UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Técnica Superior de Ingeniería de Sistemas Informáticos

Advanced Deep Learning Models for Precise Medical Image Analysis and Diagnosis

D O C T O R A L T H E S I S

Submitted for the degree of Doctor by:

Guillermo Iglesias Hernández
Máster Universitario en Inteligencia Artificial

Madrid, 2024


UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Técnica Superior de Ingeniería de Sistemas Informáticos

Doctoral Degree in Computer Sciences and Technologies for Smart Cities

Advanced Deep Learning Models for Precise Medical Image Analysis and Diagnosis

D O C T O R A L T H E S I S

Submitted for the degree of Doctor by:

Guillermo Iglesias Hernández
Máster Universitario en Inteligencia Artificial

Under the supervision of:
Dr. Edgar Talavera Muñoz

Madrid, 2024

Title: Advanced Deep Learning Models for Precise Medical Image Analysis and Diagnosis

Author: Guillermo Iglesias Hernández

Doctoral Programme: Computer Sciences and Technologies for Smart Cities

Thesis Supervision:

Dr. Edgar Talavera Muñoz, Profesor Contradato Doctor, Universidad Politécnica de Madrid(Supervisor)

External Reviewers:

Thesis Defense Committee:

Thesis Defense Date:

This thesis has been partially supported by *"Convocatoria de ayudas dirigidas al personal investigador en formación predoctoral para realizar una estancia de investigación internacional igual o superior a tres meses"* for the international internship in the King's College London.

Y... es algo que se tiene o no se tiene, hermano;

hay dos tipos de personas en la vida:

Los que hablan

y los que hacen, primo,

y nosotros está claro lo que hemos decidido.

Fernando Gálvez Gómez a.k.a. Yung Beef

Agradecimientos

Cada vez que hago un trabajo siento que, por mucho que sea yo quien lo haya realizado, de alguna manera este tiene un poco de toda la gente que me rodea. Sea de una forma u otra, las personas que tengo son tan responsables como yo de lo que hago.

Por eso, si a alguien tuviese que dedicarle la tesis, sería a todas las personas a las que quiero y que sin ellas la tesis no tendría sentido. Podría estar muchas páginas agradeciendo a mucha gente todo lo que han hecho y hacen por mi, y aún así probablemente me quedase corto. Por desgracia no suelo decirle mucho a mi gente lo mucho que les quiero, pero nada de esto tendría sentido si lo hiciese aislado.

Mis padres Antonio y Nieves, sé que lo habéis dado todo por mi, sin vosotros no podría estar donde estoy. No tiene que ser fácil criar a alguien, espero que estéis orgullosos de lo que hago, yo intentaré estar a la altura. También a mi hermano por quererme y apoyarme. Muchas gracias por todo.

Mis abuelos Aurelio, Dioni, Martina y Maruja, me habéis apoyado y querido desde que soy un niño, soy consciente de que sin todo eso no podría haber llegado a nada. Porque vosotros sois los que me habéis animado a ser quien soy y habéis visto desde el principio el potencial que puedo llegar a tener. Si algo he aprendido de ellos es a esforzarme y a intentar sacarlo todo por mi mismo. Espero que estéis orgullosos, porque yo os llevo como un orgullo en toda mi vida. En general muchas gracias a toda mi familia por haberme querido.

Mis amigos a los que quiero con locura y son mi vida. Alba, Álvaro, Bringas, Dani, David, Javi, Jorge, Marta, Pablo y Sandra, creo que sabéis lo importantes que sois para mi, gracias por todo. Si de algún lado he sacado fuerzas para seguir con todo esto es porque me motiva vivir una vida con vosotros. David y Jorge, lleváis casi toda la vida conmigo y sé que de una manera u otra no vais a desaparecer. Bringas, eres una persona increible con la que siempre disfruto. Marta y Alba, os quiero muchisimo, me haceis sentir como en casa. Pablo, en muchos sentidos eres mi compañero de vida, sé que haremos cosas grandes juntos, y si no, al menos estaremos juntos para caer. Os quiero mucho a todos, gracias por hacer lo mismo.

