UNIVERSITAT POLITÈCNICA DE VALÈNCIA

School of Informatics

Using Deep Learning for Semantic Organ Segmentation in CT Images of the Abdomen

End of Degree Project

Bachelor's Degree in Data Science

AUTHOR: Abdoun , Hajar

Tutor: Gómez Adrian, Jon Ander

ACADEMIC YEAR: 2022/2023

Acknowledgement

Thank you to my supervisor, Jon Ander Gómez Adrian, for providing guidance and feedback throughout this project. Thanks also to the Grand Challenge website for providing us with the data that made this study possible. And also, a huge thanks to the Polytechnic University of Valencia for giving us access to potent machines that helped with the training of the Deep Leaning models which was crucial for this project.

Abstract

The precise segmentation of abdominal organs has critical importance for several clinical procedures. However, anatomic variability in the abdomen poses a major challenge to automated segmentation, and manual correction is almost always required. For that reason, the main goal of this project is to try to mitigate that problem by working on a large dataset of more than a thousand annotated cases that come from twelve different medical centers. At the implementation level, the main purpose is to train several supervised deep learning models in order to detect each one of these 4 organs: Liver, Kidneys, Spleen and Pancreas in the CT-scan images of the abdomen.

The dataset has been extracted for the website Grand-Challenge https://grand-challenge.org/, where multiple use cases related to Health are available. The deep learning models which are used are derived from the U-Net architecture. This architecture has been widely used in semantic segmentation tasks yielding good results in terms of the Intersection over Union (IoU) metric.

Key Words: Deep learning; semantic segmentation; medical images; image processing; neural networks

Abstract (Castellano)

La segmentación precisa de los órganos abdominales tiene una importancia crítica para varios procedimientos clínicos. Sin embargo, la variabilidad anatómica en el abdomen plantea un gran desafío para la segmentación automatizada y casi siempre se requiere la corrección manual. Por esa razón, el objetivo principal de este proyecto es tratar de mitigar ese problema trabajando con un gran conjunto de datos de más de 1000 casos anotados que provienen de 12 centros médicos diferentes. A nivel de implementación, el objetivo principal es entrenar varios modelos de aprendizaje profundo supervisados para detectar cada uno de los 4 órganos siguientes: hígado, riñones, bazo y páncreas en las imágenes de tomografía axial computarizada del abdomen.

El conjunto de datos se ha extraído desde el sitio web Grand-Challenge https://grand-challenge.org/, donde se encuentran disponibles múltiples casos de uso relacionados con la salud. Los modelos de aprendizaje profundo que se utilizarían se derivan de la arquitectura U-Net. Esta arquitectura se ha utilizado ampliamente en tareas de segmentación semántica y ha arrojado resultados suficientemente buenos en términos de la métrica de intersección sobre unión (IoU).

Abstract (Valenciá)

La segmentació precisa dels òrgans abdominals té una importància crítica per a diversos procediments clínics. No obstant això, la variabilitat anatòmica a l'abdomen planteja un gran desafiament per a la segmentació automatitzada i gairebé sempre es requereix la correcció manual. Per això, l'objectiu principal d'aquest projecte és intentar mitigar aquest problema treballant amb un gran conjunt de dades de més de 1000 casos anotats que provenen de 12 centres mèdics diferents. A nivell d'implementació, l'objectiu principal és entrenar diversos models d'aprenentatge profund supervisats per detectar cadascun dels 4 òrgans següents: fetge, ronyons, melsa i pàncrees a les imatges de tomografia axial computada de l'abdomen.

El conjunt de dades s'ha extret des del lloc web Grand-Challenge https://grand-challenge.org/, on hi ha disponibles múltiples casos d'ús relacionats amb la salut. Els models d'aprenentatge profund que es farien servir es deriven de l'arquitectura U-Net. Aquesta arquitectura s'ha utilitzat àmpliament en tasques de segmentació semàntica i ha donat resultats prou bons en termes de la mètrica d'intersecció sobre unió (IoU).

Content Table

Contenido

Figures and Tables Index

1. Introduction

Motivation

The Big Advancements in Technology

In our current day and age, technology has showed a peak in its general improvement in all history, in several parts of the globe. This is clearly displayed with the submergence of the newest AI tools, having ChatGPT and its extensions at the top-level of this improvement. It showcases the extent at which we, as a society, are bound to grow if the AI technologies continue to improve. This new wave of tangible advancements is to be used in facilitating the mondain tasks humans once had to do. It also creates a guide to follow in order to improve the quality of the work we are executing, acting as a helping system that gives some guidance on how to fulfill tasks in a proper and more optimal way. Those tasks can range from making queries, writing code, to creating tailored recipes and workout plans.

These advancements can be used to create greater things that have a bigger impact on the society. One of them being of help to the health sector. The field that is responsible of taking care of the most precious thing: The Human Life. Being more accurate in the detection of anomalies, the treatment planning, and the surgeries, can help reduce the suffering of the patients, while at the same time saving more lives, and maybe even reducing the cost of the medical treatments making it possible for low wage workers to access this basic right. That is why it is essential to start conducting some research on how to incorporate these new technologies and Artificial Intelligence into the medical field.

The more this topic is researched, the better the algorithms get when working with any specific task related to the sector. The professionals also tend to understand and value it more, having more security and certainty on the model’s prediction without having any unnecessary doubts about its performance. That means that these models should also have a degree of interpretability when it comes to why the model makes the decision it makes. This can be important in some cases more than others.

AI in the Medical Field

Currently, after noticing the big advancements in the technologies surrounding the medical field, it has become relevant for doctors and experts to try and work hand in hand with Artificial Intelligence in order to be assisted in clinical diagnosis and treatment planning.

This became of great relevance for two main reasons. The first one is based on the fact that experts need some assistance because of the human limitations. Meaning that as human, they are more prone to experience fatigue if the task is repetitive or they have to work long shifts in the hospital, which would render their diagnosis and can introduce some error in their judgement. So, by accepting that help, it gets easier to maintain the same levels of accuracy during the whole process, limiting the decreasing focus and sharpness of the attention over time.

Another big reason for incorporating AI in this field has to do with speeding up the process and getting overall better results and accuracy. Once the model is trained, its prediction usually wins over the doctor's prediction when it comes to the time of the execution. This seems to be the case because of how cheap and fast the computers are becoming these last decades with the incorporation of the recent advancements in the technologies.

All of these reasons combined create a big demand on the market for the incorporation of AI in the medical field. This big need calls for different steps of the cycle to be invested in. It goes from the collection of data, the preprocessing and the feature engineering of the variables, the training of the models, the evaluation of the algorithms to the presenting of the results in scientific papers.

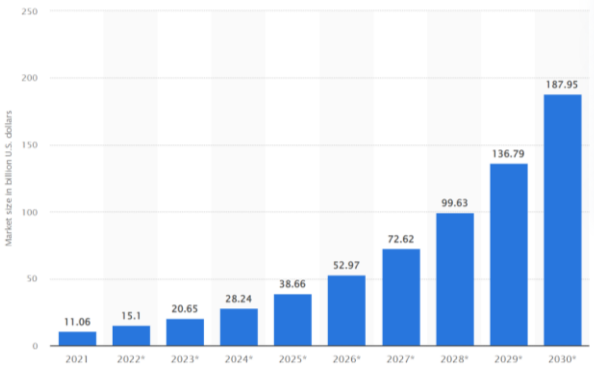

Figure 1. Artificial intelligence (AI) in the healthcare market size worldwide from 2021 to 2030. [a]

In order to better understand this demand, here we display a figure that highlights the previous talk points. It shows that in the year 2021, the worth of Artificial Intelligence in the healthcare sector was worth more than 11 billion U.S dollars, which is huge compared to the previous years.

That growth is most likely going to multiply and reach enormous levels in the years to come. It is predicted that the global healthcare AI market is going to grow to reach an amount of 188 billion U.S dollars by the year 2030, in only one decade.

After realizing the importance of the use of AI in the medical field, we have decided to conduct a study surrounding this talk point. This project is focused on the topic of automatization of the sematic segmentation of the abdomen. Which means that a model is trained in order to detect specific organs or tumors from CT-scans of the abdomen without any external help. This topic is currently growing in popularity because of how helpful it is to the health sector. It can be used in AI guided surgeries, the detection of tumors and abnormalities, or even in the longer run, to look through the improvement of the patient during a specific interval of time and evaluate his progress.

The Difficulty of Incorporating AI in the Medical Field

After understanding how valuable AI in the medical sector is, we also need to understand the fact that the automatization of the segmentation of medical images is a delicate subject, meaning that a higher level of accuracy is needed because of how important and risky this sector can be. Making the slightest error can have a big impact on the patient depending on the task carried out.

The cost of error in AI guided surgeries is very high. The patient is in a very vulnerable position where every error, as slight as it can be, can have deadly repercussions. The error in the detection of the placement of a certain organ or tumor in the body can have very hard consequences if no external help is around to double check the result before executing a risky move. For that reason, in early stages of research, there has to be an external input of a professional, that can be a medic or a doctor, before the execution of any task, no matter how accurate the model is. Moreover, the automatization should only be used in low level risk situations with a very high-performance model, because of the risky nature of the sector of Health.

The Scarcity of Data in the Medical Field

One limitation this field faces is the very low amount of data accessible to do these types of research. This is due to assorted reasons, one of them being that the data is dispersed in various places.

Every hospital and healthcare center collect its own data from their patients, but they do not communicate with each other, sharing their data sources in order to make more complete and accurate studies. Adding to that, in the unit level, the departments usually do not communicate data with each other neither. The department of radiology for example has its own database as well as the department of pharmacy, but these two do not share the data of their patients with each other. They might think it is not relevant and would not be of any help because of how different these practices are to each other, but that cannot be further from the truth.

Another reason medical data does not get used in studies is because of how unstructured it can be. The professionals generally document clinical facts about their patients during the consultation without bearing in mind how useful that data is in future studies. Meaning that in most cases the notes are taken with no regard for their structure, their sole purpose is to keep track of the past and present pathologies of each patient, which can perfectly be written in a text and does not require to follow any structure.

The medical field can also be tricky when collecting the data because of how inconsistent and changing the definitions can be. Two professionals may diagnose two different pathologies in the same patient depending on the approach each one of them takes while consulting the patient. The definitions can also be variable and change with time because of how fast the medical field progresses over time compared to other sectors.

With the advancements AI is currently making, regulations in the healthcare systems are changing constantly. Balancing between the privacy of the patients and the scientific research for the greater good is a very complex task. That is why regulation and requirements are only going to continue to change and increase because of how unknown the new advancements in technologies are. That makes to extraction of the data even more difficult for research purposes.

Limitations of Working with Images

When working with medical images, we can be faced with even more limitations, this time, based on the complexity of the images as data. Images tend to be difficult to analyze because of the way they store the data. These tend to be represented by big vectors that not only focuses on the content of its elements, being the level of intensity, but also the position of each element of the vector, being the pixels. Moreover, we can unveil more limitations that comes from using the images as data, which are discussed in the later paragraphs.

One on them being the intensity inhomogeneity. This problem is common when working with images. It generally does not affect drastically the predictions when the human perception is used because of how complex the visual image understanding is in the brain. Humans have the ability to extract the content of the image they are faced with regardless of how distorted it is, but it is not the case in automated systems.

Intensity inhomogeneity can render the performance of the methods that assume that the intensity of a tissue is constant over the range of the images. So, one solution can be to try and correct the predictions using histogram matching methods. This is a preprocessing tool which acts like a calibration technique and is used to normalize two images that generally have the same local illumination over the same location but have undergone discrepancies because of the difference in the sensors or the global illumination. These techniques are used especially if the images are taken from different sources or undergone varied conditions.

Another important limitation we can be faced with while working with medical images is the closeness in gray levels. This problem takes place when the images have a low contrast in their intensities, making it harder to identify and detect the objects in the image that act as the input of the prediction models. Having no differences in the levels of grey between the tissues of the organs, makes it near impossible for a model to pick up the patterns from the images in order to be trained and to be used in the classification of new images that are not annotated.

This makes the automatization of semantic segmentation of images even more difficult. The reason being that not only we are training the model to do a normal classification task, which is in itself difficult, but we are also aiming to classify each pixel in one of the classes while at the same time demanding for it to recreate the original image from scratch, with the pixels having regained their initial position while still holding on the result of the classification. This task shows up to be of great complexity unless some preprocessing techniques of images are used to help heighten the contrast. This can be done using some variant of the method Histogram Equalization. This method is considered a special case in Histogram Matching. This time the image’s histogram is matched with a preset histogram which is uniformly distributed.

Using any of the methods previously discussed resolves the problem by spreading the intensities and increasing the image’s contrast. This way, it gets easier and clearer for the model to identify each one of the objects present the image. This is due to how clear the border gets between the objects of the image, and how easy it gets to pick up the difference in the tissues of the organs, when working with medical images.

U-Net in Medical Case Studies

Lately, when it comes to Artificial Intelligence in supervised cases, Artificial Neural Network based algorithms can be considered one of the most complex yet best result generators when it comes to working with images. These deep learning algorithms know how to handle the complexity of image’s structure without the need of any external transformation or feature extraction techniques. They are known for their ability to adapt to the training data, to self-organize and to perform in real time thanks to their parallel configuration.

U-Net is a special Artificial Neural Network Architecture that has gained a lot of popularity in the field of semantic segmentation of the images. It is a U-shaped encoder-decoder architecture that consist of several encoder and decoder blocks connected to each other via a bridge and skip connections. These blocks generally consist of convolutional layers, ReLU activation Functions, MaxPooling and UpSampling. All of these components work together to first, classify each pixel to one of the possible outcome categories, and then to recreate the previous image with the now classified pixels. This variation of Artificial Neural Networks had showed, in previous studies, a high level of accuracy in the results independently on the size of the training data. Adding to that, it generally has a good performance when dealing with images. Its capability of computing a pixel-wise output makes it well fitted for problems that roam around tasks of semantic segmentation. For all the reason discussed above, we've decided to use this architecture on our project, since it deals with the task of semantic segmentation of the organs in the abdomen.

