uc3m | Universidad Carlos III de Madrid

Master Degree in Information Health Engineering at UC3M
Academic Year 2019-2020

Master Thesis

Automatic Diagnosis Of Renal Pathologies Using Kidney Ultrasound Imaging

María Postigo Fliquete

Iván González Díaz
Fernando Díaz de María
Madrid, October 2020

Automatic Diagnosis Of Renal Pathologies Using Kidney Ultrasound Imaging

María Postigo Fliquete

Abstract

Renal pathologies have become a major challenge for health systems. To avoid end-stage diagnosis, early detection and prevention is required. This paper proposes a CAD system for kidney ultrasound images that will help to different stages of the health system (primary care and specialists). The proposed system incorporates several approaches including: (i) A healthy/pathological binary classification as a mechanism for filtering patients in primary care, (ii) a multi-pathology diagnosis by ultrasound images classification, (iii) a multi-pathology diagnosis through local pathology detection and (iv) a multi-pathology diagnosis through a fusion mechanism that integrates the previous two.

1. Introduction

The kidneys are a pair of bean-shaped organs located at the bottom of the rib cage which form part of the urinary system. They are mainly responsible for filtering blood impurities. Therefore, they are essential to have a healthy body. When the kidneys don't function properly, the body retains toxins and can fail to make enough red blood cells, which carry oxygen to organs and tissues. If the kidneys stop working completely, dialysis will be required to purify the blood, but it won't cure the disease (Yang et al., 2020). Renal pathologies have become a major challenge for health systems, particularly in lower-middle-income countries and almost 2 million people worldwide are in urgent need of dialysis to survive. Moreover, the number of patients on renal replacement therapy has doubled every decade since 1980. The World Health Organization (WHO) estimates that only 10% of those in need undergo kidney transplantation annually (Król et al., 2008). To avoid end-stage kidney disease diagnosis, early detection and prevention is required (Kokil & Sudharson, 2019b). The identification of early stages of the condition means that primary care clinicians play an essential role. Good primary care judgment is also critical in making decisions about referral for specialist nephrology opinion (Fraser & Blakeman, 2016).

Ultrasound imaging (US) is routinely used as the first line of medical imaging and is one of the core diagnostic imaging modalities. The key advantages of this technique include real-time imaging, non-ionizing radiation, better cost effectiveness, portability and the use of conventional electrical power sources. However, US presents several challenges, such as noise, artifacts, limited field of view and operator dependence. This last one is particularly limiting since the lack of skill in acquiring and interpreting the images in primary care leads to significant challenges in clinical decision making. As a result of the high intra-operator variability, automated US image analysis promises to play a critical role facing early stages of renal pathologies (Brattain et al., 2018).

To help clinicians to make decisions, Computer-Aided Diagnosis (CAD) has received increased attention in recent years. In fact, with the rapid development of COVID-19 into a global pandemic, an important CAD application is to automatically detect COVID-19 from Lung Ultrasound Imaging or Chest Radiological Imaging, such as Computed Tomography and X-ray (Born et al., 2020; Ozturk et al., 2020). For early renal pathologies detection, a CAD system can work as a "second opinion" and assist primary care, or even specialist nephrologists.

CAD technology includes machine learning, deep learning, computer vision and medical image processing. Machine and deep learning algorithms are rapidly growing in medical imaging research. As soon as it was possible to load medical images into a computer, researchers have built systems for automated analysis (Litjens et al., 2017). In this context, biomedical imaging techniques together with computer vision approaches have a great potential to complement the conventional diagnostic techniques (Latif et al., 2019).

This paper is focused on automated diagnosis of renal pathologies from kidney ultrasound images. The problem is addressed with a CAD system that brings together several approaches: (i) A binary global image classification (healthy vs pathological), (ii) A multi-label global image classification (healthy vs 6 different pathologies), (iii) A multi-label classification based on Local Object Detection, (iv) An hybrid classification by merging (ii) and (iii) approaches. Each system is designed to assist clinicians in different scenarios. In the experimental section the performance of the different approaches in the different scenarios will be discussed.

The rest of the paper is organized as follows. The related

work is presented in Section 2 from two points of view: deep learning for visual recognition (2.1) and deep learning for medical ultrasound imaging together with CAD systems (2.2). Materials and methods are exposed in Section 3, where we present the dataset (3.1) and the different approaches for the proposed CAD system (3.2). In section 4 the experimental setup and results for each approach are made available. Moreover, we provide a comparison among the different proposed methods. Finally, in Section 5 we conclude with the proposed CAD system impact and lines for future work.

2. Related Work

In this section the relevant work related with the given topic (Automatic Diagnosis Of Renal Pathologies Using Kidney Ultrasound Imaging) will be described.

2.1. Deep Learning for visual recognition: image classification and object detection

In various research fields, especially in image analysis and computer vision, deep learning has emerged as the leading machine learning tool (Liu et al., 2019). Deep learning is a representation learning approach that can automatically learn mid-level and high-level abstract features from raw data (e.g. US images). Over the past few years, deep learning has led to an improvement of the performance on a variety of problems in computer vision, such as image classification, segmentation, image retrieval and object detection (Guan & Huang, 2018).

Starting from AlexNet (Krizhevsky et al., 2012), a Convolutional Neural Network (CNN) that won the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC), deep learning began to stand out in the area of machine learning (Russakovsky et al., 2014). A few years later, it was consolidated as the leading machine learning tool in various research domains, fundamentally in general image analysis and computer vision. Up to now, deep learning has gained rapid development in terms of network architectures or models, such as GoogleNet (Szegedy et al., 2014), ResNet50 (He et al., 2015) and EfficientNet-B0 (Tan & Le, 2019) among others, demonstrating that is a state-of-the-art tool for automated analysis tasks.

