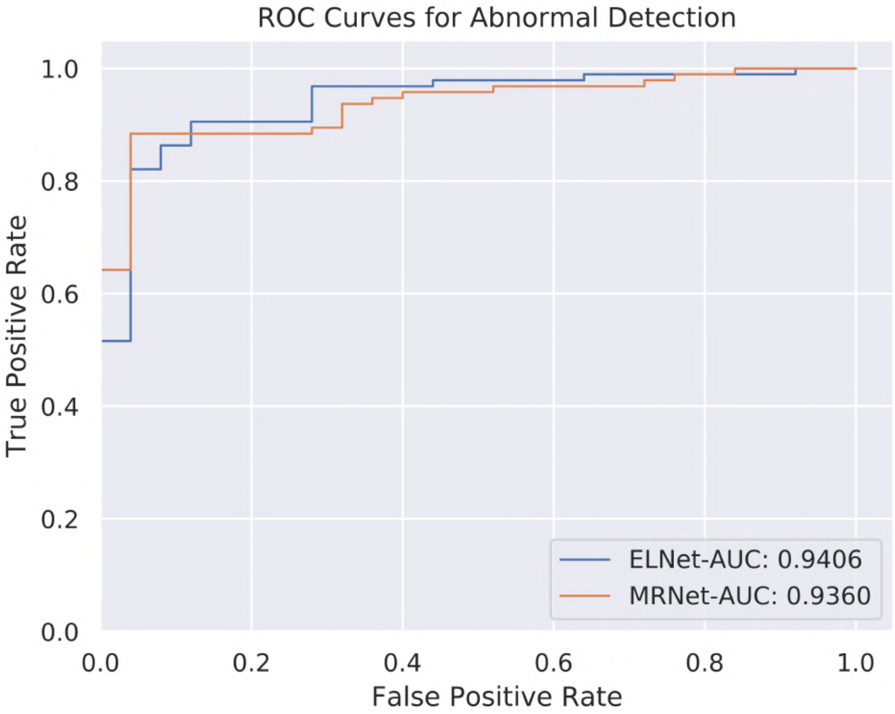


Figure 4.3: ROC for Abnormalities detection

KneeMRI Dataset. We evaluate ELNet using a 5-fold cross-validation scheme in detecting injuries in the ACL. The evaluation metrics following the 5-folds are shown on figure 4.4; we highlight the lowest value for each metric in red. The ROC's for the folds are shown in Figure 4.5. In the original paper, Štajduhar *et al.* trained an SVM and reported an AUC of 0.894 using 10-fold cross-validation. Bien *et al.* reported an AUC of 0.911 on a particular train/valid/test set split using a pre-trained MRNet. In our experiment, we obtain an average AUC of 0.913 from the 5-folds, with three of the five folds exceeding 0.92 and the highest being 0.924. Moreover, we observe just minor variations in multiple performance metrics across folds; this demonstrates our model's robustness despite the limited data and a highly unbalanced distribution.

4.2 Ablation Studies

Here, we compare ELNet performance when multi-slice normalization and BlurPool are replaced with batch normalization and max-pool. The objectives are the three pathologies presented in the MRNet dataset, and the best results following the modified ELNet designs are listed in Table 4.2.

Notice that the best performing models throughout the ablation studies are models that have utilized multi-slice normalization. One of the reasons why multi-slice normalization performs better than typical batch normalization stems from the fact that batch normalization induces an undesired standardization for each channel of feature representations across all slices. In this situation, the feature extractor (designed to extract per-slice features) would no longer process each slice independently, and a drop in performance can be expected.

The second reason batch normalization doesn't perform as well as multi-slice normalization relates to the independent and identically distributed (IID) assumption associated with batch normalization. In the case of natural images, batches are generated from a random selection of image samples within a given dataset. We assume that neighboring batch samples are independent from one another (zero correlation), and hence, when batch normalization computes the mean and variance for each channel across the entire batch, the computed mean and variance

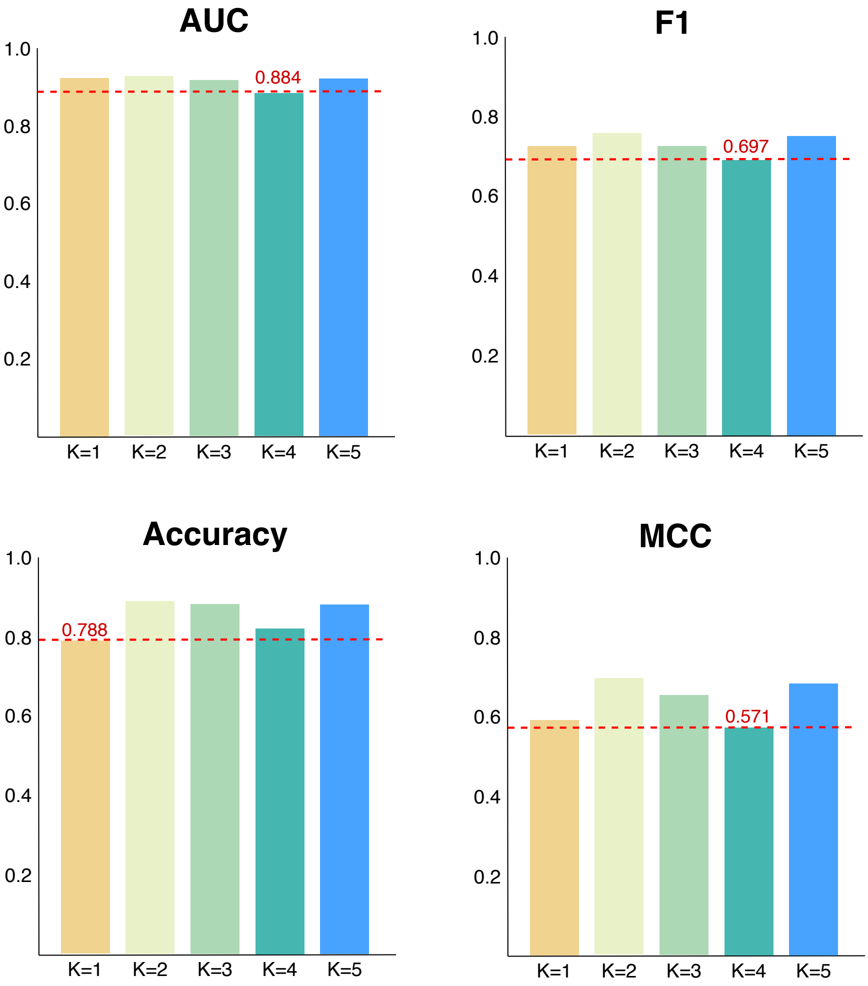


Figure 4.4: ELNet metrics following 5 fold cross-validation on the KneeMRI Dataset

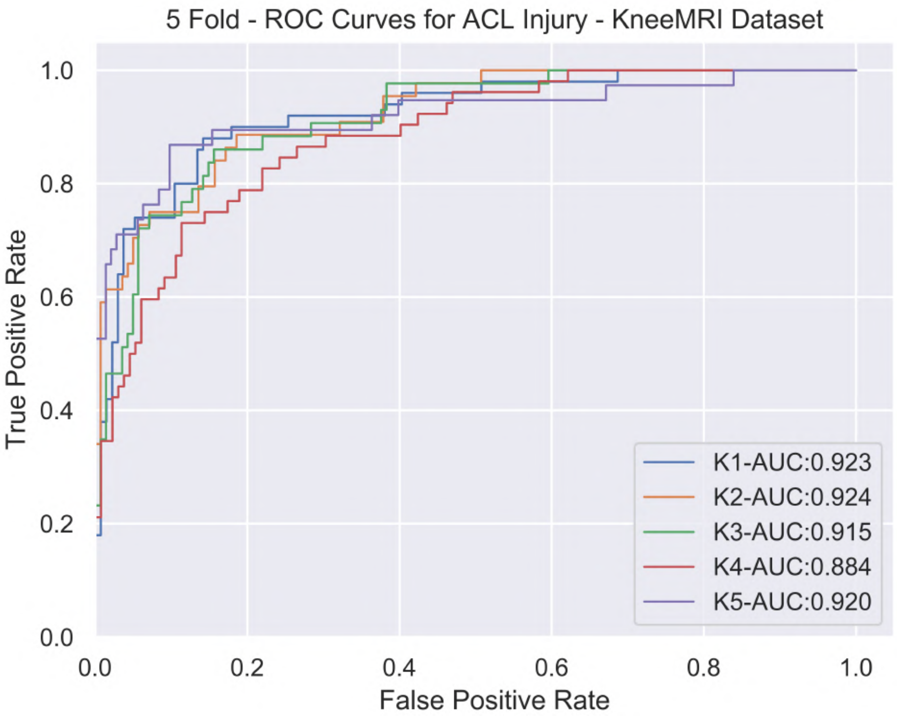


Figure 4.5: ELNet ROC's obtained from the 5 fold cross validation

resembles an approximation for some global statistic inherent in the dataset distribution. Therefore, the re-mapping parameters γ and β are closely related to that of the global statistic as well.

However, in the case of MRI images, the same IID assumption does not hold between adjacent slices. As a simple explanation, one can reasonably reconstruct a certain slice in an MRI volume by interpolating neighboring slices of the selected slice. In other words, there exists a high correlation between adjacent slices. This violates the assumption of batch normalization, and in our experiments, we observe network divergence after 10-15 epochs of training.

To our surprise, ELNet with batch norm and max-pool delivered slightly improved performance when compared with ELNet with batch norm and BlurPool, but when BlurPool is paired with the intended multi-slice normalization, we observe an overall improvement in performance compared to max-pooling.

ELNet Configuration	Meniscus Tear		ACL Tear		Abnormalities	
(K=4)	ROC-AUC	MCC	ROC-AUC	MCC	ROC-AUC	MCC
Multi-Slice Norm + BlurPool	**0.904**	**0.745**	**0.960**	**0.815**	0.941	**0.736**
Batch Norm + BlurPool	0.751	0.391	0.871	0.530	0.841	0.440
Multi-Slice Norm + MaxPool	0.848	0.534	0.923	0.633	**0.943**	0.557
Batch Norm + MaxPool	0.7972	0.403	0.906	0.693	0.880	0.312

Table 4.2: Comparison of ELNet performance when multi-slice normalization and BlurPool are replaced with batch normalization and max-pool. The ROC-AUC and MCC of the best performing model (one for each pathology) of each ELNet configuration is reported.

