INFORMATION EXTRACTION FROM TELEMEDICINE CONSULTATION IMAGES

ALEJANDRO CAMPAYO FERNÁNDEZ

Thesis supervisor: MARTIN TRIPIANA GORGOJO (ABI HEALTH SPAIN,S.L)

Tutor: GERARD ION GALLEGO OLSINA (Department of Signal Theory and Communications)

Degree: Bachelor's Degree in Data Science and Engineering

Thesis report

Facultat d'Informàtica de Barcelona (FIB)
Escola Tècnica Superior d'Enginyeria de Telecomunicació de Barcelona (ETSETB)
Facultat de Matemàtiques i Estadística (FME)

Universitat Politècnica de Catalunya (UPC) - BarcelonaTech

This thesis is submitted to the Computer Science Department, Universitat Politècnica de Catalunya in fulfilment of the requirements for the degree in Data Science and Engineering.

Alejandro Campayo Fernández, June 2023

Acknowledgements

I would like to express my sincere gratitude to the individuals who have played a crucial role in the completion of my thesis and supported me throughout my Bachelor's Degree.

First and foremost, I want to thank Abi Global Health for hiring me as an intern. The dynamic environment of this company is extraordinary, and I am grateful for the opportunity to have experienced it. This being my first internship, you have set remarkably high standards.

I am particularly grateful to my boss and mentor, Martín Tripiana, for providing me with this exciting project. Your guidance and support have been vital throughout my thesis. It is always an honor to learn from you.

I am also indebted to Candelaria Mosquera, who, despite joining the company just a month before my thesis submission, has played a vital role in its development. Thank you for generously sharing your expertise with me, and I apologize for bombarding you with numerous questions regarding medical images.

Furthermore, I would like to express my gratitude to Maria Zyatyugina, my friend and colleague, who has been by my side (quite literally) for six hours a day. Moreover, I want to congratulate you on your own thesis as well; you have done an exceptional job, and I am confident that Abi will greatly benefit from the model you have developed.

In addition, I would like to thank my tutor for correcting my thesis and to extend my thanks to UPC and its esteemed faculty members. You are responsible of my academical and personal growth.

Finally, I am truly grateful to all my friends and family. I can not express how much I owe you without exceeding the limit number of pages, but you already know you are my biggest treasure.

Abstract

A picture is worth a thousand words. This is no different in the context of telemedicine consultations, where images play an essential role in helping the patients explain their reasons for seeking medical advice. An automatic processing of these images could enrich the consultation profiling and improve the overall quality of the medical service delivered. What if we could actually exploit those thousand words?

The aim of this thesis is to explore and develop advanced Artificial Intelligence-based solutions to extract information and medical insights from the images that patients share with healthcare professionals during their digital health consultations.

The project has been developed with the Artificial Intelligence (AI) Lab of Abi Global Health, a telemedicine company based in Barcelona that is currently operating at global scale. Abi has collected a significant amount of images from their patients that can be used to train machine learning (ML) algorithms to extract insights and complement the description of the medical consultation.

The different algorithms developed during this work comprise a full-scale image processing pipeline. It begins with a first-level image classification, whose result is used to determine the following process that is applied, namely: captioning, text extraction or further classification.

On images depicting parts of the body, a short description is generated with captioning methods. Text is extracted from pictures of documents or medicine packaging. Lastly, pictures of medical imaging studies are classified further according to their modality and anatomical site.

By undertaking this research, Abi aims to improve the overall telemedicine experience, facilitating better communication between patients and doctors, and providing valuable data internal and external stakeholders.

Overall, this thesis is the first step towards unlocking the full potential of the visual information captured during Abi's telemedicine consultations to improve its healthcare outcomes.

Resumen

Una imagen vale mas que mil palabras. Esta expresión es válida también en el contexto de las consultas de telemedicina, donde las imágenes juegan un papel esencial en ayudar a los pacientes a explicar los motivos por los que buscan atención médica. Un procesamiento automático de estas imágenes podría enriquecer la elaboración de perfiles de consulta y mejorar la calidad general del servicio médico ofrecido. ¿Y si pudiéramos sacar provecho de esas esas mil palabras?

El objetivo de esta tesis es explorar y desarrollar soluciones avanzadas basadas en Inteligencia Artificial (IA) para extraer información médica de las imágenes que los pacientes comparten con los profesionales sanitarios durante sus consultas de salud telefónicas.

El proyecto se ha desarrollado con el Laboratorio de IA de Abi Global Health, una empresa de telemedicina con sede en Barcelona que actualmente opera a escala global. Abi ha recopilado una cantidad significativa de imágenes de sus pacientes que se pueden usar para entrenar algoritmos de aprendizaje automático con el objetivo de extraer información y complementar la descripción de la consulta médica.

Los diferentes algoritmos desarrollados durante este trabajo comprenden una *pipeline* de procesamiento de imágenes a gran escala. Comienza con un primer paso de clasificación de imágenes, cuyo resultado se utiliza para determinar el siguiente proceso que aplicar: *captioning*, extracción de texto o subclasificación.

En las imágenes que representan partes del cuerpo, se genera una breve descripción con métodos de *captioning*. El texto se extrae de imágenes de documentos o envases de medicamentos. Por último, las imágenes de estudios médicos se clasifican según su modalidad y sitio anatómico.

Emprendiendo esta investigación, Abi tiene como objetivo mejorar la experiencia general de la telemedicina, facilitando una mejor comunicación entre pacientes y médicos, y proporcionando datos valiosos a las partes interesadas internas y externas.

En general, esta tesis es el primer paso para hacer uso de todo el potencial capturado en la información visual de las consultas de telemedicina de Abi y mejorar sus resultados de atención médica.

Resum

Una imatge val més que mil paraules. Això s'aplica també al context de les consultes de telemedicina, on les imatges tenen un paper essencial en ajudar els pacients a explicar els seus motius de consulta. Un tractament automàtic d'aquestes imatges podria enriquir el perfil de la consulta i millorar la qualitat global del servei mèdic prestat. I si poguéssim fer-ne ús aquestes mil paraules?

L'objectiu d'aquesta tesi és explorar i desenvolupar solucions avançades basades en la intel·ligència artificial (IA) per extreure informació i coneixements mèdics de les imatges que els pacients comparteixen amb els professionals sanitaris durant les seves consultes de salut virtual.

El projecte s'ha desenvolupat amb el Laboratori d'IA d'Abi Global Health, una empresa de telemedicina amb seu a Barcelona que actualment opera a escala mundial. Abi ha recopilat una gran quantitat d'imatges dels seus pacients que es poden utilitzar per entrenar algorismes d'aprenentatge automàtic per extreure coneixements i complementar la descripció de la consulta mèdica.

Els diferents algorismes desenvolupats durant aquest treball comprenen un *pipeline* de processament d'imatges a gran escala. Comença amb primer pas de classificació d'imatges, el resultat del qual s'usa per determinar el procés següent que s'aplica, com ara: *captioning*, extracció de text o subclassificació.

Per les imatges que representen parts del cos, es genera una breu descripció amb mètodes de subtítols. El text s'extreu d'imatges de documents o envasos de medicaments. Finalment, les imatges dels estudis d'imatge mèdica se subclassifiquen segons la seva modalitat i lloc anatòmic.

Amb la realització d'aquesta investigació, Abi pretén millorar l'experiència global de la telemedicina, facilitant una millor comunicació entre pacients i metges i proporcionant dades valuoses a les parts interessades internes i externes.

En general, aquesta tesi és el primer pas per desbloquejar tot el potencial de la informació visual capturada durant les consultes de telemedicina d'Abi per millorar els seus resultats assistencials.

Contents

Chapter 1

Introduction

1.1 Context

This thesis is the result of a project proposed by Abi Global Health, a telemedicine company that aims to revolutionize the sector through innovative solutions and efficient services supported by Artificial Intelligence (AI).

The main objective of the present work was to extract information from the images provided by patients as attachments during a chat consultation. This information can be used to improve the performance of current AI applications within the company, such as identifying the type of consultation (e.g., prescription requirement, diagnosis, pharmacological advice, etc.), or determining the most appropriate specialist (e.g., dermatologist, general practitioner, gynecologist, etc). Moreover, it can open the door for new applications based on visual data data like captioning attachments to complement the medical case reported to the doctors or extracting the text from the images for anonymization or characterization purposes.

In this thesis, we explore multiple computer vision methods to achieve feasible pipelines that could be integrated into Abi's AI processes.

1.2 Solution setting

In order to establish an image insight extraction pipeline, the initial step involved understanding the problem at hand, which entailed determining the types of images that clients submit through the telemedicine platform.

For this task, visualizations and unsupervised learning techniques were combined. Unsupervised learning is a discipline of machine learning where algorithms

intend to discover patterns and structures within an unlabeled dataset without any guidance.

Once the data was explored, it was necessary to define the scheme for image classification and identify suitable processes to extract further information from each image type:

- From images of patients showing affected parts of their body and requesting a diagnosis, our goal would be to generate a brief description of the images. This task is known as captioning, and it can be solved using state-of-the-art AI models. Using this description we might be able to determine the doctor specialty suitable for each consultation.

- For medical images studies (e.g., X-rays, ultrasounds, computed tomographies, etc) a second-level classification can be implemented to determine the type of imaging we are dealing with. This information could then be attached to the images before being sent to the doctors.

- In the case of images containing text, the objective was to extract the text contained in them. To achieve this, an Optical Character Recognition (OCR) task was defined. This task consists of segmenting the part of the image that contains text and later applying text retrieval techniques using AI. This enables us to extract and process the text, making it more manageable for further analysis.

1.3 Approach

The image collection of the company did not have any metadata attached to it. So for the initial understanding and framing of the problem it was required to apply unsupervised learning and visualization techniques. To achieve this, the images were transformed to vector representations using a pretrained encoder (ResNet50 [6]). The multidimensional vectors were projected into a 2D space using PCA and K-means algorithm was applied to visualize each cluster separately.

For the initial first-level classifier, three different architectures (VGG16 [27], MobileNetV3 [7], and ResNet50 [6]) were fine-tuned and compared.

To generate a short description of the human pictures, we experimented with three different open-source image captioning models: GPT2 [23], BLIP [12] and

PromptCap [9]. We evaluated the performance of these models in this specific domain and compared different prompting strategies.

In the second-level classification of medical images, we trained a model using publicly available datasets to identify the modality and anatomical site. VGG16 was chosen as the architecture for this task due to its demonstrated high performance and low latency in the previous classification study.

Lastly, for the OCR task, we applied segmentation techniques to images with sparse text distribution or prominent noise. Two candidates considered for segmentation were CRAFT [1] and EAST [29]. After segmenting the images, we explored the Tesseract tool on printed text, and TROCR on handwritten text.

Figure 1.1 shows a schema of the full pipeline designed.

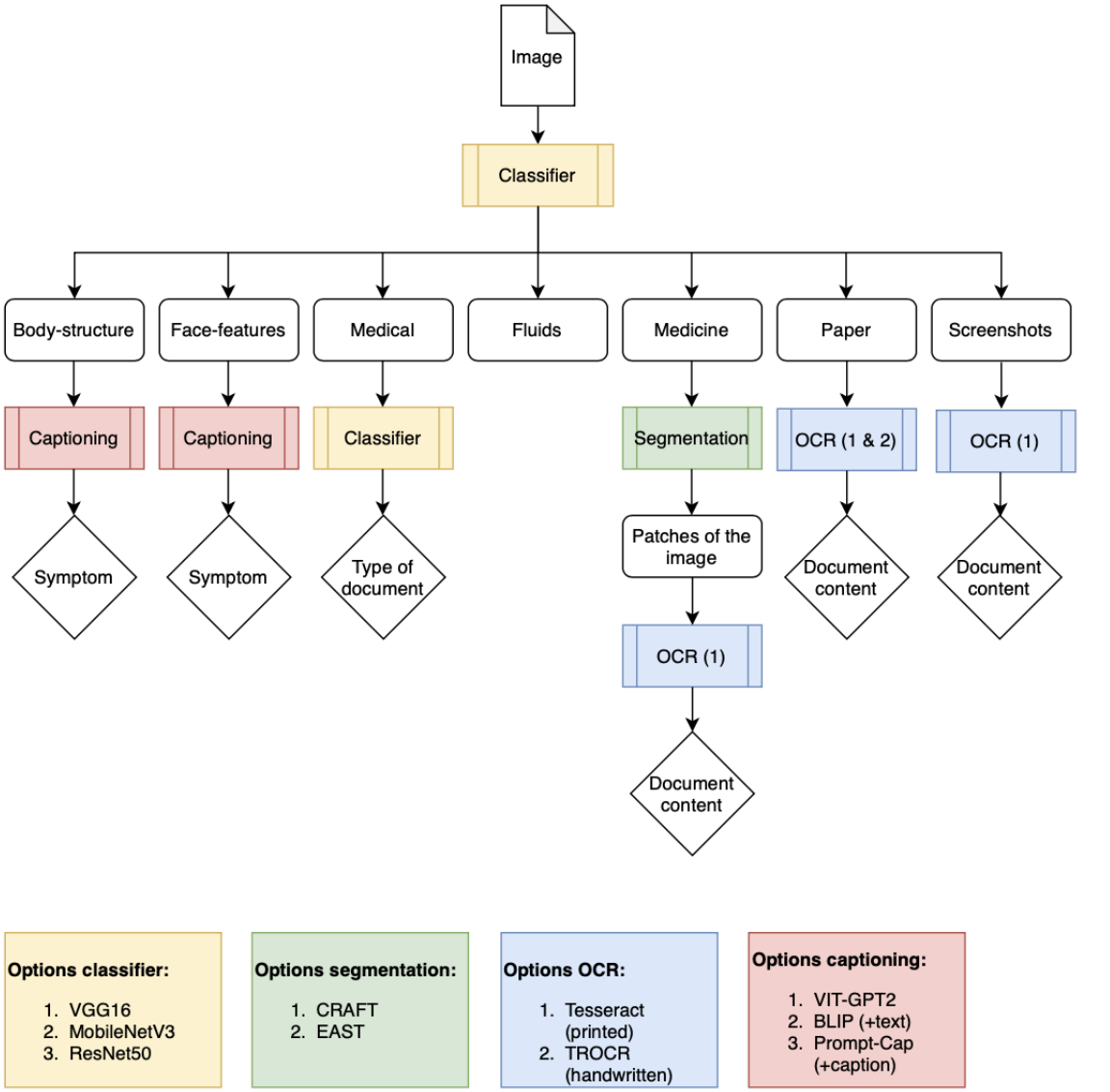


Figure 1.1: Schema of the image processing pipeline

Chapter 2

Objectives and specifications

As explained above, the main objective of this thesis is to create a functional prototype of an image processing pipeline for Abi's images. The first step towards this pipeline is to build a classifier suitable for the visual data. This model must be more fine-grained than the previous model used as baseline (which distinguishes between human, documents and medical). To develop this model, the following is required:

1. **Task understanding:** All the history of Abi's images should be revised and understood in order to determine the number of classes needed.

2. **Dataset creation:** A suitable dataset for this task should be created from scratch either by labeling Abi's images or by extracting data from public sources.

3. **Pretrained models:** Define a set of architectures to be trained.

4. **Fine-tuning parameters:** The pretrained models should be optimized to solve the task.

5. **Evaluation metrics:** The fine-tuned model should be carefully evaluated on validation and test set and insights of its accuracy and errors should be extracted. An understanding of what will happen if an error occurs while classifying is also necessary.

After this initial classification, further processing will be explored for each of the classified categories try to maximize the insights extracted from all of them (e.g. parsing text from documents and test results, classifying medical imaging files or removing personal information).

Chapter 3

First-level data annotation

The first step in our project has been to prepare our dataset for analysis and modelling. An initial set of 90.000 images were at our disposal, but without any accompanying information or labeling. Therefore, our first challenge was to categorize the images and add the metadata necessary to train the models and build the whole analysis pipeline.

In this chapter, we will discuss the practical aspects of data preparation and categorization. By establishing this groundwork, we lay the foundation for subsequent chapters where we delve deeper into our study and analyse the results obtained.

3.1 Quantity over quality

We followed a pseudo-labeling approach to speed up the process and minimize the manual tagging effort. For this we first encoded all images into a two-dimensional feature space and performed a high-level visual categorization of the projected data. Given the nature of the data, the images were mapped into clearly distinct clusters that facilitated the labeling upon minimal inspection.

The initial feature extraction was made by transferring the representation learning from a pre-trained computer vision model (ResNet50 [6], see Section 4.1). The resulting vectors (D=91) were reduced to two dimensions via Principal component analysis (PCA), as shown in Figure 3.1.