The mentioned architectures are used to address different computer vision problems, such as image classification. For object detection, which is more challenging than image classification, Faster R-CNN is a widely used architecture. It combines backbone networks (e.g. ResNet-50) with *Region Proposal Networks* (RPNs) that simultaneously predicts object bounds and objectness scores at each position in the image (Ren et al., 2015). This region proposals are used by Fast R-CNN's Region of Interest (RoI) pooling layer to offer a solution that detects and locates objects in an image (Girshick, 2015).

2.2. CAD and Deep Learning in Medical US Imaging

Computer-aided detection (CADe) or computer-aided diagnosis (CADx) in medical imaging is the computer-based system that helps clinicians to make decisions. Medical image analysis is crucial to diagnose disease in early stages non invasively. Some CAD systems have been used to assist doctors to detect breast cancer (Dromain et al., 2012), lung nodules (Shariaty & Mousavi, 2019), vertebral fractures (Kasai et al., 2006) or intracranial aneurysms (Arimura et al., 2004).

Current applications of deep learning in US imaging mainly involve classification, detection and segmentation of different anatomical structures and tissues, including breast (Hiramatsu et al., 2017; Bian et al., 2017), prostate (Shi et al., 2016; Yang et al., 2016), liver (Wu et al., 2014), heart (Pereira et al., 2017) and fetus (Qi et al., 2017; Gao & Noble, 2017).

To date, several articles have been written about deep learning applied to medical imaging. However, few focus on medical US (Liu et al., 2019). Moreover, as we can appreciate in Figure 1, deep learning has been scarcely applied in kidney US images. The most common application in kidney US images is kidney segmentation (Yin et al., 2019).

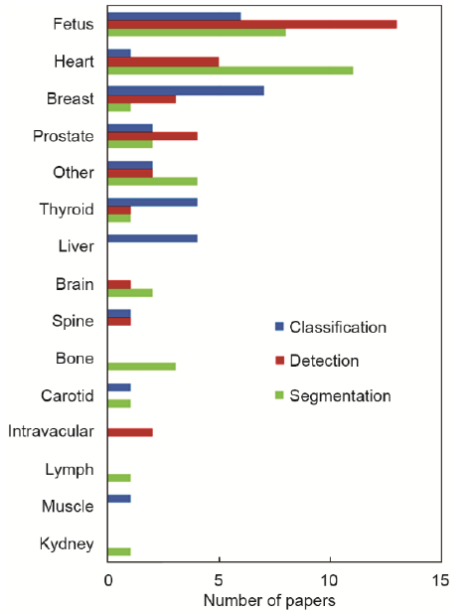

Figure 1. Applications of deep learning in medical ultrasound analysis (Liu et al., 2019)

Nevertheless, there are some methods that have been used for detecting kidney abnormalities. The identification of

kidney stone in (Verma et al., 2017) where the features are extracted with principal component analysis and classified by K-Nearest Neighbour (KNN). Renal caluli detection using meta-heuristic Support Vector Machines (SVM) in (Selvarani & Rajendran, 2019). An automatic abnormality detection system by off-the-shelf CNN features in (Kokil & Sudharson, 2019b) where a SVM classifier is used to classify kidneys in three categories. Automatic detection of renal abnormalities using ensemble Multiple Support Vector Machines (MSVM) classification model in (Kokil & Sudharson, 2019a). The application of an ensemble of Deep Neural Networks (DNNs) for kidney ultrasound image classification in (Kokil & Sudharson, 2020) where the predictions of multiple DNNs are combined to detect four categories of kidney images.

The state-of-the-art methods tackle a multi-class problem (the different classes are mutually exclusive) considering between 2 to 4 classes in the diagnosis (healthy and 1-3 pathologies). Our work provides a multi-label diagnosis between 7 classes (healthy and 6 pathologies) where the pathologies are not mutually exclusive. Moreover, our multi-label classification method is an hybrid combination of image classification and object detection scores. To the best of our knowledge, we are the first to validate this approach in seven category kidney US images.

3. Materials and Methods

In this section, we introduce the data used for this research, as well as its preprocessing and the object detection and classification methods applied.

3.1. Data Description

The Nephrology Department from *Hospital Ramón y Cajal* provided us with a set of 1985 renal ultrasound images, consisting of 450 healthy and 1535 pathological kidneys. The pathological ones presented different pathologies manually annotated by two experienced nephrologists from the hospital. The distribution of the different pathologies in the dataset is visible in Figure 3. Each image had its corresponding segmentation mask of the kidney and the bounding box coordinates of the local lesions. An example of the annotation is visible in Figure 2.

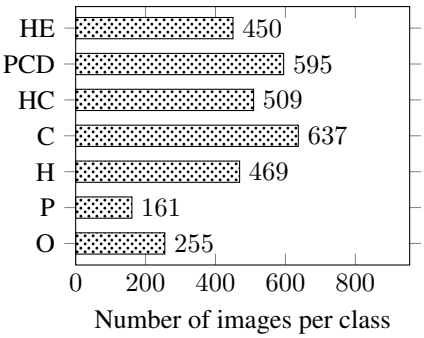

Figure 3. Number of images per class/pathology in the given dataset. (**HE**-Healthy, **PCD**-Poor Corticomedullar Differenciation, **HC**-Hyperechogenic Cortex, **C**-Cyst, **H**-Hydronephrosis, **P**-Pyramid, **O**-Others)

3.2. Proposed CAD System

This study aims to propose a CAD system that integrates several approaches: a binary image classification (healthy vs pathological kidney), a multi-label global image classification (involving healthy kidney and 6 different pathologies), local pathologies detection, and an hybrid classification by merging multi-label classification and local detection. Each of these approaches will be explained in detail in the following sections. An overall view of the classification methods is visible in Figure 4.