Aims

The precise segmentation of abdominal organs has critical importance for several clinical procedures. However, anatomic variability in the abdomen poses a major challenge to automated segmentation, and manual correction is almost always required.

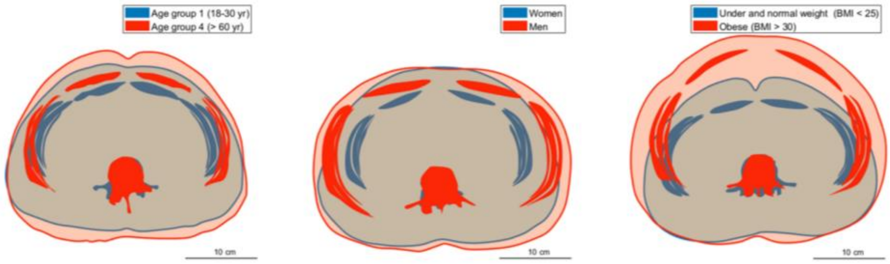

Figure 2. Abdominal wall morphometric variability based on computed tomography: influence of age, gender, and BMI. [1]

For that reason, the main goal of this project is to try to mitigate that problem by working on a large dataset of more than 1000 annotated cases that come from 12 different medical centers. Because of the privacy closure, we cannot access more personal information on the patients, like the age, the sex, or the BMI index, which can be of big help if used in this study. So, the variability introduced comes from how large the dataset is, the presence of healthy and cancerous patients, and different anatomical properties like the variability of the length of the abdomen. At the implementation level, the main purpose is to train several supervised deep learning models in order to detect each one of these 4 organs: Liver, Kidneys, Spleen and Pancreas in the CT-scan images of the abdomen.

The dataset has been extracted from the website Grand-Challenge https://grand-challenge.org/, where multiple use cases related to Health are available. That website provides interesting challenges with reliable datasets related to biomedical imaging, which can be used to train machine learning models that are, afterwards, deployed in the classification and prediction case problems. The deep learning models which are going to be used are derived from the U-Net architecture. This architecture has been widely used in semantic segmentation tasks yielding good enough results in terms of the intersection over union (IoU) metric.

Intersection over Union (IoU) is used as a metric for object detection, it evaluates the overlap of the Ground Truth and Prediction region, taking into consideration both the location and the area of the bounding box. A high IoU score indicates that the predicted bounding box is well-aligned with the ground truth mask. IoU shows up to be a good metric when it comes to measuring the accuracy of the model used, especially when working with the complex nature of images.

Knowing the accuracy of the model shows, firstly, if the model is a good fit to the resolution of the task, making it possible to evaluate the model in its own. Then it can also be used to compare the performance of multiple models and choose the one that yields the best results and fits best the problem of our case study. The accuracy of the model is a summarized version of the results, which makes it possible to reduce the complexity of the output while reducing at the same time the memory and time resources cost.

When it comes to the preprocessing of the medical images, several methods are used. One of them being a variant of Histogram Equalization, and the other being Contrast Limited Histogram Equalization (CLAHE), which can also be considered a special variant of the previous algorithm. This latter technique is an enhancement of the Histogram Equalization method. The Histogram Equalization method is used in images as an aim spread out the most frequent intensity values which translates into increasing the global contrast of the images when the intensity values of the pixels have a similar range of values. The application of these methods makes the areas of the image with low contrast gain a higher level of contrast.

However, the method discussed previously has an important limitation. The Histogram Equalization technique cannot be implemented in images with large intensity variation. So, in order to overcome this limitation, Contrast Limited Adaptive Histogram Equalization (CLAHE) algorithm is used. This new method uses Histogram Equalization in small patches rather than the whole image, then proceeds to combine the neighboring tiles using bilinear interpolation to remove the boundaries created.

These preprocessed images then get used by the U-Net algorithm in the training process. It can be of great help in the identification and segmentation of the organs in the abdomen CT-scan images. The trained model would later be used in the semantic segmentation task of new non-annotated cases.

Academic Aims

- Learn how to work with images as data. Understand how the collection, preprocessing and the modeling of this particular type of data works.
- Healthcare is a sector that is growing to use AI more and more. So, getting used to working with medical images is considered valuable for a data scientist.
- Automatic Abdomen Semantic Segmentation created a new path of studies in the medical field. This topic lacks research because of its novelty but is very relevant and important for the advances that are to occur lately in this field.
- U-Net architectures have proven their performance when working with the semantic segmentation of images. Their architectures prove to be very efficient in these types of tasks. So, learning about how they work can be very helpful and interesting for a data scientist who wants to learn more about the process that Deep Learning algorithms go through.
- Be able to clearly explain the study conducted writing a scientific paper and presenting the results to an audience of professionals.

Structure

This paper is organized as follows: Section 2 reviews the state-of-the-art and references other studies on the automatic semantic segmentation of medical images and U-NET models. Section 3 details the theorical foundation that helps understand the theory behind some specific terms. Section 4 describes the methodology used going from the preprocessing of the dataset, the topologies, the depths, the variants and the hyperparameters used in the U-Net architectures, to then finally explaining the software and hardware used in the project. Section 5 describes the experiments carried out. Sections 6 and 7 present and discuss the results, respectively, and Section 8 concludes by considering the objectives and possible future research.

2. Related Work

Image Preprocessing

Contrast Limited Adaptive Histogram Equalization (CLAHE) Approach for Enhancement of the Microstructures of Friction Stir Welded Joints [1]

This paper presents a study conducted on the possible improvement of the quality of images displaying the microstructure of the Friction Stir joints. These images play an important role in the identification of surface defects in the components produced while also used in the studying of the variations of the materials manufactured.

Images are represented by elements called pixels, those elements have two spatial coordinates, and an amplitude function called the intensity or gray level. When the value of the intensity level is a discrete number, that's when we talk about digital images. Digital images are represented, in most cases, in the form of a matrix whose elements are the pixel coordinates transformed into natural numbers.

Looking into the case presented in the article, we can observe that the histogram used to represent the digital image lies in the brighter region, which means that there is not a big contrast between the levels of gray of the image. In order to fix this issue, Histogram Equalization is used. The main goal of this transformation is to make the bins of the histogram more dispersed which would make the image hold a higher contrast between its elements.

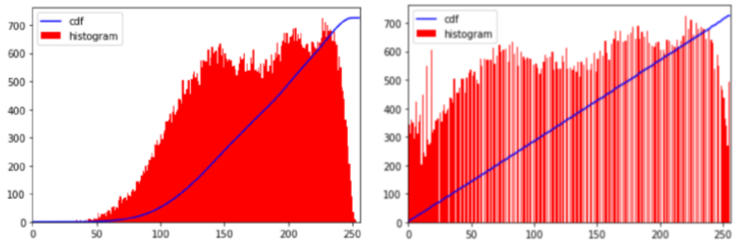

Figure 3. Application of CLAHE on the histogram of an image. [1]

However, this method has limitations when it comes to working with images that have high intensity variations. For that reason, Contrast Limited Adaptive Histogram Equalization is used, yielding better results when it comes to the entropy and the RMS Contrast metric.

In conclusion, applying CLAHE in this case enhanced the quality of the microstructure of images. It made the values of both metrics Entropy and RMS higher when compared to other microstructures used in various color spaces.

Contrast Limited Adaptive Histogram Equalization Image Processing to Improve the Detection of Simulated Spiculations in Dense Mammograms [2]

In this study CLAHE is used in manual detection of cancer in dense mammograms. Observers are faced with 40 mammograms with different backgrounds and positions in order to try and detect the spiculation. CLAHE was used in nine configurations to look for any probable improvement in the accuracy of the task.

There were a lot of combinations possible of the parameters of the CLAHE algorithm, which is why a pilot study was conducted before in order to minimize the pool of possible combinations. Two radiologists reviewed and scored dense mammograms with real lesions based on the improvement or absence of improvement in the quality and clarity images. The number of combinations went from 70 to 10, keeping the ones who actually made a positive difference. The parameters that were used are the region size and the clip levels.

Three out of the nine configurations yielded better results than not using the CLAHE technique, while there was no big difference in the other configurations. That being said, many radiologists reported having more difficulties in trying to detect the speculation after using CLAHE because of the amplification of the noise that results by applying the wrong parameters. For that reason, it is critically important to conduct more laboratory analysis in order to look through the parameters and enhance their performance rather than decrease it.

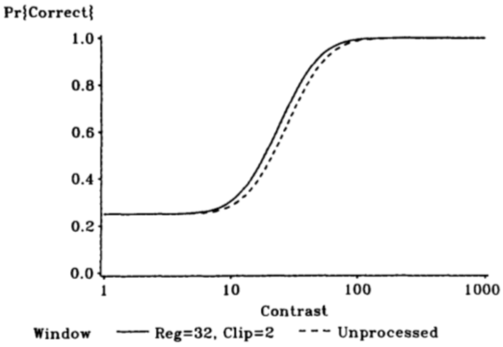

Figure 4. Estimated detection probability. The shift in the curve to the left for the processed image reflects improved detection. [2]

Medical Images

Automated medical image segmentation techniques [3]

During the last decades, automated image segmentation has grown in popularity, especially in the medical field. Once the experts understand its use and the benefits it provides, they become more lenient in using it in different aspects, like in the diagnosis of the patient, the localization of the pathology, treatment planning or even computer-integrated surgery.

The images are generally MRI or CT scans. In our study we used CT scans for various reasons mentioned in this paper. First it is because it is less expensive and widely available while it also having a very high spatial resolution which is valuable in our case study. However, we need to take into consideration that it usually has a lower contrast when it comes to comparing the different tissues when faced to the MRI scans.

When it comes to the representation of the medical images, we are faced with the complexity of the three-dimensional nature of the scans, so in our study we opted for the option that relies on the representation as a sequential series of 2-D slices because it requires lower computer complexity and less memory.

One of the techniques that is implemented in the segmentation of medical images is the Model Based Segmentation, and that is the method used in our study, specifically Artificial Neural Networks. The aim of this technique is to create a model that learns from the repetitive geometry of the organs in order to do the semantic segmentation of new images. Supervised Models have the ability to adapt and to self-organize without external help, while at the same time having the capacity of real time performance.

Medical Image Segmentation: A Review [4]

In this study, we look into the different techniques used in medical image segmentation in order to extract information valuable for the decision-making process. First there appear to be two ways that are used in segmentation: Detection of Discontinuities, which main purpose is to partition the image considering the abrupt change in the intensity using algorithms like Edge Detection, and Detection of similarities, which as the name reveals, divides the image into homogenous partitions using algorithms like Thresholding and Region Based Segmentation.

Edge Detection is an algorithm that detects the pixels where there seems to be a rapid transition in the intensity, then they get linked to one another in order to draw the edge resulting in a binary image. This method can also be used as the base of other techniques, specially to help and draw the boundaries in the Region Based Segmentation. The latter method is considered relatively simple and more immune to noise and can be used in one of two ways: Region Growing method or Region Splitting and Merging method. The first one starts with singular pixels and then proceeds to group them based on a certain criterion; however, the next technique does the opposite. It starts by the dividing the image into a set of arbitrary unconnected regions and then starts merging them following a set of conditions.

The Thresholding Method on the other hand needs some previous knowledge to set the levels of the threshold in order to be able to successfully do the task of segmentation of the image. It can be a powerful tool when the goal is to separate the object of study from its background because of its accuracy when it comes to converting a multilevel image into a binary image. It does that by looking at the intensity of each pixel and deciding if it fits the threshold or not. However, it can bump into problems when the illumination is uneven. The problem gets solved by going from global to local Thresholding, using different threshold levels for different regions. All that said, this method still has some limitations which makes it not suitable for multichannel images because of its binary nature, adding to it the fact that it does not consider the spatial characteristics of the image due to its sensitivity to the noise.

UNET

UNET: Automatic semantic segmentation of the lumbar spine: Clinical applicability in a multi-parametric and multi-center study on magnetic resonance images [5]

The main objective of this study is to do an accurate segmentation of the lumbar region using some algorithms of deep leaning. For this doing, different variants of the U-Net architectures were evaluated and compared to each other in order to come up with best fitted model for this case study.

U-Net is a specific network that belongs to the family of Artificial Neural Networks. Its architecture is what differentiates it from other networks. This model has two branches, one for the encoder and the other for the decoder. Those two are linked by a bottleneck block. These neural networks have surpassed a lot of models when it comes to performance and accuracy when working with semantic image segmentation.

In this study, first they look into the different topologies and compare their performance. Different ensembles of neural Networks have been trained, and when compared to the FCN and the original U-Net, they observed that the variations have outperformed the original architecture. Moreover, they also added convolutional blocks between the encoder and the decoder and looked into the results. There was a slight improvement in the performance. However, combining all complementary block types did not yield optimal results.

UNET++: A Nested U-Net Architecture for Medical Image Segmentation [6]

The purpose of this study is to compare the performance of the normal U-Net architecture and the new variant called U-Net ++. For that reason, four datasets were used displaying organs or lesions and the final results were compared. The comparison was based on two evaluation metrics: Dice coefficient and Intersection over Union (IoU).

UNET++ is another variation of U-NET. The main difference between the two is number of convolution blocks between the encoder and the decoder. This new architecture uses Dense Convolutional Blocks on skip pathways bridging the semantic gap between the encoder and the decoder. Adding to that, it has dense skip connections on skip pathways in order to ease the optimization of the features. Then it uses deep supervision in order to improve the performance of the model using model pruning.