4.3 Clinical Comparison Interpretation

Following the setup described in Section 3.5, we computed a heatmap for each of the 16 selected cases using the Full-Gradient representation [19]. Examples of the generated Full-Gradient heatmaps for cases containing an ACL tear are shown in Figure 4.6.

Among the 9 cases containing an ACL tear, ELNet was able to locate the correct slice and the correct region containing the tear for 8 of the cases. Of the 7 cases that do not contain an ACL tear, ELNet correctly located the regions demonstrating intact ligaments in all 7 cases.

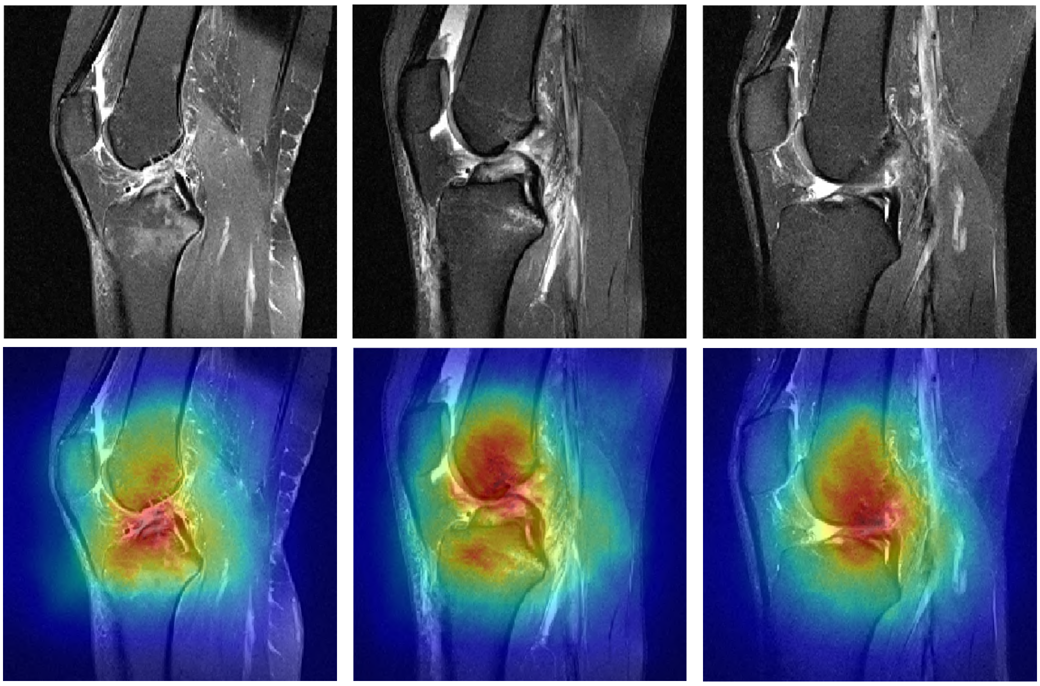


Figure 4.6: Top: Sample MRI slices containing ACL tears. Bottom: Full-Grad visualization computed using the above slices. "Hotter" areas indicate regions containing an ACL tear.

5 Conclusions

In this research, we explore the development of an automated knee diagnosis algorithm suitable for triage systems. By integrating a triage system into the workflow for which knee examinations are carried out, clinicians may first assess cases that are more probable of containing an injury and prevent further damages to a potentially injured knee.

The algorithm is based on ELNet, a unique CNN architecture optimized for knee injury detection. The input to ELNet is a 3D MRI image of a knee captured from a single plane orientation, and the output is a prediction for whether the knee exhibits signs of a certain injury. Compared with the previous SOTA MRNet, ELNet performs on a competitive level while remaining lightweight ($\sim$0.2M parameters), largely attributed to the multi-slice normalization and BlurPool modules built in.

Multi-slice normalization standardizes and re-maps the feature representations throughout ELNet in a slice-independent manner, and the optimal choice depends on the selected imaging orientation (e.g. axial, coronal, or sagittal). BlurPool, on the other hand, performs sub-sampling of the feature representations by first applying a low-pass filter on the input to avoid aliasing during down-sampling. The effects of the two novelties have also been explored in the ablation studies, and we show that the main contributor for observed performance gain is largely attributed to the choice of multi-slice normalization. The addition of BlurPool then further enhances the classification performance, and these two novelties allow for better optimization of the model during training.

Evaluation of ELNet was performed on two publicly available knee MRI datasets: the MRNet dataset and the KneeMRI dataset. For the three classification objectives provided with the MRNet dataset (i.e. ACL tear, meniscus tear, or other abnormalities), ELNet models (with $K = 4$) matched or surpassed the previous-best MRNet models by a significant margin. In the KneeMRI dataset, ELNet (with $K = 2$) demonstrated consistent performance throughout the five-fold cross validation in detecting ACL tear. In fact, four of the five ELNet models trained during cross-validation slightly exceeded the previous best ROC-AUC obtained by the MRNet model.

Further clinical evaluation was also carried out to compare our model's ability in locating tears in the knee to that of a board-certified MSK radiologist (with 17 years of experience). The Full-Gradient representation was computed with a trained ELNet for the input MRI image, and we compare the generated "heatmap" to the radiologist's determination. Of all the 16 cases (with and without ACL tear) that were randomly selected for evaluation, ELNet only mislocated the region containing the tear in a single case. This experiment demonstrates that ELNet is, in fact, capturing the correct features in performing classification. The addition of such a heatmap to the triage system may also benefit the radiologists by having the most informative slice (our algorithm predicts) presented first during manual examination.

5.1 Future Work

Following this research, two interesting directions may be further explored. First, this research demonstrated a plausible approach to develop a model for diagnosing knee injuries by training on a large number of class labels (the class labels are essentially binary labels in this work). Since class labels are much simpler and cheaper to obtain (compared with segmentation maps or manually labeled bounding boxes), one direction would be to extract a large number of class labels from existing medical reports, and use these labels to train a similar diagnostics model for both classification and localization. In our paper ”Labeling of Multilingual Breast MRI Reports” (see Appendix), we explore this topic extensively, and we propose a natural language processing (NLP) algorithm to extract such class labels from existing breast MRI reports.

The second interesting direction would be to extend ELNet models to other pathologies. Since the ELNet architecture was designed for 3D MRI images, it is very likely that ELNet models would perform just as well on other forms of 3D medical images such as computed tomography (CT) or positron emission tomography (PET) scans. In fact, several ongoing studies in our lab have adopted the ELNet architecture as a part of the respective algorithms, and promising performance has been observed.

In conclusion, we focus this research on the development of the ELNet architecture for knee injury diagnosis. Our work may assist in improving clinical workflows by prioritizing patients with potential knee injuries for the doctor's assessment. With the promising findings thus far, we believe our research may also serve as a basis for future works involving diagnosis with 3D medical images.

References

[1] Venkateshwaran Arumugam, Ganesan Ram Ganesan, and Paarthipan Natarajan. MRI Evaluation of Acute Internal Derangement of Knee. *Open Journal of Radiology*, 05(02):66–71, 2015.

[2] Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *ArXiv*, abs/1607.06450, 2016.

[3] Nicholas Bien, Pranav Rajpurkar, Robyn L. Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, Bhavik N. Patel, Kristen W. Yeom, Katie Shpanskaya, Safwan Halabi, Evan Zucker, Gary Fanton, Derek F. Amanatullah, Christopher F. Beaulieu, Geoffrey M. Riley, Russell J. Stewart, Francis G. Blankenberg, David B. Larson, Ricky H. Jones, Curtis P. Langlotz, Andrew Y. Ng, and Matthew P. Lungren. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLOS Medicine*, 15(11):e1002699, Nov 2018. doi: 10.1371/journal.pmed.1002699.

[4] N. R. Boeree, A. F. Watkinson, C. E. Ackroyd, and C. Johnson. Magnetic resonance imaging of meniscal and cruciate injuries of the knee. *Journal of Bone and Joint Surgery - Series B*, 73(3):452–457, 1991. doi: 10.1302/0301-620x.73b3.1670448.

[5] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21, 2020.

[6] J. V. Crues, J. Mink, T. L. Levy, M. Lotysch, and D. W. Stoller. Meniscal tears of the knee: Accuracy of MR imaging. *Radiology*, 164(2):445–448, 1987. doi: 10.1148/radiology.164.2.3602385.

[7] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[8] Daichi Hayashi, Ali Guermazi, and C. Kent Kwoh. Clinical and translational potential of MRI evaluation in knee osteoarthritis. *Current Rheumatology Reports*, 16(1), Jan 2014. doi: 10.1007/s11926-013-0391-6.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. doi: 10.1109/CVPR.2016.90.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, page 1026–1034, USA, 2015. IEEE Computer Society. doi: 10.1109/ICCV.2015.123.

[11] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 448–456. JMLR.org, 2015.

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F Pereira, C J C Burges, L Bottou, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[14] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Neural Information Processing Systems*, 2018.

[15] Fang Liu, Zhaoye Zhou, Alexey Samsonov, Donna Blankenbaker, Will Larison, Andrew Kanarek, Kevin Lian, Shivkumar Kambhampati, and Richard Kijowski. Deep learning approach for evaluating knee MR images: Achieving high diagnostic performance for cartilage lesion detection. *Radiology*, 289(1):160–169, Oct 2018. doi: 10.1148/radiol.2018172986.

[16] Alexandre Pedro Nicolini, Rogério Teixeira de Carvalho, Marcelo Mitsuro Matsuda, Jorge Sayum, and Moisés Cohen. Common injuries in athletes' knee: experience of a specialized center. *Acta Ortopedica Brasileira*, 22: 127 – 131, 2014.

[17] László G. Nyúl and Jayaram Koteshwar Udupa. On standardizing the mr image intensity scale. *Magnetic resonance in medicine*, 42:1072–81, 1999.

[18] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015.

[19] Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[20] Ivan Štajduhar, Mihaela Mamula, Damir Miletić, and Gözde Ünal. Semi-automated detection of anterior cruciate ligament injury from MRI. *Computer Methods and Programs in Biomedicine*, 140:151–164, Mar 2017. doi: 10.1016/j.cmpb.2016.12.006.

[21] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, page III–1139–III–1147. JMLR.org, 2013.

