UNIVERSITY OF GRONINGEN

Bachelor Thesis Report

Student:
Maxmillan RIES - s3118134

Supervisors:
Estefanía TALAVERA MARTÍNEZ
Dimka KARASTOYANOVA

January 29, 2023

Deep Learning for Kidney Stone Classification of Endoscopic Images

Maxmillan Ries

Abstract—Accurate morphological examinations of kidney stones can aid the diagnosis and treatment of renal calculi. However, despite recent advances in minimally invasive ureterorenoscopy, the use of a laser to cut samples from larger kidney stones frequently damages the specimen and hinders the morphological examination and subsequent analysis. As an alternative to sampling, this paper investigates the classification of a kidney stone image dataset using the deep learning network ResNet152V2, which consists of images taken *in vivo* through a ureterscope and manually labeled using a classification structure provided by Vincent Estrade from the University of Bordeaux. Two binary classifiers are trained to explore the possible differentiation between different calcium oxalate kidney stone morphologies, and two multi-class classifiers are trained to consider the possibility of applying this new method on a more realistic dataset.

The results show a strong performance with the binary classifiers, with the best classifier achieving an overall 97% accuracy and the hierarchical multi-class classifier achieving a 86% accuracy. The details of the classifier implementation, pre-processing steps, as well as their qualitative and quantitative performance and accuracy are presented and discussed.

1 Introduction

Kidney stones occur when hard deposits made of minerals and salts aggregate in the urinary tract. These stones typically form in the kidney and leave the body through the urine stream [1]. With small stones, the passage occurs without symptoms. In cases of larger stones, the blockage of the ureter can result in severe pain in the lower back and abdomen, and can lead to infections if left untreated [1]. Larger kidney stones are destroyed during an intervention called an ureterorenoscopy [2]. Through an endoscope, a low-frequency laser is used to fragment and remove the stones from the urinary tract. Subsequently, the chemical and morphological composition of the collected fragment are used to gain a better understanding of the metabolic causes of kidney stone formations [3].

Daudon and Cloutier found a noteworthy property in kidney stones, in that stones with the same chemical composition often exhibit distinct morphological characteristics [4]. With this discovery, they emphasized studying kidney stone morphologies for an aetiological diagnosis of the disease, rather than solely relying on the chemical composition [4].

Unfortunately, the laser used in the ureterorenoscopic procedures frequently destroys the morphology of the targeted stone [3]. The process is irreversible; once a stone has been destroyed, the morphological examination is no longer possible. To solve this dilemma, image analysis of endoscopic footage can be used for morphological classification. For this, computer vision algorithms such as deep learning can be trained on large datasets to learn and classify different objects within an image. These algorithms excel in the field of image processing and have shown to be superior to dermatologists in melanoma image classification and diabetic retinopathy [3].

The aim of this study is to help doctors improve the speed of the kidney stone morphological examination process. To do so, Vincent Estrade created a continuously growing dataset, from which 318 kidney stone images were used, after being carefully extracted, cropped and hand-labeled. The goal of this paper is to investigate the possibility of classifying kidney stones using a state of the art deep learning network and to assess the use of Saliency Maps and Class Activation Maps (CAM) to visually support the resulted classification [5].

In this paper, the ResNet 152V2 Artificial Neural Network (ANN) was trained on processed images derived from Vincent Estrade's dataset. Firstly, the images were segmented using a contour finding and growing square-based segmentation algorithm, before being over-sampled and augmented as to have a balanced and varied training set.

The rest of the document is organized as follows: The next section introduces the prior works and related research. Section 3 describes the network design, data pre-processing and augmentation, while Section 4 describes the experimental setup, implementation details and validation used to evaluate classification performance. Section 5 qualitatively and quantitatively discusses the results of the experimentation and Section 6 provides a conclusion and a direction for future work.

2 Related Work

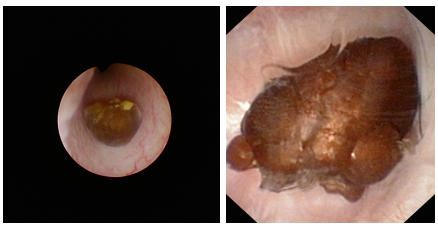
In the field of medicine, there is currently no encompassing therapeutic procedure which can help treat all types of kidney stones. Each kidney stone formation is caused by a variety of factors such as diet, genetics or obesity [6]. Moreover, each kidney stone is not singularly linked to a specific cause, but can come from a combination of the above-mentioned factors, and can lead to chronic kidney disease and kidney failure if left untreated. Primarily, kidney stone classification performed through sample observation on fragments obtained as a result of the ureterorenoscopic procedure or non-invasive procedures such as an ultrasonography or a

CT scan [7].
Martínez et al. aimed to demonstrate in their preliminary results the feasibility of an automated kidney stone classifier on images obtained *in vivo* from an ureteroscope [8]. Martínez et al. compared the use of a random forest classification to a KNN classification model on kidney stone texture and color using the HSV (Hue, Saturation, Value) color space. The random forest model resulted in an average classification accuracy of 89% on cross section classification and an accuracy of 79% on surface classification. The KNN classifier returned an accuracy of 84% on cross section classification and an accuracy of 65% on surface classification [8]. In her work, Martínez focused on finding a more realistic application of *myStone*, a project developed by Serrat et al. [9] and later improved by Torell [10].
In 2020, Black et al. released a pilot study with the objective to assess the recall of deep learning algorithms on the detection of kidney stones composition from digital photographs [11]. Using ResNet-101, Black et al. applied a multi-class classification model to a set of images obtained from 63 kidney stones using a DSLR camera fitted with a 55mm macro lens. The overall weighted prediction recall of the deep learning algorithm was 85%, with commonly encountered stones such as calcium oxalate monohydrate (COM) and uric acid (UA) having higher classification accuracy. In this pilot research, Black et al. stated that Brushite stones specifically were one of the more difficult stone compositions to classify due to the high level of intraclass variability [11].
The main noted limitations of the study was the use of "pure stones" (stones only belonging to a single class) and the use of still images on a clean backdrop. Black et al. used a green non-reflective background and removed any visual noise before passing the data into the deep learning algorithm [11]. When considered in a more realistic application, body movement, blood and debris must all be taken into account. Additionally, the imaging procedure is done through an endoscope in video format, with motion blur and unpredictable lighting making it more difficult to get consistent data.
Later in 2020, Estrade et al. [3] released a paper on the manual classification of renal calculi by analyzing the morpho-constitutional properties of 399 urinary stones bassed on both a endoscopic and microscopical examination. The study demonstrated the feasibility of using endoscopic morphology for the most frequently encountered types of urinary stones [3]. The paper further indicated the feasibility of computer-aided *in situ* recognition and provided a didactic board of confirmed images to do so.
Currently, there is little research done in the field of computer-aided kidney stone classification using modern deep learning networks. As such, one of the goals of this paper is to explore and show the potential of the current state-of-the-art tools on the classification of kidney stones images obtained from a live endoscopic procedure.