A preliminary analysis was then performed to understand the global and local structure of our dataset. To achieve this, a random partition of the entire collection was obtained and mapped in a 2D feature space. From this overview it became more evident the hierarchical distribution of the data, which defined our

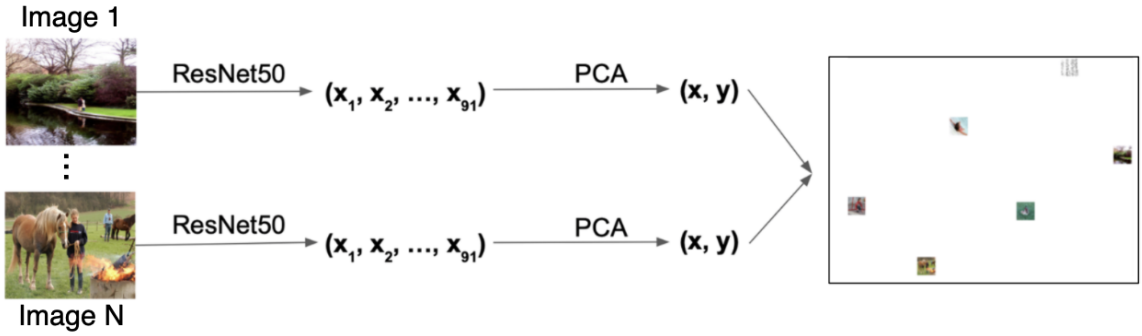


Figure 3.1: Schema of the image encoding applied in the first step of the pseudo-labeling pipeline.

final categorization schema and some of the algorithmic decisions described in the following chapters.

At high-level the images were arranged in three clear groups that we used as targets for the first labeling round:

- HUMAN: pictures of the patients themselves, usually to show an affected body part and ask for a diagnosis.
- DOCUMENTS: document scans, pictures of sheets of paper (handwritten or printed) and screenshots of text messages and websites.
- OTHER: everything that can not be classified as "documents" nor "human" (medication boxes or bottles, medical images studies such as X-rays, computed tomographies, among other pictures such as those of human fluids, drawings, etc).

3.1.1 Category labeling based on K-means

To carry out the labeling, we developed a custom user interface that allows to categorize all the images of a selected region of the embedded space (see Figure 3.3). The tagging was done in batches, to reduce the memory requirements of the serving instance in our cloud infrastructure.

To batch the image dataset into relevant groups for tagging, we divided the data in smaller analysis jobs of 8000 images each, and applied a k-means clustering (with k=8) on each one of them. Each cluster of images was then fed into the tagging UI as displayed in Figure 3.2 and classified into the three high-level categories. The clusters, of approximately 1,000 images each, were generally of high purity on a single class and could be easily selected and labeled in a single pass.

(a) Human **(b)** Documents **(c)** Other

Figure 3.2: Examples of clustered images for each group.

Smaller clusters were recategorized or tagged as OTHER to be cleaned afterwards if a considerably mixing was observed.

The process was then repeated on each pre-classified cluster, aimed at tagging subcluster regions and add extra metadata to be used in the prediction pipeline later on. For example, we identified subregions containing "faces" or "hands" within the HUMAN main cluster, and further discriminated between "paper documents" and "screenshots" from the generic DOCUMENTS category. Mixed clusters were curated and classified during this second round of annotations in similar way.

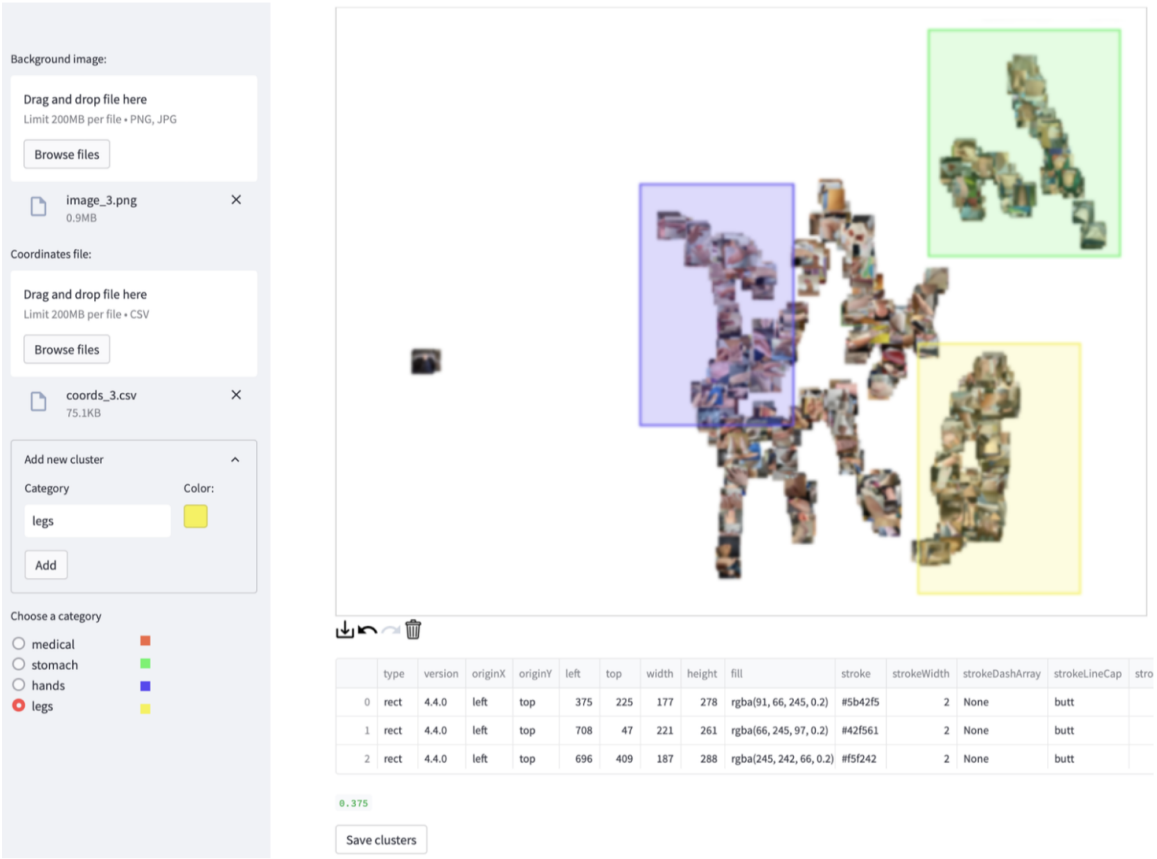


Figure 3.3: Custom annotation interface developed for image pseudo-labeling. The user can add new categories dynamically, select one for tagging and just select a rectangular region in the 2D projected space to propagate the labels.

During the annotation round 85,924 images were processed and pseudo-labeled. Out of these, 48,146 (56.03%) were assigned a subcategory as well. Moreover, 3,790 were classified as useless and spam and 3,260 needed to be checked and were not used (due to a mix of categories). Noisy images from internal tests and duplicates (i.e. with very similar encoding vectors) were removed (443 in total).

The resulting 78,431 images were considered useful and constitute the final dataset, which was further split into two disjoint sets for models training and validation as detailed in Table 3.1.

Set	Human	Documents	Other	Total
Train	46,986	15,642	12,028	74,656
Test	1,002	1,529	1,244	3,775

Table 3.1: Number of images for each category.

3.1.2 Test dataset curation

To guarantee the quality of the test set, an additional curation step was followed to detect and correct any misclassified image. A dedicated tool was developed for this purpose that allows a quick relabeling of individual images (Figure 3.4).

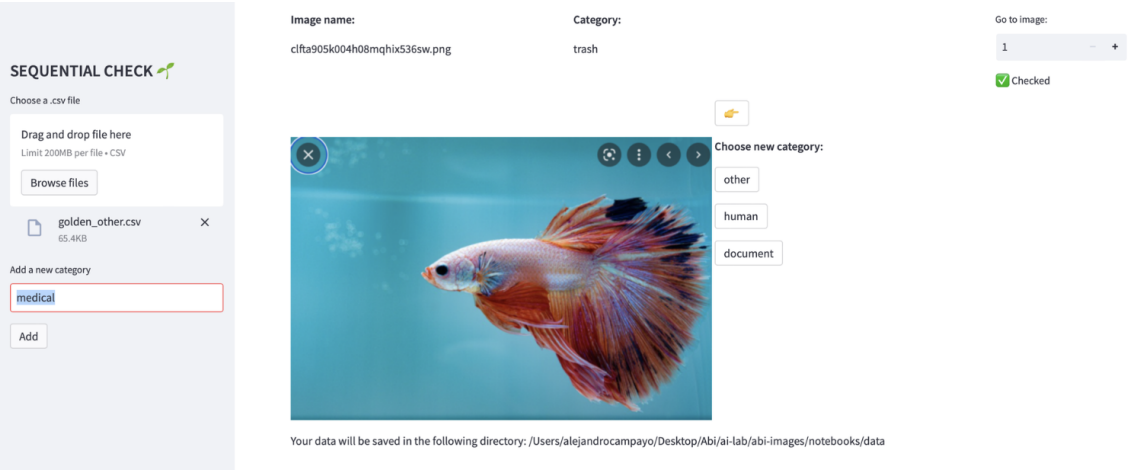


Figure 3.4: Tool UI used for checking and correcting images in the test partition.

The relabeling results are summarized in Table 3.2.

	Category after curation			
Category before curation	**Human** (%)	**Documents** (%)	**Other** (%)	**[Discarded]** (%)
Human	956 (95.41%)	0 (0%)	46 (4.59%)	0 (0%)
Documents	3 (0.20%)	1,437 (93.98%)	89 (5.82%)	0 (0%)
Others	299 (24.04%)	126 (10.13%)	770 (61.90%)	49 (3.94%)

Table 3.2: Test set relabeling results.

The best resolved category in the *test* set was HUMAN, where the first annotation round only had errors in images belonging to two settings simultaneously (e.g. a patient holding a prescribed medicine or a medical device). The misclassified DOCUMENTS were mostly images of pharmaceutical packaging or medical imaging files.

In the OTHER category, a large fraction of images were screenshots of screens showing DOCUMENTS or HUMAN. This common pattern resulted in their feature vectors being mapped in the vicinity of the OTHER clusters and were pseudo-labeled as such.

3.2 Quality over quantity

The relabeling of the test set showed that the high-level categorization might have been too generic, specially at the OTHER class, which contained many images belonging to HUMAN and DOCUMENTS classes.

To tackle these labeling errors observed during the first annotation pass (see Section 3.1.2) we decided to add granularity by breaking the main classes into smaller categories. During the curation of the *test* set it was realized that there were many images of medicine, medical images and human fluids, which could be easily differentiated and could constitute new categories. Additionally, a decision was made to create separate categories for screenshots and pictures of sheets of paper. This choice was based on the observation that these two types of images were adequately distinguishable in their 2D representation during the pseudo-labeling process. Following this same criteria, HUMAN category was further divided into two subcategories: one for images containing facial features and another for images depicting parts of the body.

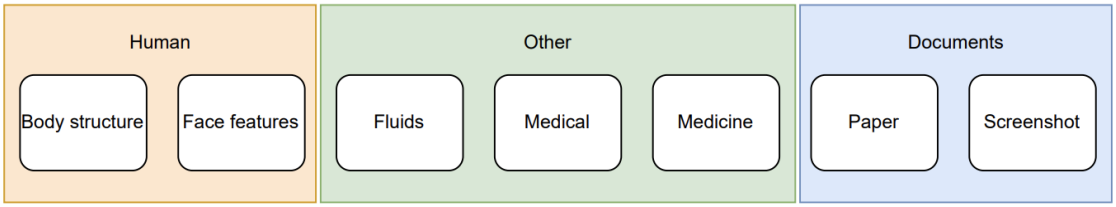


Figure 3.5: Hierarchical labeling schema.

Therefore, the new categories (Figure 3.5) are:

- FACE-FEATURES: pictures taken from above the neck or pictures where the whole person appeared and his face was recognizable.
- BODY-STRUCTURE: pictures of the rest of the human body.
- FLUIDS: blood, feces and other fluids (spit or sperm, etc).
- MEDICINE: pictures of pharmacological packaging (medicine boxes, blisters, etc) and medical devices (thermometers, blood pressure monitors, etc).
- MEDICAL: medical documents such as X-rays, MRIs, radiographies, etc.
- PAPER: photos or screenshots of paper (printed or handwritten documents), this includes sheets of paper containing electrocardiograms (ECGs).

- SCREENSHOT: digital documents and text messages, this includes screenshots of electrocardiograms (ECGs).

Examples of these categories are shown in Figure 3.6.

Figure 3.6: Examples of the 7 categories.

To assign this extra metadata, three different approaches were considered.

The first was to make use of the subcategories assigned using the tagging UI (see Section 3.1.1), matching them to the new categories. However, this method had two drawbacks. Firstly, we were not sure of the quality of these subcategories. Secondly, adopting this approach would result in certain images lacking representation, as it was not always feasible to differentiate them using the tagging UI.

The second approach was to use the coordinates of the clusters obtained in Section 3.1, choose the best pair of coordinates to represent each new category and classify the images that were closer to them as belonging to the respective category (i.e., label propagation).

The third was to generate descriptions of the images by using an AI technique called captioning. This would be done for each category (HUMAN, DOCUMENTS and OTHER) and the text generated would be used to filter and relabel (Figure 3.7).

Finally, the third option was chosen as it was deemed risky to use the coordinates obtained in the first processing stage, which had given an imprecise labeling.

3.2.1 Categories labeling with image captioning

The novel approach consisted on running the BLIP captioning model [12] (see Section 7.1.3) over all the images in the *training* set with 3 categories. The images were then filtered by keeping only those where the generated caption contained words that were representative of one the seven categories.

To gain a comprehensive understanding of the captions and select relevant words for filtering, the word count of these captions was examined. The chosen words were sufficiently frequent to encompass several tens or hundreds of images, while ensuring precision and avoiding the introduction of noise into the dataset.

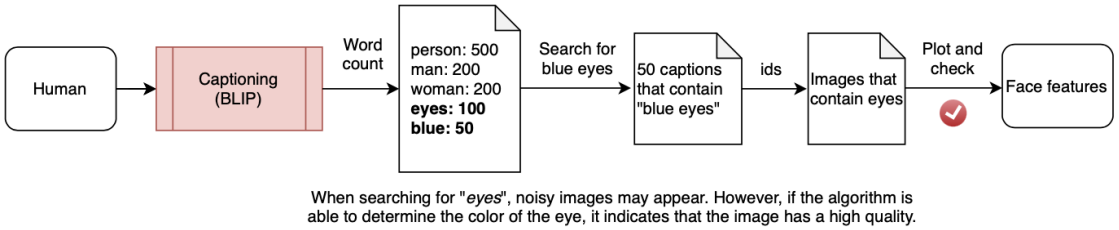


Figure 3.7: Schema of the method used for assigning precise labels.

For example, when looking for pictures of FACE FEATURES, searching for "person", "woman" or "man" was too generic, since many captions contained phrases such as "a persons leg". However, choosing "beard", "children" or combinations with adjectives such as "happy man" was more precise and yielded more accurate results. The reasoning behind this is simple: if the captioning model is able to determine not only the person but also its facial expression, it means that the face in the picture is clear.

In the implemented approach, it was possible to select specific words for searching as well as excluding certain words. For instance, if we wanted pictures of eyes but we notice that there are a few images of animals with detected eyes were present, we would exclude all the sentences containing words such as "cat", "dog", "fish" to avoid introducing noise into the dataset.

Moreover, the word searches were double checked by a visual inspection of all the selected images. Even though some noise was still present, most of the images in the dataset were accurately labeled.

3.2.2 Data addition

After the caption filtering, the resulting dataset contained roughly 10.000 images. In order to increase the size of our *training* set, 3 other sources were consulted.

First, 3260 images that were left aside for later checking in Section 3.1.1 were labeled.

Secondly, images from the *test* set were reassigned to the *training* set. This decision was made due to the recognition that a test set comprising nearly 4,000 images was unnecessarily large for a training set of that size. There was no need to relabel the *test* set images because the original subcategory metadata was precise enough to map it from the 3 original categories into the new seven categories.

Even after the previous steps, some concepts did not have enough representation in the new dataset so extra pictures were collected from public datasets. Images from Eye Dataset [26], Faces Dataset [5]and Hands dataset [11] were processed and added to their respective categories.

In the cases where no public dataset was available (such as images of human fluids), extra images were retrieved from Google public sources [20].

The resulting distribution of classes is summarized in Table 5.1. *Validation* is a subset containing 10% of the images in the full dataset.

	Data partition		
Category	**Train**	**Validation**	**Test**
Body structure	3,757 (25.18%)	326 (26.44%)	95 (13.57%)
Face features	2,653 (17.78%)	193 (15.65%)	106 (15.14%)
Fluids	963 (6.45%)	32 (2.60%)	100 (14.29%)
Medical	1,581 (10.60%)	109 (8.84%)	45 (6.43%)
Medicine	2,300 (15.42%)	202 (16.38%)	113 (16.14%)
Paper	2,015 (13.51%)	195 (15.82%)	140 (20.00%)
Screenshot	1,651 (11.07%)	176 (14.27%)	101 (14.43%)
Total	**14,920**	**1,233**	**700**

Table 3.3: Distribution of images for each category in the partitions of the dataset.

Chapter 4

First-level image classification

Once the dataset had been created (Section 3), a classification task could be defined. The model responsible for this categorization will be referred as ”first-level classifier”.