[22] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *ArXiv*, abs/1607.08022, 2016.

[23] Richard Zhang. Making convolutional networks shift-invariant again. In *ICML*, 2019.

A Knee Injury Detection using MRI with Efficiently-Layered Network (ELNet)

C.-H. Tsai, N. Kiryati, E. Konen, I. Eshed and A. Mayer, "Knee Injury Detection using MRI with Efficiently-Layered Network (ELNet)". Medical Imaging with Deep Learning (MIDL), Montreal, Quebec, Canada, 6-8 July 2020, Proceedings of Machine Learning Research (PMLR), Volume 121.

Knee Injury Detection using MRI with Efficiently-Layered Network (ELNet)

Chen-Han Tsai MAXWELLTSAI@YAHOO.COM
School of Electrical Engineering, Tel Aviv University, Israel

Nahum Kiryati NK@ENG.TAU.AC.IL
The Manuel and Raquel Klachky Chair of Image Processing, School of Electrical Engineering, Tel-Aviv University, Israel

Eli Konen ELI.KONEN@SHEBA.HEALTH.GOV.IL
Iris Eshed IRIS.ESHED@SHEBA.HEALTH.GOV.IL
Arnaldo Mayer ARNALDO.MAYER@SHEBA.HEALTH.GOV.IL
Diagnostic Imaging, Sheba Medical Center, affiliated to the Sackler School of Medicine, Tel-Aviv University, Israel

Abstract

Magnetic Resonance Imaging (MRI) is a widely-accepted imaging technique for knee injury analysis. Its advantage of capturing knee structure in three dimensions makes it the ideal tool for radiologists to locate potential tears in the knee. In order to better confront the ever growing workload of musculoskeletal (MSK) radiologists, automated tools for patients' triage are becoming a real need, reducing delays in the reading of pathological cases. In this work, we present the Efficiently-Layered Network (ELNet), a convolutional neural network (CNN) architecture optimized for the task of initial knee MRI diagnosis for triage. Unlike past approaches, we train ELNet from scratch instead of using a transfer-learning approach. The proposed method is validated quantitatively and qualitatively, and compares favorably against state-of-the-art MRNet while using a single imaging stack (axial or coronal) as input. Additionally, we demonstrate our model's capability to locate tears in the knee despite the absence of localization information during training. Lastly, the proposed model is extremely lightweight (< 1MB) and therefore easy to train and deploy in real clinical settings.

Keywords: Knee Diagnosis, MRI, Deep Learning, ACL Tear, Meniscus Tear, Knee Injury, Medical Triage

1. Introduction

Magnetic Resonance Imaging (MRI) has long been considered the most robust knee examination tool available (Saeed, 2018). Its widespread use is partly due to its capability to capture detailed structures in the knee joint while remaining a non-invasive procedure (Crues et al., 1987; Boeree et al., 1991). Given its profound capabilities to capture the knee in three dimensions, MRI has become the tool-of-choice for radiologists in an extensive range of examinations such as knee osteoarthritis and internal derangement of the knee. (Hayashi et al., 2014; Arumugam et al., 2015). Considering the ever growing workload of musculoskeletal (MSK) radiologists, automated tools for patients' triage are needed, leading to shorter delays in the reading of pathological cases. Several techniques have been proposed for this purpose. Štajduhar et al. (2017) presented a semi-automated approach that used support vector machines (SVM) to diagnose anterior cruciate ligament (ACL) injuries in the knee. In their work, an ROI is first manually extracted before being fed into the SVM for prediction. Liu et al. (2018) introduced a fully-automated cartilage lesion detection system by employing a CNN for segmentation followed by another CNN for patch classification. Although their network is trained end-to-end, the amount of manual labeling required to create the patch training set makes it an overwhelmingly cumbersome task. Bien et al. (2018) proposed an architecture that consists of three individual MRNets whose output are combined using logistic regression. An MRNet extracts a distinctive feature vector for each slice of the scan, stacks the vectors into a 2D array, max-pools the array to obtain a single vector, and performs classification by a fully connected layer with softmax activation. The backbone of the feature extractor is a pre-trained AlexNet (Krizhevsky et al., 2012).

In this work, we present an Efficiently-Layered Network (ELNet) architecture optimized for knee diagnosis using MRI. The main contribution of this work is a novel slice feature extracting network that incorporates multi-slice normalization along with BlurPool down-sampling. The proposed methods will be detailed in Section 2, followed by quantitative and qualitative experimental results in Section 3. Conclusion and future work will be given in Section 4.

2. Methods

The ELNet architecture is illustrated in Figure 1 and the details are listed in Table 1. The backbone of ELNet's design centers around *Block* modules. Inspired by ResNet (He et al., 2016), we define a *Block* as a sequence of:

$$[\text{2D Convolution} \rightarrow \text{Multi-slice Normalization} \rightarrow \text{ReLU activation}]$$

Blocks are designed to allow for non-linearities in the network, and they may be repeated while ensuring equal input and output dimensions. A skip connection is added between the input and output, allowing better optimization of the network. The first two *Blocks* are repeated twice with $4K$ and $8K$ channels, and the remaining *Blocks* are fixed with $16K$ channels.

Each *Block* is followed by another 2D Convolution and ReLU activation, and they serve to increase channel dimension. The spatial height and width are reduced using a BlurPool layer. Eventually, in the final layer of the feature extractor, 2D max-pooling is applied to obtain a $16K$-dimensional feature vector for each MRI slice. Max-pooling is consecutively applied to obtain a single $16K$-dimension feature vector that combines feature information across slices. Dropout is performed before feeding into a fully-connected layer with two output logits, and the final probability $p(y|x)$ is computed by softmax (Goodfellow et al., 2016).

In the following two subsections, we detail two innovative features of ELNet: the use of multi-slice normalization, and BlurPool.

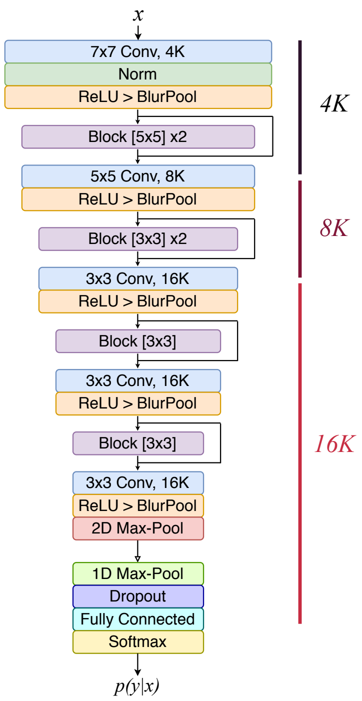


Figure 1: ELNet Design

Output Size	Layer Operation	Trainable Parameters
$s \times 4K \times 128 \times 128$	7×7 Conv, $4K$	$196K$
	Normalization	$4K$
$s \times 4K \times 62 \times 62$	ReLU → BlurPool	
	Block $[5 \times 5]{\times}2$	$800K^2 + 16K$
$s \times 8K \times 62 \times 62$	5×5 Conv, $8K$	$800K^2$
$s \times 8K \times 29 \times 29$	ReLU → BlurPool	
	Block $[3 \times 3]{\times}2$	$1152K^2 + 32K$
$s \times 16K \times 29 \times 29$	3×3 Conv, $16K$	$1152K^2$
$s \times 16K \times 13 \times 13$	ReLU → BlurPool	
	Block [3×3]	$2304K^2 + 32K$
	3×3 Conv, $16K$	$2304K^2$
$s \times 16K \times 5 \times 5$	ReLU → BlurPool	
	Block [3×3]	$2304K^2 + 32K$
	3×3 Conv, $16K$	$2304K^2$
$s \times 16K$	ReLU → BlurPool → 2D Max-Pool	
$16K$	1D Max-Pool → Dropout	
2	Fully Connected → Softmax	$32K + 2$
	Total Trainable Parameters	$13120K^2 + 348K + 2$

Table 1: ELNet architecture in detail

Multi-Slice Normalization

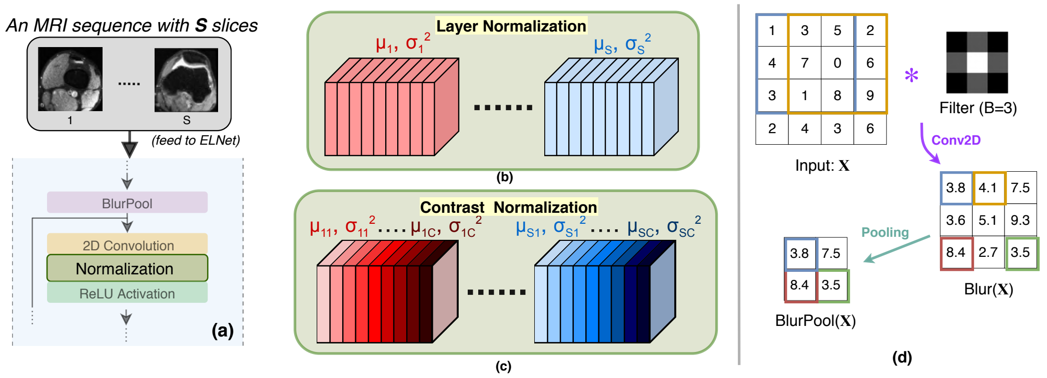
We propose two possible variants of multi-slice normalization: a first one based on *layer normalization* (Ba et al., 2016), and a second one based on *contrast normalization* (Ulyanov et al., 2016). Let's assume a feature representation $x^{(i)} \in \mathbb{R}^{S \times C \times H \times W}$ from some layer i in the network (usually a 2D-convolution), where S is the number of slices in the MRI sequence, C is the number of channels in the representation, and H, W are the spatial height and width of the representation. The network applies a normalization on x (omitting i for simplicity) by computing the appropriate mean and variance.