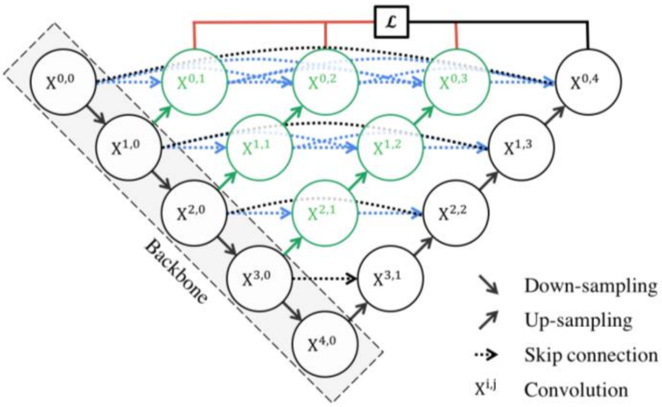

Figure 5. Architecture of the model U-Net ++. [6]

The results showed an improvement of 0.5 in the IoU metric from using U-Net to using U-Net++ without deep supervision, and an improvement of 0.6 going from not using supervision to using it. Meaning that changing the architecture did enhance the performance of the artificial neural network, especially when working with the segmentation of the liver and the lung nodule.

RA-U-Net: A Hybrid Deep Attention-Aware Network to Extract Liver and Tumor in CT scans.[7]

In this study, another variation of the U-NET architecture is used, this time to tackle the complexity of working with 3D imaging. The model is trained to work with 3D images of the liver, for both detection and segmentation of the tumors. In the image preprocessing part, a global windowing step of 100 to 200 was used in order to isolate the liver from any external noises. Meaning that irrelevant organs and tissues were removed in order to make the images clearer to help with the later training of the model. The architecture used is named RA-UNET and works with 3D structures in a pixel-to-pixel fashion. It works with residual learning and attention residual mechanism in order to do it.

Residual Learning is the method used in order to help with the problem of gradient vanishing when working with very deep neural networks. It utilizes residual blocks in all the levels except the first and the last layer and uses identity maps as skip connections. This technique has shown improvements in the performance of the models previously. However, in this case, a particular type is used: Attention Residual Learning. This technique, adding to all thing mentioned previously a different kind of staking is used. The attention module gets divided into two blocks, the first block is called the trunk branch, and is used to process the features, and the other one is called soft mask branch, which is used to construct identity mapping.

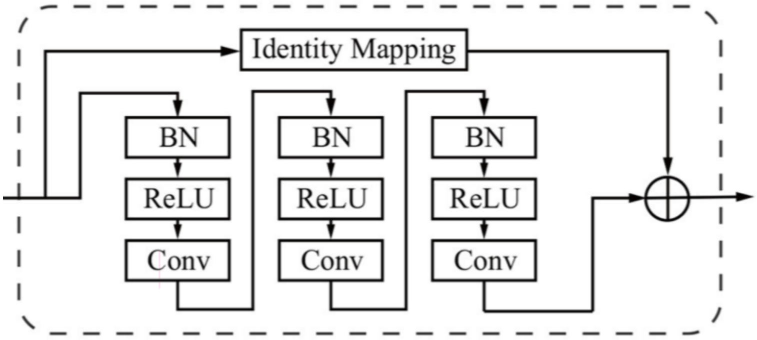

Figure 6. Sample of a residual block in the dashed window. An identity mapping and convolutional blocks are added before the final feature output. [7]

In order to optimize the parameters of the neural network, the loss function is used. The Dice Coefficient is used to calculate the latter and try to minimize its value.

These alterations seem to yield better results and can be generalized to fit other tumor segmentation datasets. The limitation that needs to be addressed while using this algorithm is the high training time required to work with the lager number of parameters that come with 3D convolutions. However, the model’s performance increased using this specific architecture and yielded competitive results at the end.

3. Theoretical Foundation

Preprocessing of the images:

Histogram of an Image

The histogram of an image is a plot in which in horizontal axis pixel intensities are represented and the vertical axis represents the frequency of each intensity. In our case, we are working with CT scans which are represented in Grayscale, which means that intensity varies in the range between 0 and 255.

Generally, the histograms are used, when working with image segmentation, to isolate the background from the object. To do that a threshold of intensity is specified, which leaves us with two classes, the pixels that are lower than the threshold, and the pixels higher than the threshold specified previously. That way one class would represent the object and the other the background.

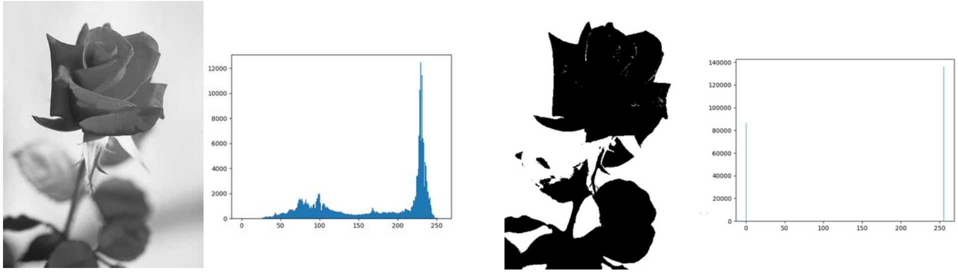

Figure 7. Histogram of an image. [b]

Histogram Equalization

Histogram Equalization is a method used to increase the contrast of the images. It does that by spreading out the most frequent intensity values all along the range of the possible intensity values, in our case it would be between 0 and 255. This technique is generally used when the image has close contrast values.

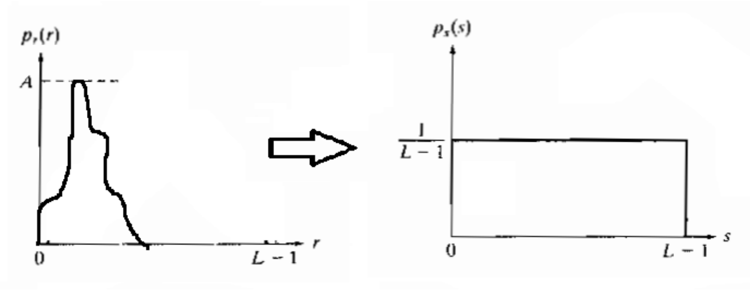

Figure 8. Histogram Equalization. [c]

Looking at the first image bellow, we see that the intensities range from 100 to 200. The values between 0 and 100 are not present, meaning that the objects in the image that should be black show as a darker shade of gray. This can be a huge limitation to the training of a model whose job is to learn patterns from the images. The solution is to try to make the values spread between 0 and 255, this way the contrast of the image increases and the detection of the borders of the objects becomes clearer.

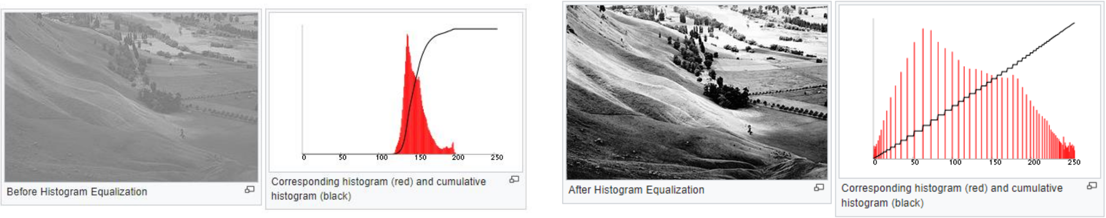

Figure 9. Global Histogram Equalization. [d]

However, this method works on the image as a whole, and this can have some limitations. If we are dealing with an image that already has a high contrast, meaning it has frequent intensities in both sides of the spectrum, then applying this technique would get rid of some small details. Those details might be of big importance in the process of training of the algorithm.

Adaptive Histogram Equalization

This technique is based on the previous one, but instead of working on the whole image, it works on small tiles of the original image. It, also, enhances the contrast of the image making the detection of objects in the images an easier process.

The algorithm takes the blocks forming the image and applies Histogram Equalization on each one of them making it easier to redistribute the lightness values of the image. This results in the improvement of the local contrast and therefore the clarity of the image.

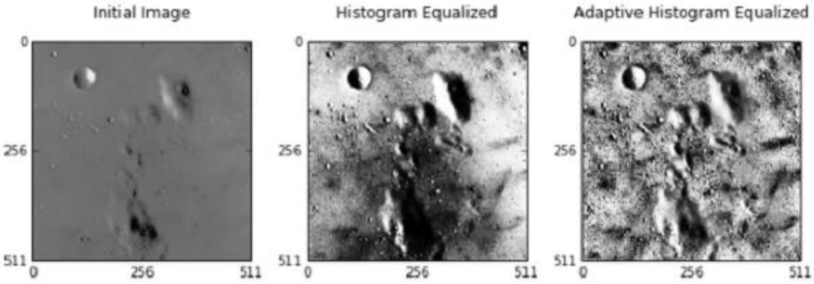

Figure 10. Difference between Histogram Equalization and AHE. [d]

This method enhances the local contrast while at the same time preserving the edges of each one of the sections. However, the downfall of this algorithm is the fact that it overamplifies the noises in the image. Meaning that Adaptive Histogram Equalization overamplifies the contrast in the borders between the sections of the image. That might render the performance of the segmentation model, so another variation of the technique is needed.

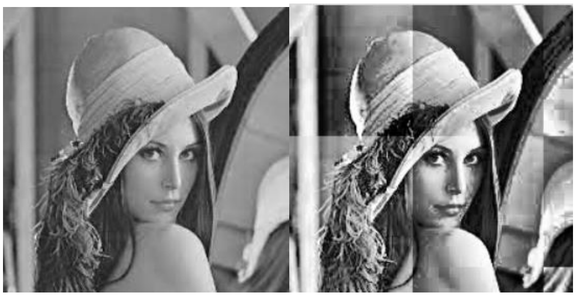

Figure 11. Adaptive Histogram Equalization. [e]

Contrast Limited Adaptive Histogram Equalization (CLAHE)

When the image treated has high intensity variations, the previous models don't usually work well. So, to overcome this limitation Contrast Limited Adaptive Histogram Equalization is used. This algorithm can be divided into three processes: Tile Generation, Histogram Equalization and Bilinear Interpolation. The first step consists of dividing the image in partitions called "tiles", then Histogram Equalization is performed, ending up with as many histograms than number of tiles. Each histogram has the bins of intensity clipped at a particular limit, set previously, and then redistributed into other bins. Lastly, the resulting tiles are joint together using bilinear interpolation to get the final image with the levels of contrast modified. The use of this algorithm helps to prevent the overamplification of the noise, resulting in better performances in tasks such semantic segmentation.

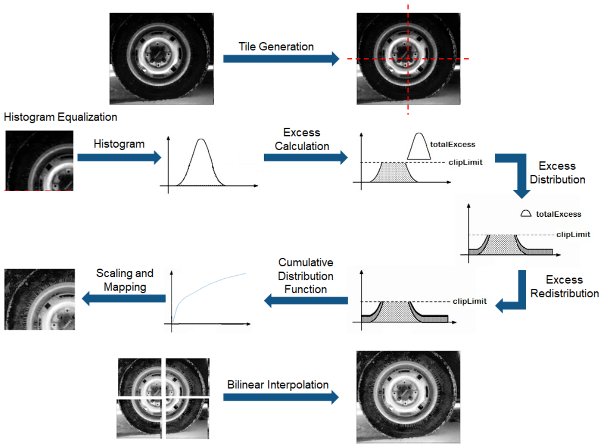

Figure 12. CLAHE with 4 sections. [f]

Bilinear Interpolation:

Bilinear Interpolation is a method used when trying to predict a value in a 2-dimensional space, using four of its adjacent points that form a rectangle. To do so, it calculates the average of the data of the corners of the rectangle. The weights are fixed depending on the distance of each one of the corners to the value of interest. In the case of working with images, those data values constitute the level of intensity of the pixel in a particular position. This method is used when it comes to reconstructing an image that was previously divided into several partitions and maintain the harmony of the image.

U-NET architecture model:

Currently, after the advancements of Artificial Intelligence specially with deep learning, there appears to be solutions to several problems that we were not able to resolve before, one of them being Semantic Image Segmentation. Its main purpose is to identify specific segments from the image. For this reason, Convolutional Neural Networks grew in popularity giving good enough results in simple image segmentation. However, once the images grew in complexity, this algorithm faces a lot of limitations lowering the accuracy of its performance. This made another Neural Network Architecture grow in popularity, in particular when dealing with medical images, this new architecture is called U-NET.

When working with classification tasks with images, CNN works decently. Its architecture makes it possible to learn the feature mapping of an image and convert it into a vector which would be used further for classification. Nevertheless, it is not enough when it comes to segmentation tasks, because after having the vector we would need to convert it again to have the original image in order to classify its segments. This task is much more complex and that’s why the idea of UNET emerged.

This new architecture uses the same feature map to convert the image into a vector and undo the process and restore the original image. This reduces the distortion of the image and conserves its integrity.

Architecture

UNET is a U-shaped Neural Network Architecture that has generally four encoder blocks and four decoder blocks connected by a bridge. So, unlike the classification where the end result is the most important thing to extract, in semantic segmentation the algorithm also needs to be able to project the features learnt in the encoder blocks onto the pixel space and reconstruct the image.

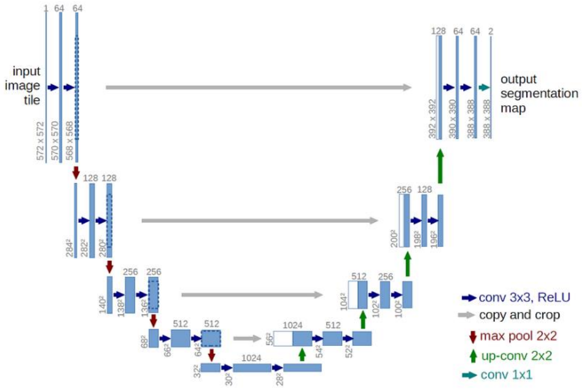

Figure 13. Example of a U-Net Architecture. [g]

Encoder Network:

Each encoder block uses a 3x3 convolutional layer followed by the activation function ReLU. The convolutional layers are used in order to learn from the data, in our case it reduces the dimensionality of the image turning all the pixels in its 3x3 window into a singular value. Afterwards an activation function is used to introduce non-linearity into the network in order to introduce generalization of the training data. ReLU is an activation function used in neural networks with the aim of converting all negative values into 0. It is used instead of sigmoid functions and hyperbolic tangent because of its ability to accelerate the training speed because of how simple its derivative is. That derivative being 1 for any positive input.