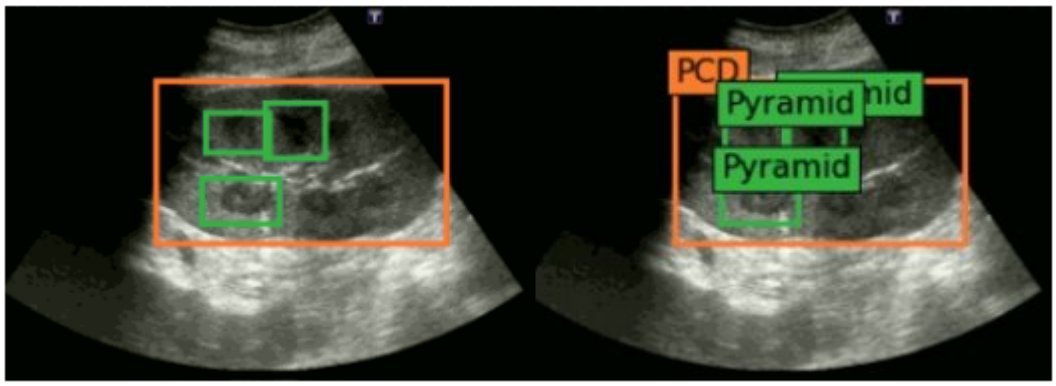

Figure 2. Annotation example. The image presents Poor Corticomedullar Differenciation (PCD) and Pyramids.

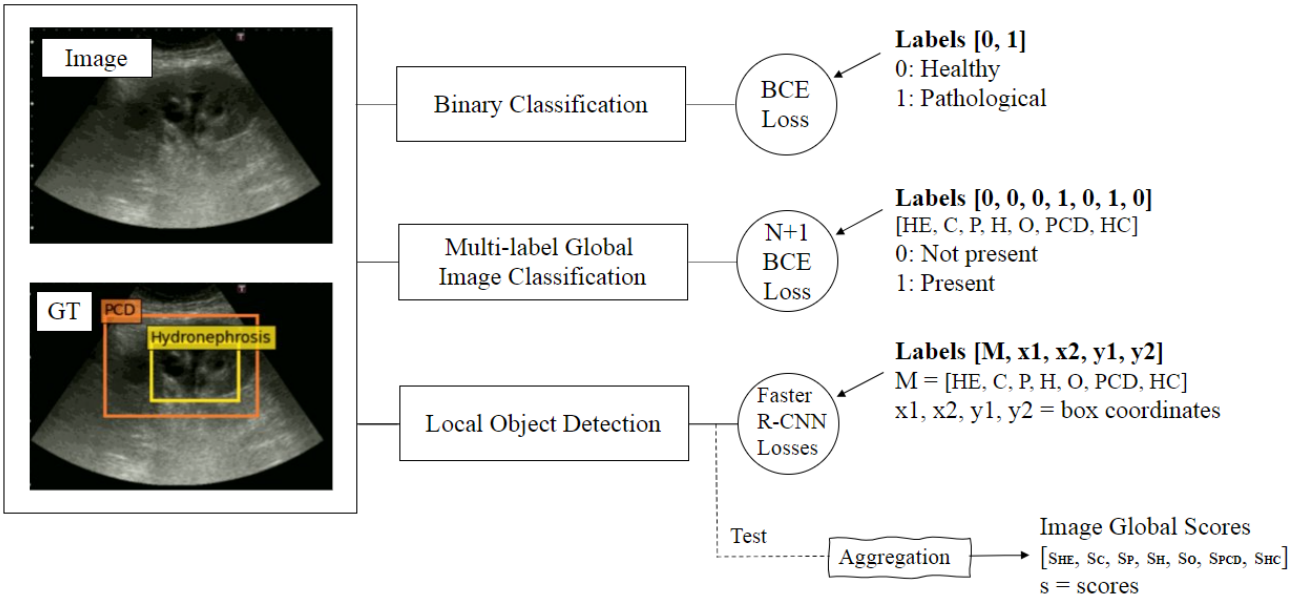

Figure 4. Schemes of binary classification, multi-label global image classification and local object detection for the proposed CAD system. Each approach is trained with the set of images, specific labels and different losses. The local object detection outputs go through an aggregation mechanism resulting in multi-label global image scores.

3.2.1. Binary Classification

The main goal of this binary classification is to provide the primary care with a simple filtering mechanism that allows only those patients with a high probability of pathology to be referred to the specialist.

All images $X = \{\overrightarrow{x}_1, ..., \overrightarrow{x}_N\}, \overrightarrow{x} \in \mathscr{X}$, (where $\overrightarrow{x}_i$ represents each single image and $i \in \{1, ..., N\}$ where N is the number of images) are associated with a ground truth label y_i, with $\overrightarrow{y} \in \{0, 1\} = \mathscr{Y}$. The healthy kidneys are labelled as 0 whilst the pathological ones are labelled as 1. Due to the fact that the size of the samples is small, transfer learning is a preferable strategy to train the deep CNNs. Transfer learning is a methodology applied when training data is expensive or difficult to collect. It consists of obtaining high-performance learners trained with more reachable data from different domains. This approach makes possible the retention of the knowledge extracted from one task to perform an alternative function (Weiss et al., 2016).

We will use this first approximation to evaluate the performance of several backbone networks to establish a baseline to train the other two methods (multi-label classification and local detection). The trained models were AlexNet, ResNet-18, ResNet-50, ResNet-101, EfficientNet-B0, EfficientNet-B1 and EfficientNet-B2. The different models were trained with different image transformations (for data augmentation) and different hyperparameters setting (batch size, learning rate and epochs). As we will show in the experiments section (Section 4), Resnet-50 offers an optimal compromise between performance and complexity to classify the images.

Since the dataset is highly umbalanced, we used a weighted random sampler to prevent class overfitting. The model was trained with the Binary Cross Entropy (BCE) Loss criterion, the Stochastic Gradient Descent (SGD) optimizer and the `StepLR` scheduler. The results of some of the given models will be exposed in Section 4.

3.2.2. Diagnosis of multiple pathologies based on Global Image Classification

The multi-label classification will provide a more comprehensive diagnostic of the different renal pathologies which may be useful in both primary care and consultation with the specialist.

