



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA



Escola Tècnica
Superior d'Enginyeria
Informàtica

Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

Diagnosis Support CNN for Renal Cysts & Tumours in CT Scans

Trabajo Fin de Grado

Grado en Ingeniería Informática

Author: Álvarez Llopis, Nicolás

Tutor: Onaindia de la Rivaherrera, Eva

External cotutor: Quille, Keith

Academic Year: 2021/2022



TU DUBLIN, TALLAGHT CAMPUS

BSC PROJECT

**Diagnosis Support CNN for Renal Cysts &
Tumours in CT Scans**

Nicolás Álvarez Llopis

Department of Computing

Supervised by

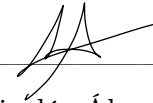
Dr. Keith Quille

Department of Computing

1 May 2022

Declaration

I hereby certify that the material, which I now submit for assessment on the programmes of study leading to the award of Master of Science, is entirely my own work and has not been taken from the work of others except to the extent that such work has been cited and acknowledged within the text of my own work. No portion of the work contained in this thesis has been submitted in support of an application for another degree or qualification to this or any other institution.



Nicolás Álvarez Llopis

1 May 2022

Acknowledgements

To my tutor Keith Quille, who was a great support along the way. For the teachings he has given me, the guidance during the development of this project and the many meetings we had to clarify my doubts. Thank you for your kindness and support during my Erasmus studies.

To the volleyball team TUD Tallaght, which has become very important to me. For all the laughter and fun in training and matches that gave me the strength to go on. Thanks for all the good times we had together on and off the volleyball court.

To the friends I made during my stay in Dublin, who became like a second family to me. Especially Liam and Lucia, who were always there for me when I needed them. For all the parties, the dinners, the trips, the chats and, in short, everything we have been through and lived together. Thank you for being my support and letting me be part of your lives this year.

To my lifelong friends, who, despite the distance, I always feel close to them. To Barbet and Miguel, with whom I don't need to be in contact to know that they will always be there for me. To Lusin and Baonz, with whom I can talk about anything even if it has been weeks since the last time. And above all, to Anita, for all the conversations we have had that allowed me to get things off my chest. Thanks to all of you for the love and support you give me.

Finally, to my parents, who have given me the opportunity to live the Erasmus experience, being able to grow as a person, live unique experiences and meet amazing people. For everything you have invested in me, the education you have given me, and, above all, the unconditional support throughout my life. I couldn't be luckier to be your son. Thank you very much for everything.

List of Figures

| | | |
|----|--|----|
| 1 | CT scan planes | 12 |
| 2 | Cyst axial CT scan | 13 |
| 3 | Cyst coronal CT scan | 13 |
| 4 | Tumor axial CT scan | 14 |
| 5 | Tumor coronal CT scan | 14 |
| 6 | Healthy axial CT scan | 14 |
| 7 | Healthy coronal CT scan | 14 |
| 8 | Abdominal angle coronal CT scan | 15 |
| 9 | Thoracic angle coronal CT scan | 15 |
| 10 | Original image | 19 |
| 11 | Augmented image | 19 |
| 12 | Small model diagram | 20 |
| 13 | Small model accuracy plot | 21 |
| 14 | Medium model diagram | 23 |
| 15 | Medium model accuracy plot | 24 |
| 16 | Large model diagram | 26 |
| 17 | Large model accuracy plot | 27 |
| 18 | Modified large model accuracy plot | 28 |
| 19 | Training images confusion matrix | 30 |
| 20 | Unseen cyst | 31 |
| 21 | Unseen normal | 31 |
| 22 | Unseen tumor | 31 |
| 23 | Unseen tumor CT scan | 31 |
| 24 | Unseen images confusion matrix | 32 |
| 25 | Correctly classified tumor | 33 |
| 26 | Incorrectly classified tumor | 33 |
| 27 | Modified large model | 35 |
| 28 | Modified model confusion matrix | 36 |
| 29 | 500 normal images confusion matrix | 37 |
| 30 | 3-subset confusion matrix | 42 |

List of Tables

| | | |
|---|---|----|
| 1 | Small model modifications & results | 22 |
| 2 | Medium model modification & results | 24 |
| 3 | Large model modifications & results | 27 |
| 4 | Modifications & results of all models | 29 |
| 5 | Modified large model results | 35 |
| 6 | Normal training images reduction test results | 37 |
| 7 | Training images reduction test results | 39 |

Contents

| | | |
|----------|---|-----------|
| 1 | Introduction | 9 |
| 2 | Literature Review | 10 |
| 3 | Data Selection | 12 |
| 3.1 | Cyst | 13 |
| 3.2 | Tumor | 13 |
| 3.3 | Normal | 14 |
| 4 | Data Collection & Preparation | 15 |
| 5 | Data Preprocessing & Augmentation | 16 |
| 5.1 | Data Preprocessing | 17 |
| 5.1.1 | Rescale | 17 |
| 5.1.2 | Grayscale | 17 |
| 5.1.3 | Normalisation | 17 |
| 5.2 | Data Augmentation | 18 |
| 6 | Model development | 19 |
| 6.1 | Small | 20 |
| 6.2 | Medium | 22 |
| 6.3 | Large | 25 |
| 7 | Model Performance | 28 |
| 7.1 | Training data | 28 |
| 7.1.1 | Performance results | 29 |
| 7.2 | Unseen data | 30 |
| 7.2.1 | Performance results | 32 |
| 8 | Discussion (Including Ethics & Bias) | 33 |
| 8.1 | Evaluation | 34 |
| 8.1.1 | Model | 34 |
| 8.1.2 | Data set | 40 |

| | | |
|-----------|---|-----------|
| 8.2 | Ethical & bias background | 43 |
| 8.2.1 | Human Agency & Oversight | 43 |
| 8.2.2 | Technical Robustness & Safety | 43 |
| 8.2.3 | Transparency | 44 |
| 9 | Conclusions | 46 |
| 10 | Future Work | 47 |

Abstract

Cancerous carcinomas are the most common kidney tumours, accounting for more than 90% of clinical cases. Studies show that more and more cases are being diagnosed with cancerous masses, with more than 400,000 new cases reported worldwide in 2020. The difficulty of diagnosis between renal tumours and renal cysts implies the need to develop an accurate diagnostic system. This paper proposes a CNN model for detecting pathological kidneys containing cysts or cancers. Several topologies were studied, focusing mainly on the comparison of the models versus their size, selecting the large one as the best of them. The model, trained and validated with a 66% - 33% percentage split, is able to classify tumours, cysts and healthy kidneys with 99.91% accuracy. When tested on clinical cases, the model obtains a maximum accuracy of 31.58%. The study conducted to justify such variation in performance indicates that the training images are not representative of the real cases. Either because of the alterations they may have undergone or because of the specific characteristics they follow, the model is limited and conditioned by the images it can classify.

1 Introduction

According to estimates from the World Health Organization in 2019, cancer is the first or second most common cause of death before the age of 70 in 112 out of 183 countries and ranks third or fourth in 23 other countries [Sung et al. 2021]. Cancer incidences are growing rapidly worldwide, many of which are associated with socio-economic development in the more advanced countries. Studies estimate that there were more than 400,000 new cases of kidney cancer in 2020, as well as more than 170,000 deaths. Kidney cancer is in the 16th position out of 36 with the highest number of cases, as well as in the 15th position in terms of mortality. [ibid.]

Renal cell carcinomas, or RCCs, account for more than 90% of all kidney cancers [Hsieh et al. 2017]. RCCs can be removed by conventional treatments. However, metastatic cases are resistant to such techniques. Metastasis is the spread of cancer cells to other parts of the body, which occurs in approximately 33% of cases [Flanigan et al. 2003].

Renal cysts are challenging to diagnose as they often co-exist with RCCs. In fact, studies suggest that renal carcinomas may develop through cysts [Moch 2010]. This implies an additional difficulty in the interpretation of medical tests, as some cases of cysts lead to the development of cancerous tissues, which may result in negligence due to misdiagnosis.

Advances in computational processing power, memory and storage open up the possibility of tackling challenges in the field of healthcare. Previous studies demonstrate the effectiveness of using convolutional neural networks, or CNNs, in tasks such as medical image analysis [Deo 2015]. The increase of cancer cases encourages the use of image recognition and feature extraction from CNNs to identify such diseases, avoiding the generation of metastasis in the patient, which gives more options to improve their health and provides timely opportunities for intervention.

In this paper, a convolutional neural network developed for the classification of kidney tumours and cysts is proposed. The main objective of this project is to achieve the highest possible accuracy in the classification of cysts and tumours so that it can be used as a medical aid in making diagnoses.

2 Literature Review

Vasanthselvakumar, Balasubramanian and Sathiya 2020 proposed a method for the recognition of chronic kidney disease. The Histogram of Oriented Gradient, or HOG, is used to detect particular characteristics of the kidney region. This technique applies gradients to the image by comparing the contrast of the pixels around each pixel in the image, determining the basic structure of the images. The extracted HOG features are then used as input to the AdaBoost classifier to localise kidney diseases. A convolutional neural network is finally used as a means of recognising the diseases defined in its training. The accuracy in detecting kidney diseases is 89.79%, and their classification with the CNN achieves an accuracy of 86.67%.

Türk, Lüy and Barışçı 2020 proposed a new hybrid model based on the existing V-Net models, a convolutional neural network for 3D image segmentation. V-Net enables seamless segmentation of 3D images, with high accuracy and performance, and can be adapted to solve many different segmentation problems. The objective is to detect more features of the same image using fewer modalities, i.e. types of information. The V-Net fusion model introduces multiple parameters in the encoder part of the neural network, the one in charge of converting the data into a required format for further decoding, so that a high level of learning is obtained without large amounts of data. The proposed model showed better segmentation performance than existing image models, with average Dice coefficients, a measure of similarity between the V-Net fusion model output and the ground truth annotation, of 97.7% and 86.5% for kidney and tumour segmentation, respectively.

Hussain et al. 2017 developed a collage-based convolutional neural network for the detection of pathological kidneys containing renal cell carcinomas. CT images are placed side by side, which forms a collage composed of all images from the same CT scan. The idea behind this is that not all CT slices necessarily show tumours; however, all images are labelled based on the status of the kidney as a whole. As it is not correct to use conventional CT slices, as the volume-based labelling is not applicable to all constituent axial slices, they are regrouped so that

they constitute a single, correctly labelled image. The model, validated with a CT dataset of 160 patients, classifies cancerous and normal kidneys with 98% accuracy.