Following the convolutional layers and the ReLU activation function we find a max pooling of dimensionality 2x2. It is used to reduce the dimension of the previous feature maps by half, which reduces largely the computational cost.

Skip connections:

The encoder blocks are connected to the decoder blocks using long skip connections. They are used to make the reconstruction of the image in the decoder part possible. It is used to pass the features from the encoder to the decoder in order to recover spatial information used during down sampling.

Bridge:

The bridge is formed by 3x3 convolutional layers and ReLU activation function and is used to pass the information from the deep part of the encoder to the deep part of the decoder.

Decoder Network:

After going through the bridge, a 2x2 transposed convolution layer is used in each one of the four blocks. Then the feature maps in each one of the levels is used to reconstruct the image in each level of depth until reaching the original size of the image using, like in the encoder part, 3x3 convolutional layers and the ReLU activation function.

The main purpose of using this algorithm is semantic segmentation. In other words, the aim is to classify each pixel in the class it most likely belongs to. For that reason, the last output goes through a 1x1 convolution layer with sigmoid activation. This way, it assigns a probability to each one of the pixels in the image. That probability is the one that makes the binary classification task possible. However, in our case we would use SoftMax because of the fact that we are dealing with a multiclass problem.

Proprieties of U-Net

When working with high-resolution images, U-Net preforms very well in the creation of segmentation maps because of its encoder-decoder structure. It also works well in multiclass segmentation because of its capability on making a pixel-level segmentation map for each one of the classes.

Nevertheless, the Neural Network has some limitations. One of them being the high computational cost produced by the skip connections, those being more costly than other types of connections. Adding to that the high number of parameters required for the architecture to function because of the skip connections and the additional layers in the expanding path. This particular limitation makes the architecture more prone to overfitting than other Neural Networks like CNN for example. Another problem U-Net faces is its high sensitivity to initialization. Meaning if the Network makes an error, it would be amplified because of how the skip connections work.

One challenge that does not target U-Net particularly, but the task of multi-class image segmentation as a whole is class overlap. When it comes to working with medical images, the boundary between two organs may be difficult to distinguish, because the pixel can belong to both classes at the same time. One solution to this problem would be to work with a probabilistic segmentation map.

Selection of Hyperparameters:

Depending on the images we are working with, the accuracy we want to reach and the computational cost we are ready to deal with, we would work with certain parameters or other. The selection of hyperparameters is crucial in the model development process because it can significantly increase or decrease its performance. While working with U-Net, there are four parameters to keep special attention to: The number of filters, the size of the Kernel, the Stride, and the Pooling size.

The number of filters determines the number of feature maps used. The higher the number is, the more the model learns from the complex features of the image, the more prone to overfitting the model is. The Kernel size determines the size of the sliding window used by the convolutional layers to generate the feature maps from the input image. A higher size implicates capturing more information which would make the model more accurate, but at the same time it makes it more prone to overfitting.

Another hyperparameter that can be adjusted is the Stride. It represents the step size of the convolutional layers when the filters are applied to the image. It reduces the size of the feature map when it gets increased, helping with computational cost of the algorithm while at the same time discovering fine details of the image that would normally go unperceived. The pooling size also determines the amount of down sampling allowed, meaning that the larger the size the more summarized the information about the input image is, lowering the cost while at the same time may reduce the accuracy of the model. When it comes to selecting the values for these hyperparameters, it is suggested to start with low levels and then increase it gradually until the performance required is reached.

Intersection over Union metric (IoU)

Intersection over Union is a metric that computes the accuracy of a given model when working with images. It measures the overlap between the prediction of the model and the ground truth. To do this it calculates the ratio between the area of intersection and the area of union, meaning that we get a number between 0 and 1, 0 reflecting that there is no overlap and 1 reflecting that they perfectly overlap.

When working with object detection, bounding boxes make perfect sense when it comes to calculating this metric. However, like in our case, when the task is based on the segmentation of the image, then we are working with irregular shapes, which means that a pixel-by pixel analysis is to be done.

In order to calculate this metric, we calculate the ratio between the True Positive and the summatory of the tree classification types: True Positive, False Positive and False Negative. True Positive being the area of intersection between Ground Truth and Segmentation Mask, False Positive being the predicted area outside the Ground Truth and False Negative being the number of pixels in the Ground Truth that the model failed to predict.

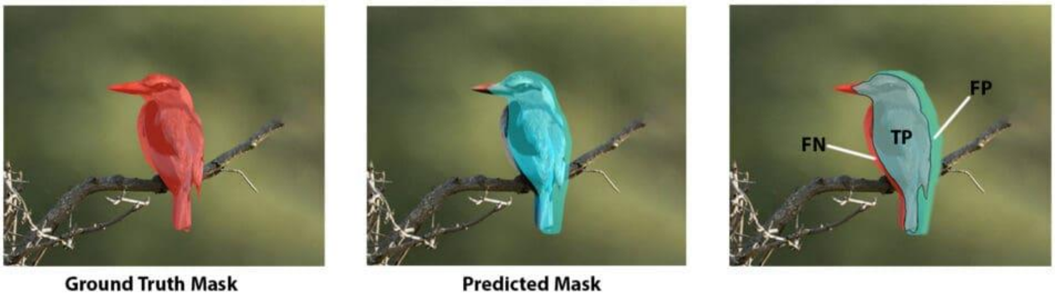

Figure 14. IoU metric application. [h]

4. Methodology

Analysis of the case

The main goal of this study is to train a supervised model to be able to do semantic segmentation on CT images of the abdomen with the highest accuracy possible, considering the diversity of the patients. For that reason, we get to work with healthy patients as well as with patients that have cancerous tumors in the colon, the pancreas, or the liver. This injects some diversity in the sample, because we are not guaranteed that the new patient that needs the semantic segmentation of its abdomen is healthy or suffers from a pathology. Even if we are aware of such information, training a model with these variations is going to make it possible, not only for healthy patients to get this procedure done, but also for the patients that do suffer from tumors in those specific organs.

Moreover, we boost the diversity of the samples by collecting the samples from different clinics and hospitals. This type of diversity is used to mitigate any biases that might come from having a one-center collection of the data. Depending on the sensors and how the professionals are formed, there might be some differences in the CT-scan images collected which might hinder the performance of the model trained in this project.

Nevertheless, injecting diversity in the samples makes the case study more complex. Having images that differ so much one from the other make it harder for the models to learn from a pattern, which in itself hinders the performance of the algorithm and lowers the accuracy of its results. In addition to that, we also face the difficulty of having a variable number of slices of CT-scans from one patient to another because of anatomical variations.

The goal of this research paper is not solely to have a very accurate model that only works with a specific type of patient but to generalize it to a wide population that can be very diverse. For this reason, although injecting some diversity might limit the performance of the model, we deem it to be an important component that help reach the aim of the study discussed above.

Several approaches are used to reach the main objective previously mentioned. Two preprocessing tools are used to increase the contrast of the image in order to help with the identification of the organs. Two U-Net variants are used in order to find the architecture that best suits the dataset. Finally, three levels of depth of the U-Net algorithms are used in order to look if there can be any increase of the accuracy of the results when there is an increase in the number of layers of the Neural Network.

Legal and Ethical Framework:

We are working with medical data in this project, meaning that the data has to obey to some rules in order to be used in an ethical way. In this study, the annotated cases were provided by the platform Grand Challenge with the following origin: Ma, Jun, Zhang, Yao, Gu, Song, Zhu, Cheng, Ge, Cheng, Zhang, Yichi, An, Xingle, Wang, Congcong, Wang, Qiyuan, Liu, Xin, Cao, Shucheng, Zhang, Qi, Liu, Shangqing, Wang, Yunpeng, Li, Yuhui, He, Jian, & Yang, Xiaoping. (2021). AbdomenCT-1K: Fully Supervised Learning Benchmark [Data set]. In IEEE Transactions on Pattern Analysis and Machine Intelligence: Vols. 10.1109/TPAMI.2021.3100536. Zenodo. https://doi.org/10.5281/zenodo.590303, under the license of Creative Commons Attribution 4.0 International.

When it comes to the privacy of the data, the patients do not have any of their personal information displayed. The only data available are the CT-scan images. Adding to that, the patients cannot be identified. For that reason, we can claim that the privacy of the data used in this project is intact.

Preprocessing of the Dataset

As discussed in the previous section, we have used several configurations of the preprocessing of the images and the architectures of the U-Net algorithm as a way to evaluate the best performing models and compare their results to each other. The main objective of this study is to reach a high Intersection over Union value when applying the semantic segmentation of the organs in the CT-scans of the slices of the abdomen.

This goal can be reached more accurately if a preprocessing tool is used. In order to verify this assumption and find the technique that works best in our case, we evaluated and compared three methods. The first method does not use any algorithm to do the preprocessing of the image. The second method is a slight variation of the Histogram Equalization methodology. The last technique used is Contrast Limited Adaptive Histogram Equalization (CLAHE) which also can be considered as a variation of the Histogram Equalization methodology.

The first approach uses the images without using any algorithm to prepare them for the U-NET network. This approach uses the raw CT scans and feeds them to the model, without making any changes to the original images. We have decided to use this configuration in order to create a base case for the benchmarking of the other two mechanisms.

The second approach uses a variation of the Histogram Equalization methodology on the images before using them in the training of the model. Histogram Equalization is a popular method used to increase the contrast of the images. It is generally used when the intensities are close to each other and the objects in the image can't be detected easily. This technique spreads the most frequent intensity in the scale of grey, between 0 and 255, to get rid of the limitations that come with closeness of the intensity levels. This way, the boundaries in the image are clearer, and the different objects can be detected more easily.

However, in this study we use a slight variation of Histogram Equalization. First, all the pixels that have intensities close to 0, meaning they are on the darker spectrum, get tuned to 0. Then the Histogram Equalization algorithm gets to be deployed to increase even more the contrast of the image. This is used in order to highlight the objects to be detected by turning black, all the pixels that are most likely to be representing the background or the boundaries between the organs. Adding to that, the increase of the contrast helps identify and separate one type of tissues from the others.

The third approach consists of using Contrast Limited Adaptive Histogram Equalization in the preprocessing of the images. This technique works significatively better with images of high intensity variations. It does that by, instead of working on the whole image, dividing the image into partitions, performing Histogram equalization on each one of the partitions and then conducting Bilinear Interpolation. The first step consists of generating tiles and then applying the method of Histogram Equalization on each one of the partitions. This way, each portion of the image has its own Histogram of intensities, which makes it easier to spread the highest intensities as a way to heighten the contrast of that specific tile. After this process, the image gets sewed together using bilinear interpolation. This method uses the neighboring pixels to the border in order to reassemble the image into its original form.

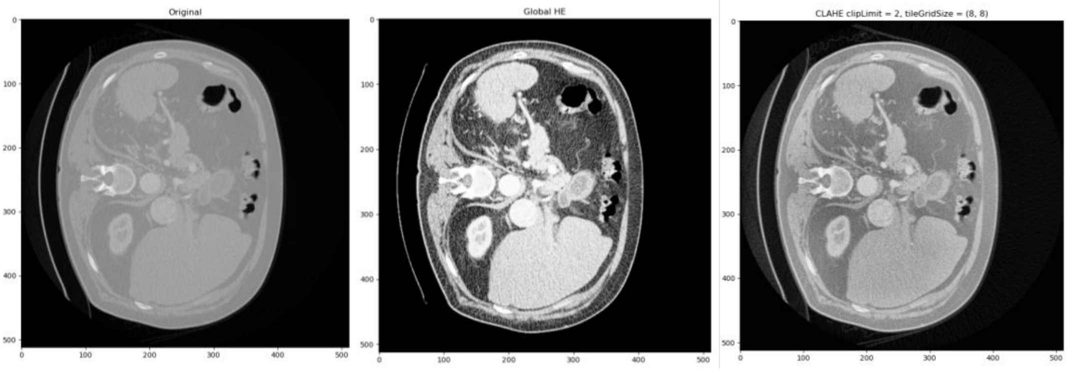

Figure 15. Comparison of the three preprocessing techniques used.

The General Topology of the U-Net architecture

The U-Net architecture is a variant of artificial neural networks that has the shape of a “U”. Having the encoder in one side and the decoder on the other. The two sides are connected by a “bridge” and some skip connections. The length of the encoder and the decoder is what is known by the depth.

The architecture of our model considers that each level of depth is formed by two 2D Convolutional Networks, a ReLU activation function and 2D MaxPooling or 2D UpSampling depending on the side of the “U”. So, if we are working with a depth of 4, that means that the previous process gets to be repeated 4 times before reaching the bridge, and four times before the final classification.

We need to bear in mind that the architecture is always symmetrical. That way, first, the encoder learns from the input images, forming feature maps and reducing the dimensionality of the images. The objective of this first process is to learn from the data and train the model in order to be able to classify each of the pixels in one of the output categories. Then afterwards comes the second process. The decoder’s aim is to recreate the original image maintaining the classification of the pixels. This is when the U-Net architecture gains attraction. Because of how symmetrical the architecture is and with the help of the skip connections linking one side of the “U” to the other, the reconstruction of the image is possible, and the segmentation can occur after training the model.

In our case study, we are using levels of depths ranging from four to eight in order to find the best performing model. However, after looking at the IoU values we reduced this pool to three depths: five levels, six levels and seven levels. Furthermore, two variants of architectures are used, under the names U-Net 1 and U-Net 2. Those two models are going to be explained in the next section.