4.1 Models proposed

The first step towards finding an optimal classifier was to find the most suitable architecture. With this in mind, the three backbones proposed for this task were MobileNetV3 [7], ResNet50 [6] and VGG16 [27]. These backbones serve as the main component of the model and are responsible for extracting discriminative features from input images.

MobileNetV3 was considered because it is the latest version of the image classifier previously employed in the company, MobileNetV2 [25]. Research papers suggest that MobileNetV3 not only delivers improved performance but also has lower latency.

ResNet50, on the other hand, is a more powerful backbone with a larger number of parameters. It provides more potential for problem-solving but at higher risk of overfitting.

VGG16 is the simplest among the three architectures, offering the advantage of the lowest latency. Additionally, its simplicity makes it less susceptible to overfitting.

The final architecture connects each of these backbones to a classification head with two fully connected layers. These layers have 128 and 64 neurons, with a dropout of 20%, are are the only ones which are fine-tuned during the training loop.

Its activation layer is a softmax that will aim to classify into 7 categories: BODY STRUCTURE, FACE FEATURES, FLUIDS, MEDICAL, MEDICINE, PAPER and SCREENSHOT. The model is trained with categorical cross entropy loss and evaluates accuracy based on the category that achieves the highest score in the softmax activation layer.

To avoid overfitting, a data augmentation layer was added during training. The transformations applied consist of random shear, zoom, contrast, brightness and horizontal and vertical flips.

4.2 Models comparison

The three architectures were run for 20 epochs with a batch size of 32. Learning rate was reduced when no significant improvement in validation loss was observed in the last three epochs. Its training loss, training accuracy, validation loss, validation accuracy and training time were analyzed.

MobileNetV3 achieved the worst performance at the highest latency and therefore was discarded as a final model candidate.

ResNet50 and VGG16 achieved similar performance. VGG16 started from a better initialization and ResNet50 improved its accuracy considerably along the epochs. In the end, they both ended in similar situations in terms of accuracy and loss. ResNet50, however, kept improving its training loss while seemed to get stuck in validation accuracy and loss (showing signs of overfitting). Moreover, VGG16 latency was much lower than ResNet50, which was also determinant in the final choice of the model.

Models comparison and results such as training and validation curves for accuracy and loss can be observed in Appendix A.

In the end, VGG16 was chosen as our final model and was further trained up to 24 epochs until early stopped.

4.3 Environment

The training process was done in AWS Sagemaker, a cloud-based machine learning platform provided by Amazon Web Services (AWS). Images were downloaded from an S3 bucket to the disk of a Sagemaker notebook instance. As images are a heavy type of input, data generators were used to optimize memory use, by creating a flow of batches from the local directory.

Cost-effectiveness was evaluated for three different types of instances by running an epoch of 2100 training steps using a batch size of 32 images. The results are presented in Table 4.1.

Instance type pricing			
Instance name	**Price ($/h)**	**Time (time/epoch)**	**Price ($/epoch)**
ml.inf1.xlarge	0.254$/h	2h/epoch	0.51$/epoch
ml.m5.xlarge	0.257$/h	5h/epoch	1.285$/epoch
ml.g5.xlarge	**1.57$/h**	**12min/epoch**	**0.314$/epoch**

Table 4.1: Instance type pricing (considering the pricing of AWS instances in the region where the they are created).

The instance type selected for training was a *ml.g5.xlarge* with 5GB and an NVIDIA-A10G GPU. For non-training workloads (for example when loading or processing the datasets) the instance used was a *ml.t3.medium* (CPU-based) with a cost of 0.05$/h.

4.4 Results and discussion

VGG16 results on validation set were 88,72% of accuracy on validation and 82,84% accuracy on test.

As observed in Figure 4.1, the most common errors were between the BODY STRUCTURE and FACE FEATURES classes and the PAPER and SCREENSHOT classes. These mistakes can be tolerated as the confused classes will follow the same post-processing pipeline: image captioning on the first pair and Optical Character Recognition (OCR) on the second pair.

From that point of view, if we consider BODY STRUCTURE and FACE FEATURES as a single category in the pipeline, the identification accuracy of this new class translates into 99%. The same applies to PAPER and SCREENSHOT for which the combined prediction accuracy would be 96%.

The next most common mistake is misclassifying FLUIDS as BODY STRUCTURE (8%) or FACE FEATURES (4%). In terms of post-processing this would mean applying unnecessary captioning on pictures of human fluids.

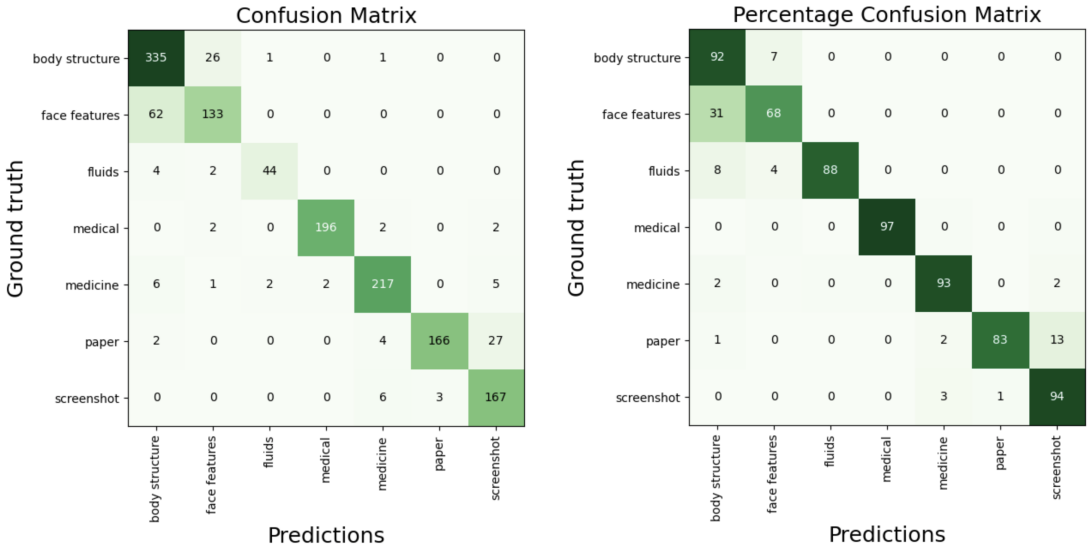


Figure 4.1: Confusion matrices on validation set

On the test set, results were similar than in validation with a 6% decay in total accuracy (Figure 4.2).

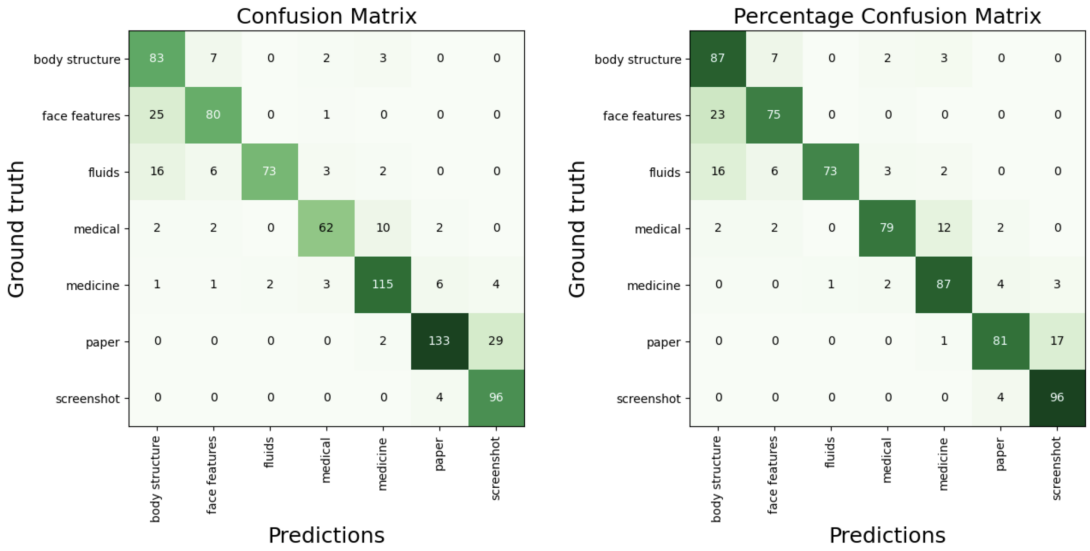


Figure 4.2: Confusion matrices on test set

We evaluated whether classification performance could be improved by introducing a decision-pipeline on the set of seven scores (i.e., the complete softmax output), instead of applying a simple maximum-likelihood criteria.

We evaluated whether classification performance could be improved with an alternative strategy for accuracy evaluation. Instead of relying on simple maximum-

likelihood criteria for classification we could exploit the complete softmax output. These scores can be interpreted as the confidence of the model for classifying an image as one of the seven classes. We first studied the distribution of confidences for each of the predicted categories, as presented in Figure 4.3. Each histogram shows the confidence values for true and false positives of each predicted category.

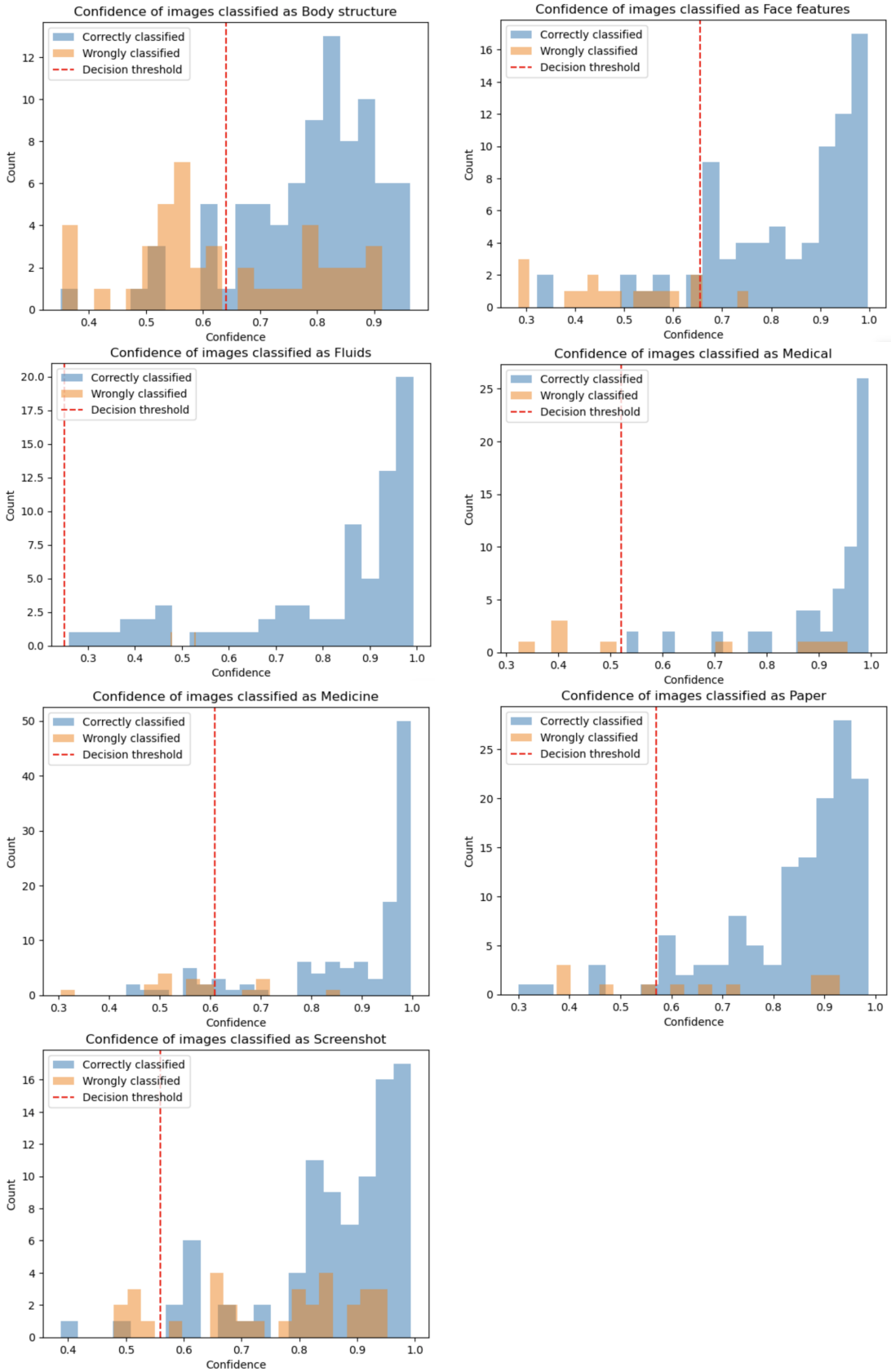


Figure 4.3: Confidence histograms for each category predicted and its decision threshold

We defined a set of minimum-score thresholds for the categories, that is used in an iterative way as follows:

1. Select the category with maximum score.
2. Evaluate if it is higher than the minimum-score threshold for the corresponding category.
3. If so, classify into this category. Otherwise, repeat steps 1, 2, and 3 with the next higher score.
4. If none of the category scores was higher than their respective minimum-thresholds, the image is categorized as UNKNOWN class.

In this way, the system only assigns a category if the model output is "confident enough" on the prediction. The thresholds that have been chosen are 0.64 for BODY STRUCTURE, 0.655 for FACE FEATURES, 0.25 for FLUIDS, 0.52 for MEDICAL, 0.61 for MEDICINE, 0.57 for PAPER and 0.56 for SCREENSHOT.

After the addition of these thresholds, the precision of the classification increased at the cost of sending 13.94% of the images to UNKNOWN category as shown in Figure 4.4.

With these thresholds, precision improves and the model obtains 90.25% accuracy on testing. As future work, performance might be improved by incorporating a top-level-category thresholding criteria: in images where no score was over its minimum-threshold, the sum of category scores belonging to a top-level family could be further compared against a threshold. For example, for those predictions where the sum of the FACE FEATURES and BODY STRUCTURE scores is over a particular "human" threshold - although a particular category cannot be assigned - the system could output a general label such as HUMAN class rather than an unknown label.

Same reasoning could be applied on PAPER and SCREENSHOT adding GENERIC DOCUMENT as a new category.

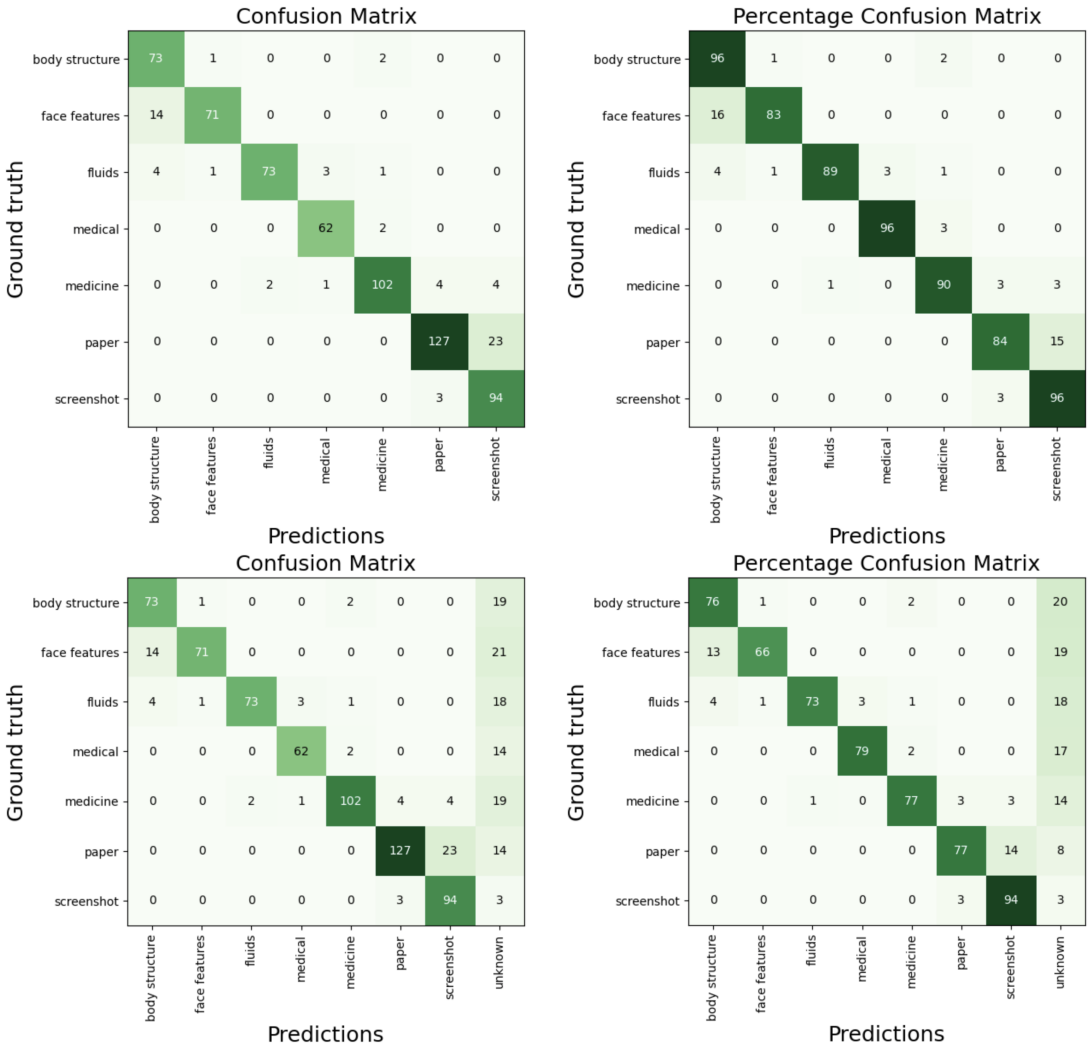


Figure 4.4: Confusion matrices on test set with decision-pipeline based on minimum-thresholds. Top row shows results without considering data sent to UNKNOWN. Bottom row shows an extra column explaining UNKNOWN category.