In the *layer normalization* variant, the mean μ_s and variance σ_s^2 are computed from x for each slice s $(1 \leq s \leq S)$. In *contrast normalization*, the mean μ_{sc} and variance σ_{sc}^2 are computed for each slice s and also for each channel c $(1 \leq c \leq C)$ (Figure 2 a-c). Using the computed mean and variance, x is standardized into $\hat{x}$. An affine transform is applied to $\hat{x}$ to obtain the normalized output y. The normalization process is expressed by Equation (1) for *layer normalization* and Equation (2) for *contrast normalization* respectively:

$$\hat{x}_s = \frac{x_s - \mu_s}{\sqrt{\sigma_s^2 + \epsilon}} \rightarrow y_s = \gamma \hat{x}_s + \beta \qquad \forall s : 1 \rightarrow S \tag{1}$$

$$\hat{x}_{sc} = \frac{x_{sc} - \mu_{sc}}{\sqrt{\sigma_{sc}^2 + \epsilon}} \rightarrow y_{sc} = \gamma \hat{x}_{sc} + \beta \qquad \forall s : 1 \rightarrow S, c : 1 \rightarrow C \tag{2}$$

Parameters γ, β (C dimensional vectors) are learned independently for each normalization layer. Typically, γ, β, ϵ are initialized to **1** , **0**, and 1e-8 respectively.



1	3	5	2
4	7	0	6
3	1	8	9
2	4	3	6

3.8	4.1	7.5
3.6	5.1	9.3
8.4	2.7	3.5

3.8	7.5
8.4	3.5

Figure 2: **(a)** An MRI sequence fed as input to ELNet, and an illustration of an ELNet Block. **(b&c)** Our proposed multi-slice normalization: Layer normalization and Contrast normalization (multi-slice norm standardizes slice-wise unlike batch norm which standardizes channel-wise) **(d)** BlurPool example: Input X is convolved with binomial filter (kernel $B = 3$) to obtain an anti-aliased representation $\mathrm{Blur}(X)$. Pooling is then applied to obtain $\mathrm{BlurPool}(X)$.

BlurPool

In the work of Zhang (2019), a BlurPool operation was proposed to mitigate the shift-variance phenomenon observed in modern CNN architectures where max-pooling is often utilized. BlurPool functions by first applying an anti-aliasing filter (binomial filter with kernel size B and stride 1) to the input representation, then strided pooling is applied to obtain the pooled feature map (see Figure 2d). The resulting representation is therefore a pooled version of the blurred input representation, and a more detailed analysis is available in the paper (Zhang, 2019).

2.1. Training Pipeline

As suggested by Nyúl and Udupa (1999), we perform histogram-based intensity standardization according to the training set statistics, thus enabling similar-valued pixels to be associated with the relevant tissue type. In addition, we perform randomized data augmentations to each series which includes translation, horizontal flip, scaling, and minor rotations up to ±10 degrees around the center of the volume. For volumes captured in the axial and coronal orientations, we apply an additional random rotation of a multiple of 90 degrees to the volume. Finally, all the images are resized to 256 × 256 before entering the network.

Aside from data augmentation, we implement oversampling to compensate for dataset imbalance. For each pathology, we select the minority class samples (allowing repeats) from our training set and apply augmentations on them until the number of minority class samples (along with their augmented copies) equals the number of majority samples.

We train ELNet using standard cross-entropy loss (Goodfellow et al., 2016). Optimization can be done using a simple grid-search over relevant hyperparameters such as learning rate, choice of multi-layer normalization, BlurPool kernel sizes, dropout rate, etc.

3. Experiments

3.1. Datasets

MRNet Dataset. The MRNet Dataset contains 1,370 knee MRI examinations that were carried out at the Stanford University Medical Center. Each case was labeled according to the presence/absence of an anterior cruciate ligament (ACL) tear, a meniscus tear, or other signs of abnormalities in the corresponding knee. Each exam was randomly assigned either to the training, validation, or test set (Bien et al., 2018). It should be noted that each exam may contain multiple labels (e.g. an exam labeled positive for abnormality and ACL tear indicates other forms of abnormality in addition to an ACL tear).

The provided dataset includes, for each case, corresponding axial, coronal and sagittal MRI acquisitions. As reported by Bien et al., a sagittal T2-weighted series, a coronal T1 weighted series, and an axial proton density weighted series were selected for this dataset. Each image is of size 256 × 256 and the number of slices ranges between 17- 61 (mean 31 and standard deviation 7.97). The MRNet Dataset is currently the largest public labeled knee MRI dataset.

KneeMRI. The KneeMRI dataset collected at the Clinical Hospital Centre Rijeka, Croatia by Štajduhar et al consists of 917 exams labeled with ACL conditions in the corresponding knee. For each exam, the ligament condition was classified as either healthy (690 exams, 75.2%), partially injured (172 exams, 18.8%), or completely ruptured (55 exams, 6%). Each assessment corresponds to a T1-weighted sagittal MRI series, containing 320 × 320 or 290 × 300 images. The number of images in each series ranges between 21-45 (mean 31 and standard deviation 2.27). The dataset was divided into 10 strata with similar distributions, and we perform stratified sampling for evaluation.

3.2. Training

MRNet Dataset. In the MRNet dataset, we were provided with three imaging orientations per examination. For the three pathologies, we trained three separate ELNet's with $K = 4$, and the network weights were initialized uniformly by choosing the best random seed between 0-4 (He et al., 2015). Based on experiments, we selected coronal images for detecting meniscus tears, and axial images for detecting ACL tears and abnormalities. Contrast normalization yielded the best results for detecting meniscus tears, and layer normalization for detecting ACL tears and abnormalities (notice the correspondence between the selected multi-slice normalization and image modality.) Each model was trained using Adam with a learning rate between 1e-5 and 3e-5 for 200 epochs, taking roughly 1.5 hours (Kingma and Ba, 2014).

KneeMRI Dataset With the KneeMRI dataset, we perform 5-fold cross validation using eight out of the ten strata, and validation using the remaining two. Similar to the MRNet Datset, we train an ELNet with K=2 using SGD+Momentum for 200 epochs and the training time is roughly an hour for each fold (Sutskever et al., 2013).

Architecture	Pathology	Accuracy	Sensitivity	Specificity	ROC-AUC	MCC
MRNet	Meniscus Tear	0.735	0.827	0.662	0.826	0.489
	ACL Tear	0.9	0.907	0.894	0.956	0.769
	Abnormality	0.883	0.947	0.64	0.936	0.628
ELNet	Meniscus Tear	0.88	0.86	0.89	**0.904**	**0.745**
	ACL Tear	0.904	0.923	0.891	**0.960**	**0.815**
	Abnormality	0.917	0.968	0.72	**0.941**	**0.736**

Table 2: Evaluation Statistics between MRNet and ELNet on the MRNet validation set

By choosing K=2, and K=4 for the ELNet architectures, our trained model involves 53,178, and 211,314 trainable parameters respectively. In relation to AlexNet (∼61M trainable parameters), ELNet (with K=4) contains 288× fewer parameters than AlexNet. In comparison with MRNet, ELNet with K=4 contains 866× less parameters, and ELNet with K=2 contains 1147× less parameters. Each trained model was saved using standard PyTorch format. Model sizes are 850kB and 435kB for K=4 and K=2 respectively. Our experiments were perfomed on an NVIDIA GTX 1070 8GB GPU.

3.3. Evaluation

MRNet Dataset. We evaluate ELNet's performance using the validation set provided by the MRNet dataset (since the test set is not publicly available), and we compare it with the MRNet model proposed and trained by Bien et al. Although they evaluated their models primarily using the ROC-AUC, we perform a more thorough analysis by considering additional metrics that are just as significant, such as Sensitivity and the Matthew Correlation Coefficient (MCC). The evaluation results are presented in Table 2 and the ROC is plotted in Figure 3 (a-c), where we can observe noticeably higher MCC of the ELNet model.

KneeMRI Dataset. We evaluate ELNet using a 5-fold cross-validation scheme in detecting injuries in the ACL. The evaluation metrics following the 5-folds are shown on figure 3 (d-g); we highlight the lowest value in for each metric in red. In the original paper, Štajduhar et al trained an SVM and reported an AUC of 0.894 using 10-fold cross-validation. Bien et al reported an AUC of 0.911 on a particular train/valid/test set split using a pre-trained MRNet. In our experiment, we obtain an average AUC of 0.913 from the 5-folds, with three of the five folds exceeding 0.92 and the highest being 0.924. Moreover, we observe just minor variations in multiple performance metrics across folds; this demonstrates our model's robustness despite limited data and a highly unbalanced distribution.

3.4. Ablation Studies

This section aims to compare ELNet performance when multi-slice normalization and BlurPool are replaced with batch normalization (Ioffe and Szegedy, 2015) and max-pooling. The objectives are the three pathologies presented in the MRNet dataset, and the best results following the modified ELNet designs are listed in Table 3. Stemming from the fact that batch normalization induces an undesired standardization for each channel of feature