3 Methodology

This section describes the preparation of the data along with the Convolutional Neural Network (CNN) which was employed in the experimentation.

3.1 Data Pre-processing



(a) Two unprocessed images manually extracted from recordings of a live endoscopic procedure.

(b) Images obtained from (a) after finding the cropping procedure.

Fig. 1: Figure showing in (a) the original images and in (b) the cropped images obtained from finding the image contour and extrapolating the largest square within using a growing square-based method. In the final images, the black pixels are entirely removed from the image at the cost of the loss of a small section of data. Each crop was manually inspected to ensure no crucial data was removed as to create a falsely labeled image.

Figure 1 shows two typical data samples obtained from a mix of differently cropped frames extracted out of the recordings of an endoscopic procedure. To prevent the ResNet152V2 Residual Neural Network (RNN) from attributing any weight to the black pixels, a pre-processing step was used. This step consisted of obtaining the contour of the colored pixels after a gray-scale conversion and extrapolating the largest possible square within the contour. The OpenCV Python library [12] was used to gray-scale the image and determine the contour, and the final pre-processing step was automated through Python scripting.
Unfortunately, finding the largest rectangle in a fixed polygon is a very tedious and difficult process with up to $O(n^3)$ time complexity [13], and as such, a simpler method was employed. First the center of the contour was obtained using the OpenCV library, before a square was progressively expanded and shifted as to get a good fit. In the case of the images of Figure 1(a), the image obtained is nearly identical to the optimal crop, as shown in Figure 1(b).

3.2 Data Augmentation

Convolutional Neural Networks (CNN) are powerful paradigm that can be applied to a wide variety of visual tasks. As a powerful processing tool, CNNs have shown to be very capable in the field of computer vision, as a tool to mimic the way a human perceives and learns pattern recognition [14]. To train a network however, large amounts of hand-labeled data is required, something which is often not fully available. With a limited dataset, it is still possible to provide enough data for a CNN to be trained using a technique known as data augmentation [15]. Through this process, the diversity of the data can be increased without requiring new information, and has the intended benefit

of reducing overfitting and increasing the generalization capabilities of a classifier.
In this paper, the images were augmented to randomized extents based on a uniform distribution random integer generator. The augmentations consisted of possible **horizontal/vertical flipping**, **channel shifting** and **blurring** and aimed to mimic some of the differences caused by the shifting of the endoscope between frames.

3.3 Convolutional Neural Network (CNN)

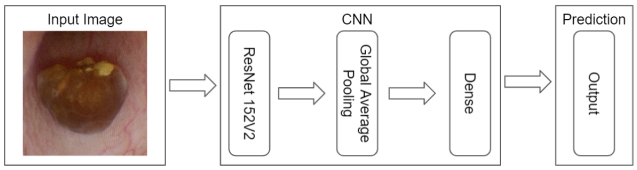


Fig. 2: Visualization of the model used for the experimentation evaluated in this paper. After the ResNet152V2 pre-made model, a global average pooling is applied before a softmax output is created.

The main convolutional model used in this paper consists of an untrained ResNet152V2 model, due to its reported performance and recent release frame. ResNet, short for Residual Network, is an Artificial Neural Network (ANN) used as a backbone for many computer vision tasks due to it's ability to train deep learning networks of over 150 layers while mitigating the vanishing gradient problem [16], a machine learning dilemma which occurs due to the gradient becoming too small to be altered [17]. Residual Networks mitigate the vanishing gradient problem through the use of skip connections, a system which allows gradient information to pass through layers, preventing the later layers from having too small a gradient to work with.
The output of the ResNet152V2 model is a $7 \times 7 \times 2048$ feature map which is fed into a GlobalAveragePooling layer, a layer which replaces the fully connected layers of the ResNet152V2 model with a feature map for each classification category (1×2048). The resulting feature vectors are directly fed into a Dense layer, which generates an output based on a softmax (1×3) or sigmoid (1×2) activation function, depending on the classification task at hand. In the case of multi-class classification, a softmax activation function is used, and in the case of binary classification, a sigmoid activation function is employed.
Finally, the loss function used to train the final model consists of the categorical cross-entropy function, defined below [18]:

$$CE = -\sum_{i}^{C} t_i log(f(s_i))$$

where t_i is the ground truth prediction for each class i in C and $f(s_i)$ is a softmax activation function applied to the CNN score of each class C.

4 Experimental Framework

This section describes the dataset provided, the experiments performed and method of implementation, and the quantitative and qualitative validation metrics used to analyze the classification performances.

4.1 Dataset

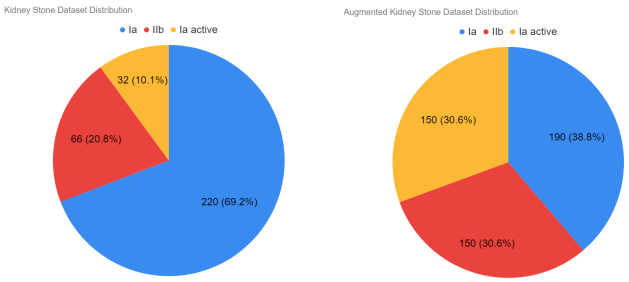


Fig. 3: On the left, the distribution of the original dataset provided by Vincent Estrade (CHU de Bordeaux). Of the total 318 images, 220 belong to class Ia, 66 to class IIb and 32 to class Ia active. Each class has distinct morpho-constitutional properties. On the right, the distribution of the augmented data. Through oversampling, the discrepancy between each class is reduced, minimizing the bias attributed to a higher data count for a single class.

The dataset provided consists of 318 kidney stone images belonging to three kidney stone classes defined by their morpho-constitutional classification, with the distribution shown in Figure 3. The images were extracted from videos obtained through several endoscopic procedures, which were then hand labeled by an expert urologist.
Of the 318 images provided, 220 were labelled *Ia*, which are defined by Dr Estrade to have *a smooth or mamillary dark brown surface, and frequent papillary umbilication with a piece of Randall's plaque* [3]. Out of the remaining 98 images, 66 were labeled *IIb*, which are defined as *yellowish or light brown surfaces with smooth long bipyramidal crystals resembling small desert roses* and 32 were labelled *Ia active*, a variant sub-class to the *Ia* kidney stones [3].

Image Class	Image Count	Train	Validation	Test
Ia	190	137	34	19
Ia active	150	108	27	15
IIb	150	108	27	15

TABLE 1: Distribution of the augmented data. The dataset was split with a 9:1 ratio into two sets using the Sklearn Python library [19]. 5-fold validation was applied on top of the training split, to ensure generalized results, with the same test data being used to evaluate each fold.

4.2 Experimental Setup

In order to investigate the use of CNNs in the classification of kidney stone images, four models were trained and subsequently evaluated.