Nithya et al. 2020 proposed the detection of kidney stones and cancers by means of a neural network and segmentation using a multi-core k-means clustering algorithm. The ultrasound images are preprocessed, removing the speckle noise present, i.e. the noise that arises due to the effect of environmental conditions on the imaging sensor during image acquisition, and the main features are extracted from them. After classifying the images as normal or abnormal by means of an artificial neural network, the abnormal ones are introduced in the segmentation stage, where the stone or tumour present in the image is extracted. The results obtained show that the proposed system achieves a maximum accuracy of 99.61%.

Tuncer and Alkan 2018 provided a decision support system for detecting renal carcinomas using abdominal images of healthy kidneys and tumour tissues. Such a process is composed of two steps: the segmentation of the kidney areas in the images by means of the K-means algorithm; and the subsequent classification of these segmented images into healthy or tumourous based on the Support Vector Machines, or SVM. A total of 130 images were used, obtained from the Radiodiagnostic Department of Firat University Medical Faculty. The proposed system obtained a Dice coefficient in segmentation of 89.3%, while the classification achieved 88% accuracy.

3 Data Selection

Finding a suitable dataset was not trivial. The anonymity and privacy of hospitals data hindered the access to the required data for the training of the proposed neural network. This made the data selection process challenging.

An alternative open source dataset was identified, the "CT KIDNEY DATASET: Normal-Cyst-Tumor and Stone" dataset from the Kaggle website. It consists of a total of 11069 images, with a class breakdown of 3709 of kidney cyst CT scans, 2283 of kidney tumours and the remaining 5077 of normal or healthy images. The three classes are composed of axial and coronal CT scans, which provide different angles of the human body for a more accurate interpretation. While the coronal CT scans are composed of different image sizes, the axial ones follow a size of 512x512 pixels, the standard size for CT images [Sureshbabu and Mawlawi 2005].

CT scans produce signals that are processed to generate cross-sectional images, or "slices", of the body. Thus, this dataset contains a series of image sequences that traverse the subjects' bodies in the two planes mentioned above: the axial and coronal planes.

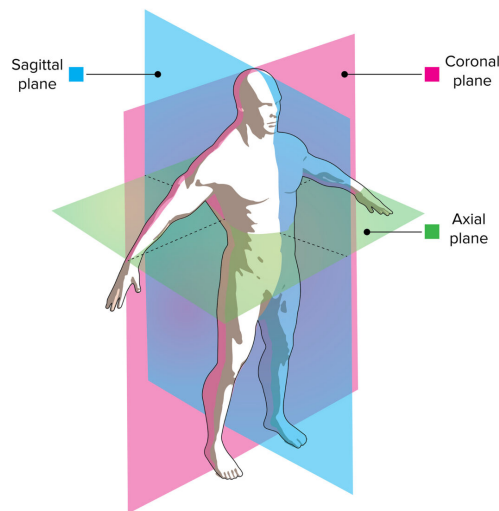


Figure 1: CT scan planes

An axial plane describes a vertical trajectory of the human body, producing, as a result, a sequence of images that section it from vertex to base.

A coronal plane, on the other hand, traverses the human body horizontally, or what is the same, from the front to the back of the user. A progression of the human torso is thus obtained, providing an alternative perspective which presents additional visual information on the state of the kidneys.

3.1 Cyst

Simple renal cysts are a persistent finding on CT scans, and are seen in more than half of patients over 55 years of age. Simple cysts are benign, fluid-filled structures [Herring 2016]. Furthermore, they tend to have well-defined borders on CT images [ibid.].

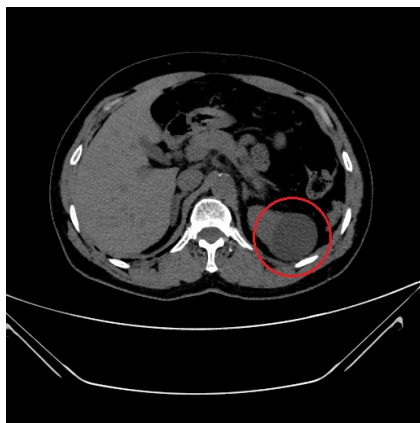


Figure 2: Cyst axial CT scan

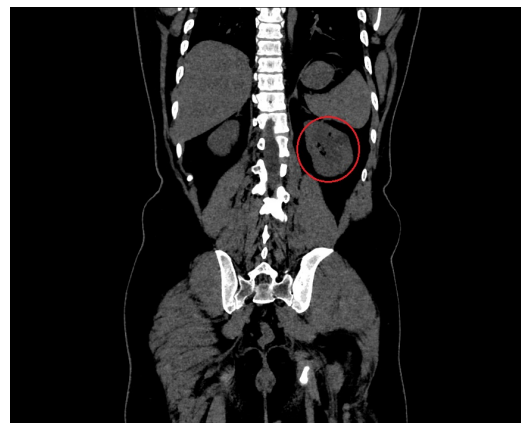


Figure 3: Cyst coronal CT scan

Figures 2 and 3 show different perspectives of cystic masses present in the kidneys. Both images show a well-defined round shape and a darker colour, which characterises the liquid inside it.

3.2 Tumor

Renal cell carcinoma is the most common primary renal cell carcinoma in adults. They may be completely solid or completely cystic, but are usually solid lesions

that may contain areas of low-attenuation necrosis [Herring 2016]. In either case, the wall defining the mass is thicker and more irregular than that of a simple cyst [ibid.].

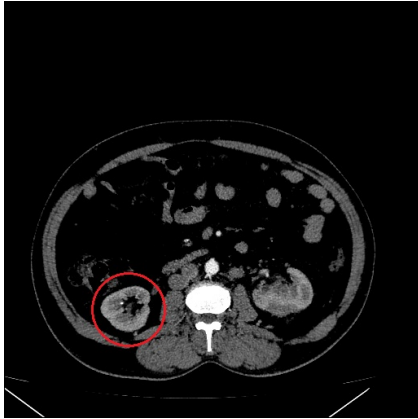


Figure 4: Tumor axial CT scan

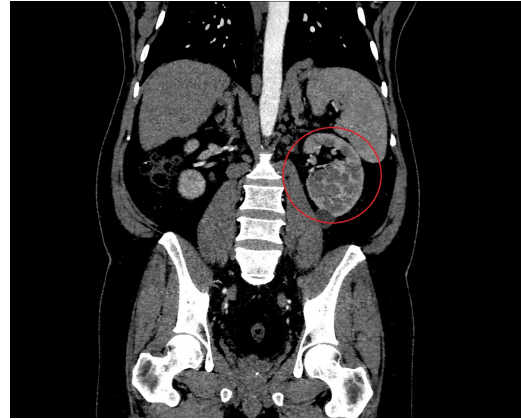


Figure 5: Tumor coronal CT scan

The left kidney in figure 4 shows an irregular dark mass, while the left kidney in figure 5 is characterised by necrosis, which justifies the faint areas within the mass.

3.3 Normal

Healthy kidneys are identified by the correct definition of their shape, similar to that of a bean. In contrast to the other two classes, no peculiar mass is identified in them.



Figure 6: Healthy axial CT scan

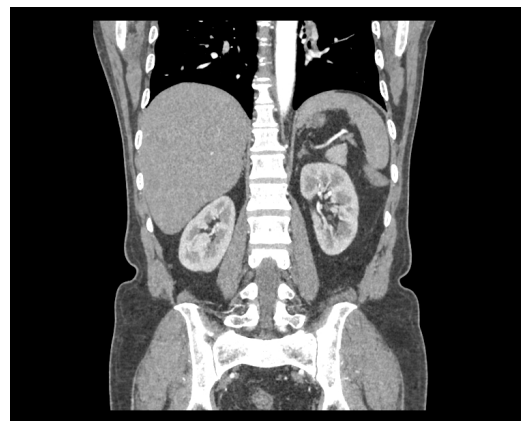


Figure 7: Healthy coronal CT scan

As can be seen, no abnormal masses or figures are visible in the above images. Figures 6 and 7 show patients with healthy kidneys that are defined by their characteristic shape, without deformation.

4 Data Collection & Preparation

The dataset obtained already went through a preparation process. It was collected from PACS, Picture Archiving and Communication System, from several hospitals in Dhaka, Bangladesh. After axial and coronal slices were selected from the contrast and non-contrast studies, i.e. two different approaches with more or less contrast in the examined area, the images were processed [Islam 2021]. Each diagnosis was studied, selecting those images that pictured the region of interest, i.e. the abdominal region, to visualize the kidneys. Then, all information from the patients was excluded, as well as the metadata from the DICOM images, which were eventually converted to a JPG format. After the preparation, all images were again checked by a radiologist and a medical technologist to confirm the accuracy of the data [ibid.].

An additional preparation process was subsequently applied in this project. Through a thorough analysis of the class set of the selected dataset, it was found that the coronal images have different angles.



Figure 8: Abdominal angle coronal CT scan

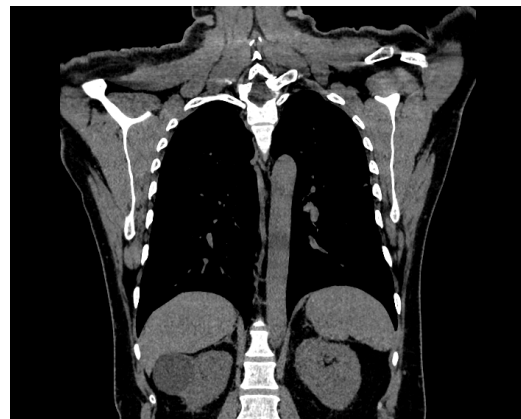


Figure 9: Thoracic angle coronal CT scan

As can be seen in the figures, despite being the same plane for the CT scan, the defined area is not the same. Figure 8 corresponds to the abdominal area of the user, while figure 9 shows the thoracic cavity. Furthermore, it was found that the Normal class does not have thoracic images but only has coronal images of the abdominal cavity.

A further investigation was carried out where the images of each class were grouped by the type of plane used. Thus, it was documented that out of a total of 11065 images, 4208 are coronal scans, of which 771 are of the thoracic cavity. Due to their low number and the fact that one of the classes does not have such images, it was decided to remove them from the dataset. This leaves 3437 coronal images, of which 805 are cysts, 1803 are normal, and 829 are tumours. It was also decided not to use such images for training the model, as it was concluded that having a low number of these images could impair and confuse the extraction of features from the axial images. Therefore, 6857 images are available from this point on, being 2243 cysts, 3274 normal and 1340 tumours, all with an axial plane.

5 Data Preprocessing & Augmentation

Data Preprocessing is an essential task in every machine learning project. Usually, raw data cannot be directly used. It requires some previous transformation in order to be more suitable for modelling. The cyst and tumour dataset is composed of RGB images of different sizes and ranges of values. These different characteristics could affect the correct performance of the model, as it expects input images with equal proportions and the differences they present could lead to over-fitting. Therefore, the raw data must be preprocessed before being used to fit and evaluate a deep learning algorithm.