The classification of the different pathologies was tackled with ResNet-50. Since the pathologies are not mutually exclusive, i.e. several abnormalities may appear at the same time in a clinical case, we address the problem as a series of binary classifications (one per pathology) instead of a multiclass classification. Therefore, the ground truth labels are now encoded as a binary vector $\overrightarrow{y} \in \{0, 1\}^M = \mathscr{Y}$. This is a classification task with $M = 7$ labels from 7 possible classes (healthy and 6 pathologies) where the possible classes are not mutually exclusive except the healthy label.

In order to adapt the ResNet-50 pretrained model to the new task, we replaced the last dense layer of the original architecture with a new dense layer matching the number of labels.

Since the classification is multi-label, the loss criterion employed during the training was the BCE with Logit Loss, which combines a *Sigmoid* layer and the *BCELoss* in one single class. As in the binary classification, the optimizer was SGD followed by a `ReduceLROnPlateau` scheduler.

3.2.3. Diagnosis of multiple pathologies based on Local Object Detection

Object detection provides labels that locate the pathology in the kidney US image. We believe that this information can help to improve the diagnosis, especially in very local pathologies present in small areas of the kidney, which could go unnoticed (being hidden or blurred) in global approximations like the previous ones.

This is a more complex task since it needs to locate and classify existing objects in any image. Therefore, apart from the classification difficulty of the previous approaches (binary and multi-label classification) we have to locate the lesion areas (Zhao et al., 2019).

3.2.3.1. Local detection of pathologies

To target the detection task we applied the most common object detection architecture, known as Faster R-CNN (Ren et al., 2015). This system is conformed by two modules: a deep fully CNN for region proposal and the Fast R-CNN detector (Girshick, 2015) that uses the proposed regions, and has proved to be efficient in many different tasks, for example face detection (Jiang & Learned-Miller, 2017), vehicle detection (Fan et al., 2016) or malaria parasytes detection (Hung et al., 2018).

The Region Proposal Network (RPN) takes an image as input and outputs a set of rectangular object proposals, each with an objectness score. Then the Fast R-CNN is trained to predict class-specific scores, generating as output some labeled and scored bounding boxes. To obtain the results of the object detector, a set of steps have been followed:

1. Train the Faster R-CNN model and obtain the detection outputs. The detection outputs are composed by bounding boxes coordinates, with a class label and a confidence score associated.
2. Discard low confidence outputs. Different scenarios have been considered as low confidence outputs:
   (a) For the healthy class, since it is mutually exclusive, if any of the pathological scores is greater than 0.7, the healthy output is discarded. Moreover, if the previous condition is not met, the pathological outputs will be discarded if the healthy score is greater than 0.9.
   (b) Since we could appreciate that PCD is hard to detect, those outputs for this global pathology with a score greater than 0.2 will be considered. The rest of pathologies need to overcome the 0.5 confidence level to be contemplated as a detection.
3. The mean Average Precision (mAP) per class has been computed to measure the detector performance. It is an average precision (AP) score, where the true positives (TP) are those bounding boxes that have an Intersection over Union (IoU) greater than a given threshold with the ground truth (GT) bounding box and the correct class label. The false positives (FP) are those boxes that don't overcome the IoU threshold or don't match the GT class label.

3.2.3.2. Detections' aggregation into a global diagnosis

In addition to the local detection, a global image-level diagnosis score is proposed by applying an aggregation mechanism. The goal of this step is therefore to aggregate, for every image, the scores of the detected bounding boxes into a vector, providing a global image classification.

The equations of the aggregation methods are shown below. Here, k represents each class $k \in \{\text{HE, C, P, H, O, PCD, HC}\}$; B is the number of bounding boxes detected; s, w and h are the bounding boxes' scores, width and height respectively; H and W are the image's height and width.

- The *Max* aggregation: (Equation 1) sets the maximum bounding box's score per class as class specific score. With this mechanism we consider the highest confidence scores from Faster R-CNN to address the multi-label classification based on local detections.

$$\text{Max}_k = \max_i (s_{ik}) \tag{1}$$

- The *Area* aggregation: Since small local detections have a higher probability of being false positives whilst larger objects are more reliable, we opted to consider the bounding boxes' areas (Equation 2) as another aggregation. It establishs the sum of the bounding boxes areas multiplied by their scores as class scores.

$$\text{Area}_k = \frac{1}{HW} \sum_i^{B_k} s_{ik} w_{ik} h_{ik} \tag{2}$$

- The *Sum* aggregation: Many images can present the same pathology several times. For example, if an image presents several cysts, it must be labeled as a cyst image with a high probability. The *Sum* aggregation (Equation 3) determines each class specific score as the sum of all the bounding boxes' scores per class.

$$\text{Sum}_k = \sum_i^{B_k} s_{ik} \tag{3}$$

- The *LSE* aggregation: (Equation 4) calculates the Log-SumExp function of the boxes' scores per class as class specific scores. This allows to reinforce the high confidence detection scores. It can be seen as a soft approximation to the max operator.

$$\text{LSE}_k = \log\left(\sum_i^{B_k} \exp(s_{ik})\right) \tag{4}$$

- The *Mean* aggregation: (Equation 5) sets the mean of the bounding boxes' as the class score for each category.

$$\text{Mean}_k = \frac{1}{B_k}\sum_i^{B_k} s_{ik} \tag{5}$$

A summary of the different aggregation methods is shown in Figure 5.

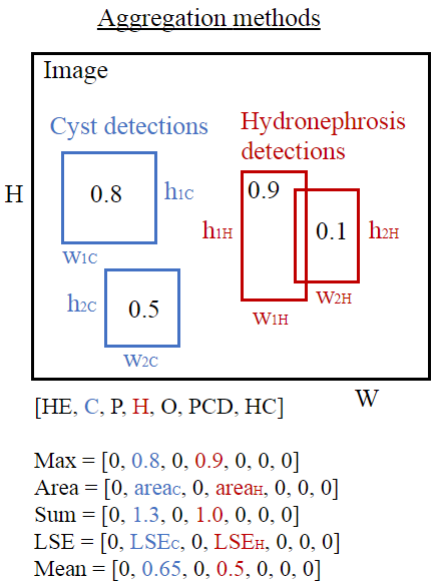

Figure 5. Example of how each aggregation method works.