The Use of Varying Depths of the U-Net Architecture

As previously mentioned, we are using three levels of depth in the U-Net architecture. These three levels seem to yield the best results and the highest IoU scores. We are using architectures with the depth of five levels, six levels and seven levels. These configurations are going to be trained with the dataset of the CT-scan images in order to be, later on, evaluated and compared to each other.

The next figure displays the architecture of a five-level U-Net model. As we can see there is five MaxPooling blocks, five UpSampling blocks and five skip connections that go from one side of the level to the other side. Each level follows the structure that we have explained in the previous section. However, this time, Skip connections bear one convolutional block followed by the ReLU function and then connects the encoder to the decoder.

The bridge contains two convolutional blocks, one ReLU function and two other convolutional blocks which are followed again by a ReLU function. After the input and before the output layer, there are two convolutional blocks, one having a ReLU function following those blocks while the other having a SoftMax activation function. The SoftMax layer is put in place in order to do the five-class classification.

This is the structure of a five-level U-Net model. We have decided to train U-Net architecture that go from having five levels to seven in order to compare their performance using the IoU metric and decide which depth is the most optimal depth in our case.

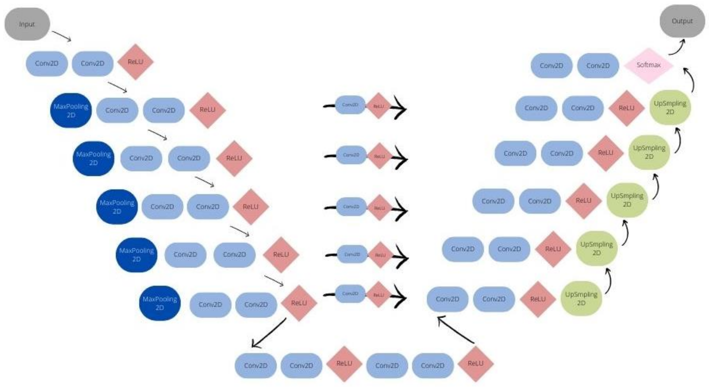

Figure 16. Architecture of U-Net 1 with five levels.

The Use of Other Variations of the U-Net Architecture

This project uses two variants of the U-Net architecture. The first is the one explained previously. It uses the basic structure of the U-Net model but adds on one convolutional bock and one ReLU function to each one of the skip connections. The architecture is represented by the figure above.

The second model is also a variant of the U-Net algorithm, it uses the base of the U-Net 1 and applies some changes. This new architecture is displayed in the figure bellow.

The new changes are mostly applied to the first and last blocks of the networks. Three convolutional blocks are added before the first layer and a ReLU function is placed before the first MaxPooling block. The output blocks also adhere to some changes. The SoftMax layer is preceded by only one convolutional block, while a chain of one convolutional block, one ReLU function, one convolutional block and another ReLU function are added before the last blocks of the architecture.

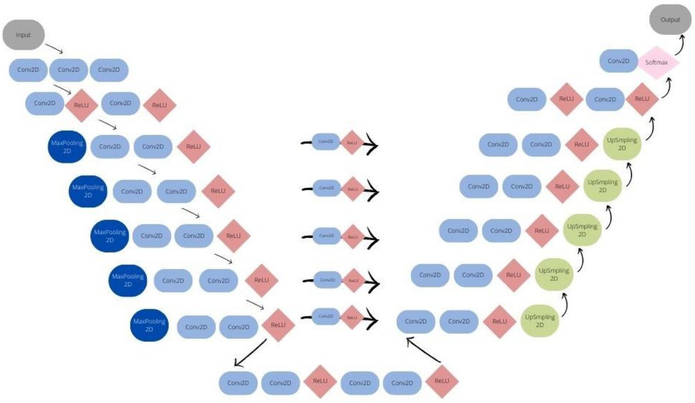

Figure 17. U-Net 2 Architecture with five levels.

The Use of Various Epochs in the Training of the Model

The epoch parameter is an integer that is used to determine the number of times the algorithm goes through the training dataset before stopping and showing the results. It is used to optimize the learning process when working with gradient descent, which is an iterative process. Meaning that when repeating the process x times, the algorithm updates the weights over each step-in order to optimize the learning. In our study, we evaluated and compared the models using epochs ranging from 1 to 200. We have calculated the IoU values of the validation set in order to look for the most optimal size of epochs in our case study.

The Batch size is another hyperparameter used in deep learning. It indicates the size of the samples to use in the training of the model before the next iteration. In each iteration the model updates its internal parameters as a way to figure out the best optimization of the parameters. In each one of the loops, the predictions are compared to the real values and an error rate is calculated, and then the parameters are modified if the error rate does not fit into the threshold fixed. In our project, we’ve selected the batch size to be 30 which means that we are working on a Mini-Batch Gradient Descent algorithm.

In our case, we are using both hyperparameters simultaneously. Each epoch uses a total of 30 training samples in each iteration. The number of iterations needed in each epoch has to do with the number of batches needed in order to cover the whole training dataset.

Software and Hardware used:

In this project, we have used several programs and hardware in order to reach the final aim of the project. We have used Python as the programing language of choice because it perfectly fits projects that revolve around Data Science, Machine Learning and Deep learning. Several built in libraries were used, like Nibabel[I] to open the nii extension files and TensorFlow [II] to train the Artificial Neural Network (U-Net).

Other than that, we also had to have a very potent hardware in order to follow up with the experiments. We are working with thousands of patients, each patient having hundreds of slices that are crucial for the training of the model. This makes the dataset very complex to deal with, especially for a single laptop that tries to process it. Moreover, Deep Learning models are generally more complex than Machine Learning models. They require a larger number of resources in both the need of space memory and the execution time. For this reason, we have worked with a powerful Machine equipped with NVIDIA RTX49 GPU with 24 GB of RAM that was handed to us by the Polytechnical University of Valencia (UPV).

5. Experimentation

In this section, we are going to discuss the various experiments conducted in order to generate the best performing U-Net prediction model that matches the dataset we are working with. For that doing, we worked with various depth of U-Networks, various architectures, hyperparameters and, most importantly, various preprocessing image tools in order to find the model that yields the most accurate results.

Dataset Selection

Medical images are used constantly by healthcare providers to diagnose patients and subscribe prescriptions to treat their pathologies. However, the reading of these images has proven to cost a lot of the doctor's valuable time and is highly dependent on the experience and the subjectiveness of the radiologist in charge. In order to surpass these limitations, various studies have emerged trying to make the task of medical image semantic segmentation automatic, with little to no help from the healthcare providers. That would help in the diagnosis of the patient, the localization of the pathology, the treatment planning and may even be considered as an entry gate to computer-integrated surgeries.

The task of collecting images that are good enough for the training of the model can be tricky. That is due to the scarcity of the data that has to be collected, the lack of diversity in the sample of patients, the noise generated by the sensors or the closeness in gray level of the different soft tissues. These are just a very few of the limitations that we can be faced with dealing with medical images. Adding to this the fact that the Healthcare sector can be very sensitive to errors since we are dealing with human lives. For this reason, the models generated need to have a good performance and high levels of accuracy if they are later to be deployed in hospitals and care centers.

In order to start the study on a firm note, the first most important task is to look for a dataset that is reliable, holds at least some of the diversity of the population and has a large number of images already labeled by professionals. After investigating, we found out that the website Grand Challenge had a good reputation when it comes to delivering data, challenges and scientific papers based on the medical field.

Grand Challenge is a platform that gives the opportunity to organizations in the medical field to share an unresolved case study and create a competition between its users to look for the best model for the problem shared. The users are generally familiar with the techniques that help preprocess the data, do the exploratory analysis, help with the building of the predictive models and the validation of the results. By accepting a challenge, these users get to practice their skills in real-world problems, helping them gain knowledge and experience, while at the same time improving some of their soft skills. These can range from learning how to compete and how to work in groups to learning how to communicate your results to a wider audience and how to write a scientific paper that have the possibility to be published in some of the leading journals such as IEEE Transactions on Medical Imaging and Medical Image Analysis.

In our study, we are dealing with a case of multi-organ segmentation that includes four different organs. These organs are annotated as the following: liver (label=1), kidney (label=2), spleen (label=3), and pancreas (label=4).

The dataset collects more than 1000 3D CT scans from five existing datasets: LiTS (201 cases), KiTS (300 cases), MSD Spleen (61 cases) and Pancreas (420 cases), NIH Pancreas (80 cases), and a dataset from Nanjing University (50 cases). The last dataset contains 20 patients with pancreas cancer, 20 patients with colon cancer, and 10 patients with liver cancer. This last dataset is introduced to give more diversity to the study, and work with patents with pathologies as well.

The CT scan have a resolution of 512x512 pixels with the slice thickness ranging from 1.25mm to 5mm. The number of slices vary from one patient to another depending on the length of their torso. The annotations were done first by trained single-organ models, then passed to fifteen junior annotators which then would be verified by a 10-years' experience senior radiologist.

The dataset is formed by two folders, each one containing archives of nii extension, one of them storing the CT images of the slices of the abdomen of the patients. The other folder containing their ground truth. A fraction of 60% of the dataset then gets used to train the model, 20% gets used in the testing phase and the other 20% gets used as the validation set.

Visualization of the CT-scan images

CT-scan images along with their Ground Truths of more than a thousand patient are used in the study. Each of the patients has a three-dimensional archive of images. The first and second shape values display the number of pixels that form each one of the slices. All the slices have a resolution of 512x512. The third shape value represents the number of slices the patient has collected of the

2D CT-scan images from his abdomen. However, as previously mentioned, every patient has a different number of slices depending on their own anatomy.

When it comes to the Ground Truth images, five different classes can be observed. The class with the label 0 forms the background, the label with the value equal to 1 refers to the liver, the value 2 to the kidney, the value three to the spleen and finally the value 4 refers to the pancreas. These labels are not present in all the slices of the patients, it only shows in the slices that capture them. Meaning that depending on where the slice is situated in the abdomen, it would display all the organs or a few of them.

In order to better understand the images, we've decided to overlap both the CT-scan and the Ground Truth on top of each other. To do that, first we have created a three-color channel: Red, Green, and Blue. We did that so every organ can be differentiated, giving it a color different to the rest that helps identify it in the image amongst the other organs.

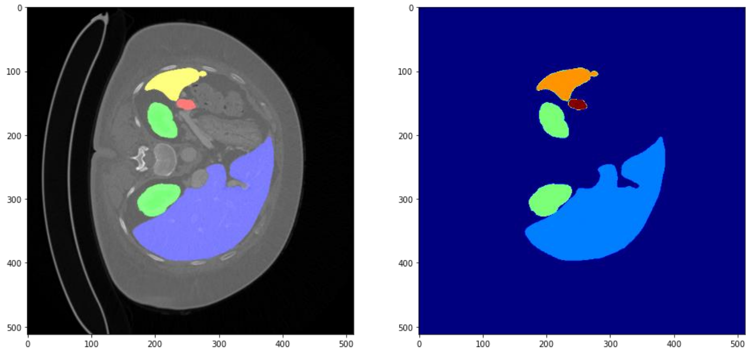

Figure 18. Visualization of the slice 350 of the patient number 12.

As we can see in the image above, the CT-scan image is displayed on a gray scale, having values that range between 0 and 255. Then on top we can distinguish the two green areas representing the kidneys, the blue area representing the liver, the yellow area representing the spleen and the red area representing the pancreas. On the right side we can see the Ground Truth displayed by itself. The images are taken from the patient "Case_00012", using the slice number 350.

Methods for the preprocessing of the images

When it come to the preprocessing of the images, three methods are used. The first one consists of not using any technique to enhance the contrast of the image. This gives us the base case that would be used to compare the other two preprocessing tools used. The images we are dealing with have close intensities between the pixels, which means that there is low contrast between the organs in the image, which is not ideal. This represents a limitation to the training of the model.

The difficulty of identifying different tissues of the organs increases. This is clearly displayed in the poor quality of the results and the low IoU metric calculated for the model not using any preprocessing method. The model trained is the U-Net1 with five levels and no preprocessing of the image.

The second technique applies a variant of Histogram Equalization. Histogram Equalization, as explained before, is a method used to increase the contrast of the image globally. This variant adds to it the fact that all the intensities that fall in the low range are changed to 0. This way, the background and the borders are clearer to identify, which helps the semantic segmentation of the organs. Then the Histogram Equalization technique is used. This method works on the image as a whole and spreads the intensities of its pixels all along the 0 to 255 range of grey. This way the contrast of the image is increased even more, and the organs are easier to detect and locate in the CT-scan images of the abdomen.

The third method used in the comparison is Contrast Limited Adaptive Histogram Equalization (CLAHE), which consist of a technique that, like the previous method, increases the contrast of the image separating the most frequent intensities. However, this time it does that by focusing on small parts of the image. It partitions the image into tiles, applies the contrast algorithm, and then stitches the image back to its original form.

The method of the variant of Histogram Equalization has shown to be the most effective, yielding the best IoU results compared to the other two preprocessing algorithms. For this reason, this method is the most optimal which means that it is going to be chosen to be used in the final model. The CLAHE algorithm has yielded better results than not preprocessing the image but did not have a better performance than the variant of Histogram Equalization. For this reason, the preprocessing technique that is going to be used in this project is the second one.

Evolution of the metrics considering the number of epochs:

Another hyperparameter that is to be chosen is the number of epochs used. Epochs, as discussed in the Methodology section, represents the number of trainings loops the model goes through before coming up with the final the results. It is used in order to optimize the internal hyperparameters of the Artificial Neural Network used. Generally, the more epochs in a system, the better the accuracy of the model. In our case, we decided to plot the loss function and the IoU metric values along with the number of the epochs in order to choose the mot optimal number of epochs for the model. We visualized the values for the U-Net 1 and -Net2 with a level of depth of five while using the variant of Histogram Equalization as the image preprocessing tool.