Tensorflow was the preferred tool for this purpose. It provides support for Keras, an open-source Neural Networks library written in Python.

Tensorflow 2.7 was the version used to preprocess the images, as it offers a wide range of functionalities that allowed to properly prepare the data for its further

use, along with Python 3.7.

5.1 Data Preprocessing

5.1.1 Rescale

Neural Network algorithms perform better with images in the same dimensions. Rescaling the images was then the first task to do. Images loaded in the dataset were iterated and resized, maintaining the original aspect ratio to avoid data modification.

A size of 224 x 224 was set for all images, at which the files are of an adequate size to work with while still clearly appreciating the information they show. Some of the studies covered in section 2 use a size of 256x256 pixels [Türk, Lüy and Barışçı 2020], even 512x512 [Hussain et al. 2017], obtaining good results in their analysis. However, a larger image size requires more computational operations per layer, as well as more computational requirements that are not available for this study.

5.1.2 Grayscale

Grayscale involves turning the images from RGB into a grayscale by applying a dimensionality reduction. Coloured images are composed of three different arrays for red, green and blue values, whereas grayscale ones only have one array. This means complexity is added when working with RGB images, as convolutions must be performed on the three different arrays, while only one is needed with gray-scaled ones. By removing all colour information and just leaving the luminance of each pixel, the redundant information is discarded, and therefore the model's calculations are improved.

5.1.3 Normalisation

Once grayscale, the input parameters, i.e. the pixel values, are in the same range of values from 0 to 255. Normalizing these values means that they become in the range of 0 to 1, which implies a considerable improvement in the training speed of the model.

5.2 Data Augmentation

Data augmentation benefits the neural network's image recognition by exposing it with perturbed versions of the existing data. This makes it less likely that the neural network recognizes unwanted characteristics in the cyst and tumour images.

As a result, the following effects were randomly applied to the preprocessed images:

- Horizontal flipping:
A horizontal flip was applied to the image, so that there is a greater variety of cysts and tumors in the training images.
- Contrast:
Images extracted from contrast and non-contrast studies suggest the existence of different contrast settings used in CT scans. This implies the need to cover different contrast cases in order to train the model in a wider range of possible situations.
- Rotation:
A random rotation was applied as an added difficulty to the feature extraction process of the model, in order to provide more variation and avoid overfitting to the training images. With a factor of 0.1, the output rotates by a random amount in the range $[-10\% * 2\pi, 10\% * 2\pi]$.
- Zooming:
A random zoom allows, once again, to bring variety to the training phase. With a zoom of 0.1, more images were obtained without resulting excessive or unrepresentative.

Keras has several built-in tools consisting of layers sequentially applied to the images to modify their characteristics. Once applied, transformations are present in the dataset, such as in the following case:

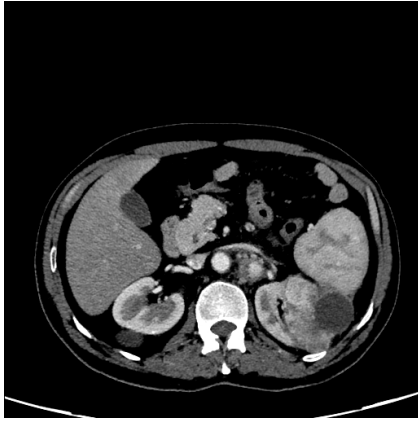


Figure 10: Original image

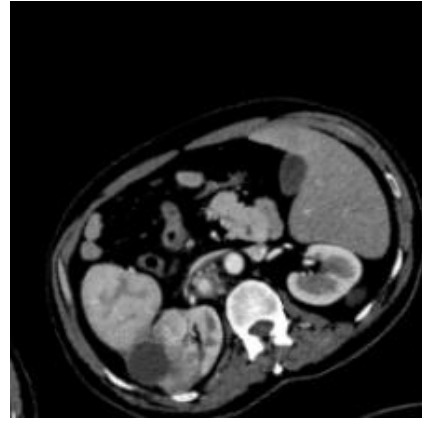


Figure 11: Augmented image

Figure 11 is a good example of the data augmentation that figure 10, the original image, has been through. Not only has it been rotated, but also been flipped horizontally and randomly zoomed. As a result, the model now has more variety in its training data, which will result in being more robust and accurate.

It should be noted that the operation of Keras layers modifies the images without preserving the original ones, i.e. the data are overwritten. Therefore, to also provide the neural network with the original preprocessed images, both groups were saved in local storage and subsequently merged, resulting in two folders containing the cyst and tumour images separately.

6 Model development

This section covers the model's development process for the classification of kidney images. This section covers the model development process for kidney image classification. The main workspace used was a local environment. However, given the need for more memory capacity to run some of the larger models, the Integrated Development Environment, or IDE, Google Colab was also used.

The comparison presented is based on a model's sequential evolution, developing as its results are evaluated and changes are applied to improve it. This finally

leads to the main comparison of 3 models, each larger than the previous one, and a comparative table of their characteristics and results is presented.

6.1 Small

The first attempt of defining a model led to the one known as "small", which is composed of the following features:

- 1 x Convolution layer of 12 filters of 3x3 kernel and stride 2
- 1 x Maxpool layer of 2x2 pool size and stride 2
- 1 x Convolution layer of 32 filters of 3x3 kernel and stride 2
- 1 x Maxpool layer of 2x2 pool size and stride 2
- 1 x Dense layer of 60 units
- 1 x Dense Softmax layer of 3 units

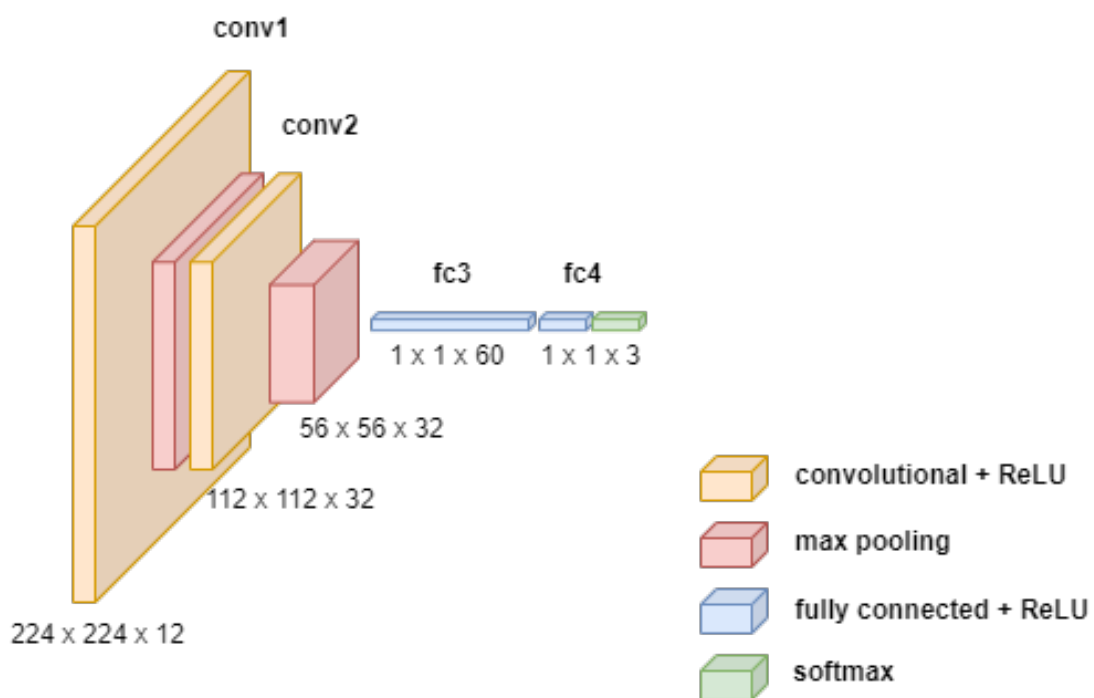


Figure 12: Small model diagram

This model generated, with a batch size of 32 and 10 epochs, the following output:

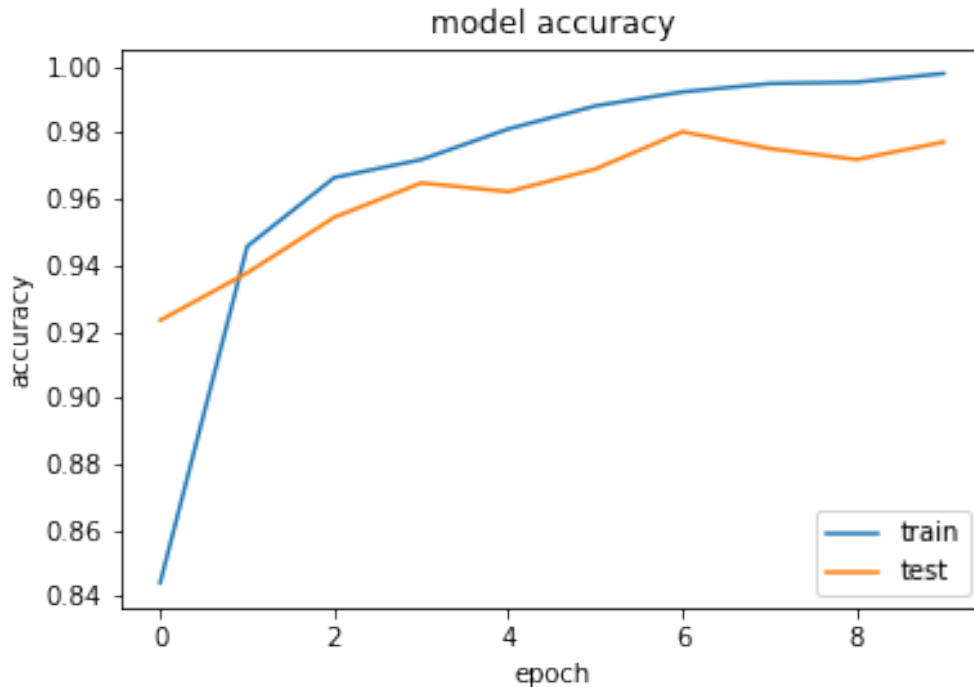


Figure 13: Small model accuracy plot

It produced a final train accuracy of 99.79% and a test one of 97.72%, as well as a train loss of 0.0082 and a test loss of 0.0628. The model presents some overfitting, as the training accuracy is slightly higher than the test one, so some regularization was applied in order to correct it. After applying a 20% dropout on the dense layer, accuracy declined and overfitting remained present. Instead, a bigger stride was used, with a size 7 x 7. Although the total accuracy improved, with a value of 99.65%, the difference between the training and testing accuracy was still significant. A last trial was done where the batch size was increased to 120. In general, smaller batch sizes result in noisier gradient estimates, so they can better escape poor local minima and avoid overfitting. However, small batches may be too noisy for good learning. With such a large amount of data available for training, a bigger batch size could improve the model's performance. As a result, overfitting was slightly reduced, as also was its accuracy.