3.2.4. Hybrid Multi-Label Classification

This hybrid method aims to enhance the different renal pathologies diagnosis. By combining the global with the local classification, the approach can benefit of both general and particular features in the images.

Therefore, the multi-label classification scores are refined by merging the global multi-label classification in 3.2.2 and the local detection aggregation mechanism in 3.2.3.2. This fusion method provides a CAD system that exploits the benefits of each approach. Both classification scores are linearly combined as in Equation 6, where α was obtained from a validation set.

$$s_{hybrid} = s_{global} + \alpha s_{local} \tag{6}$$

As it can be seen in Figure 6, the hybrid classification is a simple workflow where the outputs from the global classification are combined with the scores from the aggregation.

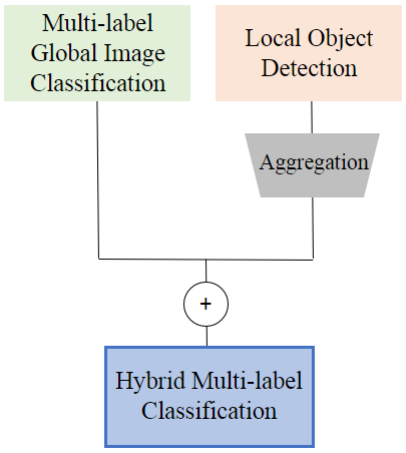

Figure 6. Proposed method for hybrid multi-label classification.

4. Experimental setup and evaluation metrics

In this section, the obtained global classification and local object detection results from different models will be discussed.

Each task was evaluated over 5 different folds. The train, test and validation sets were conformed by 1191, 397 and 397 images respectively. Even though several transformations for data augmentation were tested, the best performance was reached by only applying an Horizontal Flip on the images.

Different metrics were applied in the evaluation of the models:

- Area Under the Receiver Operating Characteristic Curve (ROC AUC): The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) and the AUC is the area under the ROC curve. It tells how much a model is capable of distinguishing between classes (Narkhede, 2018).
- Average precision (AP): summarizes a precision-recall curve as the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight. The precision is the $TP/(TP+FP)$ ratio and the recall is the $TP/(TP+FN)$ ratio where TP is the number of true positives, FP the number of false positives and FN the number of false negatives (Hui, 2018).

- Specificity 95 (SP-95): the specificity is the metric that evaluates a model's ability to predict true negatives of each available category. It is the ratio $TN/(TN+FP)$ where TN is the number of true negatives. The specificity 95 is the specificity value when the sensitivity is 0.95, which is the percentage of healthy detected as pathological (percentage of FP) when detecting 95% of the TP (Mitrani, 2019).

4.1. Healthy vs Pathological Binary Classification

Finetuning and testing pretrained models showed that ResNet-50 is the best model to address kidney ultrasound image classification. The batch size was set to 32 and the learning rate $lr = 0.01$.

Some of the tested architectures' results can be seen in Table 1. This table shows the different evaluation metrics applied to the binary classification results (each image was labeled as 0-healthy vs 1-pathological).

Backbone		Healthy vs Pathological
AlexNet	AUC	0.7883
	AP	0.9258
	SP-95	0.2681
ResNet50	AUC	**0.8472**
	AP	**0.9505**
	SP-95	0.3363
EfficientNetB0	AUC	0.8281
	AP	0.9422
	SP-95	**0.3451**
EfficientNetB1	AUC	0.8225
	AP	0.9391
	SP-95	0.3121

Table 1. Different metrics in Healthy vs Pathological binary classification. The best scores are highlighted in bold.

We achieve good results for the binary classification, with a 0.8472 AUC score. We can appreciate that the best AUC and AP is obtained with ResNet-50. Even if EfficientNet-B0 shows a slightly better specificity, we concluded that ResNet-50 is a good backbone for the multi-label and object detection approaches.

The SP-95 indicates that when detecting the 95% of the pathological kidneys, a 34% of the healthy kidneys will be diagnosed correctly as healthy. But the 66% will be misclassified as pathological (False positive). Given the nature of the problem, is preferable to obtain a FP rather than a FN, since missing a pathological detection supposes the development of the disease without treatment (exactly what we aim to avoid). Supposing that every symptomatic patient is referred to an specialist by the primary care, with this CAD system, 2 out of 5 healthy patients referrals could be avoided. But there is still room for improvement in this classification.

4.2. Diagnosis of multiple patologhies based on Global Image Classification

In this approach, the ResNet-50 model was trained receiving binary vectors indicating the presence/absence of each considered category. Here we aim to solve a more complex problem, the diagnosis of 6 different pathologies and healthy kidneys.

Table 2 illustrates the different metrics for each of the 7 categories. We can appreciate that the classification between classes with this architecture works properly, with an average AUC of 0.793.

Global	HE	PCD	HC	C	H	P	O	AVG
AUC	0.83	0.76	0.78	0.76	0.89	0.82	0.71	0.79
AP	0.59	0.57	0.58	0.66	0.78	0.48	0.31	0.57
SP-95	0.30	0.25	0.29	0.21	0.21	0.44	0.19	0.27

Table 2. Multi-label classification metrics based on Global Image Classification.

The performance of this global classification is notable for every category. Some pathologies are better classified than others. Those pathologies with bigger size, more sample images or with characteristic features are more likely to be better classified in this global approach.

The "Hydroneprosis" is well classified since is very descriptive at image level and its size is considerable. The "Pyramids" are small pathologies, but, due to their shape, they are descriptive enough to be properly classified. Moreover, this class presents the highest SP-95 in this approach (0.44). For the "Cysts" there are several images, so the AUC should be better, but probably a global image classifier missed very local small cysts. The classification of those categories involving the whole kidney (HE, PCD and HC) are favoured in this approach since they are present in a big area of the image. It seems that the healthy's features stand out more than those from either PCD or HC. The "Others" classification is the worst since this class involves different unrelated pathologies that we could not classify separately due to the low amount of samples.