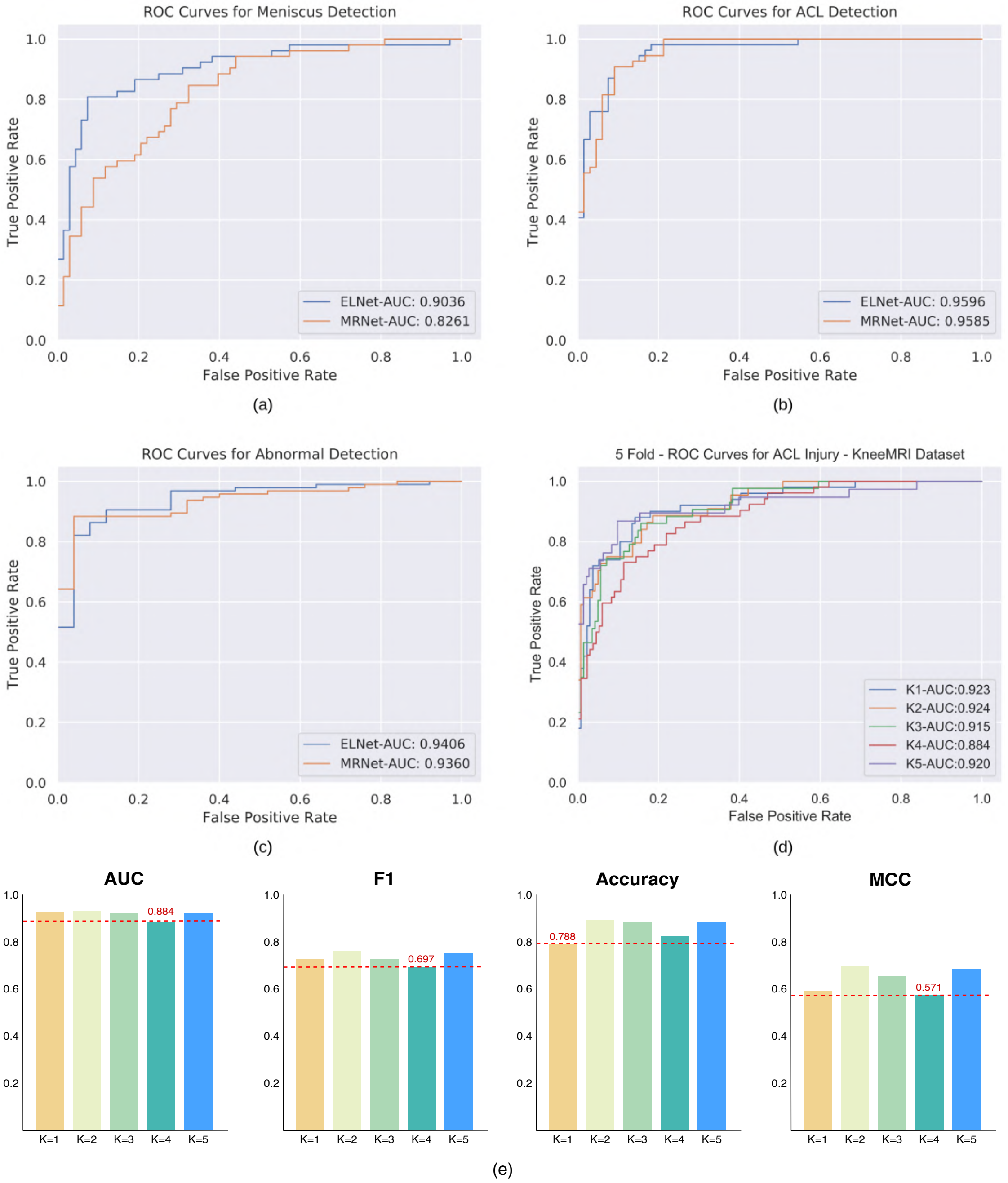


Figure 3: **MRNet Dataset:** (a-c) Comparision of ELNet and MRNet ROC **KneeMRI Dataset:** (d) ELNet ROC's obtained from 5-fold cross-validation (e) ELNet metrics following 5-fold cross-validation

ELNet Configuration	Meniscus Tear		ACL Tear		Abnormalities	
(K=4)	ROC-AUC	MCC	ROC-AUC	MCC	ROC-AUC	MCC
Multi-Slice Norm + BlurPool	**0.904**	**0.745**	**0.960**	**0.815**	0.941	**0.736**
Batch Norm + BlurPool	0.751	0.391	0.871	0.530	0.841	0.440
Multi-Slice Norm + MaxPool	0.848	0.534	0.923	0.633	**0.943**	0.557
Batch Norm + MaxPool	0.7972	0.403	0.906	0.693	0.880	0.312

Table 3: Comparison of ELNet performance when multi-slice normalization and BlurPool are replaced with batch normalization and max-pool. The ROC-AUC and MCC of the best performing model (one for each pathology) of each ELNet configuration is reported.

representations across all slices, the feature extractor (designed to extract per-slice features) would no longer process each slice independently, and degraded performance deems reasonable. Following our experiments, it is evident that the use of batch normalization aggravates ELNet performance. In practice, we observe network divergence during training after 10-15 epochs. To our surprise, ELNet with batch norm and max-pool delivered slightly improved performance when compared with ELNet with batch norm and BlurPool, but when BlurPool is paired with the intended multi-slice normalization, we observe an overall improvement in performance compared to max-pooling.

3.5. Model Interpretation

To understand how ELNet identifies certain attributes for diagnosis, we compute the Full-Gradient representation of ELNet using the FullGrad algorithm (Srinivas and Fleuret, 2019). FullGrad generates a heat-map that corresponds to parts of the input that most influence the output prediction. Conceptually, the generated heat-map should be "hotter" in areas indicating an injury and "cold" elsewhere.

To verify that ELNet is indeed performing diagnosis based on features in the given acquisition, we randomly selected one of the five cross validation splits and evaluated the trained ELNet from that split. Samples from the validation set were randomly selected from both classes, resulting in 9 cases containing ACL tear and 7 cases without. A board-certified MSK radiologist with 17 years of experience was asked to identify the most informative slice (slice containing the most area for which a tear resides) in a given series and furthermore indicate the region in the (most informative) slice corresponding to an ACL injury. The identical task was performed on the trained ELNet, and of the 9 cases that contain ACL tear, the trained ELNet's prediction of the most informative slice and tear region coincided with the radiologist's evaluation in 8 of the cases. Of the 7 cases where the ACL is intact, our model's prediction matched the radiologist's assessment in all 7 cases. In Figure 4, we present a few examples of the generated heat-maps.

4. Conclusion

In this work, we present ELNet, a unique CNN architecture optimized for knee injury detection. The novel integration of multi-slice normalization and BlurPool operations allow ELNet models to remain lightweight (∼ 0.2M parameters, requiring single imaging stack, trained from scratch) while performing favorably against MRNet models (∼ 183M

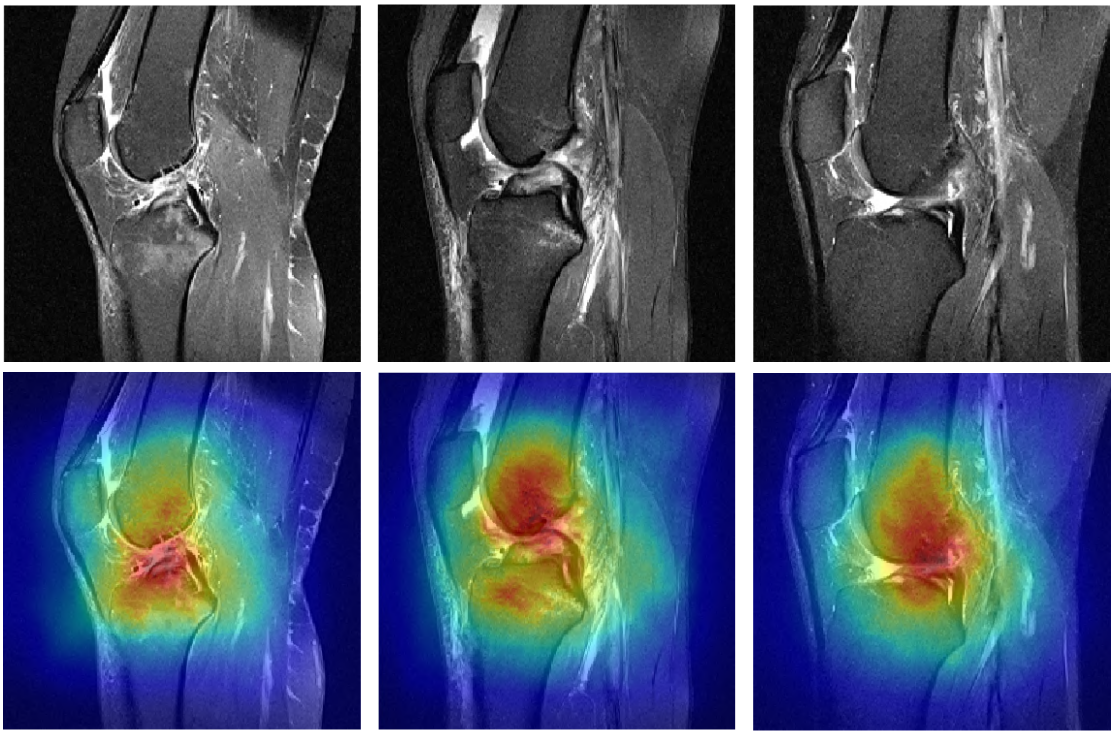


Figure 4: **Top:** Sample MRI slices containing ACL tears. **Bottom:** Full-Grad visualization computed using the above slices. "Hotter" areas indicate regions containing ACL tear.

parameters, requiring three imaging stacks, pretrained AlexNet) on the MRNet dataset. Cross-validation on the KneeMRI dataset have demonstrated consistent improved performance with ELNet models, proving the architecture to be robust regardless of a highly unbalanced distribution. In a clinical setting, where large number of cases await evaluation, our algorithm may be used for triage, improving workflow efficiency. In addition, by having our algorithm locate regions containing tears, radiologists can benefit by having the most significant slice presented first for each case.

Future work may include performance enhancement by incorporation of all three MRI volumes, axial, coronal and sagittal, if available. Further research is also needed to facilitate application of trained models on MRI data acquired using different scanners with various intensity scales. With the promising findings thus far, we believe ELNet may serve as a solid basis for future works involving knee injury triage.

References

Venkateshwaran Arumugam, Ganesan Ram Ganesan, and Paarthipan Natarajan. MRI Evaluation of Acute Internal Derangement of Knee. *Open Journal of Radiology*, 05(02): 66–71, 2015. doi: 10.4236/ojrad.2015.52011.

Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *ArXiv*, abs/1607.06450, 2016.

Nicholas Bien, Pranav Rajpurkar, Robyn L. Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, Bhavik N. Patel, Kristen W. Yeom, Katie Shpanskaya, Safwan Halabi, Evan Zucker, Gary Fanton, Derek F. Amanatullah, Christopher F. Beaulieu, Geoffrey M. Riley, Russell J. Stewart, Francis G. Blankenberg, David B. Larson, Ricky H. Jones, Curtis P. Langlotz, Andrew Y. Ng, and Matthew P. Lungren. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLOS Medicine*, 15(11):e1002699, Nov 2018. doi: 10.1371/journal.pmed.1002699.

N. R. Boeree, A. F. Watkinson, C. E. Ackroyd, and C. Johnson. Magnetic resonance imaging of meniscal and cruciate injuries of the knee. *Journal of Bone and Joint Surgery - Series B*, 73(3):452–457, 1991. doi: 10.1302/0301-620x.73b3.1670448.