The aforementioned modifications and results can be visualized in the following table:

| Small model modifications & results | | | | | | | |
|-------------------------------------|-------------|---------|----------------|------------|---------------|-----------|----------------|
| BS | Kernel size | Dropout | Train accuracy | Train Loss | Test accuracy | Test Loss | Total accuracy |
| 32 | (3, 3) | NO | 99.79% | 0.0082 | 97.72% | 0.0628 | 99.15% |
| 32 | (3, 3) | 0.2 | 99.34% | 0.018 | 97.15% | 0.0841 | 98.52% |
| 32 | (7, 7) | NO | 100% | 4.7049e-04 | 98.94% | 0.0312 | 99.65% |
| 120 | (7, 7) | NO | 99.65% | 0.0158 | 98.39% | 0.0439 | 99.29% |

Table 1: Small model modifications & results

The overall results show how a bigger kernel size provides better accuracy results. A larger batch size also reduces the overfitting present in the first version of the model.

6.2 Medium

A slightly bigger model was implemented, taking into consideration the previous features used, to evaluate how a more complex model would behave in training and testing the data. The "medium" model consists of the following features:

- 1 x Convolution layer of 32 filters of 7x7 kernel and stride 2
- 1 x Maxpool layer of 2x2 pool size and stride 2
- 1 x Convolution layer of 64 filters of 7x7 kernel and stride 2
- 1 x Maxpool layer of 2x2 pool size and stride 2
- 1 x Dense layer of 60 units
- 1 x Dense Softmax layer of 3 units

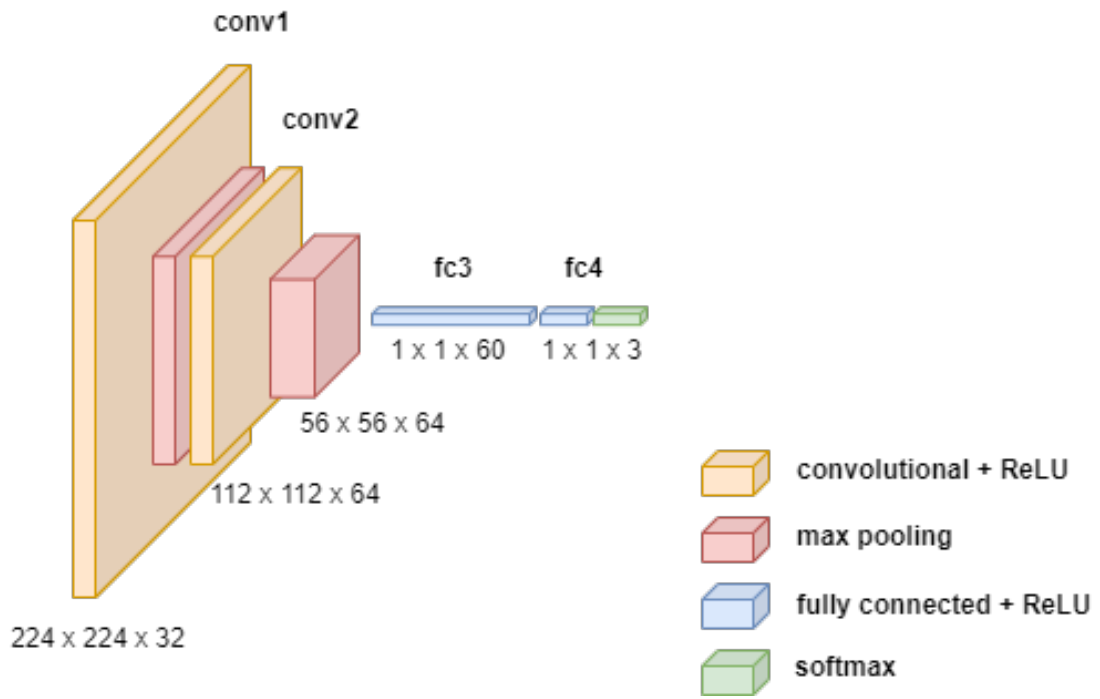


Figure 14: Medium model diagram

The proposed model, using a batch size of 120 and 10 epochs, obtained the following results:

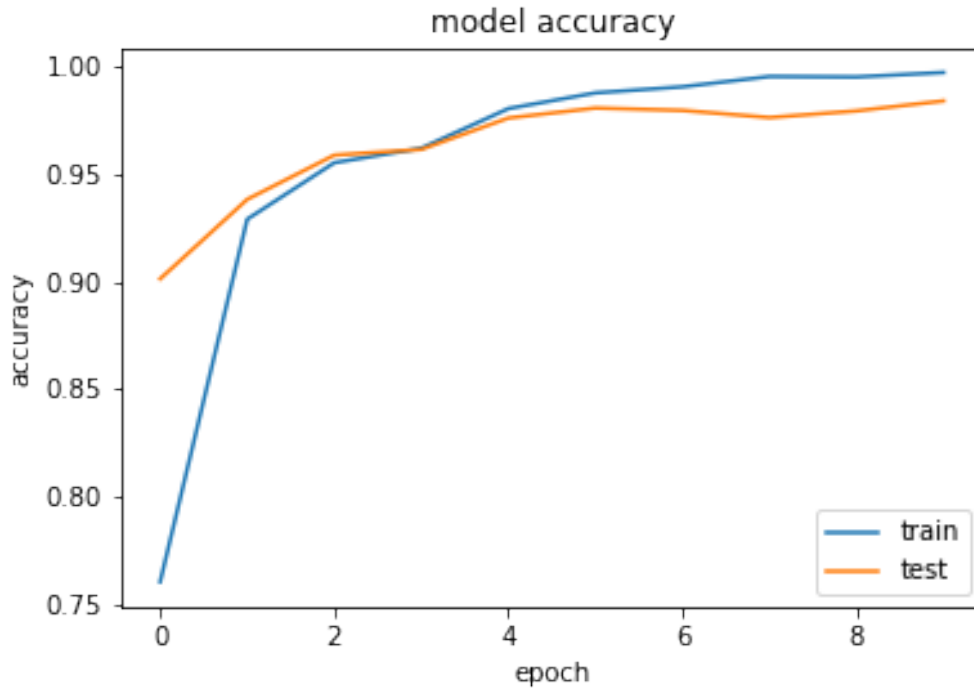


Figure 15: Medium model accuracy plot

A minor presence of overfitting was found in these results, compared to those of Figure 13; achieving a total accuracy of 99.12%, being specifically a 99.71% accuracy for training and 98.39% accuracy for testing. Nonetheless, a second attempt was performed, in which a 40% dropout was applied on the dense layer to analyse if the overfitting could be improved even further. The results showed the removal of overfitting at the cost of accuracy, with a 98.31% accuracy in training and a 98.92% accuracy in testing.

| Medium model modifications & results | | | | | | | |
|--------------------------------------|-------------|---------|----------------|------------|---------------|-----------|----------------|
| Batch size | Kernel size | Dropout | Train accuracy | Train Loss | Test accuracy | Test Loss | Total accuracy |
| 120 | (7, 7) | NO | 99.71% | 0.0102 | 98.39% | 0.0436 | 99.22% |
| 120 | (7, 7) | 0.4 | 98.31% | 0.0459 | 98.92% | 0.0326 | 99.56% |

Table 2: Medium model modification & results

The fact that the use of dropout improved the model's performance may be due to the added complexity in this model with respect to the previous one. In order to find out, a larger model is subsequently presented.

6.3 Large

The next model, known as "large", includes an additional convolutional layer than the previous ones. This test aims to determine whether adding complexity to the model is disturbing its results or, on the other hand, having more filters allows for a better image prediction. Thus, it is composed of the following characteristics:

- 1 x Convolution layer of 32 filters of 3x3 kernel and stride 1
- 1 x Maxpool layer of 2x2 pool size and stride 2
- 1 x Convolution layer of 64 filters of 3x3 kernel and stride 1
- 1 x Maxpool layer of 2x2 pool size and stride 2
- 1 x Convolution layer of 128 filters of 3x3 kernel and stride 1
- 1 x Maxpool layer of 2x2 pool size and stride 2
- 1 x Dense layer of 60 units
- 1 x Dense Softmax layer of 3 units

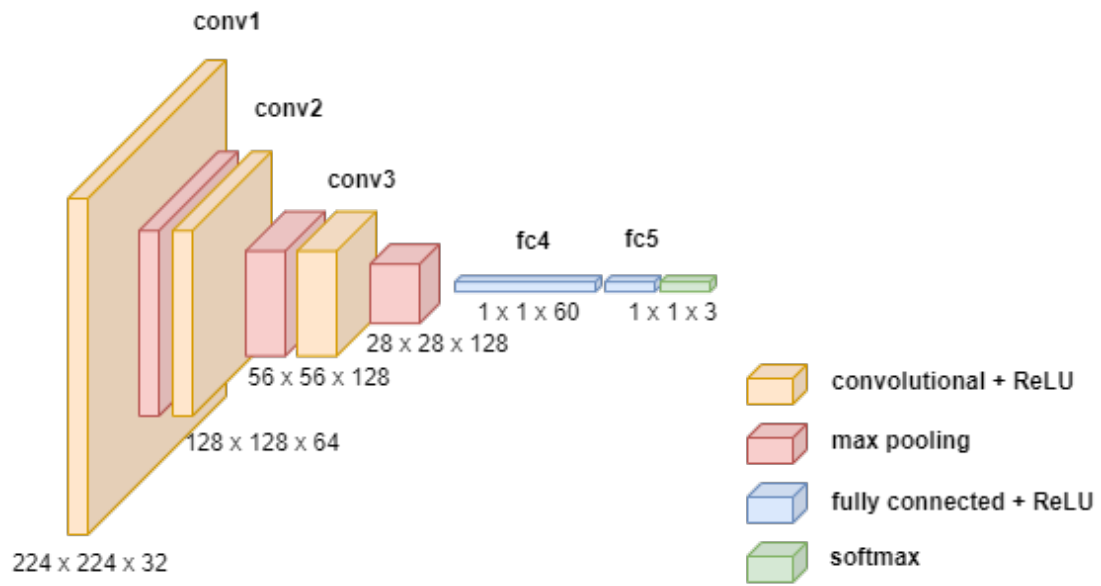


Figure 16: Large model diagram

As before, this model was configured with a batch size of 120 and 10 epochs. This produced the following outcome:

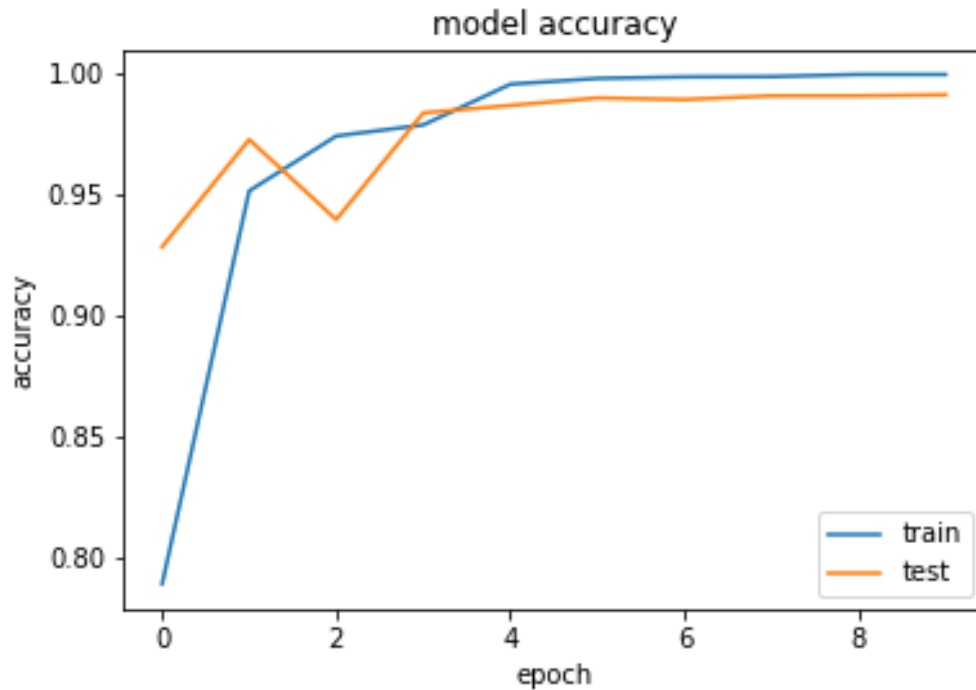


Figure 17: Large model accuracy plot

Figure 17 shows a much more balanced progression between training and testing results over the number of epochs, with a final training accuracy at epoch 10 of 100% and a testing one of 99.16%.

Another test was done with a kernel size of 7x7 and stride 2. This model gave some significant better results, which can be seen compared to the first ones in the following table:

| Large model modifications & results | | | | | | | |
|-------------------------------------|-------------|---------|----------------|------------|---------------|-----------|----------------|
| Batch size | Kernel size | Dropout | Train accuracy | Train Loss | Test accuracy | Test Loss | Total accuracy |
| 120 | (3, 3) | NO | 100% | 2.6619e-04 | 99.16% | 0.0261 | 99.72% |
| 120 | (7, 7) | NO | 100% | 8.2500e-04 | 99.73% | 0.0090 | 99.91% |

Table 3: Large model modifications & results

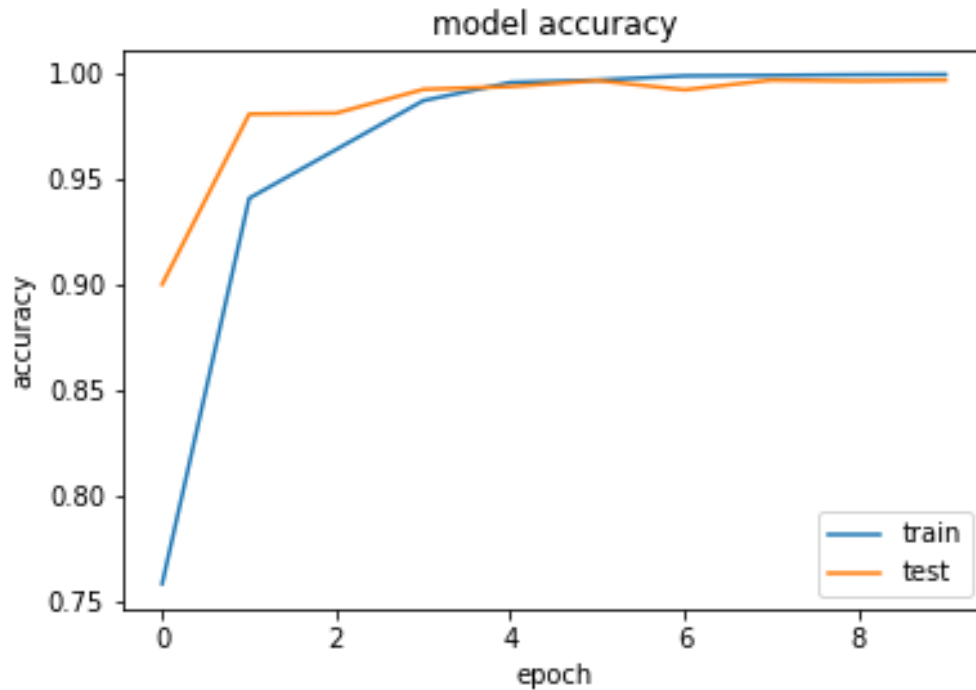


Figure 18: Modified large model accuracy plot

No signs of overfitting can be observed as the difference between the training and testing accuracy results, 100% and 99.73%, respectively, are minimal. The model provides in the second attempt some remarkable results, having a total accuracy of 99.91%.

7 Model Performance

7.1 Training data

This section will compare the results obtained from training the different models implemented. The technique used during this step to validate the model was a 66% - 33% percentage split of the randomised data, which allowed the comparison between training and validation results and helped to determine the modifications needed to improve the prediction result.

The only metric used as a method of evaluation is the accuracy obtained from each model; no other metric is required due to the results shown in the previous section. Such is the accuracy of the models during the training process with such a large data set that it is not possible for them to be biased towards a particular class. Given that the results of the model training and validation are around 100% accuracy, it can be assumed that the sensitivity and specificity values are similar and therefore do not need to be studied.

The best model will also be chosen among all the proposed options to predict new unseen data.

7.1.1 Performance results

The aforementioned outcomes are compiled and shown in the following table:

| Modifications & results of all models | | | | | | | | |
|---------------------------------------|------------|-------------|---------|----------------|------------|---------------|-----------|----------------|
| Model | Batch size | Kernel size | Dropout | Train accuracy | Train Loss | Test accuracy | Test Loss | Total accuracy |
| Small | 32 | (3, 3) | NO | 99.79% | 0.0082 | 97.72% | 0.0628 | 99.15% |
| Small | 32 | (3, 3) | 0.2 | 99.34% | 0.018 | 97.15% | 0.0841 | 98.52% |
| Small | 32 | (7, 7) | NO | 100% | 4.7049e-04 | 98.94% | 0.0312 | 99.65% |
| Small | 120 | (7, 7) | NO | 99.65% | 0.0158 | 98.39% | 0.0439 | 99.29% |
| Medium | 120 | (7, 7) | NO | 99.71% | 0.0102 | 98.39% | 0.0436 | 99.22% |
| Medium | 120 | (7, 7) | 0.4 | 98.31% | 0.0459 | 98.92% | 0.0326 | 99.56% |
| Large | 120 | (3, 3) | NO | 100% | 2.6619e-04 | 99.16% | 0.0261 | 99.72% |
| Large | 120 | (7, 7) | NO | 100% | 8.2500e-04 | 99.73% | 0.0090 | 99.91% |

Table 4: Modifications & results of all models

As discussed in the previous section, the small models show a distinguished performance; however, there is notable overfitting in the results provided as the training accuracy is significantly higher than the test accuracy, which indicates that the model is too well fitted to the training data.

The same issue occurs in the first medium model; nevertheless, after some adjustment, the model shows no signs of overfitting and overall good performance.

Finally, it is clear how the large models perform at an outstanding level. The stability between the training and test results in both accuracy and loss, and the total accuracy of 99.91% of the second large model, make it the best choice of all those suggested.

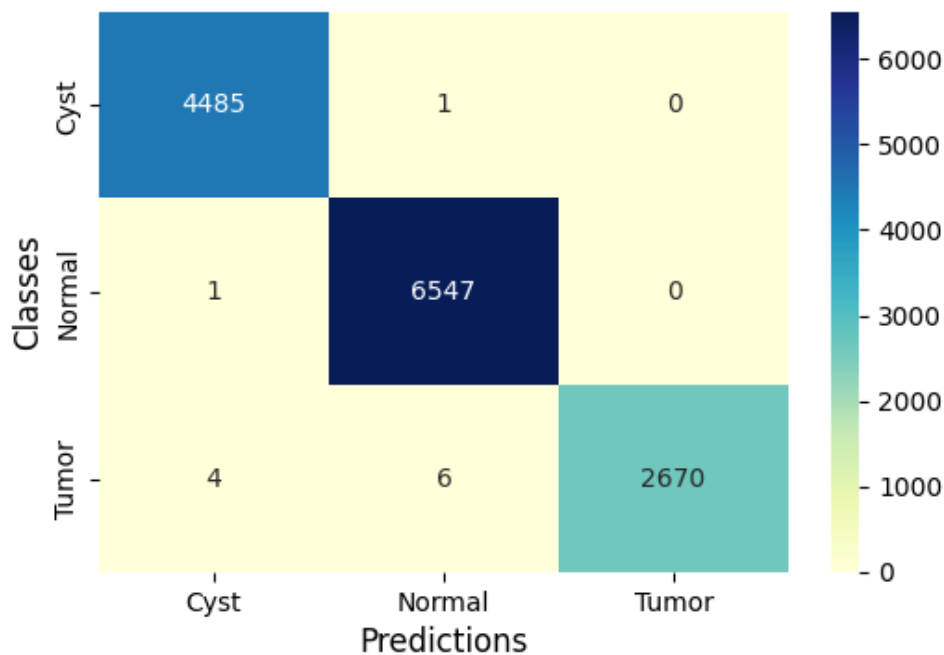


Figure 19: Training images confusion matrix

Figure 19 shows the distribution of the images to the predicted classes, with only 12 being incorrectly classified.

7.2 Unseen data

Once the best model, i.e. the large model, was determined, it will first be re-trained. It is currently trained with 66% of the data due to the percentage split applied in training, so before proceeding further, it will be trained on the entire data set.

After that, it will be evaluated by using images that do not belong to the dataset used in training. A total of 133 axial images were collected. 33 of them were obtained from Radiopaedia and correspond to the three classes, i.e. 10 images belong to the cyst class, 13 of normal or healthy kidneys and 10 of tumours.