Improvement of this classification is addressed in the following subsections.

4.3. Diagnosis of Multiple pathologies based on Local Object Detection

In this section, the mentioned neural network with "attention" mechanisms (Faster R-CNN) is used for object detection and for a multi-label classification based on local detection.

4.3.1. Object Detection approach

The Faster R-CNN was trained for 30 epochs, with $lr = 0.005$ and batch size 1. To evaluate the accuracy of the object detector the IoU threshold was set to 0.3, because some GT boxes with the same label are predicted as a single big bounding box involving the GT boxes (specially for hydronephrosis). We can appreciate that the object detection does not work as well as the multi-label classification, in part due to the bounding box overlapping requirement. The mAP scores are visible in Table 3. These scores are low because the amount of false positives is high, as we can appreciate in Figure 7. Nevertheless, the accuracies for some of the pathologies are acceptable: for example in Hyperechogenic Cortex 416 out of 509 GT bounding boxes are detected, which is an accuracy of 0.817. Without discarding the low confidence outputs, the majority of the GT bounding boxes are detected, but the ammount of false positives per class is excessive.

	HE	PCD	HC	C	H	P	O
mAP	0.45	0.13	0.52	0.45	0.59	0.23	0.01

Table 3. Per categpry mAP scores per class

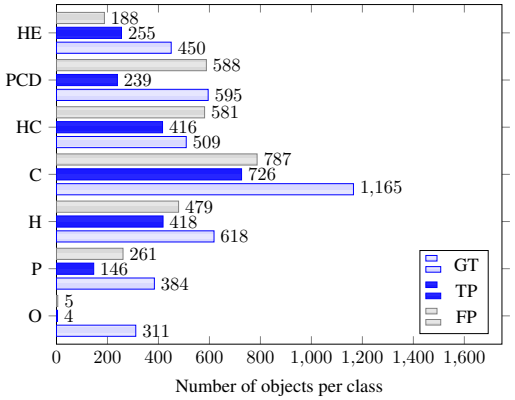

Figure 7. Ground Truth (GT), True Positives (TP) and False Positives (FP) object detections for the different pathologies

4.3.2. Multi-label classification based on Local Object Detection

Faster R-CNN architecture generates from a given image a set of rectangular object proposals, each with a class specific label and a confidence score. The aggregation of these scores provides a local classification of the different pathologies present in the kidney US images. The average AUCs for each aggregation method is visible in Table 4.

The multi-label classification based on Local Object Detection (average AUC = 0.7825) is not better than multi-label global classification (AUC = 0.793).

	Max	Area	Sum	LSE	Mean	GIC
AUC	**0.7825**	0.754	0.776	0.753	0.754	**0.793**

Table 4. Average AUC scores for the different aggregation methods for local classification. The Global Image Classification (GIC) average AUC is also present for comparison.

The average AUC of both global and local classifications are very similar, with a difference of less than a 1%.

In Table 5 we can see the different metrics for each category with the *Max* aggregation. Some classes have benefited from this approach. The "Pyramids", "Hydroneprhosis" and "Cysts" scores are better in this classification since this method pays attention to different regions in the image instead classifying the image as a whole. Moreover, even if the "Healthy" and the "HC" classes involve the entire kidney, they take advantage of this approach, especially the "Healthy", with a considerable increase in the specificity, from 0.3 to 0.56. The "PCD" and "Others" do not take advantage of this approach. Probably the model is having problems to distinguish between the 3 categories that involve the entire kidney (HE, PCD, HC). Since the "Others" category is not very descriptive, we did not expected any improvement in this approach.

Max Local	HE	PCD	HC	C	H	P	O	AVG
AUC	**0.87**	0.47	**0.80**	**0.83**	**0.92**	**0.88**	0.70	0.78
AP	**0.66**	0.29	0.58	**0.71**	**0.84**	0.46	0.28	0.55
SP-95	**0.56**	0.03	**0.32**	**0.38**	**0.59**	**0.56**	0.0	0.35

Table 5. Multi-label classification metrics based on Local Object Detection with *Max* aggregation. Those scores that overcome the previous global image classification are highlighted in bold.

Considering that both classification exploits different features from the images, the combination of both approaches is promising.

4.4. Hybrid Multi-label Classification

Since local object detection for multi-label classification is not better than global multi-label classification for all the pathologies, both output scores are combined to reach an improved categorization.

The hybrid score is obtained by the linear combination of both classification scores previously shown in Equation 6, with $\alpha = 1.58$. The mean AUC score for the different aggregations can be seen in Table 6, where a considerable improvement can be appreciated from local to hybrid classification.

The best aggregation for the hybrid approach is also the *Max* aggregation. Table 7 shows different metrics of each class.

Aggregation	Local AUC	Hybrid AUC
Max	**0.7825**	**0.8366**
Area	0.7540	0.8109
Sum	0.7762	0.8242
LSE	0.7528	0.8199
Mean	0.7539	0.8250

Table 6. Average AUC scores for the different aggregation methods for local and hybrid multi-label classifications.

We can appreciate that every category is better classified in this approach, taking advantage of both global and local classifications.

Max Hybrid	HE	PCD	HC	C	H	P	O	AVG
AUC	0.89	0.72	0.83	0.85	0.93	0.90	0.75	0.84
AP	0.68	0.52	0.64	0.76	0.86	0.58	0.37	0.63
SP-95	0.60	0.18	0.37	0.41	0.61	0.56	0.17	0.41

Table 7. Multi-label classification metrics based on Hybrid approach with *Max* aggregation.

Due to the given improvement obtained with the hybrid method in the multi-label classification, we considered that either the global, local or hybrid classifications could overcome the binary classification. Therefore, the outputs of these approaches were transformed to binary outputs as in the binary problem (healthy vs pathological). The healthy score is the healthy score from the multi-label vector and the pathological score is the maximum pathological score in the vector.