Chapter 5

Medical image classification

For Abi, classifying images under the category of MEDICAL provided crucial information, including identifying the reason for patient consultation and indicating to the doctors the specific medical image associated with each patient's case. Consequently, a secondary classification step was introduced specifically for the MEDICAL images.

The initial challenge involved determining which categories were relevant to identify within the MEDICAL images domain.

An exploratory visual inspection was performed using the same UI presented in Figure 3.4. After leveraging the results with medical experts from Abi, the following seven classes were defined:

- BRAIN: axial views of brain computed tomographies (CT) and brain magnetic resonance images (MRI).
- CHEST: frontal and lateral views of chest x-rays.
- LUNGS: axial views of lungs CT.
- TRAUMATOLOGY XRAYS: X-Rays of craniums, necks, arms, hands, hips, legs and feet.
- GASTRO: Upper gastrointestinal endoscopies.
- ECOGRAPHY: ultrasounds for fetus monitoring
- MAMMOGRAPHY: X-Ray pictures of the breast.

Examples of these categories are shown in Figure 5.1.

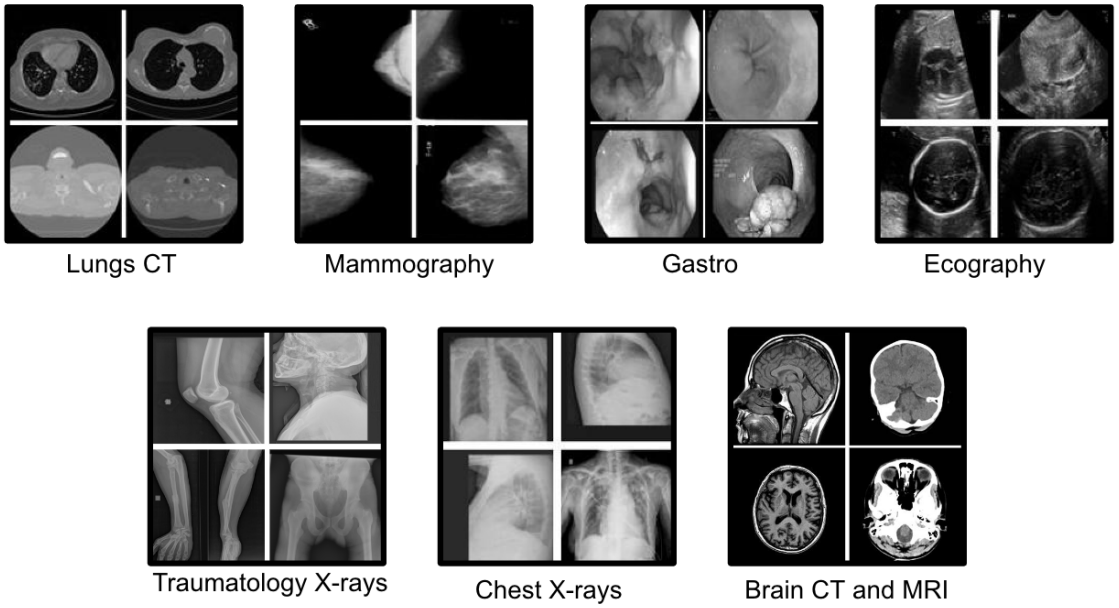


Figure 5.1: Examples of the 7 categories.

5.1 Dataset creation

The number of medical images available from Abi's databases were not enough for deep learning training. Therefore, they were used as testing set. To build a training set, we conduced an extensive search for publicly available datasets of medical images.

This dataset contains more metadata than needed for our task. This extra information has been kept for possible future improvements. In the dataset, we can differenciate between:

- BRAIN MRI, BRAIN CT within BRAIN category.
- CRANIUM, ARM, HANDS, VERTEBRAE, UPPER LEG (hips and thighs) and LOWER LEG (shanks, ankles and feet) within TRAUMATOLOGY XRAYS category.

Regarding the processing of the data, images were downloaded and transformed from medical formats (such as *.dcm* or *.tiff*) to *.png* or *.jpg*. Images were resized to 224x224 pixels and converted to grayscale.

The dataset has been uploaded to S3. Its specifications (sources, usage rights and data volume) are available in Table E.1 and E.2 of Appendix E.

5.2 Test dataset curation

In order to create our test dataset, images from the training set described in Table 3.2 belonging to MEDICAL category were manually labeled into the seven medical imaging categories, with the UI used in previous annotation tasks (Figure 3.4). Information regarding the part of the body visible in each TRAUMATOLOGY XRAYS image is included in the metadata. The number of examples in the testing set is shown in Table 5.1.

The three less common categories are MAMMOGRAPHY, GASTRO and LUNGS. This class imbalance should be considered in algorithm design and results analysis.

	Data partition		
Category	**Train**	**Validation**	**Test**
Brain	12,110 (19.73%)	467 (17.47%)	100 (12.64%)
Chest	11,630 (18.95%)	446 (16.69%)	137 (17.32%)
Ecography	12,089 (19.69%)	446 (16.69%)	293 (37.04%)
Gastro	8,547 (13.92%)	447 (16.72%)	6 (0.76%)
Lungs	8,834 (14.39%)	397 (14.85%)	40 (5.06%)
Mammography	1,743 (2.84%)	132 (4.94%)	4 (0.51%)
Traumatology X-ray	6,432 (10.48%)	338 (12.64%)	211 (26.68%)
Total	**61,385**	**2,673**	**791**

Table 5.1: Distribution of images for each category in the partitions of the dataset.

The main challenge for this classification task is that the domain of the training images differs notably from the final application domain: training data consists on original medical imaging studies, whereas Abi's data includes pictures of medical imaging studies taken by the patients themselves (ex: screenshots, photos of computer screens, etc). These pictures often exhibit issues such as light reflections, varied backgrounds, the presence of hands holding the paper, and fingers pointing at specific areas in the medical image.

5.3 Data augmentation

The primary challenge in achieving good performance on the noisy test dataset was to ensure the model's ability to generalize. Data augmentation played a crucial role in addressing this challenge.

We found that pictures from patients often consisted on a collage of various views of a medical study. To simulate this pattern in the training set, we augmented the training data by creating collages out of multiple single images. We applied varying numbers of columns and rows (ranging from 1 to 3 for each collage) and included random white padding between the images. The size of the white padding ranged from 1 to 5% of the image dimensions. The idea behind this custom data augmentation technique was to simulate the collages of medical images encountered in Abi images. The inclusion of white padding intended to replicate the effect of pictures being printed on a white sheet of paper. The schema used is shown in Figure 5.2.

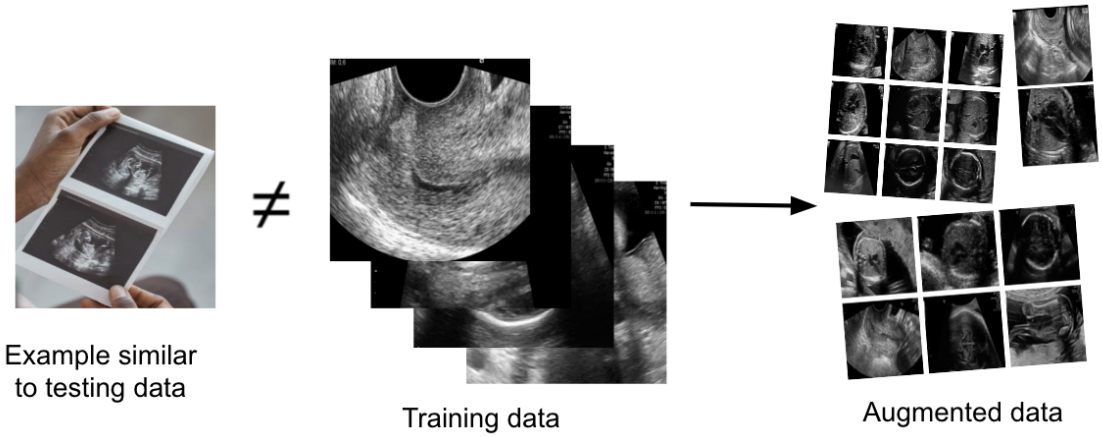


Figure 5.2: Schema of the custom data augmentation.

In addition, standard data augmentation techniques were employed, including horizontal and vertical flips, shear, rotation, zoom (in and out), contrast adjustment, brightness modification, and the addition of Gaussian noise.

This augmentation was applied to validation and training partitions.

5.4 Model training

The model consisted of a VGG16 backbone followed by two dense layers of size 128 and 64 initialized with random weights. The choice was made according to the results obtained in Section 4.1. Dropout layers with a dropout rate of 20% were added between the dense layers.

The model implementation involved an early stopper with a patience of 5 on the validation loss. Additionally, a learning rate reducer with a patience of 3 on the validation loss was employed.

Regarding hyperparameters, the initial learning rate was set to 0.001. Whenever the learning rate reducer was triggered, the learning rate was divided by 4.

For the training process, a two-step approach was adopted. Initially, the model backbone was frozen and kept unchanged until the early stopping criterion was met, which occurred after 14 epochs. Subsequently, the model backbone was unfrozen, allowing it to be fine-tuned for an additional 4 epochs. Towards the end of the training, data augmentation was gradually reduced, and the model underwent a final epoch of training.

5.5 Results and discussion

The model achieved an overall accuracy of 96% on the validation set. The confusion matrix is presented in Figure 5.3. The most common mistake in validation set was an 11% confusion of LUNG computed tomographies classified as BRAIN. Most of these mistakes were made when classifying LUNG collages, probably due to the both being axial views CT, the low quality of the image and the small size of the pictures conforming the collage.

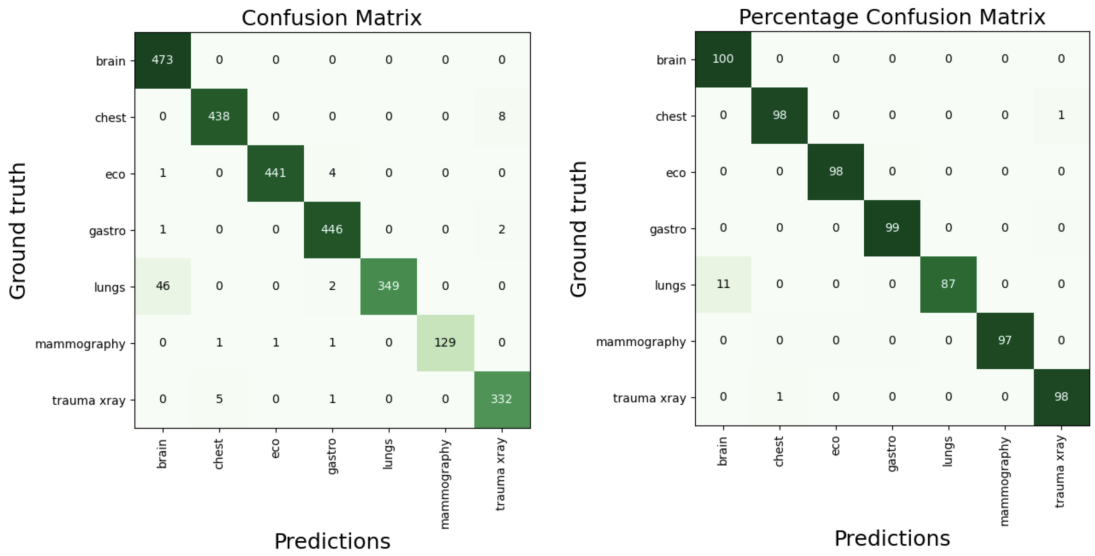


Figure 5.3: Confusion matrices on validation set

With regard to the testing set, the first results showed poor performance (Table 5.4).

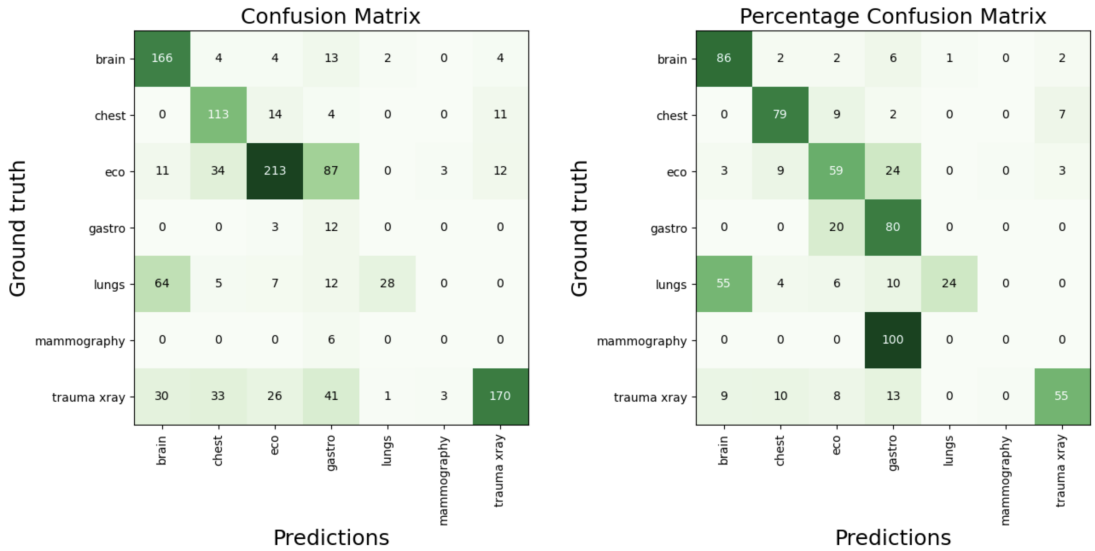


Figure 5.4: Confusion matrices on test set

To gain better understanding on misclassified images and enhance the model's precision (which takes priority over recall in this scenario), we repeated the analysis of confidence scores described in Section 4.4. We used the histograms presented in Figure 5.5 to assess differences in the confidence distribution of errors and correct classifications for each category scores.

The decision thresholds chosen for each category were: BRAIN: 0.93, CHEST: 0.8, ECOGRAPHY: 0.6, GASTRO: 0.99, LUNGS: 0.5, MAMMOGRAPHY: 0.3, TRAUMATOLOGY XRAY:0.3. The results after the addition of the decision thresholds are shown in Figure 5.6.