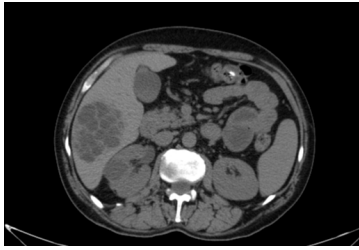


Figure 20: Unseen cyst



Figure 21: Unseen normal



Figure 22: Unseen tumor

As it can be seen, these images do not meet the same characteristics as a conventional CT scan, as they have instead different sizes and aspect ratios. Before any further testing, the images will be addressed in the same manner as the training images, discussed in sections 5.1.1 and 5.1.2.

The remaining 100 images conform to a complete CT scan of a patient with a tumour, provided by The Cancer Imaging Archive, or TCIA.



Figure 23: Unseen tumor CT scan

7.2.1 Performance results

The model was further investigated using a series of unseen image prediction tests. These images were previously processed to adjust their size to the expected 224x224 pixels and in grayscale. As a result, the prediction was far from what was expected. The model obtained an accuracy of 31.58%, which contrasts with the results obtained in the training phase.

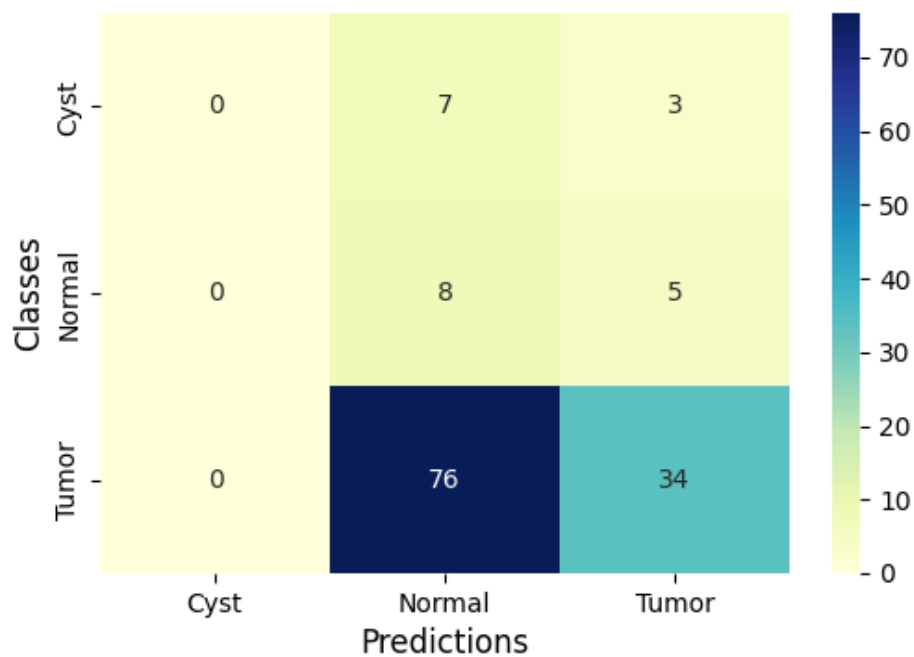


Figure 24: Unseen images confusion matrix

Figure 24 clearly shows the misclassifications made during the prediction process. It should be noted that no cyst images were correctly classified, but more interestingly, most of the 110 tumour images were labelled as normal, and only 34 of them were correctly classified.

A more precise analysis should be performed for the CT scan of the tumour. As detailed in the data selection section, a CT scan is a sequence of images along the user's body. Therefore, of the 100 images that constitute that sequence, only

a tiny portion of them correspond to the lumbar region of the body.

Of the 22 images of the CT scan where the tumour is visible, only 2 of them were classified as tumours, i.e. 9.09% of the images.

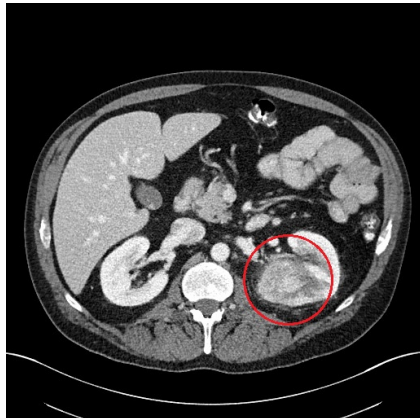


Figure 25: Correctly classified tumor

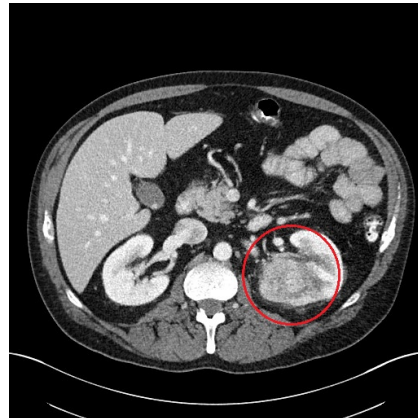


Figure 26: Incorrectly classified tumor

Figures 25 and 26 demonstrate the curious and interesting behaviour of the model. These two images belong to the same CT scan, in fact, the axial view is almost the same, but one is correctly classified while the other is interpreted as a healthy kidney.

Further research will be done in the following section regarding this behaviour, determining the cause of such interesting results.

8 Discussion (Including Ethics & Bias)

This section analyses the unexpected behaviour of the generated model's performance in classifying unseen data, trying to determine whether it is caused by the model or by the data used during training. It also discusses the ethical and bias concerns behind the use of artificial intelligence in the medical sector, specifically in the sensitive area of kidney cyst and tumour prediction.

8.1 Evaluation

In sections 8.1.1 and 8.1.2, several tests are carried out on the model and the dataset, respectively, trying to justify why there is such a variation in the results between the training and unseen data.

8.1.1 Model

The main hypothesis that arises after obtaining near-perfect results in the model training process and poor results in the unseen data test is that the model is over-fitted to the data set.

The VGG16 model, an industry-standard model that won the imagenet challenge, will be used to test the veracity of this assumption. There was, however, a lack of available memory during the model's training due to the model's size and the number of images to be processed. Not even in Google Colab it was able to proceed, so the only choice was to reduce the number of used images in training to 1000 of each class, so 3000 training images. The obtained results showed a 7.52% accuracy from VGG16, which predicted all unseen images as cysts. This, however, does not provide enough information, as this model is meant to identify more complex figures in the images, and thus it must be too convoluted for simpler objects.

Another approach is to enhance the chosen model's classification. After the feature extraction, it only has a single dense layer, so two 2000 neuron dense layers replaced it. It provided the following results:

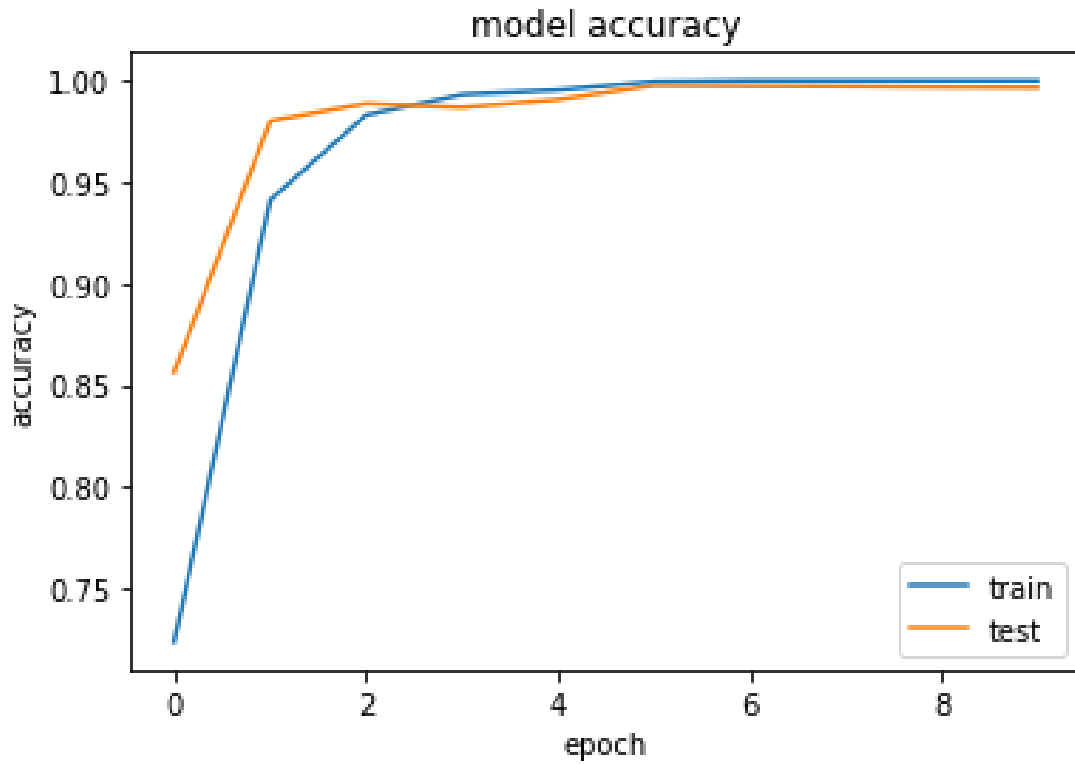


Figure 27: Modified large model

| Train accuracy | Train Loss | Test accuracy | Test Loss | Total accuracy |
|----------------|------------|---------------|-----------|----------------|
| 100% | 5.9491e-05 | 99.65% | 0.0088 | 99.88% |

Table 5: Modified large model results

Although performance is outstanding, only 22.5% of the unseen images were correctly classified.