J. V. Crues, J. Mink, T. L. Levy, M. Lotysch, and D. W. Stoller. Meniscal tears of the knee: Accuracy of MR imaging. *Radiology*, 164(2):445–448, 1987. doi: 10.1148/radiology.164.2.3602385.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Daichi Hayashi, Ali Guermazi, and C. Kent Kwoh. Clinical and translational potential of MRI evaluation in knee osteoarthritis. *Current Rheumatology Reports*, 16(1), Jan 2014. doi: 10.1007/s11926-013-0391-6.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. doi: 10.1109/CVPR.2016.90.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, page 1026–1034, USA, 2015. IEEE Computer Society. doi: 10.1109/ICCV.2015.123.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, Dec 2014.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F Pereira, C J C Burges, L Bottou, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. URL http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

Fang Liu, Zhaoye Zhou, Alexey Samsonov, Donna Blankenbaker, Will Larison, Andrew Kanarek, Kevin Lian, Shivkumar Kambhampati, and Richard Kijowski. Deep learning approach for evaluating knee MR images: Achieving high diagnostic performance for cartilage lesion detection. *Radiology*, 289(1):160–169, Oct 2018. doi: 10.1148/radiol.2018172986.

László G. Nyúl and Jayaram K. Udupa. On standardizing the mr image intensity scale. *Magnetic Resonance in Medicine*, 42(6):1072–1081, 1999. doi: 10.1002/(SICI)1522-2594(199912)42:6⟨1072::AID-MRM11⟩3.0.CO;2-M.

Ikhlas O Saeed. MRI Evaluation for Post-Traumatic Knee Joint Injuries. *IOSR Journal of Nursing and Health Science (IOSR-JNHS)*, 7(2):48–51, 2018. URL https://www.iosrjournals.org/iosr-jnhs/papers/vol7-issue2/Version-7/F0702074851.pdf.

Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Ivan Štajduhar, Mihaela Mamula, Damir Miletić, and Gözde Ünal. Semi-automated detection of anterior cruciate ligament injury from MRI. *Computer Methods and Programs in Biomedicine*, 140:151–164, Mar 2017. doi: 10.1016/j.cmpb.2016.12.006.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, page III–1139–III–1147. JMLR.org, 2013.

Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *ArXiv*, abs/1607.08022, 2016.

Richard Zhang. Making convolutional networks shift-invariant again. In *ICML*, 2019.

B Labeling of Multilingual Breast MRI Reports

C.-H. Tsai, N. Kiryati, E. Konen, M. Sklair-Levy and A. Mayer, “Labeling of Multilingual Breast MRI Reports”, 5th MICCAI Workshop on Large-scale Annotation of Biomedical data and Expert Label Synthesis (MICCAI LABELS), Lima, Peru, October 2020.

Labeling of Multilingual Breast MRI Reports

Chen-Han Tsai[1], Nahum Kiryati[2], Eli Konen[3], Miri Sklair-Levy[3], and Arnaldo Mayer[3]

[1] School of Electrical Engineering, Tel Aviv University, Israel
[2] The Manuel and Raquel Klachky Chair of Image Processing, School of Electrical Engineering, Tel-Aviv University, Israel
[3] Diagnostic Imaging, Sheba Medical Center, affiliated to the Sackler School of Medicine, Tel-Aviv University, Israel

Abstract. Medical reports are an essential medium in recording a patient's condition throughout a clinical trial. They contain valuable information that can be extracted to generate a large labeled dataset needed for the development of clinical tools. However, the majority of medical reports are stored in an unregularized format, and a trained human annotator (typically a doctor) must manually assess and label each case, resulting in an expensive and time consuming procedure. In this work, we present a framework for developing a multilingual breast MRI report classifier using a custom-built language representation called LAMBR. Our proposed method overcomes practical challenges faced in clinical settings, and we demonstrate improved performance in extracting labels from medical reports when compared with conventional approaches.

Keywords: Labeling · Medical Reports · Transfer Learning · Breast MRI · LAMBR

1 Introduction

The introduction of the Electronic Medical Record (EMR) has improved convenience in accessing and organizing medical reports. With the increasing demand for biomedical tools based on deep learning, obtaining large volumes of labeled data is essential for training an effective model. One major category where such deep learning models excel is in the area of computer assisted diagnosis (CADx), and several works (e.g. [1, 4, 12]) have demonstrated effective utilization of weakly labeled data to achieve promising performance. Since understanding medical data requires specialized training, datasets often contain a small subset of all past exams, that are manually relabeled by doctors for the target task. Not only is this a labour-intensive process, but the resulting dataset is often too small to represent the true distribution, resulting in underperforming models.

In this work, we present a framework for developing multilingual breast MRI report classifiers by using a customized language representation called LAMBR. LAMBR is first obtained by pre-training an existing language representation on a large quantity of breast MRI reports. Fine-tuning is then applied to obtain separate classifiers that can perform tasks such as: (1) determining whether the corresponding patient in the report has been suggested to undergo biopsy or (2) predicting BI-RADS [4] score for the reported lesion (see Figure 1). With such classifiers, one may avoid the manual labeling required from doctors, and instead, automatically extract a large number of weak labels from existing medical reports for training weakly supervised breast MRI CADx models.

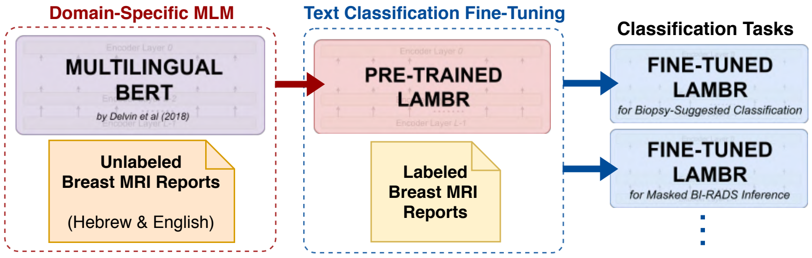
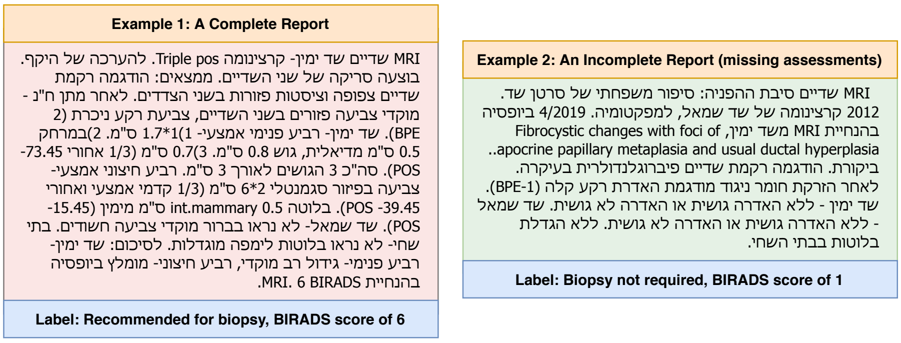

Domain-Specific MLM
MULTILINGUAL BERT
by Delvin et al (2018)
Unlabeled Breast MRI Reports
(Hebrew & English)
Text Classification Fine-Tuning
PRE-TRAINED LAMBR
Labeled Breast MRI Reports
Classification Tasks
FINE-TUNED LAMBR
for Biopsy-Suggested Classification
FINE-TUNED LAMBR
for Masked BI-RADS Inference
⋮

Fig. 1. An overview of training stages presented in our framework. Pre-training is performed on the multilingual BERT with unlabeled breast MRI reports to obtain a pre-trained LAMBR. The pre-trained LAMBR is then fine-tuned using a small number of labeled reports to obtain classifiers for specific downstream text classification tasks.

Prior to our work, text classification has been explored extensively by several studies such as ULMFIT [5] and SiATL [2]. ELMo [9], BERT [3], and XLNet [16] have also demonstrated adequate approaches towards text classification using the notion of a generalized language representation. However, the majority of these approaches require pre-training an encoder on a massive text corpora, and this is a time consuming and resource intensive procedure that is impractical for a clinical setting [10]. Moreover, the majority of prior works perform encoder pre-training on widely available natural language texts which differ greatly from the scarcely available medical texts.

To overcome the difference in distrubtion between medical texts and natural texts, BioBERT [7] introduced a pre-training objective that relied on a large collection of PubMed abstracts and PMC articles. Although BioBERT demonstrated improved performance compared with BERT, their method does not avoid the above resource intensive pre-training. Within English medical reports, ALARM [14] proposed a simple approach for labeling head MRI reports by utilizing a pre-trained Bio-BERT and this avoids the expensive pre-training often required. Yet for multilingual medical reports, such Bio-BERT does not exist, and in this work, we present a solution based on the multilingual BERT. Our novel approach introduces an inexpensive pre-training objective that yields favorable text classification performance when fine-tuned, and in our experiments, we demonstrate the robustness of the resulting classifiers even in cases where

[4] Breast Imaging-Reporting and Data System: a score between 0-6 indicating the level of severity of a breast lesion



Example 1: A Complete Report

MRI שדיים שד ימין- קרצינומה Triple pos. להערכה של היקף.
בוצעה סריקה של שני השדיים. ממצאים: הודגמה רקמת
שדיים צפופה וציסטות פזורות בשני הצדדים. לאחר מתן ח"נ -
מוקדי צביעה פזורים בשני השדיים, צביעת רקע ניכרת (2
BPE). שד ימין- רביע פנימי אמצעי- 1)1*1.7 ס"מ. 2)במרחק
0.5 ס"מ מדיאלית, גוש 0.8 ס"מ. 3)0.7 ס"מ (1/3 אחורי -73.45
POS). סה"כ 3 הגושים לאורך 3 ס"מ. רביע חיצוני אמצעי-
צביעה בפיזור סגמנטלי 2*6 ס"מ (1/3 קדמי אמצעי ואחורי
-39.45 POS). בלוטה 0.5 int.mammary ס"מ מימין (-15.45
POS). שד שמאל- לא נראו בברור מוקדי צביעה חשודים. בתי
שחי- לא נראו בלוטות לימפה מוגדלות. לסיכום: שד ימין-
רביע פנימי- גידול רב מוקדי, רביע חיצוני- מומלץ ביופסיה
בהנחיית MRI. 6 BIRADS.

Label: Recommended for biopsy, BIRADS score of 6

Example 2: An Incomplete Report (missing assessments)

MRI שדיים סיבת ההפניה: סיפור משפחתי של סרטן שד.
2012 קרצינומה של שד שמאל, למפקטומיה. 4/2019 ביופסיה
בהנחיית MRI משד ימין, Fibrocystic changes with foci of
apocrine papillary metaplasia and usual ductal hyperplasia..
ביקורת. הודגמה רקמת שדיים פיברוגלנדולרית בעיקרה.
לאחר הזרקת חומר ניגוד מודגמת האדרת רקע קלה (BPE-1).
שד ימין - ללא האדרה גושית או האדרה לא גושית. שד שמאל
- ללא האדרה גושית או האדרה לא גושית. ללא הגדלת
בלוטות בבתי השחי.