Una parte inesperada y esencial en la tesis fue estar con Goncho y Guille, que me acogieron en su casa y me hicieron sentir en familia. Guille, eres mi compañero en este proceso de tesis, he aprendido mucho de ti y espero seguir contigo cerca por muchos años.

Ana, eres parte esencial de mi vida, te quiero mucho.

Toda la gente de la ETSISI, los cuales he sentido, desde mi época de estudiante, que me apoyáis y me habéis ayudado con toda la vorágine de la tesis. Alberto, Elvira, Fernando, Gema, Pepi y Raúl que desde el primer minuto me habéis ayudado desinteresadamente con cosas que quizás no os correspondían. Gracias.

A KNODIS, por cederme el uso del servidor *Columbia*, no sé qué habría hecho sin él.

A Héctor, por acogerme en Londres y hacerme sentir que aun viviendo fuera de casa tenía a

alguien que se preocupaba por mi.

A Jesús Troya, por estar siempre dispuesto a colaborar y apostar por mi. Espero que podamos sacar cosas juntos.

Y por último a Edgar, sabes que te quiero muchisimo. Si algo agradezco infinitamente es tener a alguien como tú que esté a mi lado. Fuiste el primero que supo ver en mí algo que acabó derivando en esta tesis, sin tí no sé dónde estaría, pero escribiendo esta tesis seguro que no. Sé que te esfuerzas mucho por mí, y yo siempre intentaré estar a la altura de tus expectativas. Cuando te prometo algo intento cumplirlo porque sé la fé que me tienes, no quiero fallarte nunca. Te quiero mucho.

Probablemente me deje a mucha gente por agradecer, a los que de una manera y otra quiera. Gracias a todos por tanto, espero estar a la altura.

Abstract

Deep Learning has emerged as a strategic area in medical image diagnosis, with numerous studies exploring automation in this field. However, many existing approaches attempt to fully replace medical professionals by relying exclusively on model predictions. This thesis presents an alternative strategy, proposing Deep Learning architectures designed to support, rather than replace, physicians in diagnostic tasks.

Two distinct methodologies are explored to achieve this goal: Content-Based Image Retrieval and eXplainable Artificial Intelligence. Content-Based Image Retrieval system enables physicians to compare complex cases with similar ones from a curated database, enhancing decision-making through feature-based similarities. The eXplainable Artificial Intelligence approach provides visual explanations through heatmaps, indicating the regions of interest detected by the model.

The thesis introduces two novel architectures—Multi-Output Classification Autoencoder (MOC-AE) and Multi-Output Classification Variational Autoencoder (MOC-VAE). These models employ multi-objective training to generate rich image descriptors, capturing both pathological and anatomical information. These descriptors are then used to identify visually similar cases in a low-dimensional space. To address the challenge of quantifying visual similarity in Content-Based Image Retrieval, the Sliced Wasserstein distance is proposed, offering a robust metric for comparative diagnosis.

Furthermore, the thesis demonstrates the utility of these architectures in generating precise, non-linear explanations for diagnoses, leveraging the factorization process inherent to the models. The simplicity and enhanced performance of the proposed approaches, compared to prior methods, offer a foundation for future research and real-world applications in medical diagnostics using Deep Learning.

Resumen

El Deep Learning ha surgido como un área estratégica en el diagnóstico de imágenes médicas, con numerosos estudios que exploran la automatización en este campo. Sin embargo, muchos enfoques actuales buscan reemplazar por completo a los profesionales médicos basándose únicamente en las predicciones de los modelos. Esta tesis presenta una estrategia alternativa, proponiendo arquitecturas de Deep Learning diseñadas para apoyar, en lugar de reemplazar, a los médicos en tareas diagnósticas.

Se exploran dos metodologías distintas para alcanzar este objetivo: Sistemas de Recuperación de Imágenes Basada en Contenido e Inteligencia Artificial Explicable. Los Sistemas de Recuperación de Imágenes Basada en Contenido permiten a los médicos comparar casos complejos con otros similares de una base de datos curada, mejorando la toma de decisiones mediante similitudes basadas en características. El enfoque de la Inteligencia Artificial Explicable proporciona explicaciones visuales a través de mapas de calor, que indican las regiones de interés detectadas por el modelo.