U-Net 1

First, we are going to look into evolution of the loss function and the IoU metric when increasing the number of epochs in the model U-Net1 that is trained with images preprocessed by the variant of Histogram Equalization.

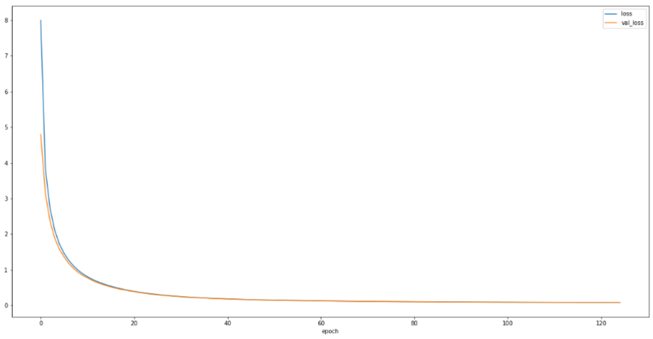

Figure 19. Loss Function considering the number of epochs used in U-Net1 of five-levels.

Looking at the loss function in the plot above, we can see that it has an asymptotic shape, meaning that when the number of epochs is low, the loss of information is really high. Increasing the number of epochs in the left side of the plot has a bigger impact than on the right side. For this reason, when reaching 20 epochs, the model can be considered as being trained correctly, and the loss function is low enough to accept the model as valid.

The plot bellow shows both the training and validation IoU values of the model trained, considering the number of epochs that ranges from 0 to 120. We can observe that the IoU metric, in both datasets, increases during the whole range. For that reason and considering the evolution of the loss function, the most optimal size of epochs would be on the higher scale. Not all number of epochs are displayed in the plot bellow. This decision was made in order to make the plot more readable. Comparing the sizes of epochs, we find that having 177 epochs yields the best IoU values when looking into the validation set.

One of the big limitations of Artificial Neural Networks is its tendency to overfit. In our case, this does not show up as being a problem. We can observe in the plot bellow that the validation set has a higher IoU value than the training set considering all the range of numbers of epochs used in the model.

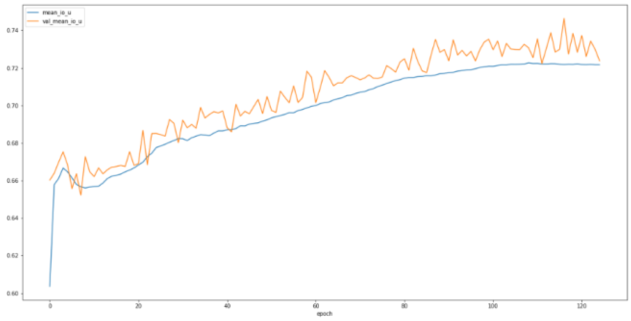

Figure 20. IoU metric considering the number of epochs used in U-Net1 of five-levels.

U-Net 2

Next, we are going to look into evolution of the loss function and the IoU metric when increasing the number of epochs in the model U-Net2 that is trained with images preprocessed by the variant of Histogram Equalization.

The plot bellow shows the evolution of the loss function of the U-Net2 architecture in a varying range of number of epochs. This plot is similar to the previous loss function plot. It has an asymptomatic tendency, which means that with values higher than ten, the model shows a low level of the loss of information metric. This means that when ten epochs are used, the model does not lose a lot of information, so this value can be used in order to optimize memory and time resources.

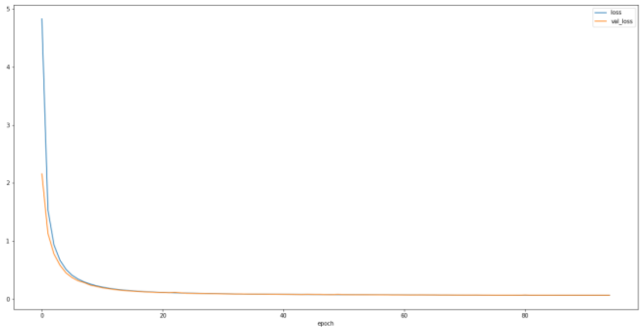

Figure 21. Loss Function considering the number of epochs used in U-Net2 of five-levels.

The figure bellow displays the evolution of the IoU metric for the U-Net2 architecture. This case is slightly different than the previous model. We can see that the plot of the IoU value increases until reaching a limit halfway, and then continues to decrease. This is the case for both the validation and the training set. However, although a pattern can be detected, the validation set's values oscillate more frequently making harder to find the most optimal number of epochs.

Finally, looking at the results, we choose to have a total number of 47 epochs in the case of the U-Net 2 model because it yields the best results in terms of the IoU metric for the validation set.

This model does not suffer from overfitting neither. The validation set gets generally better results. Although it oscillates and yields IoU values that have a high level of variation, most of them still land on top when the IoU values are compared to each other.

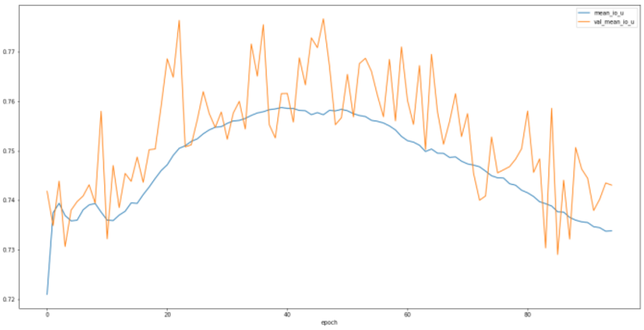

Figure 22. IoU metric considering the number of epochs used in U-Net2 of five-levels.

The same study of the loss function and IoU metric has been conducted for both models U-Net with six levels of depth. The number of epochs optimal for U-Net1 was 57 epochs yielding the best results when it comes to the IoU metric. For the U-Net2 with six levels the most optimal value for training is 29 epochs. Those optimal values are going to be used in the training and comparison of both architectures of U-Net using five and six levels of depths in each.

Model used in the case study.

Various experiments were conducted before reducing the pool of the models that would later be compared. We have chosen the architectures U-Net 1 and U-Net2 previously explained to evaluate in more depth in this project. The first model has the basic U-Net structure but adds a slight variation when it comes to the skip connections. The latter use one convolutional block and one ReLU activation function before reaching the other end of the "U" structure. Meanwhile, the second architecture adds some extra convolutional blocks and ReLU activation functions in both the start and the finish of the structure.

Another metric that is going to be taken account of is the preprocessing of the image. As discussed before, the three preprocessing tools used in the experiments are: No preprocessing, a variation of Histogram Equalization, and CLAHE. Those methods are used to modify the contrast of the image making it clearer for the model to identify and segment the organs and the tissues displayed on the image. Looking at the results of the experiment, the preprocessing techniques that yielded the best results is the variation of Histogram Equalization. This method consists of turning the darker intensities to zero and applying global Histogram Equalization on the image. This increases the contrast even more, making it possible to separate one organ from the other in an easier and more accurate way.

6. Results

Two models were trained: U-Net 1 and U-Net 2, to solve the semantic segmentation problem we are faced with. The models differ in the introductory and the final layers of the neural network. We are going to proceed to analyze the results, evaluate the models and compare them to each other.

The visualization of the result is a 2x3 grid that carries six images. The first image is the image preprocessed before applying the segmentation. The following image displays the anterior image with the final prediction of the model. Meaning the 4 organs are identified and classified. However, having the four classes in the same image may lead to the confusion of the observer. Even though we can represent each class having its own color we would still face the problem that comes from the possibility of the overlapping of the classes, making it a lot harder to differentiate between the organs, which represents the main objective of this study. For that reason, the following images show each one of the 4 organs displayed separately. First, we have the liver, then the two kidneys, the spleen and then finally the pancreas.

When it comes to the channels of colors, we currently have 3 different channels. The first color used is yellow, it represents the area that is well detected by the model. This would be the area of intersection between the segmentation of the model and the ground truth mask. The next color would be the green color, it displays the area that is predicted as being part of a given class but isn’t. This would be the area outside the intersection, where we only have the prediction of the model and not the ground truth. The last color is red, and it represents the area that does indeed belong to one of the 4 classes but is classified wrongly. This would be the representation of the area outside the intersection that belongs solely to the ground truth mask. We will see this in more details in the section bellow.

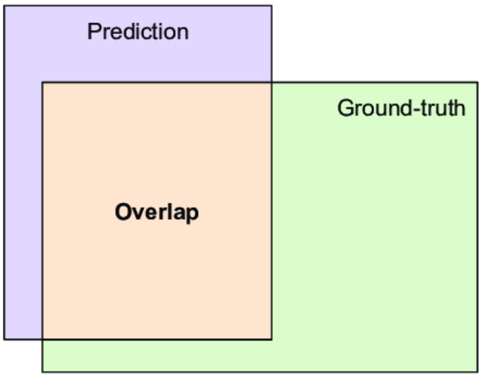

Figure 23. Representation of the accurate prediction (Overlap Area). [9]

In order to compare the results of the models to each other, we've decided to look into the results of two patients. In this case it would be the patient number 428 and 457. One of them being extracted from the training set and the other from the testing set. Bearing in mind that each patient has a different length of the abdomen, meaning a different number of slices. The first patient has a total of 326 slices, while the second patient has a total of 461.

Comparison of the results considering different depths of the U-Net model

The following figures show the results of the abdomen segmentation of the patient number 428, the slice 170, using the architecture of the model U-Net 1 with both five and six levels. These results are displayed in order to highlight the differences in the segmentation considering the depth of the architecture.

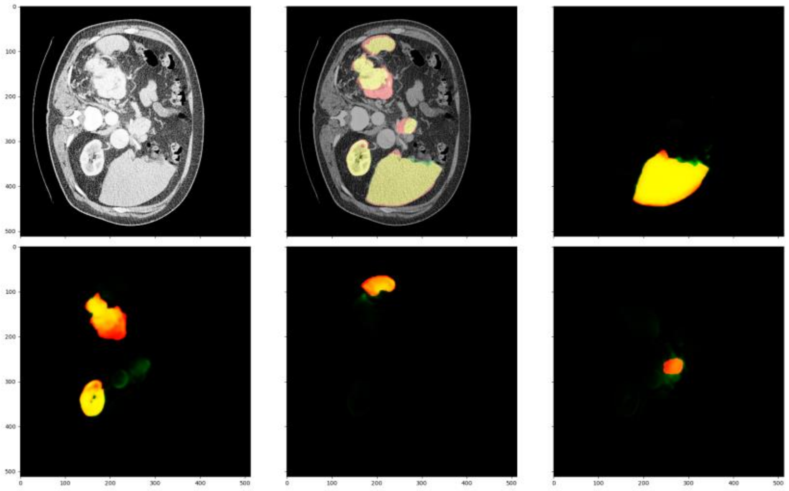

Figure 24: Case 428 slice 170 using Unet1 with five levels.

When looking at the liver's image, we can notice that the architecture with five levels segments the organ more accurately than the other architecture. This can be deduced because of the big presence of the yellow area in the image of the liver, highlighting its accurate segmentation. However, when it comes to the next organ, we can see an increase in the accuracy of the segmentation when we go from five levels of depth to six. The latter architecture detects more accurately the kidneys, specially the left one.

The spleen is an organ that follows the liver in this aspect, the five-layer model detects it more precisely because of the big presence of the yellow area in the image. The other architecture experiences more difficulty identifying the organ and locating it. It has a larger green shadow, meaning it falsely predicts part of the image as being part of the spleen when it is not the case. The pancreas is not well detected by neither of the models because of the predominance of the red area over the yellow. The six-layer U-Net architecture might be considered to have a worse performance detecting the pancreas because of the overwhelming presence of the green area which shows the false positives detected by the model.

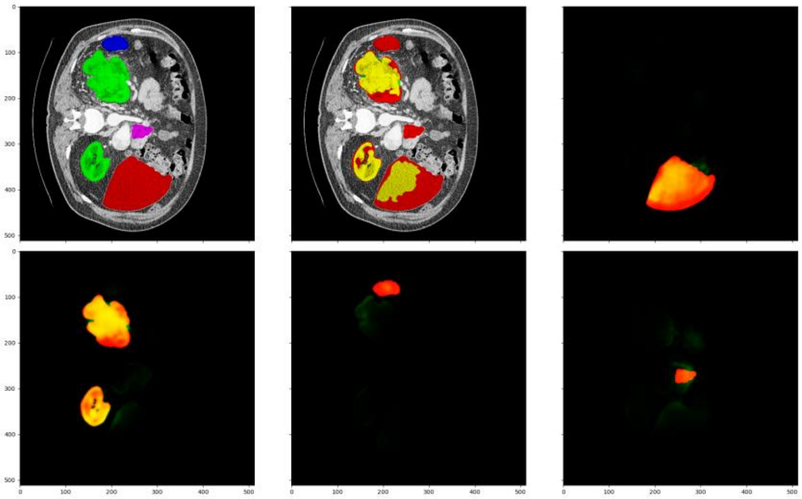

Figure 25: Case 428 slice 170 using Unet1 with six levels.

Comparision of the results considering different variantions of the U-Net model

As previously mentioned, we are training two different architectures of U-Net in order to evaluate and compare the models in order to choose the most accurate trained algorithm for our study. U-

Net1 has a basic U-Net architecture, with a slight modification in the skip connections. It adds one convolutional block and a ReLU activation function in each one of the skip connections.

U-Net 2 is the next architecture used in this study. This specific neural network is different from the previous one because of the three convolutional blocks introduced in the beginning. The two models yielded similar values in the IoU metrics, so we've decided to look further into the result of the images after the segmentation.

For that doing, we looked into the same slices of the same patients using both models. The results showed to be very similar independently of the model used, making more or less the same errors while predicting the false positives and the false negatives. However, in some specific slices, there were some slight differences. We can see this when looking into the CT-scan of the slice 160 extracted from the patient 428.