Table 8 shows a global overview of every approach discussed over the binary and multi-label problem. We can see that the improvement given by the hybrid version is noticeable better, both AUCs and SP-95 overcome the other approaches.

	Multi-label AUC	Binary AUC	Binary SP-95
Binary	-	0.8472	0.3363
Global	0.793	0.8401	0.3378
Local	0.7825	0.8904	0.4667
Hybrid	**0.8366**	**0.9017**	**0.4911**

Table 8. Global overview of the proposed methods for multi-label and binary classifications.

We can see that, for the binary classification, the performance of both local and hybrid classifiers is similar. However, the binary SP-95 from the combination is considerably better. Moreover, the hybrid method provides a significant improvement in the multi-label classification.

The combination of both local and global classifications leads to an enhancement in both binary and multi-label classifications, since each approach contributes in different aspects.

5. Conclusions and future work

The identification of renal pathologies on their early stages is essential to avoid the development of the disease. Therefore, good primary care judgment is critical. However, the lack of skill in adquiring and interpreting kidney US images in primary care leads to significant challenges in clinical decision making.

In this paper we propose a CAD system for kidney ultrasound images, which may help both primary care and specialist nephrologists, acting as a "second opinion". Within the different approaches proposed, the combination of a global and a local image classification results in an AUC = 0.9017 with a SP-95 = 0.4911. This is a high confidence CAD system that can be exploited in primary care. On the assumption that every symptomatic patient is referred to an specialist, with this system, 3 out of 5 healthy patients referrals could be avoided.

In addition to prevent the patients' overload for the specialists, the proposed CAD system is a good multi-pathology classifier that can help nephrologists in decision making, with a multi-label AUC = 0.8366.

Although the classification obtained in this paper is an outstanding contribution to first line kidney diagnosis, several improvements are possible and should be considered. When more data becomes available, training the models only with kidney US images could lead to an improvement since medical images have different characteristics from normal images. This training may help detecting ultrasound specific patterns. Moreover, since we have obtained a good image classifier from local and global classifications, taking advantage of this combination for the local object detection seems to be an appealing future work.

Acknowledgements

I would like to extend my sincere thanks to *Universidad Carlos III* and its Signal Theory and Communications department to allow this project to happen. I am very grateful to the project tutors Iván González Díaz, Fernando Díaz De María and Miguel Molina Moreno for helping me during the study. Thank you to Ignacio Serrano Llabrés and Alberto Gutiérrez Cantera for looking through this paper so meticulously. I am also thankful to Jaime Yuste Muñoz for his encouragements and wholehearted support. Finally I would like to express my gratitude to the Master in Information Health Engineering for giving me most of the knowledge that I have applied in this project.

References

Arimura, H., Li, Q., Korogi, Y., Hirai, T., Abe, H., Yamashita, Y., Katsuragawa, S., Ikeda, R., and Doi, K. Development of cad scheme for automated detection of intracranial aneurysms in magnetic resonance angiography. volume 1268, pp. 1015–1020, 06 2004. doi: 10.1016/j.ics.2004.03.102.

Bian, C., Lee, R., Chou, Y.-H., and Cheng, J.-Z. Boundary regularized convolutional neural network for layer parsing of breast anatomy in automated whole breast ultrasound. pp. 259–266, 09 2017. ISBN 978-3-319-66178-0.

Born, J., Brändle, G., Cossio, M., Disdier, M., Goulet, J., Roulin, J., and Wiedemann, N. Pocovid-net: Automatic detection of covid-19 from a new lung ultrasound imaging dataset (pocus), 2020.

Brattain, L., Telfer, B., Dhyani, M., Grajo, J., and Samir, A. Machine learning for medical ultrasound: status, methods, and future opportunities. *Abdominal Radiology*, 43, 02 2018. doi: 10.1007/s00261-018-1517-0.

Dromain, C., Boyer, B., Ferré, R., Canale, S., and Delaloge, S. Computed-aided diagnosis (cad) in the detection of breast cancer. *European journal of radiology*, 82, 08 2012. doi: 10.1016/j.ejrad.2012.03.005.

Fan, Q., Brown, L., and Smith, J. A closer look at faster r-cnn for vehicle detection. In *2016 IEEE Intelligent Vehicles Symposium (IV)*, pp. 124–129, 2016.

Fraser, S. and Blakeman, T. Chronic kidney disease: identification and management in primary care. *Pragmatic and Observational Research*, 2016:7:21–32, 08 2016. doi: 10.2147/POR.S97310.

Gao, Y. and Noble, J. Detection and characterization of the fetal heartbeat in free-hand ultrasound sweeps with weakly-supervised two-streams convolutional networks. pp. 305–313, 09 2017. ISBN 978-3-319-66184-1. doi: 10.1007/978-3-319-66185-8_35.

Girshick, R. Fast r-cnn. *CoRR*, abs/1504.08083, 2015.

Guan, Q. and Huang, Y. Multi-label chest x-ray image classification via category-wise residual attention learning. *Pattern Recognition Letters*, 2018. doi: https://doi.org/10.1016/j.patrec.2018.10.027.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. 7, 12 2015.

Hiramatsu, Y., Muramatsu, C., Kobayashi, H., Hara, T., and Fujita, H. Automated detection of masses on whole breast volume ultrasound scanner: false positive reduction using deep convolutional neural network. pp. 101342S, 03 2017. doi: 10.1117/12.2254581.

Hui, J. map (mean average precision) for object detection, 2018. URL `https://medium.com`.

Hung, J., Goodman, A., Lopes, S., Rangel, G., Ravel, D., Costa, F., Duraisingh, M., Marti, M., and Carpenter, A. Applying faster r-cnn for object detection on malaria images, 04 2018.

Jiang, H. and Learned-Miller, E. Face detection with the faster r-cnn. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pp. 650–657, 2017.

Kasai, S., Li, F., Shiraishi, J., Li, Q., Nie, Y., and Doi, K. Development of computerized method for detection of vertebral fractures on lateral chest radiographs - art. no. 61445d. *Proceedings of SPIE - The International Society for Optical Engineering*, 6144, 03 2006. doi: 10.1117/12.653265.