Ia vs IIb: The first model aimed to verify whether or not kidney stones with distinct morphological properties can be separately classified.
(Ia + Ia active) vs IIb: The second model aimed to test if the introduction of a variant of the *Ia* class would affect the binary classification.
(Ia + Ia active) vs IIb —— Ia vs Ia active: This model consisted of a hierarchical classification, whereby the second model is used to classify *Ia* and *Ia active* kidney stones from *IIb*, and the second model is used to more

finely classify between the possible two remaining classes. The same train/test split as the previous network is used to ensure generalized results.
Ia vs Ia active vs IIb: In contrast with the previous model, this model was a straightforward categorical classification. In comparison to the hierarchical classifier, this model is expected to perform slightly better, though some fine details may be lost in the process due to the limited data.

In addition to the four principle experiments, rotation and shearing were evaluated as part of the data augmentation step, along with the use of the focal loss function.

Rotation and Shearing Data Augmentation: The introduction of rotation and shearing is predicted to have a positive effect on the accuracy of the model as the data augmentation described in section 3.2 already aimed to mimic the position shifting of the endoscope. The rotation and shearing range of 30 degrees was used for testing.
Focal Loss: Focal loss is considered an improved version of binary/categorical cross-entropy which aims to compensate for class imbalance by assigning more weights to harder or more easily classified examples, as shown below [20]:

$$FL = -\sum_{i}^{C} \alpha(1 - f(s_i))^{\gamma} t_i log(f(s_i))$$

where α is a coefficient added to handle class imbalance, $(1 - f(s_i))^{\gamma}$ is a probability dependent weight and the rest correspond to the same variables as categorical cross-entropy. The chosen parameters for the focal weight were α = 0.25 and γ = 2, values chosen for their reported performance in Tsung-Yi Lin et al's paper [20]. The model is trained on the unaugmented dataset, as the need for balancing weights is not required in a balanced dataset.

4.3 Implementation Details

Data Augmentation
The models evaluated in this paper are used to investigate the classification of kidney stone images. In order to streamline and simplify the training process, the ImageDataGenerator class from the Keras Python library [21] was used in order with the following properties:

- Pre-processing function: Blur (OpenCV blur with a random kernel size between 0 and 5)
- Channel shift range: 50
- Zoom range: 0.15
- Horizontal flip: True
- Vertical flip: True

The dataset is an imbalanced dataset, with 69.2% of the image belonging to a single kidney stone class (*Ia*). In order to prevent a natural bias towards this class, the dataset was balanced through a combination of over and undersampling. First, as many images in *Ia* were duplicates, several were removed, reducing the *Ia* count to 190 images. Subsequently, using an ImageDataGenerator, additional *Ia active* and *IIb* images were augmented and added to the dataset.
The final dataset used for classification consisted of 190 images of class *Ia*, 150 images of class *IIb* and 150 images of class *Ia active*, shown in the distribution chart of Figure 3.

Model Parameters
Each model was trained with the following properties, with the highest accuracy being used to save the best training epoch and no early stopping being used.

- Batch Size: 4
- Epoch Count: 200
- Optimizer: Adam with Learning Rate of 0.001.

The Tensorflow and Keras Python libraries were used in the implementation and training of the models, with the Sklearn packages being used for subsequent analysis [19].

4.4 Validation

To evaluate the proposed CNN classification performances, both quantitative and qualitative evaluation metrics were used. Each experiment was conducted using 5-fold cross validation on top of a train/text split ratio of 9:1 to ensure successful generalization, with the number of samples shown in Table 1.

Quantitative:
Confusion Matrix: The confusion matrix is a table used to describe the performance of a classification model on testing data where the ground truth is known. In the case of our testing data, a prediction is performed on each test image, and the correct and incorrect predictions are noted as either True Positive, True Negative, False Positive or False Negative. This validation measure already indicates the general performance of the classifier.
Learning Curve: The learning curves consist of two graphs, one plotting the training and testing accuracy over the epochs, and another plotting the training and testing loss over the epochs. The use of these graphs is to evaluate the fit of the classifier. The difference between the training and testing curves can be used to infer a possible over/underfit or the unrepresentativeness of the training or testing data.

Qualitative:
T-SNE and PCA: T-SNE and PCA are two dimensionality reduction algorithms, which can be used to visualize the final fully connected layer of the CNN [22]. In visualizing the final layer on a 2D plot, inferences on the model's separation of the classes can be made.
Saliency Maps and Class Activation Maps (CAM): Saliency Maps and Class Activation Maps can be obtained from extracting the weights of the last convolutional layer of a network [23]. In doing so, one can create a heatmap, which can help visualize what the trained algorithm used in the data to determine its prediction. These maps, while visually helpful, are not a perfect indicator and cannot solely be relied upon to determine whether a classification is good or bad.

5 Results and Analysis

This section contains an analysis of the classification performances of each model described in section 4.2. Each performance is evaluated using the qualitative and quantitative metrics described in section 4.4. The comparison between each experiment was evaluated using the best performing models.

Name	Accuracy	Precision	Recall	F1 Score
Ia vs IIb	0.83 ± 0.065	0.79 ± 0.065	0.79 ± 0.065	0.80 ± 0.065
Ia+Ia active vs IIb	0.79 ± 0.071	0.71 ± 0.071	0.76 ± 0.071	0.72 ± 0.071
Ia vs Ia active (h)	0.67 ± 0.097	0.54 ± 0.097	0.56 ± 0.097	0.55 ± 0.097
Ia vs Ia active vs IIb	0.76 ± 0.027	0.65 ± 0.029	0.71 ± 0.040	0.67 ± 0.034

(a) Derivations calculated using Sklearn classification report over the ground truth and model predictions for the base models of section 4.2 trained on the unaugmented dataset, along with the standard deviation.

Name	Accuracy	Precision	Recall	F1 Score
Ia vs IIb	0.97 ± 0.045	0.97 ± 0.045	0.97 ± 0.045	0.97 ± 0.045
Ia+Ia active vs IIb	0.89 ± 0.044	0.89 ± 0.044	0.88 ± 0.044	0.88 ± 0.044
Ia vs Ia active (h)	0.86 ± 0.058	0.85 ± 0.058	0.86 ± 0.058	0.86 ± 0.058
Ia vs Ia active vs IIb	0.92 ± 0.048	0.92 ± 0.048	0.91 ± 0.048	0.91 ± 0.048

(b) Derivations calculated using Sklearn classification report over the ground truth and model predictions for the base models of section 4.2 trained on the augmented data, along with the standard deviation.