La tesis introduce dos arquitecturas novedosas: el Autoencoder de Clasificación Multi-Salida (MOC-AE) y el Autoencoder Variacional de Clasificación Multi-Salida (MOC-VAE). Estos modelos emplean un entrenamiento multiobjetivo para generar descriptores de imágenes enriquecidos que capturan tanto información patológica como anatómica. Estos descriptores se utilizan para identificar casos visualmente similares en un espacio de baja dimensionalidad. Para abordar el desafío de cuantificar la similitud visual en Sistemas de Recuperación de Imágenes Basada en Contenido, se propone la distancia de Wasserstein por secciones, ofreciendo una métrica robusta para el diagnóstico comparativo.

Además, la tesis demuestra la utilidad de estas arquitecturas para generar explicaciones precisas y no lineales de los diagnósticos, aprovechando el proceso de factorización inherente a los modelos. La simplicidad y el rendimiento mejorado de los enfoques propuestos, en comparación con métodos anteriores, ofrecen una base sólida para futuras investigaciones y aplicaciones reales en diagnósticos médicos utilizando Deep Learning.

Table of Contents

List of Figures

List of Tables

Glossary

ACGAN Auxiliary Classifier GAN. i, 15, 42

AE Autoencoder. i, xiii, 5, 6, 13–15, 20, 22, 25, 27, 30, 36, 40–43, 52, 55, 56, 58–62, 65, 67, 68, 71–73, 76, 82, 83, 104

AI Artificial Intelligence. i, 3, 4, 7, 9, 11–13, 15, 16, 18, 19, 23–25, 36, 79, 80

ANN Artificial Neural Network. i, 38

Bi-LSTM Bidirectional Long Short-Term Memory. i

BraTS Multimodal Brain Tumor Segmentation Challenge. i, xiii, 6, 36–38

CAM Class Activation Mapping. i

CBIR Content-Based Image Retrieval. i, xi, xiii, 4–8, 11, 18–25, 27–30, 36, 40, 41, 43, 44, 46, 48–52, 55–58, 60, 62, 63, 65, 66, 75–77, 79, 101–105

CDSS Clinical Decision Support System. i, 4, 5, 7, 23–25, 29, 49–51, 105

CGAN Conditional GAN. i, 15, 42

CIU Contextual Importance and Utility. i, 25

CNN Convolutional Neural Network. i, 4, 12, 19, 20, 25, 43, 105

CSGAN Cyclic-Synthesized GAN. i, 42

CT Computed Tomography. i

CV Computer Vision. i

CycleGAN Cycle-Consistent GAN. i, 14

DL Deep Learning. i, 3–7, 9, 11–16, 19, 23, 25, 26, 28–30, 35, 51, 55, 58, 75, 79, 97, 98, 102–104, 106, 107

DualGAN Unsupervised Dual Learning for Image-to-Image Translation. i, 42

ED Pleritumoural edema. i, 37

ET Gd-enhancing tumor. i, 37

GAN Generative Adversarial Network. i, 12, 14, 15, 42

GLCM Gray Level Co-occurrence Matrix. i, 18

GMM Gaussian Mixture Model. i, 13

LBP Local Binary Pattern. i, 18

LIME Local Interpretable Model-agnostic Explanations. i, 17, 18, 25, 26, 81, 92, 93, 95, 98

LLM Large Language Model. i

LS-GAN Loss-Sensitive GAN. i, 42

MISS GAN Multi-IlluStrator Style GAN. i, 42

ML Machine Learning. i, 3, 4, 11, 12, 15, 18, 19, 24, 36

MOC-AE Multi-Output Classification Autoencoder. i, xi, xiii, xv, 5–8, 30, 36, 40–45, 47–52, 55–58, 60–62, 66–69, 71–73, 75, 76, 79, 81, 101–104, 106, 107