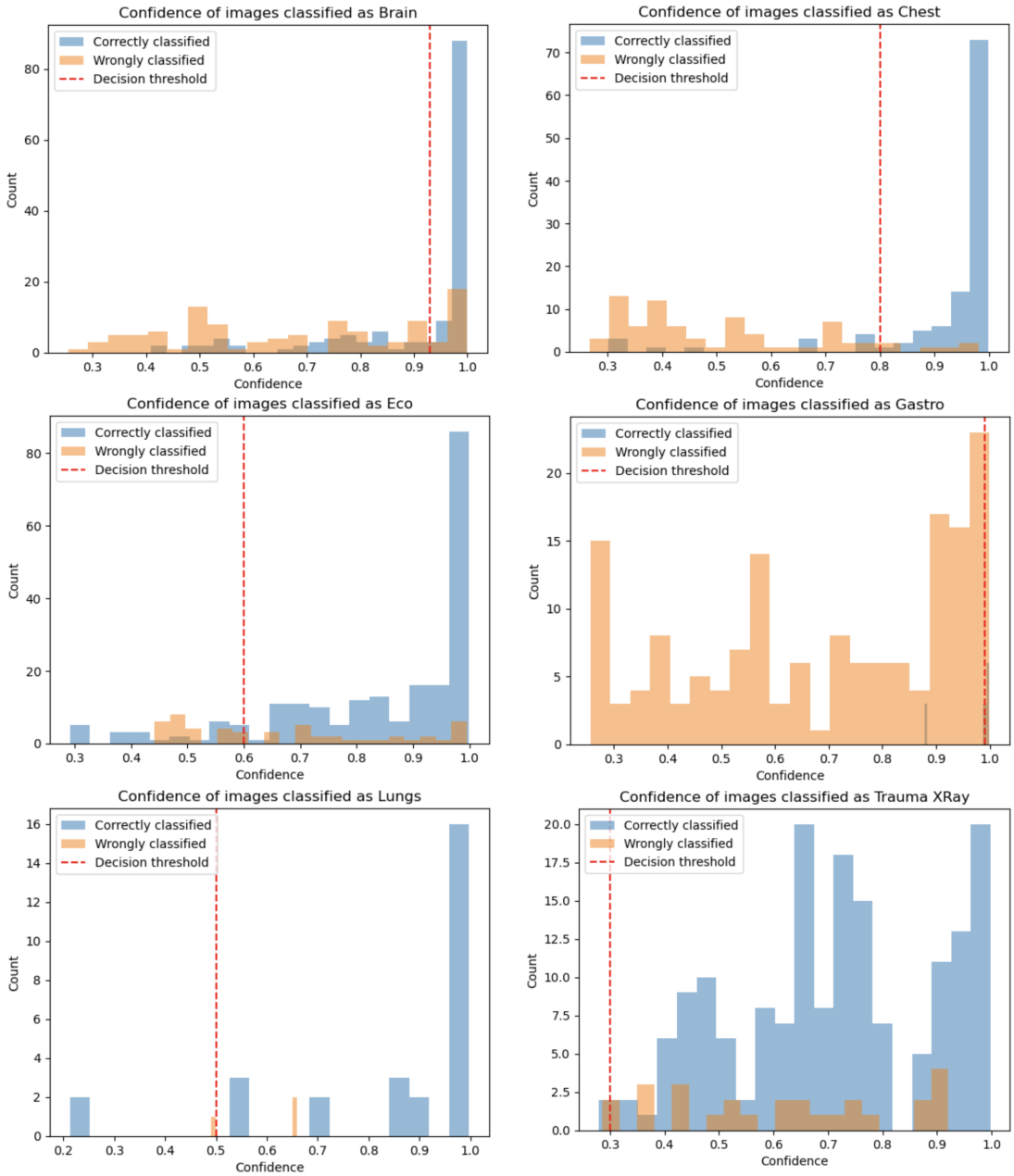


Figure 5.5: Confidence histograms for each category predicted and its decision threshold

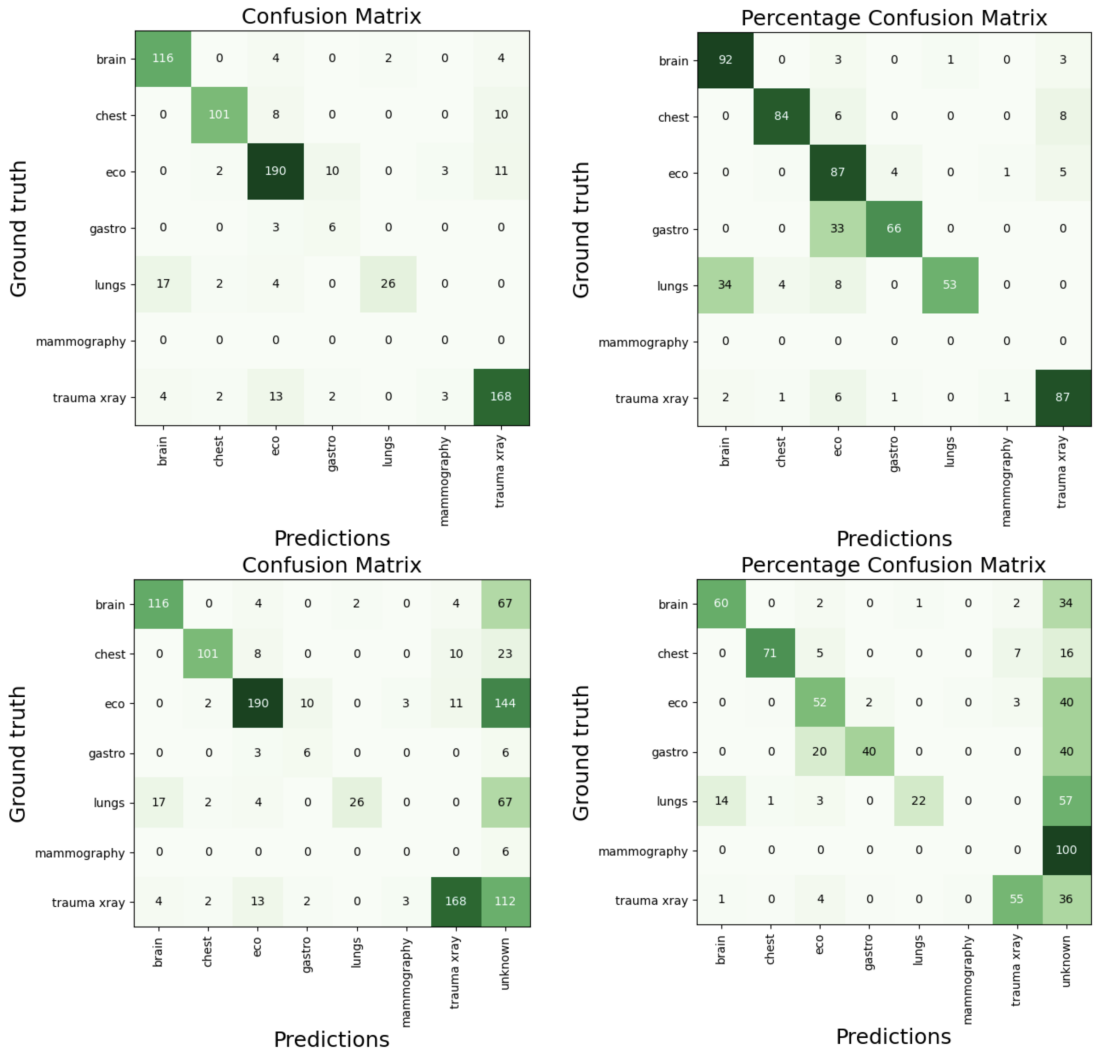


Figure 5.6: Confusion matrices on test set with decision-pipeline based on minimum-thresholds. Top row shows results without considering data sent to UNKNOWN. Bottom row shows an extra column explaining UNKNOWN category.

After the addition of thresholds, the precision of the classifier has improved considerably (achieving an overall accuracy of 86.1%) at the cost of not determining the class of 37.41% of the images. The results are reliable in BRAIN, CHEST, TRAUMATOLOGY XRAYS and ECOGRAPHY. The lowest performance was achieved in the other 3 categories:

- **MAMMOGRAPHY**: There was not enough representation of this category on the test dataset to determine whether the results are reliable.
- **GASTRO**: There was not enough representation of this category on the test

dataset to determine whether the results are reliable. Most of the false positives were actually bright ECOGRAPHY images (i.e., tissues with greater absorbing properties than liquid). This is probably because the ECOGRAPHY class in the training dataset was represented only with obstetric ultrasounds. Other types of ultrasound following other patters are incorrectly classified as GASTRO.

- **LUNGS**: Some pictures of LUNGS are incorrectly classified as BRAIN with a high confidence. Most of these images are lung MRIs, which have no representation during training due to the lack of public datasets containing this kind of images.

Regarding the TRAUMATOLOGY XRAYS class, 36% of images could not been classified. Part of these images were before wrongly assigned CHEST class. This happened when dealing with X-rays of necks or hips (which are also present in part of the images corresponding to CHEST during the training phase).

Overall, the results are promising and could be further improved if more data were available. For instance, augmenting the dataset with additional types of ultrasounds could be beneficial. Nevertheless, it is essential to have the supervision of the medical team in this process, as different types of ultrasounds may be utilized in various medical specialties. Consequently, if the classifier's objective is to determine the suitable healthcare professional for patient consultations, it is not advisable to include such diverse ultrasound types.

Chapter 6

Optical Character Recognition

Abi users send pictures of a wide variety of documents. Until now, no information was being extracted from these images. However, being able to acquire text from these sources would open a new set of opportunities for the company. Some examples of these would be improving user-doctor matching, ensuring client privacy and security (by deleting or blurring sensitive data such as names or credit card information), extracting insights for client reports, among others.

Three of the categories identified in Chapter 3.1.1 were suitable for text retrieval. Nevertheless, these images are quite different and require customized processing. Consequently, in this chapter, we will explore and evaluate various approaches in search of the most effective methods for each scenario. Specifically, we will investigate text segmentation techniques tailored to Abi images, along with two distinct text retrieval tools: Tesseract for printed text and screenshots, and TROCR for handwritten documents. Through these investigations, we aim to identify and establish optimal methodologies for text retrieval in the respective scenarios outlined.

6.1 Tesseract: text retrieval from printed text

Tesseract is the most popular open-source OCR tool. Even though its results are far from being perfect in non optimal images, there are many ways to improve its accuracy. Understading this tool and adapting its configuration according to the format of the data that needs to be extracted [24] is the first step towards better results. Tesseract has 3 parameters that can be specified.

The first one is the language we want to detect. This tool supports over one hundred languages and gives the option of selecting one or more alphabets to be detected. Default language is English. With regard to patients consultations,

language could be extracted from both the text messages in the consultation where the picture of the document was sent or from the users' profile information.

The second parameter is "PSM" referring to "page segmentation mode". Tesseract offers 13 different segmentation modes, however, we will focus on 3:

- **0:** Orientation and script detection (OSD) only. For rotating images when its orientation is not the usual.
- **4:** Treat the image as a single word. Which will be used later on segmented images (Section 6.2).
- **6:** Assume a single uniform block of text. For general text retrieval of documents.

The last configuration is "OEM" or "optical engine mode". Which offers three options:

- **0:** Legacy engine only.
- **1:** Neural nets LSTM engine only.
- **2:** Legacy + LSTM engines.
- **3:** Default, based on what is available.

"OEM" parameter was set to default in our study.

We studied performance on examples similar to Abi's images. Some examples of it are available in Appendix B in Table B.1. As it can be seen in the examples, Tesseract with PSM=6 and with language specification of the text achieves almost perfect performance on SCREENSHOT. For paper documents, the model yields a good performance even when the images are noisy. However, a post-processing step on the extracted text would be useful before using it in downstream tasks such as Named Entity Recognition.

Regarding text orientation, Tesseract is able to read flipped documents as well.

6.2 Text segmentation on medicine

For images in the MEDICINE category, we are specially interested in extracting the name of the medication present in the picture. However, the images of medication sent by patients are often not focused on the text labels of the medication package.

These pictures can be unclear and contain other elements in the background. To perform OCR, we should first segment the image region containing text.

To address this problem, two of the most popular models [22] proposed for this task are CRAFT [1] and EAST [29]. These two tools have been applied to twelve examples, representing four different levels of difficulty within the MEDICINE category, as part of a qualitative study.

The first set of images consists of clear pictures of medicine boxes. The second set contains noisier images where hands and other objects may appear. The third set showcases medical tools, such as blood pressure monitors, with the characteristic of displaying numbers against a dark background, as opposed to the bright background seen in medicine boxes. The fourth level encompasses images where no text is present, such as pills. This set aims to provide insights into the occurrence of false positives during text segmentation on textless images.

6.2.1 CRAFT

CRAFT [1] backbone is a VGG16 with batch normalization which relies solely on a concatenation of convolutional layers with different scopes (initial layers detect smaller letters, while later ones have a broader perspective of the image).

It is a character-level text detector which means that the model detects each character one by one and then an affinity is calculated between adjacent characters. Characters are joined and treated as a whole word when affinity is higher than a certain threshold. This design enhances the model's ability to handle curved text but limits its effectiveness with cursive letters.

In terms of its output, CRAFT provides two scores: the region score, which indicates the likelihood of the detected text being correct, and the affinity score, which indicates the probability of merging with the adjacent character.

Insights on the model are explained in Table 6.1.

Theoretical strengths	
Insight	**Motive**
Accepts curved text.	It can recognize curved text by detecting arbitrarily oriented characters and joining them if they are in close proximity, regardless of the characters' orientations (e.g., characters forming a semicircular word).
Resistant to different text sizes.	While the last convolutional layers have a wide scope, they may not cover the entire picture. However, individual characters can still be identified within the picture because these layers are specifically designed to detect one character at a time.
Good performance on Eastern Asian languages.	In Eastern Asian languages, characters are equidistant. Therefore, it is easier for the character segmentation to learn their patterns.
Theoretical limitations	
Insight	**Motive**
Struggles to detect cursive languages.	The symbols are written in a conjoined and/or flowing manner. For this reason, there is an extra difficulty in character segmentation step on these languages (e.g., Bengalí and Arabic)
Sentences are split in blocks of 2-3 words.	Affinity threshold defined to merge the characters is small enough to merge the characters of a single word but is hesitant about merging to consequent words.
The word order of the outputs is often flipped.	Segments are ordered them from top to bottom. Therefore, words are flipped when a word is written after (in the right of another word) but is located a bit higher up than the word on its left.
Observed limitations	
Insight	**Comments, motive and possible improvements**
Current approach only detects straight text	With the current approach, curved texts cannot be detected. To accomplish this, the 'poly' configuration should be chosen in CRAFT. However, adopting this method results in worse performance on straight text due to unnecessary distortions. Therefore, 'box' was selected over 'poly' at the cost of losing the ability to read curved texts.
Poor accuracy on medical devices	Probably due to the dark background. A custom image processing could be performed for these images but for that a previous classification should be made.

Table 6.1: Insights on CRAFT segmentation.

6.2.2 EAST

EAST [29] pipeline consists of a CNN backbone adapted for text detection, which is concatenated with a Locality-Aware Non-Maximal Suppression. The CNN backbone outputs multiple rotated rectangles that serve as candidates for text, while Non-Maximum Suppression determines the final results by utilizing Intersection Over Union.

The paper explores three different convolutional architectures: VGG16, PVANET, and PVANET2x backbone. Among these three, VGG16 achieves the worst results due to its smaller receptive field. PVANET is characterized by its small and fast nature, while PVANET2x achieves the best results despite being the slowest of them all.

Insights on the model are explained in Table 6.2.

Theoretical strengths	
Insight	**Motive**
Low latency	Because of the pipeline being straight forward, the latency for the heaviest model is sufficiently low to enable the architecture's applicability to videos (frame rate of 13.2 FPS).
No hyper-parameter optimization	Due to its simple and straight forward implementation.
Theoretical limitations	
Insight	**Motive**
Not suitable for detecting curved texts	Its prediction boxes are formulated as rectangles.
Low performance on chunks of text	The maximum size of the text that can be detected is constrained by the receptive field of the network. As a result, the text detector may fail to capture long lines of text entirely or may incorrectly segment very large words, leading to imprecise predictions.
Observed limitations	
Insight	**Comments, motive and possible improvements**
Angle of the crops is sometimes wrongly estimated.	When many different orientations are present, the angle of the predictions is not accurate. This probably happens due to a lack of representation of these kind of images in its training dataset. The only option to solve this would be to fine-tune the model.
Segmentation is done word-level.	Breaks texts in words rather than doing it in rows of text. This would mean many more 'pytesseract' executions. The only option to solve this would be to fine-tune the model.
Poor accuracy on medical devices.	Probably due to the dark background. A custom image processing could be performed for these images but for that a previous classification should be made.

Table 6.2: Insights on EAST segmentation.

6.2.3 CRAFT vs EAST

CRAFT outperformed EAST in both easy and difficult images by providing more compact segments and better accuracy in determining image angles, widths, and heights. Furthermore, in images without any text, EAST had more false positives compared to CRAFT. In the context of medical devices, EAST and CRAFT performed similarly, except for one of the three images where EAST clearly outperformed CRAFT. Results can be observed in Table C.1 of Appendix C. The performance of the segmentators on Abi images was also studied, but the results on those images can not be shown in the thesis due to privacy restrictions.

Considering the overall accuracy, CRAFT has proven to be more reliable. Hence, we have decided to utilize the CRAFT image segmentator for processing the images before feeding them into the Tesseract engine.

6.2.4 Segmentation (CRAFT) + OCR (Tesseract)

A qualitative study was conducted to assess the performance improvement achieved by utilizing CRAFT before the Tesseract engine for images of the MEDICINE class. For this comparison, the most suitable configuration of Tesseract was chosen in each case (PSM=8 on CRAFT+Tesseract and PSM=6 on raw Tesseract). Results can be found in Tables C.2, C.3, C.4 and C.5 of Appendix C.

Overall, CRAFT+Tesseract achieves higher accuracy and returns texts with less noise. However, it still has some limitations as explained above. Additionally, in terms of latency, CRAFT is slower due to the fragmentation of the pipeline. When using CRAFT segmentation, additional processing time is incurred due to the forward pass for text detection and the need to process Tesseract for each image crop.

Even though segmentation improves the accuracy of Tesseract engine, a few limitations have been observed:

- The model exhibits poor performance on small image crops, although these crops typically do not contain essential information. Two options can be considered to address this issue. First, increasing the threshold would retain only crops with higher certainty. Second, a size threshold can be defined, and crops smaller than the threshold can be excluded from processing.

- CRAFT generates vertical crops for text instances with an inclination over 45 degrees, leading to reduced accuracy when processed with Tesseract using

PSM8, which expects horizontal text. To overcome this, two potential options are available. First, Tesseract can be passed with a configuration that detects the text orientation, rotates it accordingly, and then applies PSM8. Second, vertical rectangle crops can be identified and omitted from further processing.