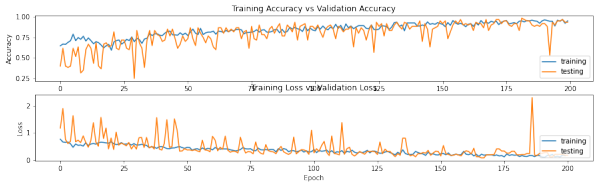
Name	Accuracy	Precision	Recall	F1 Score
Ia vs IIb	0.82 ± 0.025	0.82 ± 0.025	0.82 ± 0.025	0.82 ± 0.025
Ia+Ia active vs IIb	0.86 ± 0.025	0.86 ± 0.025	0.86 ± 0.025	0.86 ± 0.025
Ia vs Ia active (h)	0.80 ± 0.046	0.79 ± 0.046	0.80 ± 0.046	0.80 ± 0.046
Ia vs Ia active vs IIb	0.82 ± 0.034	0.82 ± 0.015	0.82 ± 0.048	0.82 ± 0.029

(c) Derivations calculated using Sklearn classification report over the ground truth and model predictions for the models trained with additional rotation and shearing data augmentation along with the standard deviation.

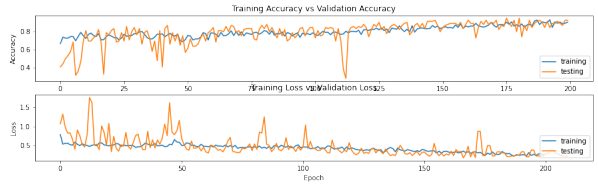
Name	Accuracy	Precision	Recall	F1 Score
Ia vs IIb	0.90 ± 0.019	0.90 ± 0.019	0.90 ± 0.019	0.90 ± 0.019
Ia+Ia active vs IIb	0.82 ± 0.035	0.82 ± 0.035	0.82 ± 0.035	0.82 ± 0.035
Ia vs Ia active (h)	0.75 ± 0.044	0.75 ± 0.044	0.75 ± 0.044	0.75 ± 0.044
Ia vs Ia active vs IIb	0.79 ± 0.050	0.88 ± 0.050	0.64 ± 0.040	0.74 ± 0.039

(d) Derivations calculated using Sklearn classification report over the ground truth and model predictions for the models trained with focal loss along with the standard deviation.

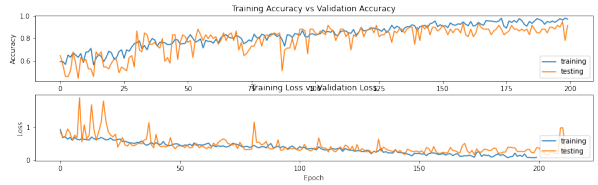
Fig. 4: Tables holding the derivations obtained from the confusion matrices after evaluation of each model and their standard deviation. These derivations represent different statistical measures of the predictions over the test dataset split. Each value is normalized, and consists of the average of the 5 k-folds. Models labeled (h) represents the hierarchical classification *Ia+Ia active* vs *IIb* —— *Ia* vs *Ia active*.



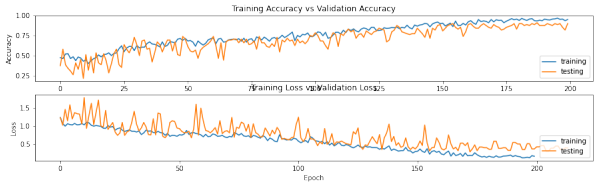
(a) Learning curves for the *Ia* vs *IIb* model.



(b) Learning curves for the *Ia+Ia active* vs *IIb* model.



(c) Learning curves for the *Ia+Ia active* vs *IIb* —— *Ia* vs *Ia active* hierarchical model.



(d) Learning curves for the *Ia* vs *Ia active* vs *IIb* model.

Fig. 5: Learning curves representing the learning process of the classifiers trained on the ideal augmented data (Figure 4 (b)). The upper graphs represent the training and validation accuracy over the epochs, while the lower graphs represent the training and validation loss over the epochs. These graphs are used to evaluate the quality of the training and fit obtained.

5.1 Ia vs IIb

Figure 6 shows two images corresponding to the confusion matrix obtained from predicting the testing split, and the reduced plot of the last fully connected layer of the network from the *Ia vs IIb* model trained on augmented data (Figure 4 (b)). The confusion matrix shows a first measure of the performance of the algorithm. With only a single miss-classification, the algorithm has a very high classification accuracy. The t-SNE plot of Figure 6 shows a successful differentiation between the two classes, with two visible clusters being formed and the miss-classification being "located" close to intersection between both clusters.

Figure 5 (a) shows the training and validation accuracy and loss over the 200 epochs the network was trained for. In this figure, the training graphs show a continuous improvement over time, with the accuracy increasing from 0.6 to 0.97

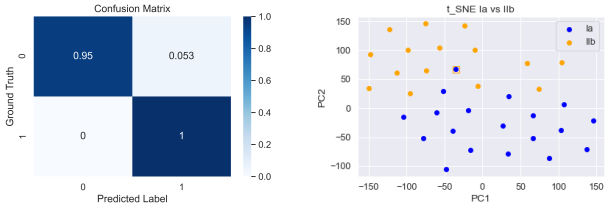


Fig. 6: On the left, the confusion matrix obtained running the *Ia vs IIb* model of Figure 4 (b). The class labels 0 and 1, correspond to the *Ia* and *IIb* kidney stone classes. On the right, an image corresponding to the reduced feature vector obtained from the last fully connected layer of the CNN. A perplexity value of 25 was used for the t-SNE reduction, with each circle the ground truth and each square the color-coded miss-classification.

and the loss decreasing from 0.85 to 0.12. The validation curves show similar final results with a however much less steady decrease. The learning curves overall show little over/underfitting, and do show measures of a successfully trained model.

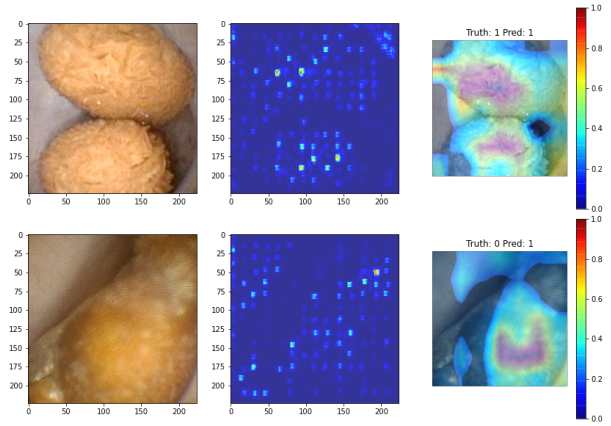


Fig. 7: On the left, the original augmented and cropped image. In the middle, the saliency map after prediction. On the right the CAM extracted from the last convolutional layer of the network. The top image shows a successful prediction, with both the saliency map and CAM highlighting the kidney stone. The bottom image shows the only unsuccessfully classified image. While the CAM shows the algorithm focused on the kidney stone in part, the saliency map shows that part of the ureter wall was taken into account for the prediction.

Figure 7 above shows two sets of images consisting of the cropped augmented image, an obtained saliency map and CAM. The top three images show a successful classification, with both the saliency map and CAM highlighting the kidney stone as the Region Of Interest (ROI). The second set of images shows the unsuccessful case in the classification, with the saliency map highlighting part of the ureter wall as a determining factor for the prediction. This noise in the decision making of the algorithm is largely due to the lack of data. As the network classifies based on recurring features, the lack of images and image variety causes undesired parts of the images to be used for the prediction.