MOC-VAE Multi-Output Classification Variational Autoencoder. i, xi, xiii, xv, 5–8, 16, 29, 30, 55–58, 60, 62, 65–68, 70–73, 75, 76, 79–84, 87, 88, 94–98, 101–107

MR Magnetic Resonance. i, xiii, 35, 37, 38, 41, 43, 45, 46, 51

MRI Magnetic Resonance Imaging. i, xiii, 39

MSE Mean Squared Error. i, 21, 22, 63, 64

NET Non-enhancing tumor core. i, 37

NIfTI Neuroimaging Informatics Technology Initiative. i, 39

NN Neural Network. i

PCA Principal Component Analysis. i, 19

QBIC Query by Image Content. i, 19

ReLU Rectified Linear Unit. i, 43

RQ Research Question. i, 28–30, 52, 76, 77, 98, 101–103

SAG Structured Attention Graphs. i, 17, 81, 92, 93, 95, 96, 98

SHAP SHapley Additive exPlanation. i, 17, 18, 25

SPEA2 Strength Pareto Evolutionary Algorithm 2. i, 81, 86, 87

SRGAN Super Resolution GAN. i, 42

SVM Support Vector Machine. i, 25

T1 Native scanner. i, 37, 43

T1Gd Post-contrast T1 weighted. i, 37, 43

T2 T2 weighted. i, 37, 43

T2-Flair T2 Fluid Attenuated Inversion Recovery. i, 37, 43

VAE Variational Autoencoder. i, 6, 13–15, 30, 55, 56, 62, 82–85, 88, 97

VGG Visual Geometry Group. i, 26

WSRGAN Weighted SRGAN. i, 42

XAI eXplainable Artificial Intelligence. i, xi, 4, 5, 7–9, 11, 15–17, 25, 26, 28–30, 76, 79–82, 84, 88, 92, 94–98, 102–104, 106

Introduction to the Thesis

Chapter 1

Introduction

Artificial Intelligence (AI) have revolutionized the science, society and industry. There have been many areas historically benefited from the use of Deep Learning (DL) [216], [231]. Machine Learning (ML) is able of extracting features from data sources, analyzing and discovering patterns in the information to produce smart decisions.

However, these concepts are not novel to the present era; almost since the beginning of humanity, we have been interested in replicating human behaviors in different ways. The golems of the Jewish folklore already captured the idea of creating synthetic creatures capable of having some kind of consciousness. Then, probably the first effort in translating human reasoning to formal mechanisms comes with Aristotelian logic that tried to translate human mechanisms of thinking to logic systems.

From that point on, humans have developed many different techniques to simulate human thinking. The concept and limits of AI have evolved with its results. Besides being a term coined in 1956 in the Dartmouth Conference [152], it is a concept almost inherent to humans. By analyzing AI through time, one can see that, in spite of what period of time is being analyzed, the most developed technology of each era was used to translate human thinking to an inanimate object.

The golems of the myths were created with mud or clay, which was the available materials of people at that age. But with further advances, the technology of the time was put in service of that *proto* AI. Heron of Alexandria designed a moving theater, powered with counterweights, that was able of reproducing a theatrical scene, inventing the automata. Years later, Ada Lovelace used the Analytical Machine invented by Charles Babbage to create programs that performed calculations in an intelligent manner. The Analytical Machine, at its time, was the most innovative device available, and Ada Lovelace leveraged it to create proto AI programs. It is easy to see the trends towards using the available technology at the time to develop methods of translating human reasoning to synthetic mechanisms. Later, with the work of Alan Turing with the Turing Machine, AI was born as it is considered nowadays, using computers at its basis.

But it can be seen that one of the most important milestones of AI in the last decade was possible because of Big Data [63]. The huge growth in available data greatly marks advances

in AI [2]. In this sense, ML is capable of benefiting from all this data to extract complex patterns from it.

The *data era* where we are immersed [31] marks the development of AI, in a similar way as the Analytical Machine or the Turing Machine, we use the available technology of our time, which is the data. In a world where data defines our behavior, this information is used to develop intelligent programs to automate different tasks.