- Images with text affected by noise, such as light reflections, result in diminished performance.
- Sometimes the text extracted contains noise even though it is still understandable. Text processing might be necessary after text extraction.

6.3 TROCR: text retrieval from handwritten documents

TROCR is a model owned by Miscrosoft [15]. It is designed to solve the task of OCR on handwritten documents. Regarding its architecture, TROCR is an encoder-decoder model. It consists of an image Transformer as the encoder, and a text Transformer as the decoder. The image encoder's weights were initialized from the weights of BEiT, while the text decoder's weights were initialized using RoBERTa.

To load the TROCT model, we referred to the instructions provided by HuggingFace [16]. It is worth noting that there are three versions of the model that vary in size. The version that have been used for the studies is the *base* version since reviews indicate that the *small* version has a significant decay in performance and the *large* version implies an increase in execution time.

Its performance on images from Abi's consultations has been studied. Some examples of it are available in Appendix B in Table B.2. Its main limitation is the fact that the model has been trained to read one row of text at a time. Therefore, a segmentation pre-possessing step is necessary. The segmentator model chosen for the task has been CRAFT [1] according to the study in Section 6.2. The model latency is high when predicting without a GPU and the segmentation step makes the process slower due to having to forward many patches of the image.

Regarding its performance, the model has shown good performance on extracting the text from the images when passing as input the patches of image extracted using CRAFT. One drawback of this method is that CRAFT tends to break sentences in blocks of two or three words, which makes the process slower. Moreover, the order of the output is switched quite often since CRAFT orders the output from top to bottom instead of doing it from left to right.

The main strength of TROCR is that it can interpret printed text as well. This would be useful when facing images containing both printed and hand written text. Nevertheless, when it comes to printed text, Tesseract achieves better accuracy. Moreover, images containing both printed and handwritten text are often medical prescriptions. In these cases, accuracy is often lower due to the poor writing style of some health professionals. In some cases, the patient consults Abi specifically to ask what is written in a prescription. Considering these cases are challenging for the average human reader, it is currently unfeasible to obtain a good performance with an automated system.

Overall, adding a layer to try to extract text from handwritten documents is not recommended for the company because of the following reasons:

1. One extra classifier step should be added to distinguish between printed documents and handwritten documents. Currently, Abi has no available labeled dataset for this task. Even though there are public databases that could be applied, a large percentage of these images contain both printed and handwritten text and would probably not achieve a high certainty in any of the classes.

2. The percentage of handwritten images in Abi's database is low compared to printed text or screenshots.

3. Handwritten documents where the writing style is difficult to read represent a visual task too complex for current deep learning models.

4. TROCR is a model that would need a GPU available if the tool was to be used for real-time applications.

Chapter 7

Image captioning for human pictures

In the following section, three captioning models are compared in their application to Abi's images from the HUMAN category (i.e., FACE FEATURES and BODY STRUCTURE combined). The models that were chosen for the study are ViT-GPT2 [23], PromptCap [9] and BLIP [12]. The reason behind this selection is that these are the most popular open-source models for image-captioning task.

7.1 Models proposed

In the following sections, we explain the architectures of the chosen models, as well as their strengths and limitations within the medical industry, including deployment costs.

7.1.1 ViT-GPT2

ViT-GPT2 is a model introduced in the work by Ziyang Luo et al. [23]. It combines a pre-trained visual encoder, CLIP-ViT, with GPT2 as its language decoder. Unlike previous approaches that relied on object detection, the visual encoder in ViT-GPT2 directly processes image patches, making it a novel and efficient architecture.

The model was fine-tuned on the COCO dataset, which comprises 200,000 images. The COCO dataset includes images of humans and provides annotations for body parts. However, it is important to note that the dataset has undergone filtering, and as a result, images containing blood or genitalia have been excluded. Therefore, the ViT-GPT2 model may not be able to accurately identify such images. Additionally, it has been observed that the COCO dataset exhibits a racial bias,

with lighter-skinned individuals achieving better accuracy in the model's predictions.

To load the ViT-GPT2 model, we followed the instructions provided by HuggingFace [28].

7.1.2 PromptCap

PromptCap is a model introduced in the work by Yushi Hu et al. [9]. It is designed to generate specific captions that are later processed by a language model (such as ChatGPT) to solve visual question answering tasks (VQA). The model's objective is to generate a prompt that describes the image while focusing on providing the necessary details for the language model to answer a specific question. The architecture of PromptCap is obtained by fine-tuning the state-of-the-art vision-language model OFA, which employs an encoder-decoder structure.

The training data for PromptCap is obtained through a modification process applied to existing VQA datasets using ChatGPT. This process involves synthesizing question-aware captions by combining general image captions with the question and answer using ChatGPT. The model is trained to generate these new artificial captions as its output. The datasets used for this process include OK-VQA, A-OKVQA, and WebQA. Furthermore, the PromptCap model incorporates the knowledge acquired prior to fine-tuning OFA.

To load the PromptCap model, we followed the instructions provided by HuggingFace [10].

7.1.3 BLIP

BLIP is a model introduced in the work by Junnan Li Dr et al. [12]. It utilizes a visual transformer for image encoding. Images are first divided into patches, and each patch is encoded into embeddings. Additionally, BLIP generates a global image embedding that represents the entire image.

The model adopts a multimodal mixture of encoder-decoder (MED) architecture. This means that the model is multi-task and can operate in one of the three different functionalities:

- **Unimodal encoder:** In this mode, image and text are encoded separately using an Image-Text Contrastive Loss. The goal is to encourage similar representations between image and text pairs.

- **Image-grounded text encoder:** Given an image-text pair, the model encodes both inputs into a single representation.
- **Image-grounded text decoder:** This mode takes the representation generated by the image-grounded text encoder and generates a human-readable sentence or caption.

BLIP has demonstrated effectiveness in various tasks, including Image-Text Retrieval, Image Captioning, Visual Question Answering, Natural Language Visual Reasoning, Visual Dialog, and Zero-shot Transfer to Video-Language Tasks.

For training, BLIP utilizes a dataset consisting of 14 million images, the same dataset used in the paper *Align Before Fuse* [14]. The dataset includes human-annotated datasets as well as web datasets. Additionally, the LAION web dataset, which contains 115 million more noisy texts, is also employed. Notably, the LAION dataset contains medical images and NSFW images, which could provide the model with an understanding of genitalia in captioning. However, the NSFW data comprises less than 1% of the dataset and can be filtered out. It should be noted that it is unclear whether the NSFW data was filtered out during training.

With regard to its training process, BLIP includes the innovative use of Captioning and Filtering (CaptFilt) to enhance the quality of low-quality data collected from the web. CaptFilt combines a pre-trained captioner and a filter (both of them pretrained). Given an image and a caption (either human-annotated or extracted from the web), the captioner generates a synthetic text describing the image. The filter then cleans both the original and synthetic captions and combines them. This technique has proven to improve the quality of both low-quality and human-annotated datasets.

To load or fine-tune the BLIP model, we followed the instructions provided by HuggingFace [13].

7.2 Qualitative study

We applied these models to images from the testing set described in Section 3.1.2, belonging to the HUMAN category. The images were classified into the following subcategories: human, hand, skin, feet, arm, leg, male genitalia, torso, face, neck, open mouth, head, mouth, back, female genitalia, eye, full person, rectal opening, ear, and nose.

The models were run on 20 examples for each subcategory, except for the following subcategories with fewer classified images: full person (15), rectal opening (15), ear (11), and nose (7).

In total the evaluation dataset contains 348 images. The 3 models were run on these images. For the models that accepted a text input, each image was run twice with different prompts. For models that accept a text input, we studied the effect of applying different prompt formats. We applied both a simple prompt (common to all images) and an elaborate prompt (which includes the body part corresponding to each image subcategory).

The idea behind this approach is to assess whether the output caption is more precise when the model has given information on the body part of interest. The logic is that if the part of the body has already been told to the model, maybe it will better focus on the symptoms located there.

The difference observed between the captions generated using the generic and elaborate prompts proved valuable in assessing the feasibility of implementing a classifier.

The prompts used for the study were:

- **BLIP simple**: *No text*
- **BLIP elaborate:** "A picture of a human (*part of the body*)"
- **PromptCap simple:** "Describe this image according to the question: what is the symptom of the patient in the picture?"
- **PromptCap elaborate:** "Describe this image according to the question: what is the symptom in the (*part of the body*) in the picture?"

These prompts were chosen after having tried many other and according to what the literature recommends for each model.

7.3 Results and discussion

Examples of the qualitative study performed to evaluate the feasibility of adding captioning to the pipeline can be found on Appendix D. Captions generated by GPT2 are not shown due to its poor performance for this task.

Overall, BLIP and PromptCap have very similar performance. They are both good enough to be considered for applications such as determining the part of the

body present in the image or the specialty of the healthcare professional required. However, particularly for the genitalia subcategory, BLIP shows poor performance while PromptCap is able to distinguish rectal opening and female genitalia but not male genitalia.

Finally, we studied whether the caption results could be use to determine symptoms from the image. We found that the models might overdetect symptoms that are not present in the picture such as black eyes or toothbrushes inside the mouth due to bias in these images. Moreover, it might hallucinate in some low quality cases returning a text unrelated to the image. For this it is important to process the text and keep only that information that it is related to the task desired. For this, Abi's Mermaid model could be used for named entity detection as shown in Figure 7.1.

Figure 7.1: Caption processed using Mermaid

These models would not be suitable for symptom detection but rather for body part classification.

With regard to the difference in performance when executing with different prompts, there is no significant improvement, at least with the prompts that have been tried until the moment.

Regarding prompts, Promptcap is specially interesting due to its application on VQA task. For example, if the model was meant to determine the specialty of the doctor required to solve the consultation, studies could be made with prompts such as "Describe this image according to the question: what doctor specialty should this patient visit?". If the objective of the model on *Face Features* was to annonymize the picture before sending it to the doctor, the prompt could be: "Describe this image according to the question: Is the face in the image recognizable?" or "Describe this image according to the question: What part of the face could be blurred while the doctor still being able to make a diagnosis?". A straight forward model for VQA would also be suitable for these tasks.

Chapter 8

Results

In this thesis, we have developed a multi-level pipeline to analyze and extract insights from the full image collection collected by Abi Global Health from the medical consultations of their users.

The first main milestone was to build a generic data processing and labelling workflow to annotate and enrich the metadata of the images dataset to train and optimize a classification model.

The final classifier is able to distinguish among seven categories: BODY STRUCTURE, FACE FEATURES, FLUIDS, MEDICAL, MEDICINE, PAPER, and SCREENSHOT. Among the several architectures explored, VGG16 demonstrated the best results with the lowest latency and therefore chosen as our final model. The resulting classifier achieved a global accuracy of 90.25%.

As a second processing stage for SCREENSHOT, PAPER and MEDICINE attachments, OCR was performed to extract the text contained in the images. Tesseract proved to be effective for extracting text from printed documents in general, but the quality of the result varied across different classes. While text extraction from SCREENSHOT require no additional post-processing, PAPER images require some cleaning before applying techniques such as Named Entity Recognition. In the case of MEDICINE images, information such as medication names and dosages could be also extracted using a variant of Tesseract running on segmented patches of the image. This segmentation is done by CRAFT model (which outperformed EAST). For handwritten text TROCR yield reasonable results using the same segmentation approach, but it is disregarded for the moment due to the high running cost and low benefit considering the small number of such instances in real user consultations.

A second classifier was built for MEDICAL attachments to categorize these images by modality and anatomical site: Mammographies, Ultrasounds, Chest X-

rays, Traumatology X-rays, Gastrointestinal endoscopies, Lung CT scans, and Brain MRIs and CT scans. A VGG16 base model was fine-tuned using a custom dataset built from public sources. However, the performance of this model was not as accurate, with the ability to classify 62.59% of medical images with 86.1% accuracy.

In the case of FACE-FEATURES and BODY STRUCTURE attachments, various AI captioning models were compared to extract a short description of the image. GPT2 yielded poor results and was excluded from further studies. BLIP and PromptCap were found useful for accurately identifying body parts (except for genitalia, where accuracy was significantly lower). However useful in some cases, their captions were not precise enough for identifying specific symptoms in general. Among the three models, PromptCap was preferred for its flexibility, as prompts could be used to guide the model in performing specific tasks.

The results of this research and the prototype delivered will help now Abi Global Health to improve the profiling of the medical cases across all products and services. The components and knowledge built during this work will contribute to future studies and the optimization of the data processing and prediction pipelines to deliver a high quality healthcare.

Chapter 9

Conclusions

This thesis has demonstrated the feasibility of extracting information from Abi's medical consultation images. The completion of this thesis has paved the way for several proposed upgrades that are now within reach for implementation. These upgrades include:

1. Improved classification of consultation types using a new first-level classifier. For instance, Abi can now identify "pharmacological advice," "diagnosis," or "prescription requirement" based on images displaying medicine, affected body parts, or documents of medical prescriptions.

2. Enhanced identification of suitable specialists for images showing affected body parts. For example, if an image caption indicates a "patch of skin with a mole," Abi can determine that a dermatologist is the appropriate specialist.

3. Improved determination of suitable specialists for images displaying specific medication. For instance, if the picture contains medication packaging identified as Donepezil, Abi can assign the consultation to a neurologist.

4. Enhanced identification of suitable specialists for images depicting medical imaging studies. For example, if the image modality is identified as "Traumatology X-Ray," Abi can recommend consulting a traumatologist.

5. Anonymization of images containing sensitive patient information, such as documents displaying patient names.

6. Anonymization of images containing sensitive content, such as facial features that could identify the patient.

By integrating the new multi-level image pipeline into their system, Abi has the potential to make their service better through the addition of new functionalities and the improvement of existing ones.

These advancements will enhance the company's services by improving communication between patients and doctors, while also providing valuable data to both internal and external stakeholders.

References

[1] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee. Character region awareness for text detection. https://arxiv.org/abs/1904.01941, 2019.

[2] I. Bombonato. X-ray body images in png (unifesp competition). https://www.kaggle.com/datasets/ibombonato/xray-body-images-in-png-unifesp-competion, 2020. Accessed: June 1, 2023.

[3] X. P. Burgos-Artizzu, D. Coronado-Gutierrez, B. Valenzuela-Alcaraz, E. Bonet-Carne, E. Eixarch, F. Crispi, and E. Gratacós. FETAL_PLANES_DB: Common maternal-fetal ultrasound images, June 2020. The research leading to these results has received funding from Transmural Biotech SL, "La-Caixa" Foundation under grant agreements LCF/PR/GN14/10270005 and LCF/PR/GN18/10310003 the Instituto de Salud Carlos III (PI16/00861, PI17/00675) within the Plan Nacional de I+D+I and cofinanced by ISCIII-Subdirección General de Evaluación together with the Fondo Europeo de Desarrollo Regional (FEDER) "Una manera de hacer Europa", Cerebra Foundation for the Brain Injured Child (Carmarthen, Wales, UK), Cellex Foundation and AGAUR under grant 2017 SGR nº 1531. Additionally, EE has received funding from the Departament de Salut under grant SLT008/18/00156.

[4] Curated Breast Imaging Subset DDSM Dataset (Mammography). CBIS-DDSM: Breast Cancer Image Dataset. https://www.kaggle.com/datasets/awsaf49/cbis-ddsm-breast-cancer-image-dataset, n.d.

[5] A. Gupta. Human faces dataset. https://www.kaggle.com/datasets/ashwingupta3012/human-faces, 2020.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam. Searching for mobilenetv3, 2019.

[8] M. Hssayeni. Computed tomography images for intracranial hemorrhage detection and segmentation (version 1.0.0). `https://physionet.org/content/ct-ich/1.0.0/`, 2019. Accessed: June 1, 2023.

[9] Y. Hu, H. Hua, Z. Yang, W. Shi, N. A. Smith, and J. Luo. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*, 2022.

[10] Y. Hu, H. Hua, Z. Yang, W. Shi, N. A. Smith, and J. Luo. Promptcap implementation. `https://huggingface.co/tifa-benchmark/promptcap-coco-vqa`, 2022.

[11] S. Kumar. Hands and palm images dataset. `https://www.kaggle.com/datasets/shyambhu/hands-and-palm-images-dataset`, 2021.

[12] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022.

[13] J. Li, D. Li, C. Xiong, and S. Hoi. Blip implementation. `https://huggingface.co/Salesforce/blip-image-captioning-large`, 2022.

[14] J. Li, R. R. Selvaraju, A. D. Gotmare, S. Joty, C. Xiong, and S. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *arXiv preprint arXiv:2107.07651*, 2021.

[15] M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, and F. Wei. Trocr: Transformer-based optical character recognition with pre-trained models. In *AAAI 2023*, February 2023.

[16] M. Li, T. Lv, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, and F. Wei. Trocr: Transformer-based optical character recognition with pre-trained models.

https://huggingface.co/microsoft/trocr-base-handwritten, 2021.

[17] LUNA16 - Grand Challenge. Luna16 - grand challenge. https://luna16.grand-challenge.org/Home/, 2016. Accessed: June 1, 2023.