5.2 Ia+Ia active vs IIb

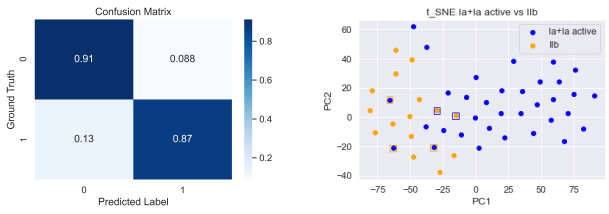


Fig. 8: On the left, the confusion matrix obtained running the *Ia+Ia active vs IIb* model of Figure 4 (b). The class labels 0 and 1, correspond to the *Ia+Ia active* and *IIb* kidney stone classes combinations. On the right, an image corresponding to the reduced feature vector obtained from the last fully connected layer of the convolutional neural network. A perplexity value of 25 was used for the t-SNE reduction, with each circle the ground truth and each square the color-coded miss-classification.

In comparison to the previous binary classification of *Ia* vs *IIb*, the addition of the *Ia active* class shows a 8% decrease in the accuracy, a 9% decrease in the F1 score and a similar decrease in the other confusion matrix derived values (Figure 4 (b)). Figure 9 shows three sample images, showcasing a possible difficulty encountered in the classification process. While most *Ia active* images display similar morphological characteristics as the *Ia* class, several images show a closer visual resemblance with the *IIb* class. The combination of the *Ia* and *Ia active* classes leads to a higher intra-class variability, with some of the data resembling the *IIb* class very closely. This can be seen in the right image of Figure 8, as a higher number of miss-classifications are present, though two clusters can still be differentiated.

Fig. 9: The image on the left is a typical example of a kidney stone belonging to the *Ia* class. The middle image belongs to the *Ia active* class and the image on the right belongs to the *IIb* class. Most *Ia active* images show similar morphological characteristics to the *Ia* class. In this case, however, with the possible addition of a blur, the color and shape information indicates a closer resemblance towards the *IIb* class, making the image more difficult for the model to correctly classify.

The learning curves of Figure 5 (b) show a generally successful training. The validation curves, while more jagged and fluctuating than 5.1 initially, show a much smoother and standard decrease after epoch 50, closely following the pattern of the training curves. The validation curves also indicate no sign of over/underfitting, with an overall accuracy of 0.89 being achieved for a loss of 0.26. The increase of 0.14 in the loss value and the small decrease in the accuracy can be attributed to the addition of the *Ia active* kidney stone images, making the classification more difficult.
Similarly to Figure 7, the top set of images of Figure 10 show a typical successful classification, while the bottom set shows an unsuccessful classification. More interestingly, in both sets of images, the saliency maps and CAM have little overlap, indicating conflicting focus for the prediction. In these two cases, the saliency maps show a subset of the CAM information, with the CAM showing a very desirable heatmap for the successful prediction. For the unsuccessful prediction, the CAM shows both a focus on the kidney stone and the surrounding intestinal walls. This specific image belongs to the *Ia active* class, and is one of the images which most closely resembles the typical *IIb* images, making it very difficult to correctly classify. A very likely cause for this focus away from the kidney stone is the size of the dataset.

5.3 Ia+Ia active vs IIb —— Ia vs Ia active Hierarchical

This classifier is a hierarchical classifier based on the previous *Ia+Ia active* vs *IIb* network. The train/test split used to separate the data is identical to the one used in subsection 5.2. The confusion matrix of Figure 11 generally demonstrates that the hierarchical classification is accurate, with a total of 7 miss-classifications, 5 between the *Ia+Ia*

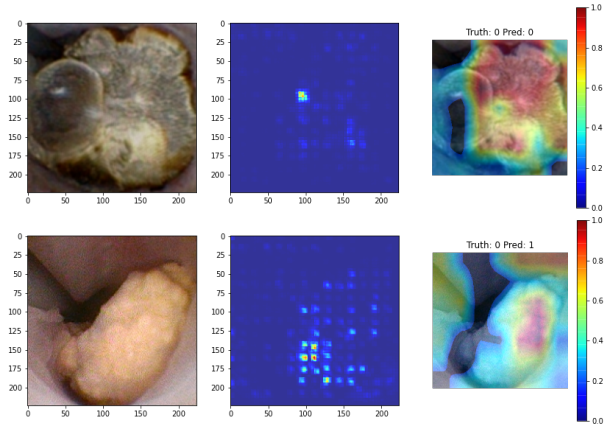


Fig. 10: On the left, the original augmented and cropped image. In the middle, the saliency map after prediction. On the right the CAM extracted from the last convolutional layer of the network. The top image shows a successful prediction, with both the saliency map and CAM highlighting the kidney stone. The bottom image shows an unsuccessfully classified image. While the saliency map shows the algorithm focused on the kidney stone in part, the CAM shows that most of the ureter wall was taken into account.

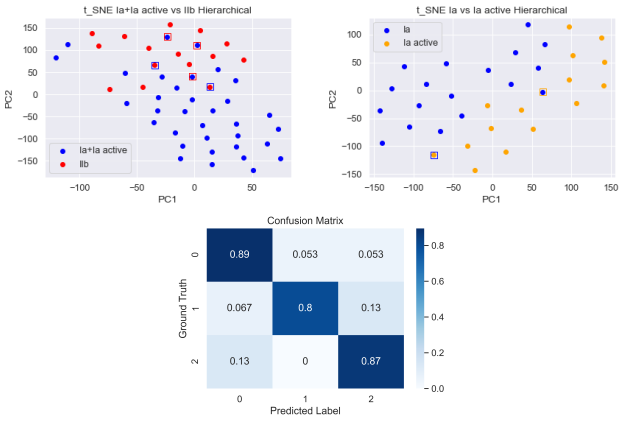


Fig. 11: On top, two images corresponding to the reduced feature vector obtained from the last fully connected layers of the hierarchical classifier. The top left shows the reduced feature vector for the outer hierarchical network classifying *Ia+Ia active* and *IIb*, and the top right shows the reduced feature vector for the inner hierarchical, which classified between the *Ia* and *Ia active* kidney stone classes. For both graphs, a perplexity value of 25 was used, with each circle the ground truth and each square the color-coded miss-classification. At the bottom, the confusion matrix obtained running the *Ia vs Ia active (h)* model of Figure 4 (b). The class labels 0, 1 and 2 correspond to the *Ia*, *Ia active* and *IIb* kidney stone classes.

active and *IIb* classes, and 2 between the *Ia* and *Ia active* classes. The miss-classifications of the *Ia+Ia active* vs *IIb* t-SNE plot are primarily located around the border between the two respective clusters, with two miss-classifications being located deeper within the wrong cluster. The likely cause of this is described in section 5.2.
The right t-SNE plot shows a strong split between the *Ia* and *Ia active* classes, with the miss-classifications being located close to the intersection between both respective clusters, with the only noteworthy observation being the likely lack of training data. Generally however, the error propagation can be seen through the decrease in performance, as an incorrect *Ia+Ia active* prediction results in the inability for the inner classifier to correctly predict the image label.
The learning curves of Figure 5 (c) demonstrate a slight overfit. The final validation accuracy is roughly 5% lower than the training accuracy, with the validation loss showing a similar overfitting pattern. The training curves show a similar pattern section to 5.1 and 5.2, with the validation curves following a similar trend as the training curves.