This context is the perfect breeding ground for the huge development of DL [110]. Many areas of science and industry have implemented AI to solve different problems [66], [98], [192].

One of the most successful areas of DL is Computer Vision. Recently, many researches have been published focusing on image processing with Artificial Neural Networks [99], [134], [188]. The use of Convolutional Neural Networks (CNNs) [58] has revolutionized DL possibilities in Computer Vision. CNNs are a type of Artificial Neural Network for image processing based on convolutional operation. The convolutions of the network learn to extract complex patterns in the images during their training. Because of their adaptability, CNNs have been used in many different fields, from Industry 4.0 [112] to drug discovery [225].

One of the main problems of DL is that the models are black-box systems, therefore their predictions are not explainable. This is specially important when they are applied in sensible areas, such as law or defense [18].

Several approximations have been followed to tackle this. eXplainable Artificial Intelligence (XAI) is an area of research focused on providing explanations to ML black-box models. It is based on finding explanations that are related with the predictions gave by the models. Content-Based Image Retrieval (CBIR) is another possibility to create models that do not fully rely on the prediction of the model. CBIR systems are Computer Vision systems that retrieve the most similar images from a database based on a query. This way the professional has the last word in the prediction and the ML system acts as a search engine over the database.

As in many other fields, medicine can be smartly used DL to alleviate human tasks. This aspect is particularly important, by reducing the time that physicians consume on doing certain tasks, their time can be heavily optimized, reducing the amount of futile tasks they have to do and making it possible to use their time for more specific and important procedures. Medical centers need to optimize their processes, and by applying these types of techniques, they can use their resources in a more intelligent manner. Nevertheless, AI is considered to never be able to replace human professionals in the medical area [83], but rather to change their procedures and protocols, to allow human efforts to adapt to the real necessities of the medical centers, and not to tedious and useless tasks. In addition, physicians workload is usually very high [146] and there is an inverse correlation between workload and productivity [168].

Most of recent research on medical diagnosis with DL focuses on predicting characteristics of the patients with neural networks [54]. But this approximation tries to substitute the human task with the prediction of an algorithm.

Among other applications of DL in medical tasks, Clinical Decision Support System (CDSS)

has emerged as one of the most disruptive and applicable techniques in the area [131], [200], [237]. CDSS are any kind of system that supports the professional task instead of completely substituting it. Many efforts have been made to ease diagnostic tasks in medical centers, from the introduction of new modern technology [94] to leverage previous data to perform comparative diagnostics or use XAI techniques.

Comparative diagnosis consists of providing the physician with a set of similar samples when receiving a new case for diagnostic. The recommendation provided to the physician serves as a second opinion to better diagnose the patient. In this way, the physician can compare the new case with previous information that can be used to decide the possible pathologies the patient is suffering. Specially for difficult cases, where comparison can greatly improve professional accuracy. In addition, comparative diagnostic makes it possible to use past knowledge to access the clinical reports of similar cases and use this information to their advantage.

In this context, the focus of this thesis is to research and develop new DL techniques for medical diagnostics. This objective will be divided into different milestones, first new models of CBIR will be developed and studied to achieve a comparative tool that physicians could use to compare difficult cases. One of the main problems that will be studied and discussed along the thesis is the difficulty of providing precise and comprehensive similar cases for a given query. The notion of precision in medical diagnosis is often associated with pathological similarities, i.e. a precise recommendation should share the same pathologies as the input query.

Besides being correct, this association has different flaws. First, when the dataset pathologies are not correctly labeled, this approximation does not work properly. Datasets poorly labeled are common in medicine, because some pathologies are not limiting to the patient and these types of pathologies are most of the time not identified. Secondly, only analyzing pathological similarities between images is not correct because its biased towards classification models.

It is necessary to also evaluate the visual similarity between images. The visual similarity between images should be focused on evaluating the ease of comparison between cases because it is easier to compare images when they are visually similar. This ease of comparison helps professional's diagnostic specially in finer details, where it is possible to compare two cases focusing on the small features that differ one case from the other.