First, we are going to compare the third image of the grid that displays the liver. We can see a larger red area in the bottom of the liver in U-Net 1, meaning that this model leaves out a bigger part of the liver without detecting compared to U-Net 2. However, U-Net 2 falsely identifies the green area on top as being part of the liver when it is not the case, while the first model draws that area closer to the organ.

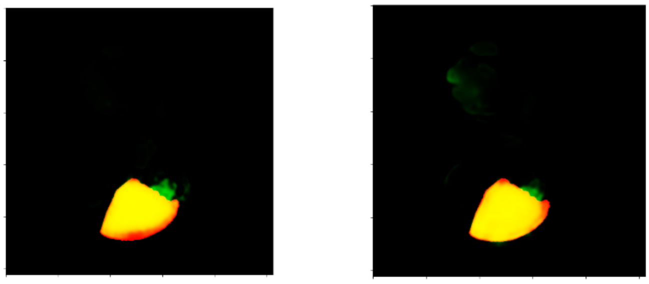

Figure 26: Segmentation of the liver using U-Net 1 and U-Net 2.

The next organ displayed are the kidneys. We will focus more on the right kidney because the other one does not show big differences in the classification. Most of the right kidney is identified by the U-Net 1 model. That observation is drawn from the amount of yellow pigment there is in the left image compared to the right image. For that reason, we can say that U-Net 1 works better on this case. However, a wider green area is also displayed on the first model compared to the second. Meaning that although the model U-Net 1 identifies and locates better the right kidney, it makes more errors because of the negatives being shown as a positive.

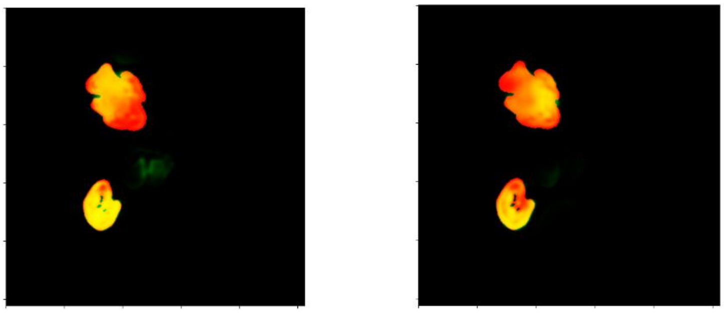

Figure 27: Segmentation of the kidneys using U-Net 1 and U-Net 2.

The spleen displayed bellow is identified poorly by both models. We can reach this conclusion looking at the images and noticing the lack of yellow areas and the prominence of the other two colours. In the second model, the colour yellow is a bit more noticible than in the other model. However this comes with a cost. There is a greater green area in the model U-Net 2 than U-Net 1. This means that although the model U-Net 2 yields better results when identifying correctly the class representing the spleen, it can aldo be worse in predicting the classes because of the amount of false positives it harvests. At the implementation level, this means that depending on the cost of the false positives, we would use one model or the other.

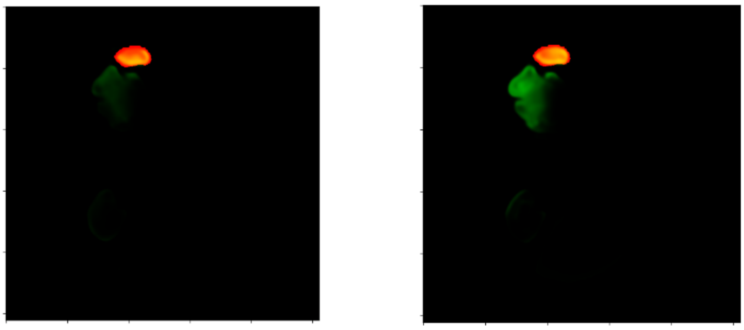

Figure 28: Segmentation of the spleen using U-Net 1 and U-Net 2

Looking at the images bellow of the pancreas, we don’t see any huge differences between the two models. Both U-Net architectures don’t detect accurately the organ in this slice because of the big presence of the red area compared to the other two colours. There is a slight difference between the two images. That being the green area that shows the false positives. In this case, we are seeing the same previous pattern: The model U-Net 2 usually identifies falsely areas that are not a part of the organ we are trying to segment.

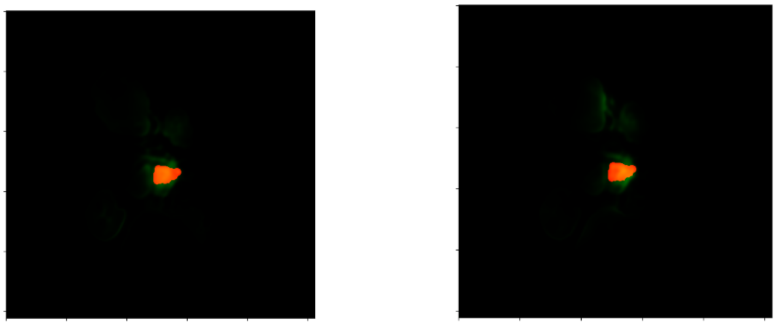

Figure 29: Segmentation of the pancreas using U-Net 1 and U-Net 2.

In conclusion, it's safe to say that the model U-Net 2 generally has a higher number of false positive classified pixels, but that enables the right detection of more of the area of the organ. For that reason, depending on the circumstances of the case and the cost linked to the false positives and false negatives, one architecture would be better suited for the problem than the other.

Results of the IoU metric:

The following table displays the value of the metric of Intersection Over Union for each one of the organs and each one of the models used. The table is of high importance to the study. Not only we get to identify the most accurate model for the task of the segmentation of the organs, but we also get to choose the model depending on the organ we are interested in identifying.

Table 1. IoU metric values for each model and each organ.

Model	**Background**	**Liver**	**Kidneys**	**Spleen**	**Pancreas**
unet-1-5-levels	0.99	0.84	0.71	0.68	0.05
unet-1-6-levels	0.98	0.73	0.62	0.55	0.02
unet-2-5-levels	0.99	0.82	0.65	0.68	0.01
unet-2-6-levels	0.99	0.82	0.68	0.67	0.03
unet-7-5-levels	0.98	0.81	0.68	0.64	0.01

The models used in the comparison are the U-Net 1 and U-Net 2 with five, six and seven levels, using a variant of Histogram Equalization as the preprocessing tool and the most optimal number of epochs for each model considering the IoU metric.

As we can observe from looking into the table above, the U-Net 1 architecture using five levels yields the best results in the identification of all the organs. The Liver yields the best results compared to all the organs, yielding a value of 0.84 in the IoU metric using the model previously mentioned.

On the other hand, the pancreas yields the worst results, having the best value being 0.05 in the IoU metric which is far from being accurate. Generally, the models are not able to detect the pancreas, and that might be due to its small size or to the fact of it not showing in a clear way in the slices taken from the abdomen.

The background in all of the cases yields very high IoU scores. This class is not a very important class to detect because of it not being an organ. However, it needs to be used in order to showcase the area that does not belong to any of the four organs. The background hinders the overall IoU value because it increases its value, making it difficult to evaluate and compare the models to each other. For this reason, separating the IoU values depending on the organ it's identifying does help with the evaluation task.

Results of the most accurate model

As shown before, the model with the highest IoU score therefore the most accurate to the study conducted is the model U-Net1 with five levels of depth, using the variant of Histogram Equalization as a preprocessing tool. In this section, we are going to analyze in more details the results yielded by the best performant model with slices from both previous patients.

Results of the patient number 428:

Slice number 150:

The image below shows the slice number 150 which is located in the middle part of the torso of the patient 428. The third image displays the liver being well segmented. We reach this conclusion because of the overwhelming presence of the yellow area compared to the red and green ones. Matter of fact the model does not recognize some of the borders but does not make any false prediction stating that a particular part of the area belongs the liver when it doesn't.

The next image shows the kidneys also being overall well segmented. This time there is a presence of the green area representing the false positives. The kidney on the right is accurately segmented but the same cannot be said about the left kidney. That kidney has a big red area meaning that the model does not detect it as being part of the kidney. That might be due to unusual shape of the kidney in question, which leads us to the possibility of assuming that the patient might suffer from a pathology.

As we discussed previously, this dataset is formed not only by healthy patients, but also with patients with cancer in the liver, pancreas, and colon. For that reason, there is a possibility that having a one of these three cancers would alter the shape of one or multiple organs in the abdomen, which would result in the kidneys having an unusual shape, thus the model is not used to these particular cases and leaves parts of the organ out of the segmentation. We need to bear in mind that the assumption of the patient having a pathology is nothing but an assumption. We would need a doctor or a medical care professional to look closely into this case in order to reinforce or debunk the assumption.

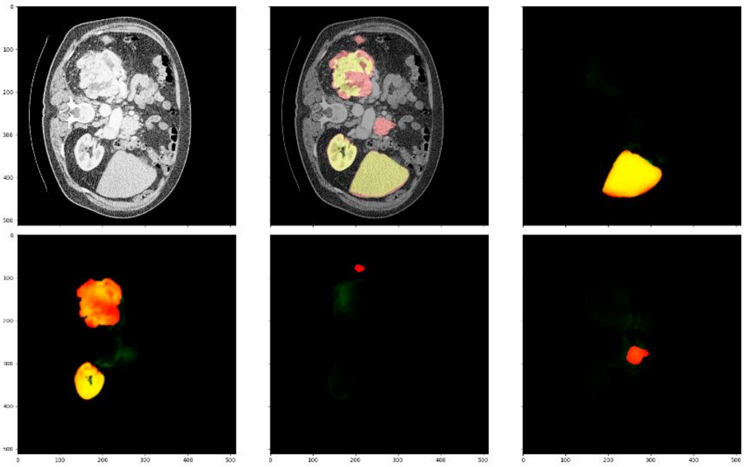

Figure 30: Case 428 slice 150.

The image following the kidneys shows the spleen not being detected accurately at all. It is represented by the red dot. The model predicts it being lower than it actually is and having a different shape that it actually has. Looking into the last image, we can see that the model detects no pancreas being displayed in the CT-scan, which is false. The pancreas not detected is represented by the red area in the center of the image.

Slice number 180:

Looking at the figure bellow representing the slice number 180, we can see that the grid follows the same structure explained before. The second images recompile all the predictions in one, having some organ areas overlapping with each other. The third image of the first row displays the liver. We can observe that the liver is very well classified because of the wide yellow area displayed.

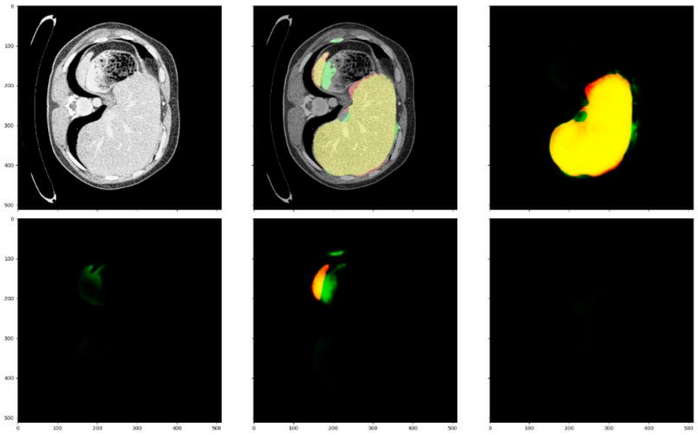

Figure 31: Case 428 slice 180.

Moving to the kidneys, we can observe some green shadows, meaning that even if there is an absence of that specific organ, the model detects wrongly the presence of the kidneys. When we look into the next figure, we observe that a small area of the spleen is accurately detected but there is also an area of the organ that is not detected, and a rather big green area showing that the model falsely detects an area to be a part of the organ when it is not the case. The last image shows that there is no pancreas to be detected, and the model does not make any false predictions in regard its presence.

Slice number 190:

This slice is one of many that do not show signs of any of the four organs we are trying to detect, the reason being its location near the end of the abdomen. In this case it is impossible to find any trace of a yellow or red area. The model makes good predictions when all of the four images are blank. In this case in particular, the kidneys and the spleen are the organs worse detected because there are some areas that the algorithm labels falsely as being those organs. However, we can say that the liver and the pancreas are well classified in this slice because of the absence of any color, especially the red and the green.

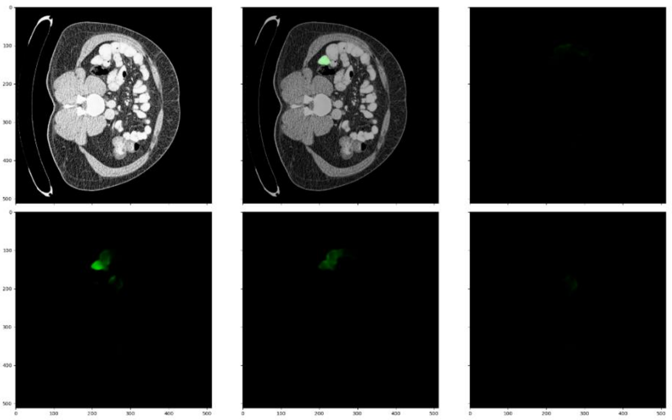

Figure 32: Case 428 slice 190.

Results of the patient number 457:

Slice number 170:

Moving into the second patient, we are starting with a slice that is situated high on the abdomen. We are working with the slice number 170 on a total of 461 slices. This is situated more or less on the same level as the first slice discussed for the previous patient. However, this patient seems to have a longer torso which means that we are at a higher level compared to the patient number 428. In this case, the slice 170 does not display any of the four organs which results in the absence of red and yellow areas. Moreover, no green area is in sight. Meaning the model is not making any false detection of the organs. For that reason, the model works very well in this slice in particular. There is nothing to be detected, and the model does not identify any of the organs.