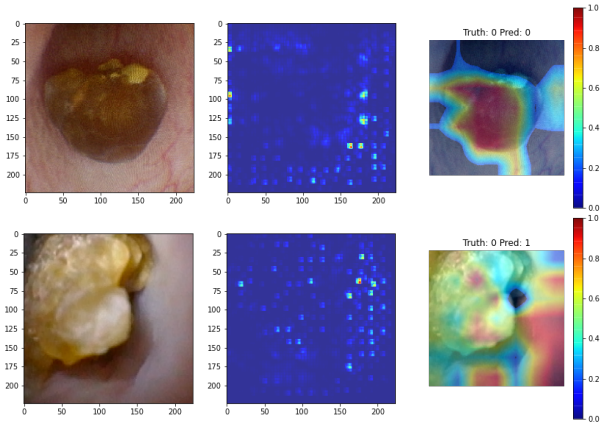


Fig. 12: On the left, the original augmented and cropped image. In the middle, the saliency map after prediction. On the right, the CAM extracted from the last convolutional layer of the network. The top image set shows a successful prediction, with only the CAM highlighting the kidney stone. The bottom image shows an unsuccessful classification. In this test case, both the saliency map and the CAM show that the ureter was the major factor in the prediction.

The bottom set of images of Figure 12 show a very unsuccessful prediction. The majority of the ROI for both the saliency map and the CAM is located on the intestinal wall, with the kidney stone itself being offset away from the center of the image. Unfortunately, as the image was obtained from a frame of a recorded video, the centering of the kidney stone is not assured, and while CNN's are designed to generalize recurring features, the lack of data induces a focus towards the wrong information.
The top set of images show a conflicting focus between the saliency map and the CAM. The saliency map indicates a complete lack of focus on the kidney stone, while the CAM indicates a complete focus on the kidney stone. As the classification is correct, one could lean towards the saliency map being incorrect. However, according to Molnar's book "Interpretable Machine Learning", there is no indicator on which of the two visualization measures is necessarily correct [23]. As such saliency maps and CAMs cannot be uniquely used to determine whether a model was trained correctly.

5.4 Ia vs Ia active vs IIb

The final base classification evaluated consisted of a simple categorical classification on the augmented data. The confusion matrix obtained shows 3 miss-classifications, 2 of them *IIb* kidney stone images and one of them being a *Ia active* kidney stone image (Figure 13 left). The t-SNE plot

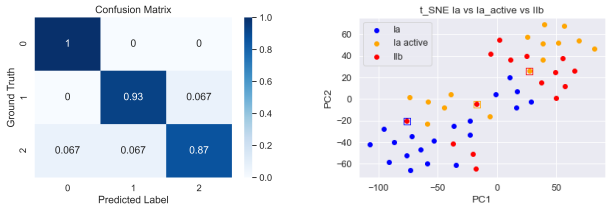


Fig. 13: On the left, the confusion matrix obtained running the *Ia vs Ia active vs IIb* model of Figure 4 (b). The class labels 0, 1 and 2, correspond to the *Ia*, *Ia active* and *IIb* kidney stone classes. On the right, an image corresponding to the reduced feature vector obtained from the last fully connected layer of the CNN. A perplexity value of 25 was used for the t-SNE reduction, with each circle the ground truth and each square the color-coded miss-classification.

further exemplifies this, as while the *Ia* and *Ia active* images are easily classified apart, the *IIb* images have considerable overlap. The most likely cause of this error is the lack of data, as 84 *IIb* and 118 *Ia active* images are over-sampled images, a parameter which was arbitrarily set, but not further evaluated in this paper.

While the categorical classification seemingly outperforms the hierarchical classification when looking at the confusion matrices, the t-SNE plots show that the hierarchical classification is more capable at separating the individual classes. Figure 11 clearly shows distinct separable clusters between both the *Ia+Ia active* and *IIb* classes and the *Ia* and *Ia active* classes while Figure 13 shows a more visible inter-melding of the three kidney stone classes.

The learning curves of Figure 5 (d) show clear signs of an overfit. The final validation accuracy is lower than the training and the validation loss is noticeably higher. Unlike the previous models however, the validation loss show slight signs of an unrepresentative training dataset. The validation loss curve very vaguely follows the same trend as the training loss for the first 150 epochs, before stagnating around the same loss value. This is further shown in the bottom image set of Figure 14, where the CAM clearly shows a lack of proper training data.

Unfortunately, unlike binary or hierarchical binary classification, categorical classification requires a larger dataset in order to uniquely identify the differences between each class. As the dataset used in this paper is limited and unbalanced, the combination of biases and a lack of data causes the categorical classifier to produce inconsistent results, notably when differentiating between the *Ia active* and *IIb* classes.

5.5 Data Augmentation —— Rotation and Shearing

The data augmentation process aimed to enhance the dataset by providing images which mimicked the changes brought forth by the movement of the endoscope between each image frame. The prior sections 5.1-5.4 show the results on data which has been augmented by means of blurring, channel shifting and horizontal/vertical flipping.

Figure 4 (c) however shows that the additional of rotation and shearing in the data augmentation process reduces the performance of all classifiers. With the addition of rotation

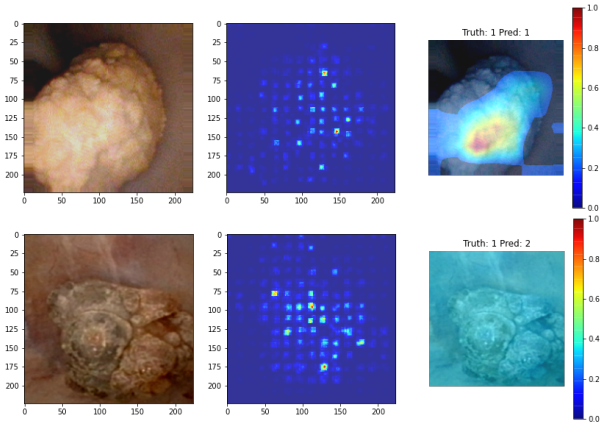


Fig. 14: On the left, the original augmented and cropped image. In the middle, the saliency map after prediction. On the right the CAM extracted from the last convolutional layer of the network. The top image shows a successful prediction, with both the saliency map and CAM highlighting the kidney stone. The bottom image shows the only unsuccessfully classified image. While the saliency map shows a very clear focus on the kidney stone itself, the CAM shows the entire image as being the ROI.

and shearing, each trained model has its performance reduced, with the *Ia vs IIb* model showing a performance decrease of 15%, the *Ia vs Ia active vs IIb* model showing a performance decrease of 10%.