Therefore, a complete CBIR system should focus on both aspects at the same time, visual and pathological similarity. Abstracting the medical diagnosis, the final objective is to develop a model that focus on geometric and semantical features of the images. By capturing geometric features of the images that are being analyzed, the model should focus on images that are easy to compare. However, capturing semantical characteristics of the images will ensure that the images are not only visually appealing, but also interesting from a professional analysis perspective.

In this context, this thesis proposes two DL architectures for medical CBIR the Multi-Output Classification Autoencoder (MOC-AE) and Multi-Output Classification Variational Autoencoder (MOC-VAE). First, the core idea of the research is proposed with MOC-AE, which consists of generating enriched embedding spaces combining the Autoencoder (AE) architecture along with a classifier. The idea is that with the combination of both techniques,

it is possible to construct descriptors of medical cases that store information about semantical and visual characteristics of the patients. The proposal of these architectures follows previous researches that generate image descriptors using AEs [118] and classifiers [161]. The research of the thesis tries to combine both approximations under the same framework, obtaining not only results that exceed previous architectures better balancing visual and semantical features, but also producing more precise recommendations in absolute values.

Therefore, the main contribution of the proposed architectures is based on enriching and filtering the contents stored in the latent space by combining different objectives in the Neural Network architecture. This idea is deeply studied and the results suggest that including more information in the latent space is beneficial for image recommendation tasks. This is probably possible because the information stored in the latent space is force to focus on relevant areas of the image because of the classification output, whereas irrelevant areas, from a clinical perspective, are less prominent in these descriptors.

At the initial stage of the research, MOC-AE is proposed and studied in a comparative manner against other work of the state-of-the-art. To ensure the performance of the architecture, it is decided that comparing with the most developed state-of-the-art architecture will shed light on the real model's performance. Thus, MOC-AE is compared against the work of Kobayashi et al. [118], which uses a similar AE architecture to produce recommendations of the Multimodal Brain Tumor Segmentation Challenge (BraTS) dataset [20], [21], [143], containing tumor brain images with segmentation labels. The results suggest that MOC-AE outperforms the previous work, not only better balancing the pathological and visual attributes but also better detecting tumors in the cases.

Once it is certified that the model performance is, at the very least, comparable with similar approximations, it is decided to deeply study the behaviour of MOC-AE. First, the model is compared against traditional DL architectures of CBIR, in particular AEs and classifiers. These architectures are the base of the construction of the MOC-AE architecture, so it is crucial to compare it with its counterparts. However, an evolution of MOC-AE is proposed to improve the latent space representation.

Because the system is based on the recommendation of similar cases from a latent space, an evolution of MOC-AE is presented in MOC-VAE. This new architecture regularizes the latent space, making it complete and continuous. This is done following the training scheme of Variational Autoencoder (VAE) [116]. With these characteristics, the recommendations that are done in this space are strengthened and improved. The results confirm that MOC-VAE outperforms MOC-AE in medical image recommendation tasks. This second iteration of studying the MOC-AE and MOC-VAE architectures is done using the Padchest dataset [30], that contains chest X-ray images.

One of the main problems with respect to the evaluation of these systems is the evaluation of the performance of CBIR. There is no consensus in the literature on how to measure the results of a CBIR system. Most of the previous works evaluate the performance of their models by using Precision@k, Recall@k or their derivatives [126], which are metrics related to correlation between the query class and the retrieved image classes. As discussed before, this approximation lacks in evaluating visual similarity between cases, which is a crucial aspect

of the recommendation. However, the current possibilities for evaluating visual similarity between images are very poor. Thus, a new metric for evaluating visual similarity between images is proposed in this thesis. The Sliced Wassersterin distance [28] is a variation of the Wasserstein distance [183] that improves the efficiency of the calculations. This metric was proposed to evaluate distances between probability distributions, but to the author's knowledge, it has not previously been used to assess similarity between images in the CBIR context. The Wasserstein distance has previously been used in the comparison of color histograms [164], but never for the full image comparison.