Kokil, P. and Sudharson, S. Abnormality detection in the renal ultrasound images using ensemble msvm model. pp. 378–382, 03 2019a. doi: 10.1109/WiSPNET45539.2019.9032737.

Kokil, P. and Sudharson, S. Automatic detection of renal abnormalities by off-the-shelf cnn features. *IETE Journal of Education*, 60:1–10, 05 2019b. doi: 10.1080/09747338.2019.1613936.

Kokil, P. and Sudharson, S. An ensemble of deep neural networks for kidney ultrasound image classification. 60: 1–9, 09 2020. doi: 10.1016/j.cmpb.2020.105709.

Krizhevsky, A., Sutskever, I., and Hinton, G. Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems*, 25, 01 2012. doi: 10.1145/3065386.

Król, E., Rutkowski, B., Czarniak, P., Kraszewska, E., Lizakowski, S., Szubert, R., Czekalski, S., Sułowicz, W., and Wiecek, A. Early detection of chronic kidney disease: Results of the polnef study. *American journal of nephrology*, 29:264–73, 09 2008. doi: 10.1159/000158526.

Latif, J., Xiao, C., Imran, A., and Tu, S. Medical imaging using machine learning and deep learning algorithms: A review *. 03 2019. doi: 10.1109/ICOMET.2019.8673502.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., and Sánchez, C. I. A survey on deep learning in medical image analysis. *CoRR*, abs/1702.05747, 2017.

Liu, S., Wang, Y., Yang, X., Li, S., Wang, T., Lei, B., Ni, D., and Liu, L. Deep learning in medical ultrasound analysis: A review. *Engineering*, 5:261–275, 03 2019. doi: 10.1016/j.eng.2018.11.020.

Mitrani, A. Evaluating categorical models ii: Sensitivity and specificity, 2019. URL `https://towardsdatascience.com`.

Narkhede, S. Understanding auc - roc curve, 2018. URL `https://towardsdatascience.com`.

Ozturk, T., Talo, M., Yildirim, A., Baloglu, U., Özal Yıldırım, and Acharya, U. R. Automated detection of covid-19 cases using deep neural networks with x-ray images. *Computers in Biology and Medicine*, 2020. doi: 121. 10.1016/j.compbiomed.2020.103792.

Pereira, F., Bueno, A., Rodriguez, A., Perrin, D., Marx, G., Cardinale, M., Salgo, I., and del Nido, P. Automated detection of coarctation of aorta in neonates from two-dimensional echocardiograms. *Journal of Medical Imaging*, 4:014502, 01 2017. doi: 10.1117/1.JMI.4.1.014502.

Qi, H., Collins, S., and Noble, J. Weakly supervised learning of placental ultrasound images with residual networks. volume 723, pp. 98–108, 06 2017. ISBN 978-3-319-60963-8. doi: 10.1007/978-3-319-60964-5_9.

Ren, S., He, K., Girshick, R., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39, 06 2015. doi: 10.1109/TPAMI.2016.2577031.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A., and Fei-Fei, L. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115, 09 2014. doi: 10.1007/s11263-015-0816-y.

Selvarani, S. and Rajendran, P. Detection of renal calculi in ultrasound image using meta-heuristic support vector machine. *Journal of Medical Systems*, 43, 09 2019. doi: 10.1007/s10916-019-1407-1.

Shariaty, F. and Mousavi, M. Application of cad systems for the automatic detection of lung nodules. 15:100173, 04 2019. doi: 10.1016/j.imu.2019.100173.

Shi, J., Zhou, S., Liu, X., Zhang, Q., Lu, M., and Wang, T. Stacked deep polynomial network based representation learning for tumor classification with small ultrasound image dataset. *Neurocomputing*, 194, 02 2016. doi: 10.1016/j.neucom.2016.01.074.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. 09 2014.

Tan, M. and Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. 05 2019.

Verma, J., Nath, M., Tripathi, P., and Saini, K. Analysis and identification of kidney stone using kth nearest neighbour (knn) and support vector machine (svm) classification techniques. *Pattern Recognition and Image Analysis*, 27: 574–580, 07 2017. doi: 10.1134/S1054661817030294.

Weiss, K., Khoshgoftaar, T., and Wang, D. A survey of transfer learning. *Journal of Big Data*, 3, 12 2016. doi: 10.1186/s40537-016-0043-6.

Wu, K., Chen, X., and Ding, M. Deep learning based classification of focal liver lesions with contrast-enhanced ultrasound. *Optik - International Journal for Light and Electron Optics*, 125, 08 2014. doi: 10.1016/j.ijleo.2014.01.114.

Yang, C.-W., Harris, D., Luyckx, V., Nangaku, M., Hou, F., Garcia Garcia, G., Abu-Aisha, H., Niang, A., Sola, L., Bunnag, S., Eiam-Ong, S., Tungsanga, K., Richards, M., Richards, N., Goh, B. L., Dreyer, G., Evans, R., Mzinga-jira, H., Twahir, A., and Tonelli, M. Global case studies for chronic kidney disease/end-stage kidney disease care. *Kidney International Supplements*, 10:e24–e48, 03 2020. doi: 10.1016/j.kisu.2019.11.010.

Yang, X., Yu, L., Lingyun, W., Wang, Y., Ni, D., Qin, J., and Heng, P.-A. Fine-grained recurrent neural networks for automatic prostate segmentation in ultrasound images. 12 2016.

Yin, S., Zhang, Z., Li, H., Peng, Q., You, X., Tasian, G., and Fan, Y. Fully-automatic segmentation of kidneys in clinical ultrasound images using a boundary distance regression network. volume 2019, pp. 1741–1744, 04 2019. doi: 10.1109/ISBI.2019.8759170.

Zhao, Z.-Q., Zheng, P., Xu, S.-T., and Wu, X. Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems*, PP:1–21, 01 2019. doi: 10.1109/TNNLS.2018.2876865.