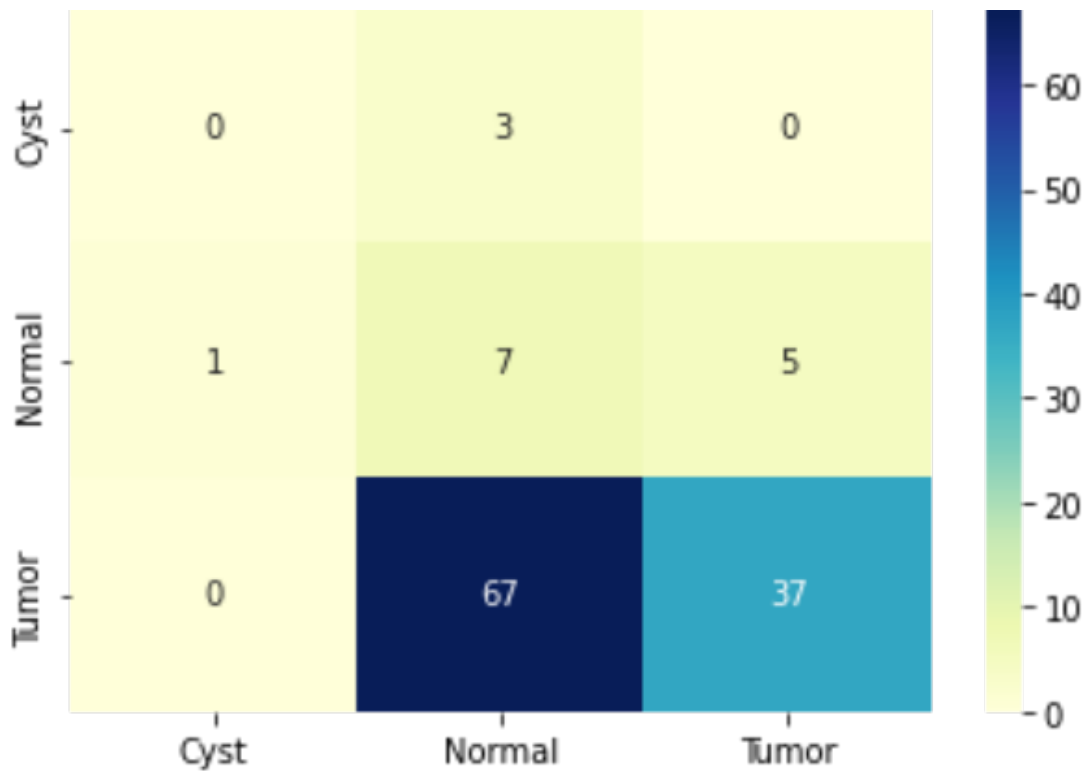


Figure 28: Modified model confusion matrix

Once again, the results differ from what was expected. It is worth noting how the main misclassification is the tumour images identified as normal ones, as occurred in the model performance section, clearly seen in figure 24.

This behaviour raised another hypothesis: The model is biased due to the higher number of normal images in the data set compared to those of cysts and tumours. Another test was performed, in which several predictions of the unseen images were made, each with fewer normal images in the training process. If an improvement in classification could be observed, it would then support the proposed assumption.

The following results were obtained from such analysis:

| Number of normal images | Accuracy |
|-------------------------|----------|
| 6548 (all) | 31.58% |
| 3000 | 36.84% |
| 2000 | 31.58% |
| 1000 | 25.56% |
| 500 | 52.63% |
| 250 | 50.38% |
| 100 | 54.13% |

Table 6: Normal training images reduction test results

As the normal images are reduced, the model’s accuracy starts to decrease as well. However, when reduced to 500 images, the accuracy increases considerably to 52.63%, slightly increasing to 54.13% in the final test with 100 normal images.

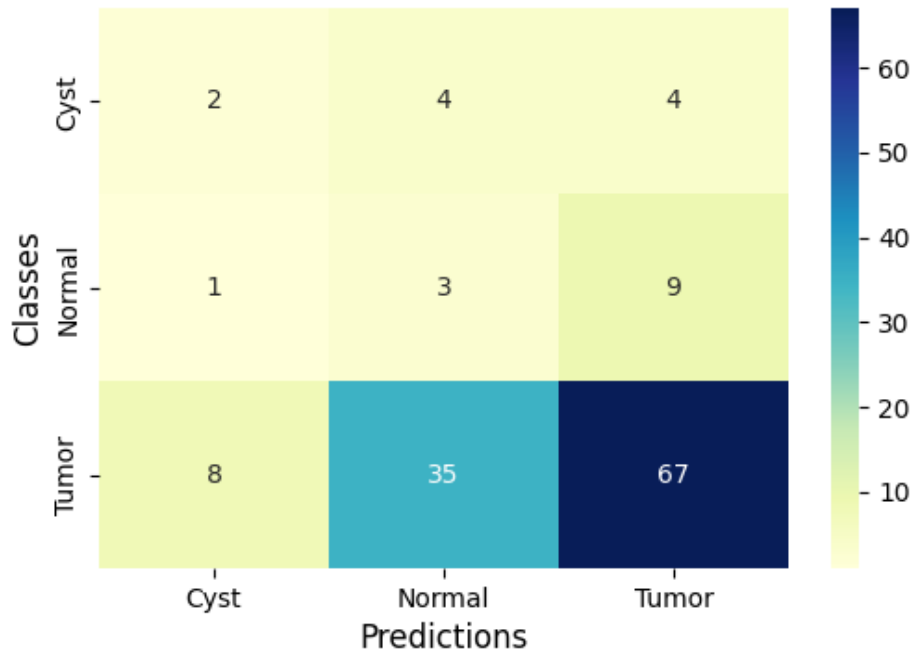


Figure 29: 500 normal images confusion matrix

Figure 29 shows the confusion matrix of the unseen images classification from the iteration with 100 normal images in the training process. An improvement in clas-

sification can be seen, as many more tumour images are now correctly classified. However, there is now a big misclassification when it comes to normal images. Due to the lack of information about that class, only 23.08% of them were correctly classified.

The following question arises in this situation: Is it a good solution to reduce the normal images to influence the model's learning? On the one hand, type 2 errors are reduced, i.e. those that indicate that a kidney is healthy when it really has either a cyst or a tumour. This change would reduce the risk of an erroneous prediction leading to a patient's worsening health, or even death. On the other hand, while it is true that there would be cases where it would be beneficial, it cannot be considered a solution. Biasing such learning favouring cysts and tumours would lead to numerous type 1 errors due to the lack of healthy kidney information available to the model. It is not a viable option to predict that a patient has a disease when he or she is perfectly healthy.

A final analysis regarding the model was performed. In order to determine which features the model learns from the images, a random subset of the training images was trained several times and then tested with the unseen images. If a significant improvement in accuracy were found, this subset's images would be studied to identify the decisive component that implies such an improvement.

A total of 9 trials were conducted, three for each of the three different numbers of training images, which were 100, 1000 and 2500. The following results were obtained:

| Number of training images | Class frequencies | Accuracy |
|---------------------------|--|----------|
| 100 | Cyst images: 36 Normal images: 42 Tumor images: 22 | 9.77% |
| 100 | Cyst images: 33 Normal images: 40 Tumor images: 27 | 9.77% |
| 100 | Cyst images: 37 Normal images: 47 Tumor images: 16 | 9.77% |
| 1000 | Cyst images: 321 Normal images: 493 Tumor images: 1186 | 20.3% |
| 1000 | Cyst images: 339 Normal images: 475 Tumor images: 186 | 18.04% |
| 1000 | Cyst images: 328 Normal images: 471 Tumor images: 201 | 36.84% |
| 2500 | Cyst images: 791 Normal images: 1224 Tumor images: 485 | 37.59% |
| 2500 | Cyst images: 786 Normal images: 1225 Tumor images: 489 | 37.59% |
| 2500 | Cyst images: 857 Normal images: 1167 Tumor images: 476 | 23.31% |

Table 7: Training images reduction test results

The difference observed in the tests for subsets of 100 images is zero. This is probably due to the small amount of information with which the model is trained. Once increased to 1000 images per subset, there is a variation in the results. The accuracy in the first two tests ranges between 18% and 20%; however, in the third test it increases to 36.84%. The images of this subset were studied, looking for particular characteristics or patterns that differentiated it from the rest, but no decisive factor was found to justify this improvement. It is possible that it was not

the type of images, but the frequency of each class, since in this third test, there is a higher number of tumour images than in the previous tests, which corresponds to a large part of the unseen images.

This assumption is rejected when observing the tests with the subset of 2500 images, where the first test consists of a large number of images of tumours, 485 to be precise, and an accuracy of 37.59% while the second test only achieved an accuracy of 27.07% with 489 images of tumours.

This implies that the model does not learn the characteristics of each class better depending on the images in the dataset. While it is true that some improvement has been observed, this is not significant with the 31.58% accuracy obtained from the model trained on all images in the dataset. Furthermore, there is no conclusive reason why the model would perform one way or another depending on which images are taken for training.

8.1.2 Data set

The above results suggest that the reason for such disparity in performance between training and unseen external data may lie in the images used in both training and testing.

A possible cause for this is that the images in the dataset differ too much from the unseen ones. As discussed in the model performance section, some images do not follow the same standard as conventional CT scans in terms of image size and quality. However, this condition was already taken into account because, as explained in the preprocessing section, a data augmentation was applied to the images, using different filters that provided the model with broader characteristics in order to adapt to different images such as these.

There is no reason why the differences between the two sets of images should affect the model's prediction in such a way.

Another possible hypothesis is that the images in the Kaggle dataset are mislabelled. As discussed in section 3, CT scans are a sequence of images that section

and pan the human body. Therefore, not all scan images show the patient's kidneys, but the before and after images show other body parts. Based on the premise put forward by Hussain et al. 2017 in section 2, the entire scans could have been integrated into the classes according to the kidneys, so there could be images in the cyst and cancer classes that do not show cystic and cancerous kidney tissues, respectively, as they are sections of the body that do not correspond to the lumbar region. However, the preprocessing applied by the author of this dataset [Islam 2021] was correct. After analysing the training images, it was verified that only the sections that explicitly show the cystic and tumour masses were added to the dataset. Therefore, the model's training could not be affected by this, as all images correspond to their assigned label.

A final attempt was made. The dataset was divided into three subsets for training (60%), validation (20%) and testing (20%) so that the model would be trained and tested with the same type of images, i.e. those in the dataset. If there is a considerable improvement in the accuracy gained, the images used are likely the main reason for the results observed so far.