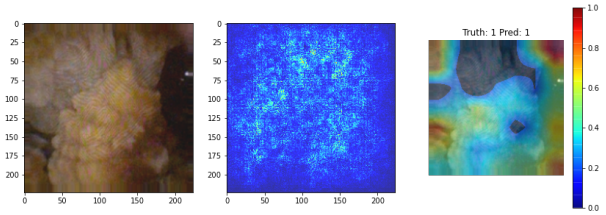


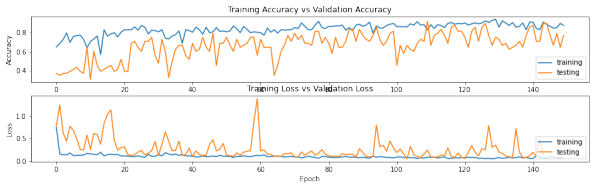
Fig. 15: On top, an image obtained after augmented the cropped image based on a uniform distribution using the Keras ImageDataGenerator. The edges of the image are warped due the shearing and rotation transformations applied to the cropped image. At the bottom, three images consisting of i) the augmented image, ii) a saliency map and iii) the CAM obtained from the last layer of the CNN. With the slight shearing on the bottom image, the network heavily focuses on these warped artifacts, biasing the prediction with undesirable data.

The heat map of Figure 15 shows a qualitative visualization of the prediction on a sheared image. While the saliency map displays a correct interpretation of the kidney stone, the CAM shows a strong focus on the shearing artifact, rather than on the kidney stone itself. The addition of rotation and shearing seems beneficial, as the endoscope is manipulated within the ureters and images are taken from multiple perspectives. However, the results show that the addition of the rotation and shearing augmentation through the ImageDataGenerator results in a drastically lower performance and requires more investigation to allow usability.

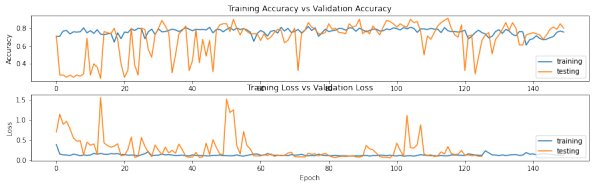
5.6 Focal Loss

The focal loss function consists of the categorical cross-entropy loss with an additional factor/weight defined by

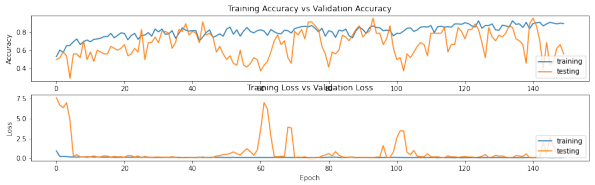
two variables, α and γ.



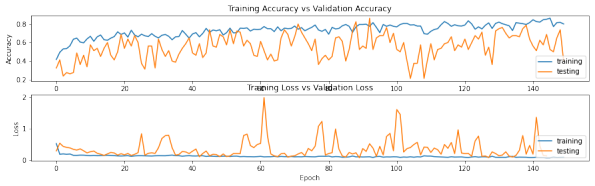
(a) Learning curves for the focal loss *Ia* vs *IIb* model.



(b) Learning curves for the focal loss *Ia+Ia active* vs *IIb* model.



(c) Learning curves for the focal loss *Ia+Ia active* vs *IIb* —— *Ia* vs *Ia active* hierarchical model.



(d) Learning curves for the focal loss *Ia* vs *Ia active* vs *IIb* model.

Fig. 16: Learning curves representing the training process for the focal loss trained CNNs evaluated in this paper. The upper graphs represent the training and validation accuracy over the epochs, while the lower graphs represent the training and validation loss over the epochs. These graphs are used to evaluate the quality of the training and fit obtained.

The general performance of the classifiers trained using focal loss is mixed, with Figure 4 (d) showing an overall lower performance compared to the base models. In comparison to the rotation and shearing augmented data, the focal loss models outperform the rotation and shearing model in the binary *Ia vs IIb* classification. In all other aspects however, the model underperforms slightly.
More interestingly, when comparing the learning curves of Figure 5 to Figure 16, a clear set of training problems can be observed. While the overall trend of the *Ia vs IIb* model is correct, signs of underfitting can be seen. The validation accuracy is lower than the training loss, with the validation loss being slightly higher.
The *Ia+Ia active vs IIb* model shows an unstable beginning, possibly due to an unrepresentative training set, before a smoother continuous improvement can be seen. The validation accuracy in the final epoch is higher than the training accuracy, with the loss showing a similar trend, indicating a successful training.
The *Ia vs Ia active* inner hierarchical learning curve show clear signs of underfitting as well, with the validation accuracy being primarily much lower than the training accuracy. The validation loss, beyond the occasional spike, shows little sign of immprovement, remaining around the loss value of 0.07.
The categorical model shows clear signs of an unrepresentative testing set, with the validation accuracy and loss remaining around a steady average while the training accuracy and loss continously improves. This could be cause by an unfortunate train/test split, with the test data containing too many unseen images out of the limited unaugmented dataset.

The general performance of the focal loss shows that a possible incorrect assignment α and γ variables. In comparison to the base models, the learning curves are lackluster and show few clear signs of minimization. In this paper, no additional parameters than the those described in section 4.2 were used, and are left as a possible avenue for future work in Section 6.

5.7 Discussion

Data Augmentation
The use of data augmentation in the sampling process aims to increase the diversity of the data without requiring any new information. In the case of kidney stone classification, the use of oversampling through horizontal/vertical flipping, channel shifting and blurring results in a 10% to 20% increase in performance (Figure 4 (a) & (b)), notably for hierarchical classification, where the imprecisions of the outer layers propagate to the performance of the inner layers.
The inclusion of rotation and shearing in the data augmentation process causes a decrease in the classification success rate when compared to the results of Figure 4 (b) with the introduction of a new artifact in the images. The small dataset size introduces the need for more augmented images, which creates a sufficient number of artifacts for a strong bias to be taken into consideration during the training. However, in comparison to the unaugmented data (Figure 4 (a)), the inclusion of rotation and shearing provides a slight improvement, as the increased pool of data still allows more training information, despite the artifact-induced bias.
Overall, the inclusion of blurring, horizontal/vertical flipping and channel shifting data augmentations result in a satisfactory performance, with the models successfully generalizing many of the common morphological characteristics between the different kidney stone images.

Saliency Maps and Class Activation Maps
The saliency maps and CAMs extracted from running the models of Figure 4 (b) on the test sets show an overall clear representation of a successful training. On correct classifications, despite some differences, both the saliency maps and the CAMs show the kidney stone as the primary ROI. On incorrect classifications, the saliency maps and CAMs often differ, with one of the two showing a subset of the other. In few cases, the two visualization techniques show unrelated ROI, with no indication of which one

is correct. Unfortunately, according to Molnar, there is currently no quantitative indicator as to which one is correct in such a scenario [23].
Generally, the visualization of the last convolutional layer is very helpful as an indicator of the ROI. Whilst not fully reliable, this technique has much use as an aid in determining whether or not an image was correctly predicted.