Label: Biopsy not required, BIRADS score of 1

Fig. 2. Examples of breast MRI reports written in Hebrew and English (read from right to left). Example 1 is complete report parsed from the EMR, and Example 2 is missing the final assessments due to incorrect parsing. The patient in Example 1 is recommended for biopsy, and patient in Example 2 is not required to perform biopsy.

parsing errors exist (see Figure 2). The remaining sections of our paper are organized as follows: the proposed framework is presented in Section 2, experimental results are reported in Section 3, and the conclusion is given in Section 4.

2 Methods

2.1 BERT Recap

BERT is a language representation based on the Transformer-Encoder [13]. The input to the Transformer-Encoder is a sequence of tokens $\{x_i\}$ generated by WordPiece Embeddings [15] from a given series of sentences. Special tokens are inserted and position encodings are added, and the output is a sequence of bidirectional embeddings that represents each input token [3]. In order to obtain the BERT language representation, Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) pre-training objectives were introduced. MLM applies random masking on the 15% of the input tokens, and BERT is trained to identify the original token of the masked token by attending to other tokens of the same sequence. The NSP objective trains BERT in understanding sentence coherence by randomly replacing the second sentence of an existing sentence pair, and BERT has to determine whether the pair are neighboring sentences.

2.2 Domain-Specific Masked Language Modeling

The Domain-Specific Masked Language Modeling (DS-MLM) we propose is a modification of the MLM pre-training objective introduced in BERT. The multilingual BERT was trained using monolingual corpora from 104 languages, and DS-MLM aims to retrain the multilingual BERT to better model the language observed in breast MRI reports written in Hebrew and English. Unlike

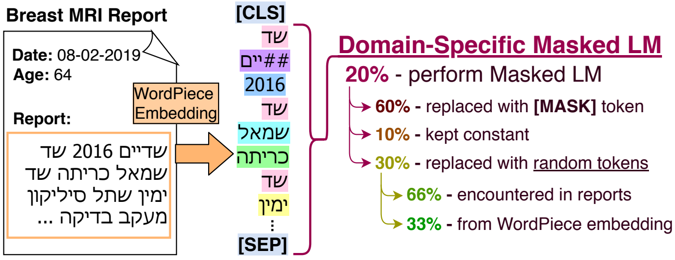


Fig. 3. An example of the tokens generated from breast MRI report using WordPiece Embeddings. During DS-MLM, a portion of the tokens are augmented and the pre-training objective is to correctly identify the original token prior to augmentation (performed only on tokens selected for augmentation).

BioBert [7], which relies on pre-training over massive biomedical corpora, we perform DS-MLM solely from the available breast MRI reports stored in the hospital's EMR.

For each medical report, tokens are generated using WordPiece Embeddings (see Figure 3). The `[CLS]` and `[SEP]` tokens are appended to the beginning and the end of the generated tokens (`[SEP]` tokens are not added between sentences). Since the multilingual BERT is already trained on a general domain corpora, we select 20% of the generated tokens for MLM. Of the selected tokens, 60% are masked using the `[MASK]` token, 30% are replaced with existing tokens and the remaining 10% are left unchanged. In order to expose our model to more frequent tokens observed in breast MRI reports, of the 30% of tokens selected for replacement, two thirds are replaced with existing tokens encountered in breast MRI reports, and one third is replaced with tokens from the complete vocabulary (may include tokens corresponding to other languages). Dynamic masking is applied to allow more exposure to a broad range of tokens.

Since most medical reports contain sentences not adhering to a strict flow of ideas, we do not incorporate NSP into the pre-training objective of our framework. In addition, RoBERTa [8] demonstrated that the removal of NSP may even improve downstream task performance, and therefore, the pre-training objective of the LAMBR language representation is simply DS-MLM.

2.3 Text Classification Fine-Tuning

Text Classification Fine-Tuning (TCFT) is a series of techniques to fine-tune a pre-trained LAMBR for performing text classification. We propose a simple classifier head to add on top of the Transformer-Encoder, and we present a method to fine-tune the complete text classifier (Transformer-Encoder along with classifier head) using a pre-trained LAMBR (see Figure 4).

Classifier Head. For a given token sequence $\{x_i\}$, we obtain the output embedding sequence $\{z_i\}$ from the pre-trained LAMBR. The average of the output

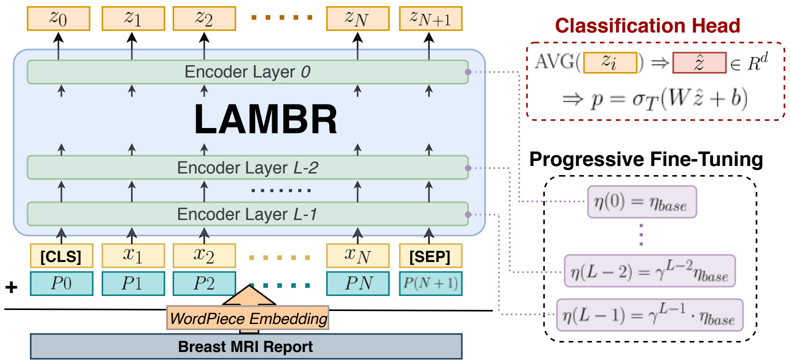


Fig. 4. An illustration of LAMBR encoding a breast MRI report. Tokens $\{x_i\}$ are generated from the report using WordPiece Embeddings, and the sum of the positional encoding and the token embedding are input to LAMBR. During TCFT, the learning rates for each layer are progressively tuned so that higher level features (layers near 0) are updated more compared to lower-level features. The classifier head takes the average of the token embeddings $\hat{z}$, applies an affine transform, and passes it into a tempered Softmax (σ_T) to generate the class probabilities.

token embeddings $\hat{z}$ is computed and is fed to an affine layer which undergoes a Tempered Softmax operation (σ_T) to obtain the outputs class probabilities p. Namely:

$$p = \sigma_T(W\hat{z} + b) \quad \leftarrow \quad \hat{z} = \frac{1}{N}\sum_{i=1}^{N} z_i \tag{1}$$

where $\hat{z}, b \in R^d$, $W \in R^{c \times d}$, and $p \in R^K$.

Progressive Fine-Tuning. Inspired by [17, 5], we propose a method for fine-tuning the complete text classifier. The learning rates are adjusted such that high-level features will be updated with a higher learning rate compared to lower level features. Specifically, for a Transformer-Encoder with L encoding layers $\{l_i\}_{i=0}^{L-1}$ (l_0 indicates the top-most layer), the layer-dependent learning rate $\eta(l)$ is formulated as:

$$\eta(l) = \eta_{base} \cdot \gamma^l \tag{2}$$

where η_{base} is the base learning rate and γ is the decay factor valued between 0 and 1. Similarly, the classifier head is updated with learning rate η_{base}.

Fine-tuning is performed by optimizing the weighted Label Smoothing Loss [11]:

$$L(x, y(x)) = -\sum_{c=1}^{K} w(c) \cdot \left[(1-\epsilon)y_c(x) + \frac{\epsilon}{K}\right] \cdot \log(p_c(x)) \tag{3}$$

where $w(c)$ are the weights for every class $c \in K$, $\epsilon \in [0, 1)$ is the smoothing term, $y_c(x) \in \{0, 1\}$ is 1 if x belongs to class c, and $p_c(x)$ is the probability of x belonging to class c as computed in Equation 1.

3 Experiments

In this section, we evaluate the proposed framework on two text classification tasks: (1) classifying whether the corresponding patient has been suggested to undergo biopsy and (2) predicting the BI-RADS score for the lesion reported.

The data is a curated list of medical reports from breast MRI examinations carried out at the Sheba Medical Center, Israel. Cases that were initially diagnosed as containing potential malignant tumors have all been suggested to undergo biopsy. Breast examinations from the years 2016-2019 were involved, and a total of 10,529 medical reports were collected. Of the 10,529 breast MRI reports, 541 reports were labeled with the relevant BI-RADS score for the (single) lesion reported, and each case was labeled with whether the patient had been suggested for biopsy.

3.1 Training Setup

Pre-Training. Pre-training was performed using DS-MLM as mentioned in Section 2.2. Of the 9,988 reports used for pre-training, 85% of the reports were randomly designated as the training set, and the remaining for validation. Cross Entropy loss was used for DS-MLM pre-training, and the multilingual BERT was trained for 70 epochs which took approximately 33 hours to complete using on an NVIDIA GTX 1070 8GB GPU.

Biopsy-Suggested Classification The goal of this task to identify whether the patient in the report had been suggested to undergo biopsy or not. We perform fine-tuning as proposed in Section 2.3. Due to dataset imbalance (26.6% of the cases were suggested for biopsy), class weights were set to the inverse of the counts per class. Evaluation was performed using 5-fold cross validation, and stratified sampling was applied to ensure equal class distribution between the training and validation sets.

Training was performed using the Adam optimizer [6] with a base learning rate of 1e-4 and a batch size of 8. Decay factor γ was set to 1/4, softmax temperature T was set to 1, and the smoothing term ϵ was set to 0. The best performing model from training for 70 epochs (approx 25 mins) was evaluated.

Masked BI-RADS Prediction Masked BI-RADS Prediction is a classification task to assign the appropriate BI-RADS score given the lesion description in the report. For reports that were parsed correctly, the BI-RADS score is written in the assessments, and a simple keyword-based tagging is often enough to label the reports with the appropriate score. However, reports might also contain

Fine-Tuning Tasks (5-Fold Average)	Accuracy	ROC-AUC	Macro Avg F1	MCC
Biopsy Suggested - LAMBR	0.965	0.989	0.935	0.913
Biopsy Suggested - BERT	0.949	0.987	0.904	0.8733
Biopsy Suggested - Baseline	0.828	-	0.718	0.6035
BI-RADS Prediction - LAMBR	0.8576	-	0.7158	0.7594
BI-RADS Prediction - BERT	0.795	-	0.572	0.672

Table 1. Several metrics following the five-fold cross validation for Biopsy Suggested Classification and Masked BI-RADS Prediction are presented. We compare the performance between LAMBR and BERT for both classification tasks (same classification head design, but different language representations). The baseline for Biopsy Suggested Classification is a keyword matching algorithm. Notice that in both tasks, LAMBR consistently outperforms their counterparts.