[18] T. M. Mohammad Fraiwan, Ziad Audat. The vertebrae x-ray images, 2022.

[19] M. Nickparvar. Brain tumor mri dataset, 2021.

[20] Y. Oh. Code was build by modifying this google image scraper. https://github.com/ohyicong/Google-Image-Scraper, 2023.

[21] K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, de Thomas Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P. T. Schmidt, M. Riegler, and P. Halvorsen. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, MMSys'17, pages 164–169, New York, NY, USA, 2017. ACM.

[22] M. R. Scene text detection in python with east and craft. https://medium.com/technovators/scene-text-detection-in-python-with-east-and-craft-cbe03dda35d5, March 2021.

[23] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019.

[24] A. Rosebrock. Tesseract page segmentation modes (psms) explained: How to improve your ocr accuracy. https://pyimagesearch.com/2021/11/15/tesseract-page-segmentation-modes-psms-explained-how-to-improve-you November 2021.

[25] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks, 2019.

[26] K. Shah. Eye dataset. https://www.kaggle.com/datasets/kayvanshah/eye-dataset, 2020.

[27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[28] Yih-Dar. vit-gpt2-image-captioning. https://huggingface.co/nlpconnect/vit-gpt2-image-captioning, 2022.

[29] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang. East: An efficient and accurate scene text detector. https://arxiv.org/abs/1704.03155, 2017.

Appendix A

Comparison of classifiers proposed

This appendix presents a concise comparison of the three proposed models, namely MobileNetV3, VGG16, and ResNet50. It highlights their performance in terms of accuracy and validation curves for both the training and validation sets. Additionally, it provides insights into the latency of each model, showcasing the number of steps per second.

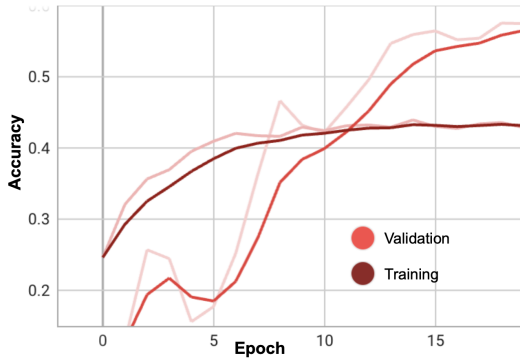


Figure A.1: MNV3 accuracy per epoch

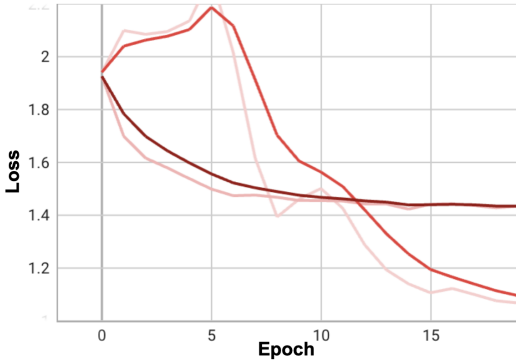


Figure A.2: MNV3 loss per epoch

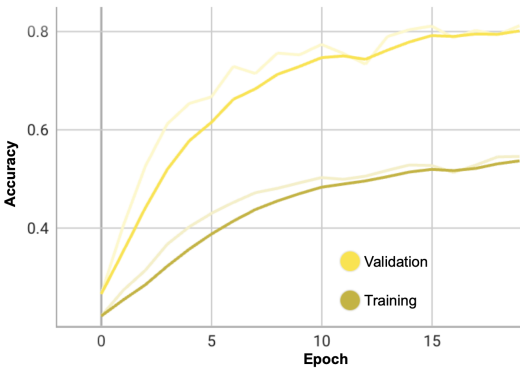


Figure A.3: RN50 accuracy per epoch

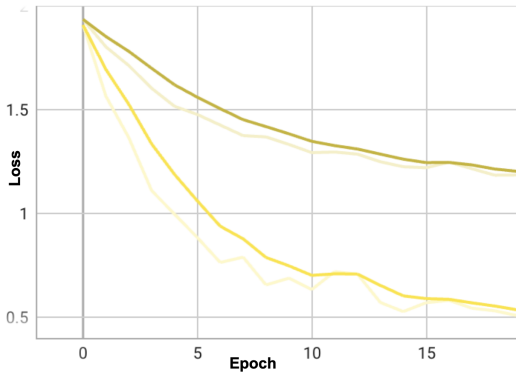


Figure A.4: RN50 loss per epoch

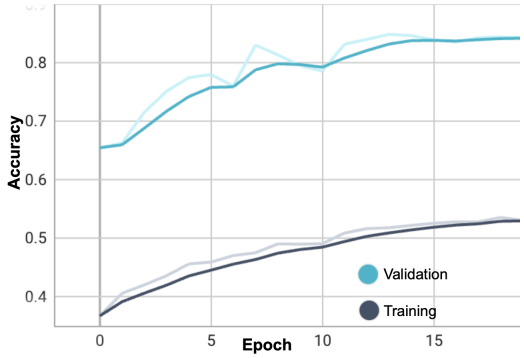


Figure A.5: VGG16 accuracy per epoch

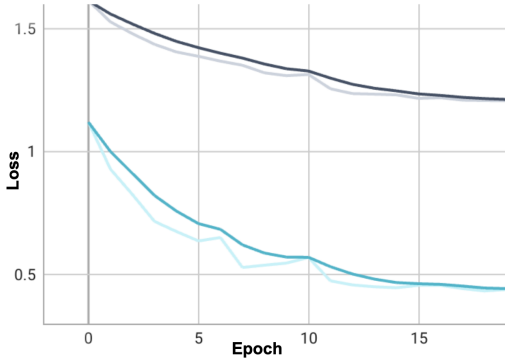


Figure A.6: VGG16 loss per epoch

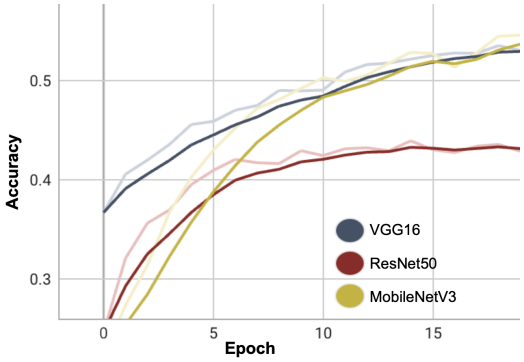


Figure A.7: Comparison of classifiers performance in training

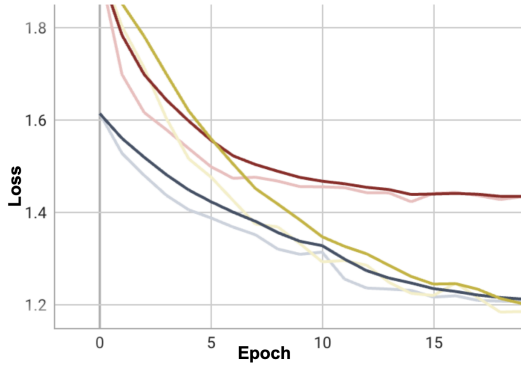


Figure A.8: Comparison of classifiers loss in training

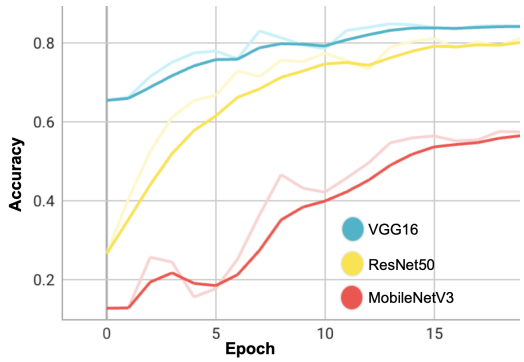


Figure A.9: Comparison of classifiers performance in validation

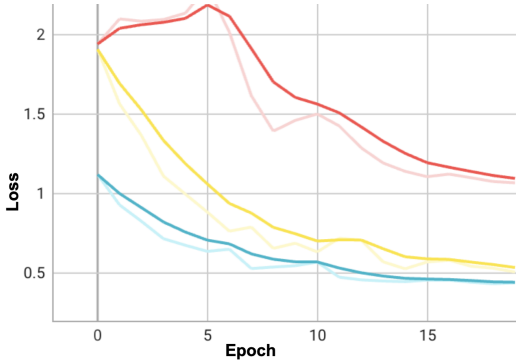


Figure A.10: Comparison of classifiers loss in validation

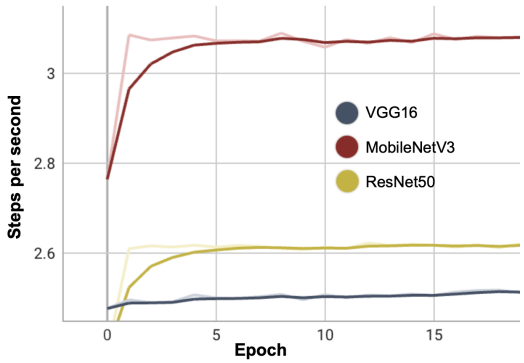
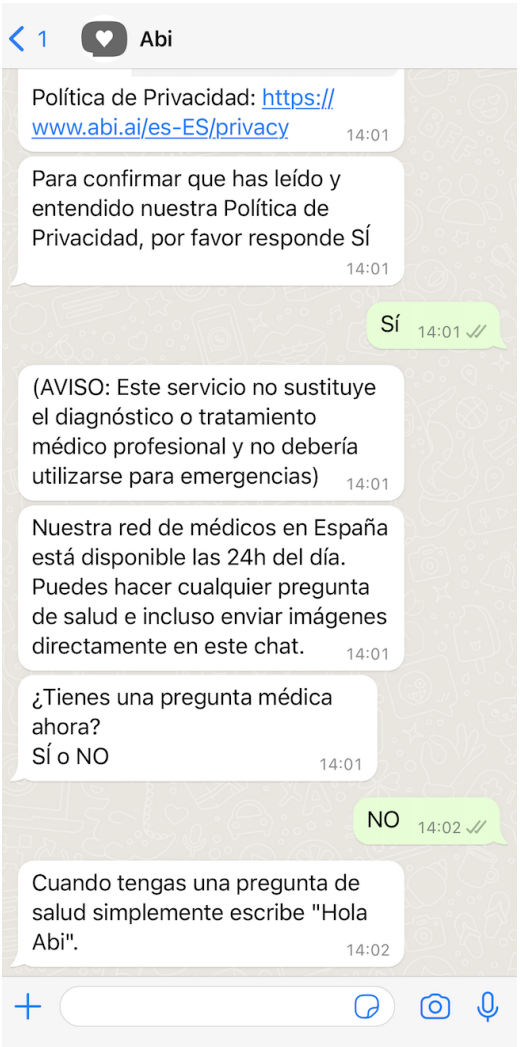
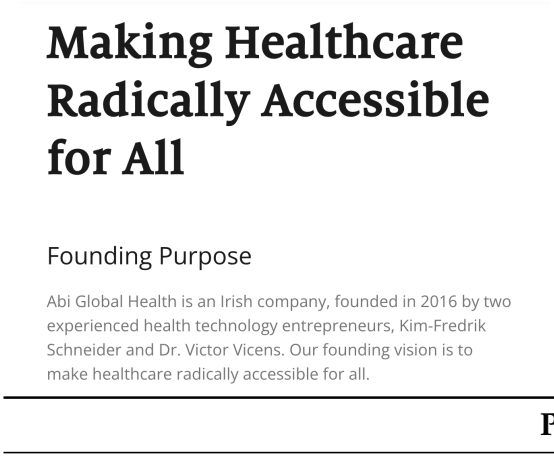


Figure A.11: Comparison of classifiers latency per epoch

Appendix B

Qualitative study on text retrieval

Tesseract performance	
Screenshots	
	¡1 Q ei Política de Privacidad: https:// www.abi.ai/es-ES/privacy 14:01 Para confirmar que has leído y entendido nuestra Política de Privacidad, por favor responde SÍ 14:01 SÍ 14:01 (AVISO: Este servicio no sustituye el diagnóstico o tratamiento médico profesional y no debería utilizarse para emergencias) ;1:0: Nuestra red de médicos en España está disponible las 24h del día. Puedes hacer cualquier pregunta de salud e incluso enviar imágenes directamente en este chat. ;¡7:0; ¿Tienes una pregunta médica ahora? SÍo NO 14:01 NO 14:02 ... salud simplemente escribe ”Hola Abi”. 14:02 + 0.00



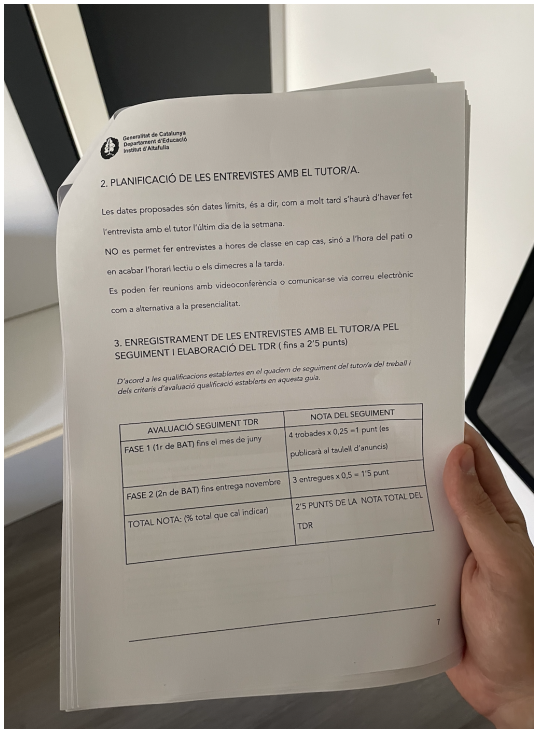
Making Healthcare Radically Accessible for All

Founding Purpose

Abi Global Health is an Irish company, founded in 2016 by two experienced health technology entrepreneurs, Kim-Fredrik Schneider and Dr. Victor Vicens. Our founding vision is to make healthcare radically accessible for all.

0) Making Healthcare 0) O Radically Accessible for All Founding Purpose Abi Global Health is an Irish company, founded in 2016 by two experienced health technology entrepreneurs, Kim-Fredrik Schneider and Dr. Victor Vicens. Our founding vision is to make healthcare radically accessible for all.

Printed text



Generalitat de Catalunya
Departament d'Educació
Institut d'Alfulla

2. PLANIFICACIÓ DE LES ENTREVISTES AMB EL TUTOR/A.

Les dates proposades són dates límits, és a dir, com a molt tard s'haurà d'haver fet l'entrevista amb el tutor l'últim dia de la setmana.

NO es permet fer entrevistes a hores de classe en cap cas, sinó a l'hora del pati o en acabar l'horari lectiu o els dimecres a la tarda.

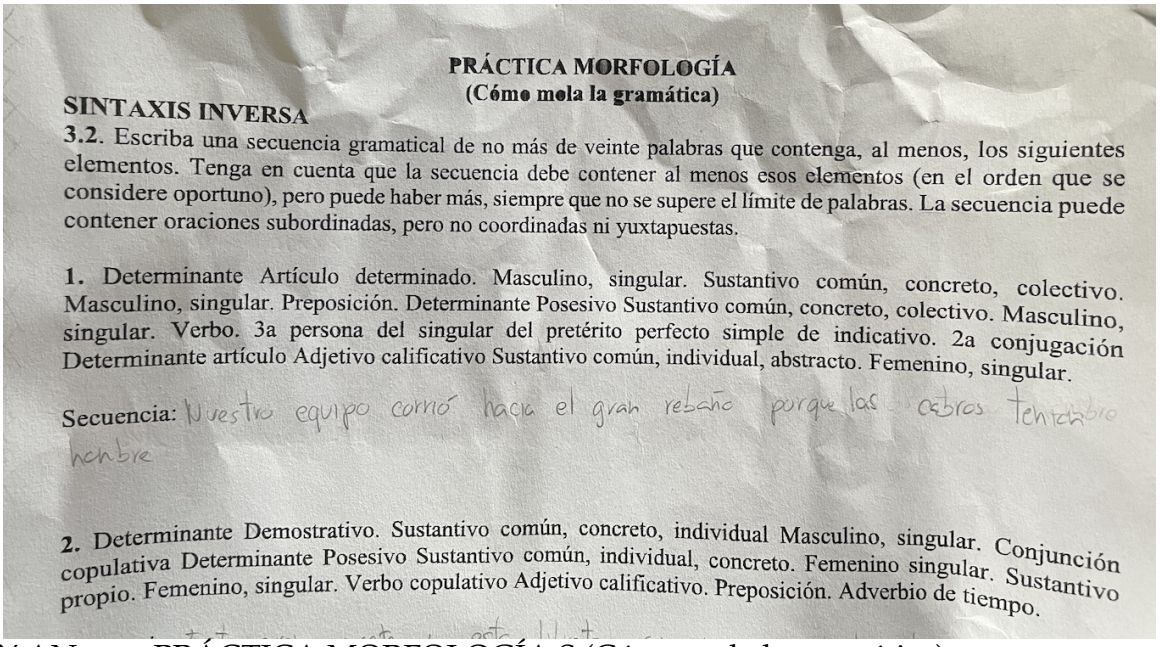
Es poden fer reunions amb videoconferència o comunicar-se via correu electrònic com a alternativa a la presencialitat.