Deep Learning for Kidney Stone Classification
The overall results of this paper show that the classification of *in vivo* kidney stone images is possible through the use of state-of-the-art models, such as ResNet152V2. The results of the binary classification show a lot of promise with *Ia* vs *IIb* and *Ia* vs *Ia active* classification achieving 97% overall accuracy. Multi-class classification however, still shows to be a more problematic challenge due to the current lack of data. With only 318 provided images, the hierarchical classification model achieved a 86% accuracy and the categorical model achieved a 92% accuracy. Unfortunately, while the accuracy values of both models is satisfactory, the multi-class classification was only performed on 3 kidney stone classes defined by Vincent Estrade [3], out of the total 21 defined classes, all belonging to the commonly encountered calcium oxalate morphology.

In comparison to the work of Black et al. [11], the work in this paper focused on the application of the same technology on images of kidney stones inside the human body. The images used in this paper were hence in a more dynamic environment, with each image being centered, lit, blurred and cropped to irregular extents. The work of Martínez et al. [8] is similar to that conducted in this paper, as an investigation of older state-of-the-art classification technologies. The dataset used in this paper consists in part of the same data used by Martínez et al., with additional calclium oxalate kidney stones images being provided by Vincent Estrade. The uric acid stones were however not trained on, preventing a complete comparison between both works. Nonetheless, Black et al. noted that both calcium oxalate monohydrate and uric acid stone classification resulted in higher classification accuracy out of all kidney stones morphologies, showcasing the improvement brought forth by use of a deep residual neural networks and the availability of all morphological information belonging to each stone. Moreover, compared to both Martínez et al. and Black et al., the dataset employed was structured for the purpose of computer-aided classification, allowing the models to train on more specialized and relevant information.

6 Conclusion and Future Work

This paper evaluated the possibility of classifying kidney stone images extracted from recordings of an *in vivo* endoscopic procedure. Four classifiers were trained after a pre-processing step, and a two additional variations to the models or data augmentation steps were tested.
The results show that the base models with minimally augmented data through horizontal/vertical flipping, channel shifting and blurring outperform the variants, with rotation and shearing introduction artifacts in the image which cannot be generalized and focal loss having a high chance of an unfortunate train/test split. Overall, the base models do show a high classification accuracy, with the *Ia vs IIb* model achieving an 97% accuracy, and the hierarchical binary classification achieving an 86% accuracy.
All models show a signs of a limited dataset, with the saliency maps and CAMs highlighting parts of the ureter as a determining factor in the classification. These minor complications in the training of the networks can easily be remedied as more data is added to the dataset and more powerful state-of-the-art networks can be trained.

This paper consisted of an overall investigation on the classification of kidney stones. While several successful experiments were conducted, the parameters used for the augmentation of the dataset were not tested, with an educated estimate being used. Additionally, the focal loss parameters α and γ were set to the default findings of Tsung-Yi Lin et al, and were not thoroughly tested.
Beyond further detailed investigations on the research conducted in this paper, a direction for future work should include the classification of the original data format, the recordings of the live endoscopic procedure, through a possible frame-based majority voting approach or through the use of an Long Short Term Memory Convolutional Neural Network (LSTM).

7 Acknowledgements

The author wishes to extend sincere acknowledgement to Estefanía Talavera Martínez for her supervision, support and advice throughout the thesis.
The author wishes to thank Dr. Baudouin Denis de Senneville of the Institut de Mathématiques de Bordeaux (IMB) for providing the idea for the project, along with the data. The guidance and the support provided by Baudouin were invaluable in completing this project.

References

[1] Malvinder S Parmar. Kidney stones. *BMJ*, 328(7453):1420–1424, 2004.
[2] V Estrade, K Bensalah, J-P Bringer, E Chabannes, X Carpentier, P Conort, E Denis, B Doré, JR Gautier, H Hadjadj, J Hubet, A Hoznek, E Lechevallier, P Meria, P Mozer, C Saussine, L Yonneau, O Traxer, and Comité lithiase de l'AFU. [place of the flexible ureterorenoscopy first choice for the treatment of kidney stones. survey results practice committee of the afu lithiasis completed in 2011]. *Progres en urologie : journal de l'Association francaise d'urologie et de la Societe francaise d'urologie*, 23(1):22—28, January 2013.
[3] Vincent Estrade, Baudouin Denis De Senneville, Paul Meria, Christophe Almeras, Franck Bladou, Jean-Christophe Bernhard, Gregoire Robert, Olivier Traxer, and Michel Daudon. Toward improved endoscopic examination of urinary stones: a concordance study between endoscopic digital pictures vs. Microscopy. *BJU International*, 2020.
[4] Jonathan Cloutier, Luca Villa, Olivier Traxer, and Michel Daudon. Kidney stone analysis: "give me your stone, i will tell you who you are!". *World Journal of Urology*, 33, 12 2014.
[5] B. Zhou, A. Khosla, Lapedriza. A., A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. *CVPR*, 2016.
[6] Nicole L Miller and James E Lingeman. Management of kidney stones. *BMJ*, 334(7591):468–472, 2007.

[7] Bailey M. Sorensen M. Brisbane, W. An overview of kidney stone imaging techniques. *Nat Rev Urol*, 13:654–662, 2016.

[8] A. Martínez, D. H. Trinh, J. El Beze, J. Hubert, P. Eschwege, V. Estrade, L. Aguilar, C. Daul, and G. Ochoa. Towards an automated classification method for ureteroscopic kidney stone images using ensemble learning. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 1936–1939, 2020.

[9] Joan Serrat, Felipe Lumbreras, Francisco Blanco, Manuel Valiente, and Montserrat López-Mesas. mystone: A system for automatic kidney stone classification. *Expert Systems with Applications*, 89:41–51, 2017.

[10] A. Torrell. Metric learning for kidney stone classification. *Escola D'Enginyeria, Universitat Autonoma de Barcelona*, 2018.

[11] Kristian M. Black, Hei Law, Ali Aldoukhi, Jia Deng, and Khurshid R. Ghani. Deep learning computer vision algorithm for detecting kidney stone composition. *BJU International*, 125(6):920–924, 2020.

[12] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[13] Daniel Smilkov. Largest rectangle in a polygon, Jul 2014.

[14] Larry Hardesty. Explained: Neural networks, Apr 2017.

[15] David A van Dyk and Xiao-Li Meng. The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50, 2001.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[17] Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 06(02):107–116, 1998.

[18] Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels, 2018.

[19] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018.

[21] Francois Chollet et al. Keras, 2015.

[22] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[23] Christoph Molnar. *Interpretable Machine Learning*. 2019. https://christophm.github.io/interpretable-ml-book/.