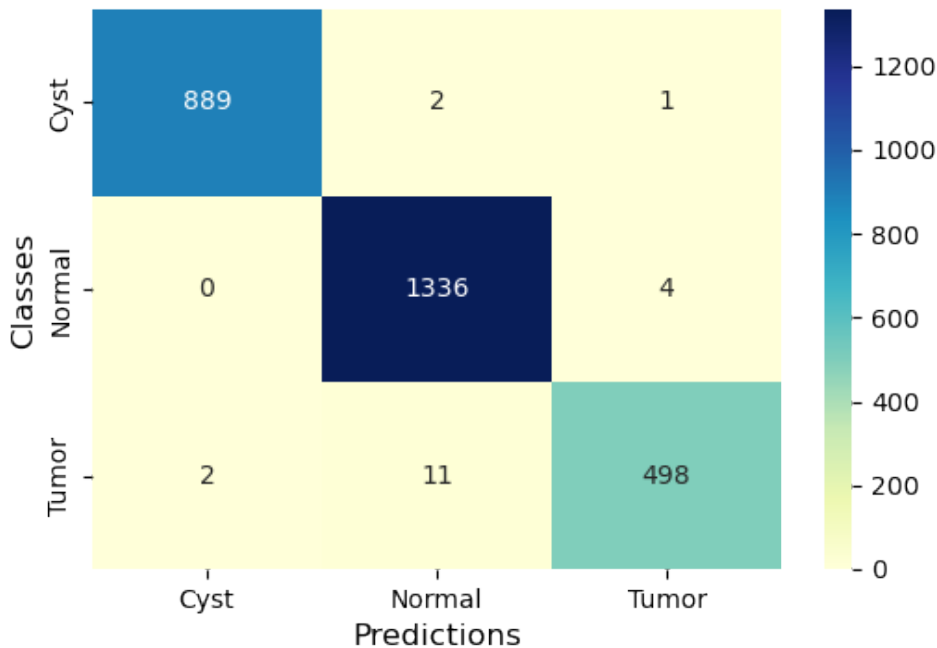


Figure 30: 3-subset confusion matrix

Figure 30 shows the classification of the test phase with 20% of the images belonging to the training data. Only 20 images were misclassified, thus obtaining a class prediction accuracy of 99.27%. Compared to the predictions of the unseen images, the model has obtained much better results.

All the tests performed lead to the conclusion that the conditioning factor for the results analyzed lies in the training images. The fact that such good results were obtained in the last test indicates that the model is over-fitted to the training images. These images may be conditioned by certain aspects, such as the scanner used or the angle of the images taken, or they may have been previously modified or altered. In other words, either because of the possible previous modifications that the training images may have undergone or because of the specific characteristics they follow, they diverge from the unseen data. This implies that the training images are not representative of the general CT standards, but rather

that the standards they follow make the model limited and conditioned by the images it is able to classify.

8.2 Ethical & bias background

This section covers several items on The Assessment List for Trustworthy Artificial Intelligence, or ALTAI [Commission, Directorate-General for Communications Networks and Technology 2020]. The importance of each of the elements covered is reflected in the ethical background of this paper.

8.2.1 Human Agency & Oversight

Medicine is a sensitive field in which the production of errors can lead to severe consequences. Expert supervision is still necessary to reinforce the possible predictions of artificial intelligence.

The generated model acts as a means to support the patient's diagnosis. It should not certify the state of health autonomously, but its function is to determine the probability that it is a case of a cyst or a tumour in the kidney. Therefore, human interaction is necessary to verify the results obtained and to make the final diagnosis.

The low results obtained in the real situation cases, i.e. the unseen images, highlight the need for human supervision. In any case, the ethics surrounding a model's prediction of a patient's health implies that a verification should be carried out, even if it has a high level of accuracy. It is possible for prediction errors to occur, leading to worsening health or even death.

8.2.2 Technical Robustness & Safety

The robustness of an AI is determined by its integrity, security and reliability.

Added to creating a model is the responsibility to preserve its integrity intact. The robustness and security of artificial intelligence are crucial points to preserving un-

biased results. Attacks on the model, such as data poisoning, i.e. manipulation of training data, or model evasion, i.e. classifying the data according to the attacker's will, could affect its results, predicting a diseased cyst as healthy. It is therefore essential to consider in the future the application of a series of security measures to avoid misuse or corruption of the model.

The accuracy of a model justifies how reliable its classifications are. As discussed above, the field of medicine does not allow for errors because of the problems they entail. This is why the accuracy of AI results, especially in this field, is so important. As far as the study is concerned, the results analysed in the previous sections show a low accuracy. In other words, the model generated is not ready to be used in real situations, as it would be potentially dangerous to determine the diagnosis based on the results it generates. The impossibility of avoiding misclassifications once again highlights the importance of human supervision over AI predictions.

The integrity of a model is determined by its reliability and reproducibility. The model built in this paper cannot be considered reliable. While it is likely to classify healthy kidneys correctly, its prediction of the other classes does not provide the confidence to be used in real cases. In terms of reproducibility, the model has performed similarly with unseen images throughout the tests performed. This may help find out more about its interesting behaviour and possibly improve it in the future.

8.2.3 Transparency

Several factors can determine the decision an AI makes. It is of great importance to be able to explain how this conclusion has been reached in order to be able to proceed with it with confidence.

The tests carried out in this paper led to the conclusion that the reason for the significant contrast in results between the training and the unseen images is that the training images are not representative of conventional CT images. Although possible causes for this were discussed earlier, this justification is not valid if one

wishes to implement the AI of this paper in medicine. As "Ethics guidelines for trustworthy AI" states: "Whenever an AI system has a significant impact on people's lives, it should be possible to demand a suitable explanation of the AI system's decision-making process".

Transparency is also determined by communication. The patient should know and understand how artificial intelligence determines his or her diagnosis. Various methods such as model cards increase transparency by providing essential information on the functioning of the models involved. In sensitive issues such as a person's health, there should not be a barrier that deprives them of the logic behind the decision taken by an AI system, especially in the early stages of its implementation in the medical field.

However, to what extent should patients be able to intervene in the decision making of their own diagnosis? The novelty of its introduction implies a lack of confidence on the part of patients to be subject to its use. It is possible that, even if the results are reliable, patients may reject them and ask for human intervention. This is not just a matter of lack of habit, but extends to the nature of human contact, to the safety that human interaction conveys in sensitive issues such as cancer detection. Transparency implies openness, but there will be cases in which that is not enough to justify the diagnosis of a machine. The introduction of new technologies will bring great benefits to health care, but such an evolution requires a process of adaptation in which human agency will still be necessary.

Artificial intelligence can reach a high level of autonomy and self-sufficiency, which raises the question of how doctors should intervene in its decision making. Some cases, such as the one studied in this paper, indicate that human intervention is essential in diagnosis, as the results are not always accurate. Nevertheless, what about those cases in which the results of an AI are close to 100% accurate? A second human opinion to verify the automated diagnosis would make no difference to the final decision. There are tasks where machines outperform humans, and if diagnosis becomes one of them, human agency will be superfluous.

9 Conclusions

In this paper, we proposed image recognition by means of a convolutional neural network-based classification. This method was applied on abdominal CT scans from Kaggle to distinguish cystic and cancerous tissues in the kidneys. Use was also made of data augmentation by applying various effects on the images, thus facilitating more efficient training over a wider range of possible cases. The paper covers the process of developing the optimal model for the defined approach, focusing on the comparison of a small versus a medium versus a large topology, as well as hyperparameter tuning and regularization techniques to obtain the best possible accuracy. The comparison resulted in the large model being chosen, having obtained 99.91% accuracy in the training phase with a split percentage of 66% - 33% from a total of 6857 axial plane images. However, testing with the unseen data did not exceed 31.58% accuracy, having occurred due to model or data error. The subsequent tests of model fitting, image reduction and division into subsets indicated that the problem lies in the images, as either the specific characteristics they follow or the preprocessing they may have undergone prior to this study make them unrepresentative of conventional CT standards, conditioning the model and limiting the images it is capable of classifying.

All this implies that the proposed approach is not prepared to face the responsibility of diagnosing cysts and cancers in actual patients. Not only would it require human agency in its application to medicine as a reinforcement to the autonomous conclusions reached, but it would be a potential danger given its low accuracy, which could lead to confounding the medical conclusion of experts and dooming the patient's health. The development of new technologies signifies an evolution of current and cutting-edge medicine. Such an advance brings with it the responsibility to be appropriately exercised in order to implement the use of artificial intelligence as an aid to patients' health.

The projection of this thesis is framed in the field of technological applications in health and medicine. This research has addressed a new methodology as a support function for healthcare in interpretation and diagnosis. Although the results obtained have not been as expected, this allows us to continue shaping the model and what may become the future of modern medicine.

10 Future Work

To advance the construction of the model, it will be trained and tested with a different dataset than the one used in this study. Exposing the model to different data could provide additional information with which to improve the approach developed in this paper.

Eventually, the proposed model will be integrated as a fully functional diagnostic support tool.

References

- Commission, European, Content Directorate-General for Communications Networks and Technology (2020). *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment*. Publications Office.
- Deo, Rahul C (2015). ‘Machine learning in medicine’. In: *Circulation* 132.20, pp. 1920–1930.
- Flanigan, Robert C et al. (2003). ‘Metastatic renal cell carcinoma’. In: *Current treatment options in oncology* 4.5, pp. 385–390.
- Herring, W. (2016). *Radiologia básica: Aspectos fundamentales*. Elsevier Health Sciences Spain. ISBN: 9788491130109. URL: <https://books.google.ie/books?id=W0gtDQAAQBAJ>.
- Hsieh, James J et al. (2017). ‘Renal cell carcinoma’. In: *Nature reviews Disease primers* 3.1, pp. 1–19.
- Hussain, Mohammad Arafat et al. (2017). ‘Collage CNN for renal cell carcinoma detection from CT’. In: *International Workshop on Machine Learning in Medical Imaging*. Springer, pp. 229–237.
- Islam, Md Nazmul (Oct. 2021). *CT KIDNEY DATASET: Normal-Cyst-Tumor and Stone*. URL: <https://www.kaggle.com/datasets/nazmul0087/ct-kidney-dataset-normal-cyst-tumor-and-stone>.
- Moch, Holger (2010). ‘Cystic renal tumors: new entities and novel concepts’. In: *Advances in anatomic pathology* 17.3, pp. 209–214.
- Nithya, A et al. (2020). ‘Kidney disease detection and segmentation using artificial neural network and multi-kernel k-means clustering for ultrasound images’. In: *Measurement* 149, p. 106952.
- Sung, Hyuna et al. (2021). ‘Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries’. In: *CA: a cancer journal for clinicians* 71.3, pp. 209–249.
- Sureshababu, Waheeda and Osama Mawlawi (2005). ‘PET/CT Imaging Artifacts’. In: *Journal of Nuclear Medicine Technology* 33.3, pp. 156–161. ISSN: 0091-4916. eprint: <https://tech.snmjournals.org/content/33/3/156.full.pdf>. URL: <https://tech.snmjournals.org/content/33/3/156>.

- Tuncer, Seda Arslan and Ahmet Alkan (2018). ‘A decision support system for detection of the renal cell cancer in the kidney’. In: *Measurement* 123, pp. 298–303.
- Türk, Fuat, Murat Lüy and Necaattin Barışçı (2020). ‘Kidney and renal tumor segmentation using a hybrid V-Net-Based model’. In: *Mathematics* 8.10, p. 1772.
- Vasanthselvakumar, R, M Balasubramanian and S Sathiya (2020). ‘Automatic Detection and Classification of Chronic Kidney Diseases Using CNN Architecture’. In: *Data Engineering and Communication Technology*. Springer, pp. 735–744.