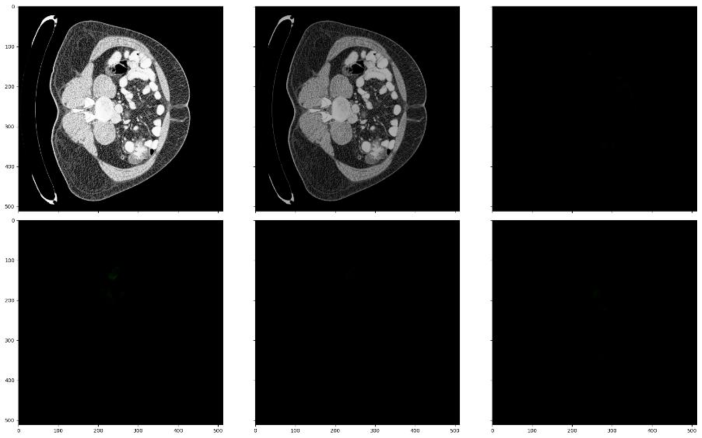

Figure 33: Case 457 slice 150.

Slice number 220:

The next grid shows the slice number 220 of the second patient mentioned previously. In this case we can observe that the liver is generally classified well, but the model detects something on top, represented by the green shadow, which is actually not part of the liver. When it comes to the kidneys, the model successfully predicts the kidney on the right but makes false predictions when it comes to the other kidney. The left kidney is not displayed in this particular slice; however, the model falsely classifies the top part of the image as being that organ.

Despite the spleen not showing in this slice, the model predicts the top area of the image as being the organ, which renders the performance of the model. Looking at the last image, we can see a small red area, meaning that the pancreas is being displayed in this slice, but the model does not pick it up, instead it draws a green surface surrounding it, which displays the falsely positive classified class by the algorithm.

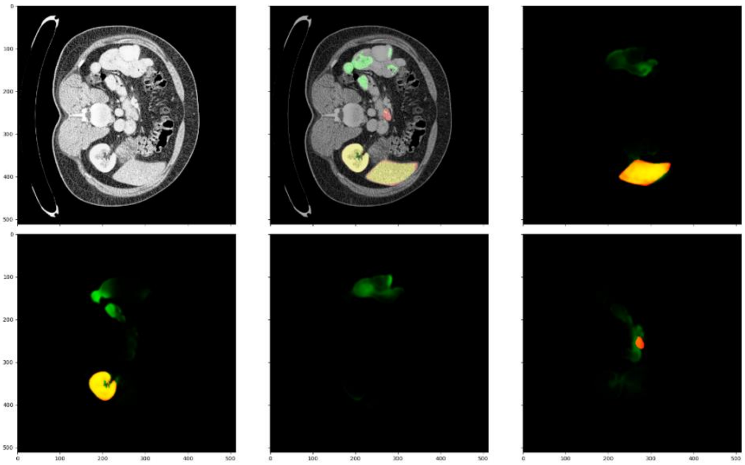

Figure 34: Case 457 slice 220.

Slice number 300:

This slice shows the best results yet for the patient number 457. The liver is classified well, having a very large yellow area displayed in the image. Both kidneys are also accurately segmented, even though the model still predicts some of the area of the image as being part of the organ, when it is not the case. The next image shows the spleen. This organ is not fully detected. The model leaves some parts of the spleen outside, and those are represented by the area in red. Then in the last image we can see that the model still didn’t grasp the segmentation of the pancreas because of the lack of the color yellow in the image. There is a large area of the other two colors, which is not a good signal. The model is still having problems segmenting the pancreas.

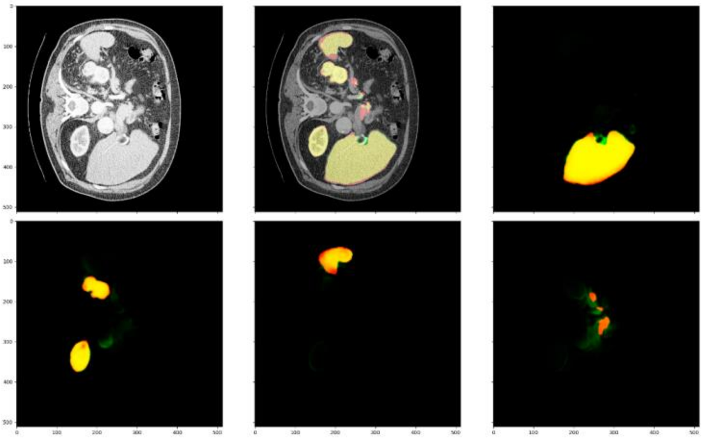

Figure 35: Case 457 slice 300.

In conclusion, we can see tht the model works very well with the recognition of the liver compared to the other organs. Independently of the patient and the slice displayed, the model leans well from the pattern in the images of the patients, and makes good segmentation of the liver.

However, the organ that is not detected properly is the pancreas. Reviewing all the slices, we observed that the model has big difficulties regonizing the pancreas. That might be due to the fact that the pancreas shows differently in the scans from one patient to another, so the model des not have a good basis or pattern to learn from.

That might be due to the natural shape of the pancreas, or maybe linked to the length of the torso. As we previously mentionned, we have different amount of slices of the imaging of the abdomen going from one patient to another. So one reasonable supposition might be that depending on the length of the torso, the pancreas shows itself as one shape or the other. Further investigation can be conducted by separating the length of the torsos in groups and retrain the model. Another supposition might be the fact that the pancreas is very small compared to the other organs and for that reason the model does not pick up the pattern and still makes non accurate predicitions.

The last supposition that might be the reason for the bad results yielded when working with the pancreas can result in the fact that we have patients that have cancerous tumors in the pancreas. This might introduce some noise in the training of the model, which would likely lower its accuracy when segmenting the pancreas.

7. Conclusion

This project follows the basic path of any study conducted by a data scientist. It starts with the collection of the data, the preprocessing of the data, the training of various models, their evaluation and comparison and then finally the analysis of the results. The aim of this study was to train a model which main purpose is to automatize the semantic segmentation of the CT-scans collected on the abdomen. We used a large diverse amount of data collected from patients from different hospitals then compared various configurations that focus on the preprocessing of the images and the architecture of the model, in order to find the best performing solution. After evaluating each configuration, we concluded that the configuration most fitted for this study is preprocessing the images with the variant of Histogram Equalization, then using the architecture U-Net1 with a depth of 5 layers. This solution yields the best results in the detection of each and every organ. The organ that has the best IoU metric is the liver, with an IoU of 0.84. The pancreas yield very low values of IoU, meaning that the organ is not being detected accurately. This might be due to its small size or the similarity of the texture of its tissue to the other organs.

Challenges

- Variant number of slices depending on the patient’s anatomy
- Closeness in grey level intensities in the CT-scan images
- The large amount of time and computational resources the U-Net architecture requires to be trained.
- The background tricks the overall IoU value of the model because of the large area it occupies compared to the organs.

Legacy

This project offers a solution to the problem of the automatization of the semantic segmentation of the abdomen, training a model that learns from patterns displayed in the CT-scans of the slices of the abdomen in order to identify the organs present in the image. It offers the code and the research paper that discusses and resolves the problem. The publication of this study can be helpful to any organization that is trying to resolve the same or a similar issue.

The Relationship between the Project and the Degree studied:

This study follows to structure of any data science project that can be done in the future. It helps the students understand the steps of working with data, training the model, and evaluating and comparing the results. It also helps practice furthermore the technologies crucial for these kinds of projects, like the libraries used during the whole process that are hosted by Python, going from the preprocessing of the data to the training of the Deep Learning model. Moreover, we get to be more familiar with the extension of files of 3D image CT-scans and understand its structure and how to work with it. Other than the technical side, we also get to learn more about the subject by investigating and researching on the topic of the study looking through the most cited and recent scientific papers that resolve a problem that is similar.

Future Works

- Focus on trying to increase the accuracy of the models specially in the semantic segmentation of the pancreas.
- Try out other U-Net architectures and other preprocessing tools for images.
- Add more information on the patients in the dataset, then divide the dataset into clusters considering the sex, age, and BMI of the patient.

8. References

Scientific Papers

[1] Jourdan Arthur, Soucasse Andrea, Scemama Ugo, GILLION JEAN, Chaumoitre Kathia, Masson Catherine, BEGE Thierry. Abdominal wall morphometric variability based on computed tomography: influence of age, gender, and BMI. Clinical Anatomy (2019). DOI: 10.1002/ca.23548

[2] AKSHANSH MISHRA. Contrast Limited Adaptive Histogram Equalization (CLAHE) Approach for Enhancement of the Microstructures of Friction Stir Welded Joints, 10 June 2021, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-607179/v1]

[3] Pisano ED, Zong S, Hemminger BM, DeLuca M, Johnston RE, Muller K, Braeuning MP, Pizer SM. Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms. J Digit Imaging. 1998 Nov;11(4):193-200. doi: 10.1007/BF03178082. PMID: 9848052; PMCID: PMC3453156.

[4] Sharma N, Aggarwal LM. Automated medical image segmentation techniques. J Med Phys. 2010 Jan;35(1):3-14. doi: 10.4103/0971-6203.58777. PMID: 20177565; PMCID: PMC2825001.

[5] Prof. Dinesh D. Patil, Ms. Sonal G. Deore. Medical Image Segmentation: A Review. International Journal of Computer Science and Mobile Computing (2013). Doi: 10.47760/ijcsmc

[6] J.J. Sáenz-Gamboa, J. Domenech, A. Alonso-Manjarrés et al., Automatic semantic segmentation of the lumbar spine: Clinical applicability in a multi-parametric and multi-center study on magnetic resonance images. Artificial Intelligence In Medicine (2023), doi: https://doi.org/10.1016/j.artmed.2023.102559

[7] Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. Deep Learn Med Image Anal Multimodal Learn Clin Decis Support (2018). 2018 Sep;11045:3-11. doi: 10.1007/978-3-030-00889-5_1. Epub 2018 Sep 20. PMID: 32613207; PMCID: PMC7329239

[8] Jin Q, Meng Z, Sun C, Cui H and Su R (2020) RA-UNet: A Hybrid Deep Attention-Aware Network to Extract Liver and Tumor in CT Scans. Front. Bioeng. Biotechnol.8:605132. doi: 10.3389/fbioe.2020.605132

[9] Fernandes Junior Francisco Erivald, Nonato Luis, Ranieri Caetano Ueyama Jó

Memory-Based Pruning of Deep Neural Networks for IoT Devices Applied to Flood Detection (2021). doi 10.3390/s21227506

Website links:

[a]https://www.statista.com/statistics/1334826/ai-in-healthcare-market-size-worldwide/

[b]https://www.baeldung.com/cs/image-histograms

[c]https://www.analyticsvidhya.com/blog/2022/01/histogram-equalization/

[d]https://towardsdatascience.com/histogram-equalization-5d1013626e64

[e]https://www.geeksforgeeks.org/adaptive-histogram-equalization-in-image-processing-using-matlab/

[f]https://www.mathworks.com/help/visionhdl/ug/contrast-adaptive-histogram-equalization.html

[g] https://towardsdatascience.com/u-net-b229b32b4a71

[h]https://learnopencv.com/intersection-over-union-iou-in-object-detection-and-segmentation/#IoU-in-Image-Segmentattion

Coding Libraries:

[I] Brett, Matthew, Markiewicz, Christopher J., Hanke, Michael, Côté, Marc-Alexandre, Cipollini, Ben, McCarthy, Paul, Jarecka, Dorota, Cheng, Christopher P., Halchenko, Yaroslav O., Cottaar, Michiel, Larson, Eric, Ghosh, Satrajit, Wassermann, Demian, Gerhard, Stephan, Lee, Gregory R., Wang, Hao-Ting, Kastman, Erik, Kaczmarzyk, Jakub, Guidotti, Roberto, … freec84. (2023). nipy/nibabel: 5.0.1 (5.0.1). Zenodo. https://doi.org/10.5281/zenodo.7633628

[II] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Rafal Jozefowicz, Yangqing Jia, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Mike Schuster, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

ANEXO

OBJETIVOS DE DESARROLLO SOSTENIBLE

Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

Objetivos de Desarrollo Sostenibles	**Alto**	**Medio**	**Bajo**	**No Procede**
ODS 1. **Fin de la pobreza.**			X	
ODS 2. **Hambre cero.**			X	
ODS 3. **Salud y bienestar.**	X			
ODS 4. **Educación de calidad.**		x		
ODS 5. **Igualdad de género.**			X	
ODS 6. **Agua limpia y saneamiento.**			X	
ODS 7. **Energía asequible y no contaminante.**			X	
ODS 8. **Trabajo decente y crecimiento económico.**	X			
ODS 9. **Industria, innovación e infraestructuras.**			X	
ODS 10. **Reducción de las desigualdades**.			X	
ODS 11. **Ciudades y comunidades sostenibles**.			X	
ODS 12. **Producción y consumo responsables.**			X	
ODS 13. **Acción por el clima.**			X	
ODS 14. **Vida submarina.**			X	
ODS 15. **Vida de ecosistemas terrestres.**			X	
ODS 16. **Paz, justicia e instituciones sólidas.**			X	
ODS 17. **Alianzas para lograr objetivos.**			x	

Reflexión sobre la relación del TFG/TFM con los ODS y con el/los ODS más relacionados.

El TFG tiene como objetivo principal la automatización de la segmentación semántica del abdomen. Este tema se considera muy relacionado con el Objetivo de Desarrollo sostenible 3 ya que este estudio pretende ayudar en el avance tecnológico del sector de la salud y su desarrollo, creando modelos capaces de acompañar a los profesionales de este sector en sus tareas generando resultados más precisos, más fiables y más rápidos.

Además, este proyecto también ayuda al crecimiento económico que sería el ODS número 8. Este modelo entrenado necesitará a técnicos para su instalación en los hospitales y su mantenimiento. Estos técnicos también deberán formar el personal del hospital y explicar el funcionamiento de la herramienta nueva para ser utilizada en los diagnósticos de los pacientes, lo que creará más ofertas de empleo y promoverá el mercado laboral.