BI-RADS keywords that refer to previous BI-RADS scores (for the same lesion or removed lesion), which would lead to incorrect inference if the keyword based approach was used. In addition, errors encountered during parsing would occasionally miss out sections containing the BI-RADS score, rendering the keyword-based approach useless. The text classifier we propose in our framework relies on the report descriptions alone, and is thus robust against such potential obstacles.

Keyword search was performed on all the reports and any revealing BI-RADS scores (in the reports) were removed. This modified report was then fed into a pre-trained LAMBR for fine-tuning, and a 5-fold cross validation was performed. There were a total of 6 classes (no reports with BI-RADS score 5). Class weights were computed as the inverse of the class counts, and stratified sampling was performed to ensure equal class distribution between training and validation sets. Optimization was performed using the Adam optimizer with a base learning rate of 1e-4 and batch size of 8. The decay factor γ was set to 1/3, Softmax temperature T was set to $\sqrt{2}$, and the smoothing term ϵ was set to 1/3. The best performing model over a training period of 70 epochs was selected for evaluation.

3.2 Experimental Results

The experimental results for our proposed framework are listed in Table 1, and we include a comparison with BERT and a baseline algorithm.

In Biopsy Suggested Classification, the keyword matching algorithm aims to label each report in accordance with keywords that hint of a potential biopsy suggestion. Of the 541 labeled reports, 90 reports were misparsed, which contributes to a 16% drop in accuracy. In contrast, the classifier trained and fine-tuned using our proposed framework performs consistently across all five folds (see Figure 5) despite misparsed reports. We also trained a classifier with the same classification head from Section 2.3 using a multi-lingual BERT, and we demonstrate a better classification performance with our approach.

In the task of Masked BI-RADS Prediction, the classifier trained using our framework was able to correctly predict the BI-RADS score for most of the reports. Unlike the previous task where the BI-RADS score was available, this task requires the classifier to attend to relevant context clues in the medical

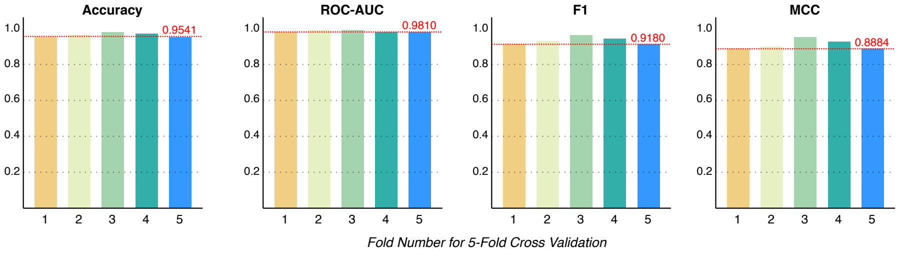


Fig. 5. Detailed visualization of the evaluation metrics for Biopsy Suggested Classification following the 5-fold cross validation. Notice the consistent performance across a 5 folds.

report for prediction (hence, the keyword-tagging algorithm does not work). An additional comparison was made between LAMBR and the pre-trained multilingual BERT, and the results in Table 1 demonstrate a clear difference the two language representations partake in training a BI-RADS classifier.

4 Conclusion

In this work, we explore the task of labeling breast MRI reports written primarily in Hebrew with occasional English texts through the use of multilingual language representations. To avoid the expensive pre-training required in obtaining a generalized language representation, the Domain-Specific Masked Language Modeling objective pre-trains a multilingual BERT on existing breast MRI reports alone to obtain the LAMBR language representation. A simple classification head is integrated onto the Transformer-Encoder, and Progressive Fine-Tuning is applied to train the classifier for its specific text classification task.

In our experiments, we train two separate classifiers to perform two classification tasks based on breast MRI reports. In the first task, we trained a classifier to determine whether the patient described in the report has been suggested to undergo biopsy. When compared with past methods, our approach demonstrates better classification performance despite parsing errors in a portion of the reports. In the second task, we trained a classifier to predict the BI-RADS score based on the lesion description in the report. Despite the absence of the BI-RADS score in the report, our classifier was able to infer the correct BI-RADS score in the majority of the cases.

Future works may include labeling medical reports for additional pathologies written in different languages. Additional tasks (apart from text classification) such as named-entity recognition for medical reports and summary generation for biomedical texts may also be investigated. In this work, we focus on the task of medical text classification, and we believe our proposed framework may assist in generating large numbers of labels for weakly supervised training tasks required for breast MRI CADx development.

References

1. Bien, N., Rajpurkar, P., Ball, R.L., Irvin, J., Park, A., Jones, E., Bereket, M., Patel, B.N., Yeom, K.W., Shpanskaya, K., Halabi, S., Zucker, E., Fanton, G., Amanatullah, D.F., Beaulieu, C.F., Riley, G.M., Stewart, R.J., Blankenberg, F.G., Larson, D.B., Jones, R.H., Langlotz, C.P., Ng, A.Y., Lungren, M.P.: Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. PLOS Medicine **15**(11), e1002699 (Nov 2018). https://doi.org/10.1371/journal.pmed.1002699
2. Chronopoulou, A., Baziotis, C., Potamianos, A.: An embarrassingly simple approach for transfer learning from pretrained language models (2019)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018)
4. Gozes, O., Frid-Adar, M., Greenspan, H., Browning, P.D., Zhang, H., Ji, W., Bernheim, A., Siegel, E.: Rapid ai development cycle for the coronavirus (covid-19) pandemic: Initial results for automated detection patient monitoring using deep learning ct image analysis (2020)
5. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification (2018)
6. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2014)
7. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics (09 2019). https://doi.org/10.1093/bioinformatics/btz682
8. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach (2019)
9. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. CoRR **abs/1802.05365** (2018)
10. Sharir, O., Peleg, B., Shoham, Y.: The cost of training nlp models: A concise overview. ArXiv **abs/2004.08900** (2020)
11. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision (2015)
12. Tsai, C.H., Kiryati, N., Konen, E., Eshed, I., Mayer, A.: Knee injury detection using mri with efficiently-layered network (elnet). ArXiv **abs/2005.02706** (2020)
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2017)
14. Wood, D.A., Lynch, J., Kafiabadi, S., Guilhem, E., Busaidi, A.A., Montvila, A., Varsavsky, T., Siddiqui, J., Gadapa, N., Townend, M., Kiik, M., Patel, K., Barker, G., Ourselin, S., Cole, J.H., Booth, T.C.: Automated labelling using an attention model for radiology reports of mri scans (alarm) (2020)
15. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Łukasz Kaiser, Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J.: Google's neural machine translation system: Bridging the gap between human and machine translation (2016)
16. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding (2019)
17. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? (2014)

תקציר

דימות באמצעות תהודה מגנטית גרעינית (MRI) הוא כלי מקובל להערכת פגיעות בברך. יכולתו לרכוש את מבנה הברך בתלת-מימד מקנה לו יכולת מצויינת באיתור קרעים אפשריים בברך. כדי להקל את עומס העבודה הגובר המוטל על רדיולוגיים מוסקו-סקלטליים (MSK), מתעורר צורך ממשי בכלי אבחון אוטומטיים למיון נפגעים, העשויים לצמצם עיכובים בפיענוח מקרים פתולוגיים. במחקר זה אנו מפתחים אלגוריתם לאבחון אוטומטי של פגיעות בברך, העשוי לסייע במיון נפגעים.

עבודתנו מתמקדת בתכנון רשת עצבית קונבולוציונית (CNN) המסוגלת לסווג תמונות MRI תלת-מימדיות בגודל משתנה. אנו מציגים רשת CNN חדשנית, המותאמת לזיהוי פגיעה בברך, אותה אנו מכנים Efficiently-Layered Network, ובקיצור ELNet. ELNet מתאפיינת במשקל קל ומשלבת מספר חידושים, כגון נורמליזציה רבת-שכבות (multi-slice normalization) ו- BlurPool. שיפורים אלה מאפשרים לאמן רשת ELNet מהתחלה בקלות.

אנו בוחנים את ELNet באמצעות שני מאגרים זמינים של תמונות MRI ברך : מאגר MRNet ומאגר KneeMRI. בשני המאגרים, ELNet השיגה ביצועים עדיפים, במגוון רחב של משימות גילוי פגיעה בברך, בהשוואה לשיטת MRNet שעמדה עד כה בחזית המחקר (SOTA). בנוסף, השווינו את יכולתה של השיטה המוצעת לאיתור המיקום של קרעים בברך ליכולתו של רדיולוג MSK, ובמרבית המקרים הממצאים שהציגה השיטה המוצעת דמו מאד לאלה שציין הרדיולוג.

רשת ELNet היא ארכיטקטורה קלת-משקל ובעלת ביצועים איתנים. היא עשויה להוות בסיס מבטיח למיון פגיעות בברך. בנוסף, ניתן להרחיב את הרשת לאבחון פתולוגיות נוספות בהתבסס על תמונות דימות תלת-מימדיות.

אוניברסיטת תל – אביב

הפקולטה להנדסה ע״ש איבי ואלדר פליישמן
בית הספר לתארים מתקדמים ע״ש זנדמן-סליינר

גילוי ואיתור פגיעות בברך באמצעות תמונות MRI תלת-מימדיות

חיבור זה מוגש כעבודת מחקר לקראת התואר ״מוסמך אוניברסיטה״ בהנדסת חשמל ואלקטרוניקה

על ידי

צ׳ן-חאן טסי

עבודה זו נעשתה באוניברסיטת ת״א בבית הספר להנדסת חשמל
בהנחיית פרופ׳ נחום קריתי ודר׳ ארנלדו מאייר

אב תש״פ

אוניברסיטת תל – אביב

הפקולטה להנדסה ע״ש איבי ואלדר פליישמן
בית הספר לתארים מתקדמים ע״ש זנדמן-סליינר

גילוי ואיתור פגיעות בברך באמצעות תמונות MRI תלת-מימדיות

חיבור זה מוגש כעבודת מחקר לקראת התואר ״מוסמך אוניברסיטה״ בהנדסת חשמל ואלקטרוניקה

על ידי

צ׳ן-חאן טסי

אב תש״פ