3. ENREGISTRAMENT DE LES ENTREVISTES AMB EL TUTOR/A PEL SEGUIMENT I ELABORACIÓ DEL TDR (fins a 2'5 punts)

D'acord a les qualificacions establertes en el quadern de seguiment del tutor/a del treball i dels criteris d'avaluació qualificació establerts en aquesta guia.

AVALUACIÓ SEGUIMENT TDR	NOTA DEL SEGUIMENT
FASE 1 (1r de BAT) fins el mes de juny	4 trobades x 0,25 =1 punt (es publicarà al taulell d'anuncis)
FASE 2 (2n de BAT) fins entrega novembre	3 entregues x 0,5 = 1'5 punt
TOTAL NOTA: (% total que cal indicar)	2'5 PUNTS DE LA NOTA TOTAL DEL TDR

l I armat de sala J Serartament d'Educació I mi J l) (4 2 PLANIFICACIÓ DE LES ENTREVISTES AMB EL TUTOR/A. / I / I Les dates proposades són dates límits, és a dir com a mol tard s'haurà d'haver fet I l'entrevista amb el tutor l'últim dia de la setmana. NO es permet fer entrevistes a hores de classe en 2P cas, sinó a l'hora del patí 0 en acabar l'horari lectiu o els dimecres a la tarda. Es poden fer reunions amb videoconferència o comunicar-se via Correu electrònic en com a alternativa a la presencialitat: / 3. ENREGISTRAMENT DE LES ENTREVISTES AMB EL TUTOR/A PEL SEGUIMENT l ELABORACIÓ DELTDR (fins a 2/5 punts) D'acorda les qualificacions establertes en el quadern de seguiment del tutor/a del treballi Gels aiteris d'avaluació qualificació establerts en aquesta guia. FASE 1 (1r de BAT) fins el mes de juny a trobades x 0,25 71 punt (es publicarà al taulell d'anunci sal CASE 2n de BAT) ins entrega Et Era 1 NOTA TOTAL DEL TOTAL NOTA: (8 total que ic i MS LA TDR 4



PRÁCTICA MORFOLOGÍA
(Cómo mola la gramática)

SINTAXIS INVERSA

3.2. Escriba una secuencia gramatical de no más de veinte palabras que contenga, al menos, los siguientes elementos. Tenga en cuenta que la secuencia debe contener al menos esos elementos (en el orden que se considere oportuno), pero puede haber más, siempre que no se supere el límite de palabras. La secuencia puede contener oraciones subordinadas, pero no coordinadas ni yuxtapuestas.

1. Determinante Artículo determinado. Masculino, singular. Sustantivo común, concreto, colectivo. Masculino, singular. Preposición. Determinante Posesivo Sustantivo común, concreto, colectivo. Masculino, singular. Verbo. 3a persona del singular del pretérito perfecto simple de indicativo. 2a conjugación Determinante artículo Adjetivo calificativo Sustantivo común, individual, abstracto. Femenino, singular.

Secuencia: Nuestro equipo corrió hacia el gran rebaño porque las cabras tenían hambre

2. Determinante Demostrativo. Sustantivo común, concreto, individual Masculino, singular. Conjunción copulativa Determinante Posesivo Sustantivo común, individual, concreto. Femenino singular. Sustantivo propio. Femenino, singular. Verbo copulativo Adjetivo calificativo. Preposición. Adverbio de tiempo.

NY AN -.o a PRÁCTICA MORFOLOGÍA S (Cóme mela la gramática)
SINTAXIS INVERSA j 3.2. Escriba una secuencia gramatical de no más de veinte palabras que contenga, al menos, los siguientes elementos. Tenga en cuenta que la secuencia debe contener al menos esos elementos (en el orden que se considere oportuno), pero puede haber más, siempre que no se supere el límite de palabras. La secuencia puede contener oraciones subordinadas, pero no coordinadas ni yuxtapuestas.
1. Determínante Artículo determinado. Masculino, singular. Sustantivo común, concreto, colectivo. Masculino, singular. Preposición. Determinante Posesivo Sustantivo común, concreto, colectivo. Masculino singular. Verbo. 3a persona del singular del pretérito perfecto simple de indicativo. 2a conjugación Determinante artículo Adjetivo calificativo Sustantivo común, individual, abstracto. Femenino, singular. secuencias ves vo equipo. comió Made el gran reco porqe eE Obres, WEN bre
Determinante Demostrativo. Sustantivo común, concreto, individual Masculino, singular. RR a alativa Determinante Posesivo Sustantivo común, individual, concreto. Femenino singular Su oción cop : : ivo Adjetivo calificati 6 y - SUstantiy propio. Femenino, singular. Verbo copulativo Adj calificativo. Preposición. Adverbio de tiempo. Mtivo

Table B.1: Qualitative study of Tesseract performance on screenshots and printed documents

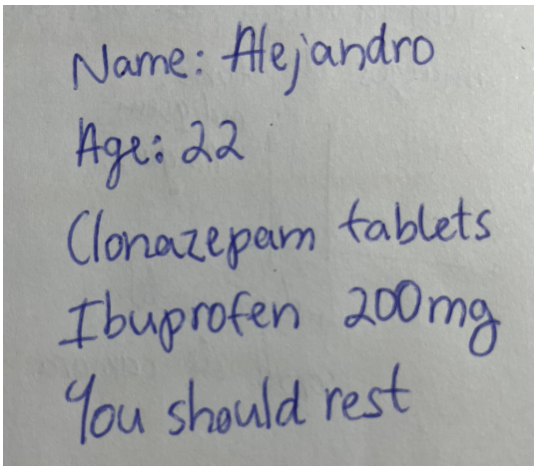
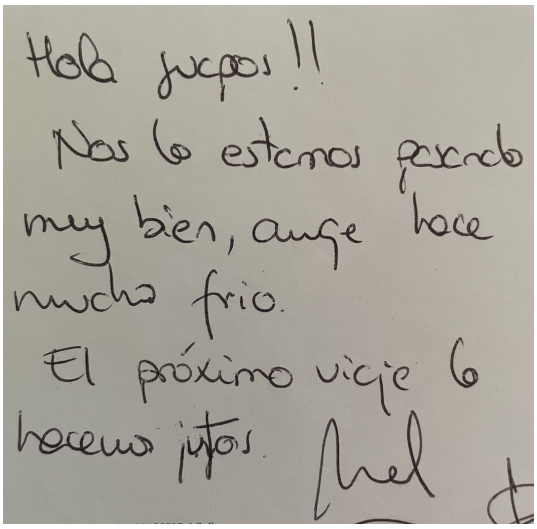
TROCR performance	
Handwritten	
	Name : Alejandro agei 22 clonazepern tablets ibuprofen. zooming you should rest
	hold juipos), nosco estends fessendo mey biem, ange hace muchs trico. el proximo viceje co heccens justors phel

Table B.2: Qualitative study of TROCR performance on handwritten documents

Appendix C

Qualitative studies on medicine segmentation

Segmentator comparison	
CRAFT	**EAST**
	
	


Ibuprofen
400mg
PAIN RELIEF
FEVER REDUCTION
ANTI-INFLAMMATORY
TABLETS
Panadol
Caplet 500mg
Effectively eases pain
Fast absorption
Gentle on stomach
500mg Paracetamol / Caplet
WOODS
ENO
SYS
DIA
PULSE
START
STOP
38.8°C

Table C.1: Comparison of EAST and CRAFT segmentation

Table C.2: Qualitative study of CRAFT+Tesseract performance (Part 1)

Easy set				
Raw image	**Raw Tesseract**	**CRAFT segmentation**	**CRAFT+Tesseract**	**CRAFT+Tesseract (sorted by height)**
	Easy To Swallow PARACETAMOL TABLETS BP 500 mg 16 . Tablets a wy B Sa		1. Easy To Swallow 2. PARACETAMOL 3. TABLETS BP 500 mg 4. Effective Pain Relief 5. 16 6. Tablets 7. GerisrTrow	1. PARACETAMOL 2. 16 3. TABLETS BP 500 mg
	See ———— Af — 4 — —— oes — Sainsburys Healthcare cba - » cy aPsules a 4 ″ 5Ke ——— x16 90h . 48 Se Effective Pain relieg as ′		1. J = 2 / ∼ $Ja)hi$ 2. POIDPDIFTID AOS LI OS 3. XK FES 4. Sainsbury's Healthcare 5. \\buprofen 6. (*Nothing*) 7. (*Nothing*) 8. : 9. 200mg, 10. Capsules 11. (*Nothing*) 12. x16 13. Effective pain relief	1. \\buprofen 2. Sainsbury's Healthcare 3. 200mg,
	te \n* \neX \′ cov — \n: Woes —a— \\a ; ¥, Lf Ks = ta ′ \\e yi) \\oa″ - (/ 7 ∼∼: | . s if ee SE a SAEZ \n# \\Ss ee ′ \′ = (* _- a \\		1. NOC XS0-2 907 2. Votrin: 3. Ibuprofen Tablets use 200 ma, 4. (*Nothing*) 5. Paln Rellever/Fever Reducer (NSAID) 6. 2A coated Caplets∼	1. Ibuprofen 2. 400mg 3. PAIN RELIEF

Table C.3: Qualitative study of CRAFT+Tesseract performance (Part 2)

Difficult set				
Raw image	**Raw Tesseract**	**CRAFT segmentation**	**CRAFT+Tesseract**	**CRAFT+Tesseract (sorted by height)**
	te \n* \neX \’ cov — \n: Woes —a— \\a ; ¥, Lf Ks = ta ‘ \\e yi) \\oa” - (/ 7 ∼∼: | . s if ee SE a SAEZ # \\Ss ee ‘ \’ = (* _- a \\		1. Ibuprofen 2. PAIN RELIEF 3. 400mg 4. FEVER REDUCTION 5. ANTI-INFLAMMATORY 7. TABLETS:	1. Ibuprofen 2. 400mg 3. PAIN RELIEF
	a ye Panadol wooDsa WO00S ODDS \\KONISA ale . ≫ w fwo0e oe \\—ii ¡ ® \% — a. — : — _# e iad — . Of o—Ee = in — — = = 9 ay sis :		1. Panadol 2. woos 3. woos; 4. WOODS 5. # 5995, 6. wO00 7. Caplet S00mg 8. Fatt absorption 9. WOm, Paracetame Capiet 10.ENO 11. 36	1. Panadol 2. ENO 3. WOODS
	A PATH OF LIGHT, ≪ D¿. = a O:,efn a: oo 8 rape y Zi ee” ai ee 8 ‘Shea Butter * f ‘Cocoa Butt” HAND CREAM CLEMENTINE —		1. A PATH OF 2. LIGHT 3. = \n4 \na \n2 \ndq \n 4. Shee Butter 5. Cocoe Butt: 6. HAND CREA.M 7. (Creme pour ies Mars 8. WOR of the re crates © a 9. CLEMENTING 10. Ietnve woe	1. = \n4 \na \n2 \ndq \n 2. LIGHT 3. A PATH OF

Table C.4: Qualitative study of CRAFT+Tesseract performance (Part 3)

Medical device set				
Raw image	**Raw Tesseract**	**CRAFT segmentation**	**CRAFT+Tesseract**	**CRAFT+Tesseract (sorted by height)**
	microlife 2% J SYS mmHg 140/90 ay DIA f mmHg be] PUL Ps i		1. Pad 2. microlife 3. SYS 4. mmHg 5. 140/90 6. DIA 7. mmHg 8. PUL 9. /min 10. Mv	1. microlife 2. DIA 3. SYS
	(*Nothing*)		1. 3887] C	
	= ∼ *é* & ‘nee % sys DIA F purse + = ustfpe — epee. Ss		1. ic \n— 2. 3 3. DIA 4. PULSE 5. START 6. sToP	1. 3 2. ic \n— 3. DIA

Table C.5: Qualitative study of CRAFT+Tesseract performance (Part 4)

Textless set				
Raw image	**Raw Tesseract**	**CRAFT segmentation**	**CRAFT+Tesseract**	**CRAFT+Tesseract (sorted by height)**
	} — sf Poe ———		1. A	
	; A \\= = $\sim\sim$ $nnWw$		(*Nothing*)	
	Y awit \\: aa \\		(*Nothing*)	

Appendix D

Qualitative study on captioning models

Captions generated on Face Features				
Image	**BLIP**	**BLIP+**	**PROMPT**	**PROMPT+**
	someone has a bruised eye and a black eye with a white sink in the background	a picture of a human eye with a small red spot on the eye	a person with a red eye	a person with a red eye
	there is a small child with a small ear that has a small bruise on it	a picture of a human ear with a small piece of skin on it	a person with a bleeding ear	a person with a bruise in the ear
	there is a close up of a baby's nose with a toothbrush in it	a picture of a human nose with a small amount of white stuff on it	a man with a bruise on his nose	a man with a bruise on his nose
	arafed woman with a tooth-brush sticking out of her mouth	a picture of a human open mouth with a toothbrush in it	a person with a sore tongue	a person with an openmouth

	there is a close up of a woman's lips with a black eye	a picture of a human mouth with a small amount of skin	a man with a pimple in his mouth	a man with a pimple in his mouth
	there is a man taking a selfie in a mirror	a picture of a human full person with a cell phone in a mirror	a man taking a selfie in a mirror	a man taking a selfie in a mirror

Captions generated on Body Structure

Image	BLIP	BLIP+	PROMPT	PROMPT+
	a man with a lot of sunburns on his back	a picture of a human back with sunburns on it	a man with a red spot on his back	a man with a red spot on his back
	someone is holding a green jacket over their stomach with a red stain on it	a picture of a human torso with a small patch of skin on it	a man with a stain on his stomach	a person with a stain on their stomach
	someone is laying down with their hand on a pillow	a picture of a human arm with a white substance on it	a man with a lot of hair on his arm	a man with a birthmark on his arm
	someone is holding a piece of food in their hand	a picture of a human hand with a small bruise on it	a person with a burn on their hand	a hand with a bug on it

	someone has a lot of skin on their legs and feet	a picture of a human feet with a lot of skin on it	a person with bruises on their arm	a person with bruises on their arm
Example of male genitalia	someone is holding a small dog's paw with a bandage	a picture of a human dick with a small black dog	a person holding a foot with a wart	a person holding a foot with a wart on it
Example of female genitalia	someone is holding a toothbrush and a tooth with a black circle on it	a picture of a human vagina with a black circle on it	a person with a wart in their vagina	a person with a wart in their vagina
Example of rectal opening	someone is holding a finger up to a man's face	a picture of a human ass with a finger on a mans finger	a hand with a wound in the ass	a hand with a wound in the ass

Table D.1: Captions generated on human images

Appendix E

Medical image dataset

This appendix serves as documentation of the dataset created for our medical image classifier. It provides information such as references to the public datasets from which visual data was collected, the modality and anatomical site, the amount of data extracted, etc.

Source Dataset	License	Image type	Images used	Category	Comment
LUNA16 - Grand Challenge	No copyright (CC0: Public Domain) [17]	CT	890	Lung	All the train images.
Chest CT-Scan images Dataset [21]	Open Data Commons Open Database License (ODbL) v1.0	CT	798	Lung	Train, test and validation images.
Brain Tumor MRI Dataset [19]	No copyright (CC0: Public Domain)	MRI	1,596	Brain	Classified as notumor.
Brain CT Images with Intracranial Hemorrhage Masks [8]	Suitable for commercial use (CC BY 4.0)	CT	5,566	Brain	All the train images.
The vertebrae X-Ray images [18]	Suitable for commercial use (CC BY-SA 4.0)	X-Ray	79	Vertebrae	Classified as NonSpond (from 224x224 partition).
FETAL_PLANES_DB: Common maternal-fetal ultrasound images [3]	Suitable for commercial use (Attribution 4.0 International)	Ultrasound	8,187	Ecography	Train images excluding those classified as other (since those were segmented ecographies).

Table E.1: Dataset created from public sources (Part 1)

Source Dataset	License	Image type	Images used	Category	Comment
CBIS-DDSM: Breast Cancer Image Dataset [4]	Suitable for commercial use (CC BY-SA 3.0)	X-Ray	294	Mammography	Subset of the train images where whole breast was visible.
Body Parts X-Ray Images in PNG [2]	No copyright (CC0: Public Domain)	X-Ray	49	Arm	Train images classified as elbow and forearm.
			200	Hands	Train images classified as hand, finger, or wrist.
			724	Chest	Train images classified as chest.
			135	Skull	Train pictures of skull, sinus and cervical spine.
			49	Arm	Train images classified as elbow and forearm.
			86	Upper leg	Train pictures classified as thigh, pelvis and hip.
Simula Datasets - Kvasir [21]	No copyright (CC0: Public Domain)	Endoscopy	3000	Gastro	All the train images.

Table E.2: Dataset created from public sources (Part 2)