Apart from CBIR, XAI is a technique very popular in the medical area for providing explanation of DL diagnostics [87], [117], [133]. As mentioned above, because DL models are black-box systems, they provide diagnostics predictions without an explanation. The criticality of many medical diagnoses makes it impossible to apply DL to them, because replacing the labor of the physician is not an option. The lack of information of what characteristics is the model using in its prediction makes impossible to apply DL in medical diagnosis, creating rejection of AI by the medical community.

Therefore XAI arises as another CDSS tool that alleviates the diagnostics of physicians without substituting them, contributing with a solution that can help the diagnostic capacity of the specialists. In this sense, the framework proposed with MOC-AE and MOC-VAE can be used for XAI tasks. The simplicity of the architecture makes it possible to apply it to provide complex explanations maintaining the network structure. The idea of leveraging MOC-AE and MOC-VAE for this purpose is to use the dimensionality reduction provided with the network to find explanations in this embedded space.

To do so, the algorithm designed finds explanations in the latent space of the MOC-VAE, because due to its regularization it better preserves features in the embedded space. To find the explanations it is studied the correlation between modifications in the latent space and changes in the classification output of the network. Finally, this changes are translated to changes in the image space by reconstructing the modified descriptors. The result of this framework is an explainable system that is tested with the Padchest dataset [30] and its results compared with the most relevant XAI systems in actuality.

Overall, the main contributions of the thesis can be summarized as:

1. A new model of medical CBIR is proposed. The so-called Multi-Output Classification Autoencoder (MOC-AE) presents a DL architecture based on storing in a low-dimensional space image descriptor that cover both semantical and visual features from the patients. The results show that the MOC-AE outperforms the previous architectures better balancing anatomical and pathological features of the patients, and in pathological similarities.

2. An improvement of the previous model is presented, the Multi-Output Classification Variational Autoencoder (MOC-VAE). The MOC-VAE regularizes the latent space where image descriptors are constructed. To evaluate the architectures in both visual and semantical terms, the Precision@k is used and the Sliced Wasserstein distance is proposed as a visual similarity measurement in CBIR. Results showcase the good results and properties of the proposals in both visual and semantical characteristics.

Furthermore, the results of the MOC-VAE are compared respect the MOC-AE and it is shown that the regularization of the latent space achieves better recommendations in both visual and semantical characteristics.

3. The use of the Sliced Wasserstein distance as a visual similarity measurement in the CBIR context is proposed. The inclusion of this metric evaluates how similar are the query and recommended cases, not by their respective labels, but rather by their geometrical similarities. This correlation is specially important to measure how similar are two cases in the context of medical comparative diagnostic, where it is important that the cases have the same pathology but also that the cases are easy to compare
4. To find explainability to the previous models it is decided to research in XAI solutions. The MOC-VAE architecture is further studied for XAI purposes, because of its regularized space where it is possible to apply linear perturbations to find image explanations, aspect that was not ensured in the MOC-AE latent space. The dimensionality reduction mechanism proposed in MOC-VAE is exploited for finding explanations in a lower dimensional space. Then, using the reconstruction output of the architecture, this explanations are translated to heatmaps of the images. By finding explanations in the embedded space, it is possible to apply linear transformations to the data that can be translated to non-linear explanations in the image domain. This framework is compare with state-of-the-art XAI systems and the results show that the MOC-VAE outperforms previous systems, providing more precise explanations.

1.1 Research Results

This thesis produced several research results, described below:

1.1.1 Articles

During the development of this thesis, several articles have been published in scientific journals. The articles are shown bellow:

- *G. Iglesias*, E. Talavera, J. Troya, et al., "Artificial intelligence model for tumoral clinical decision support systems", Computer Methods and Programs in Biomedicine, vol. 253, p. 108 228, 2024. https://doi.org/10.1016/j.cmpb.2024.108228.
- **G. Iglesias**, E. Talavera, Á. González-Prieto, et al., "Data augmentation techniques in time series domain: A survey and taxonomy", Neural Computing and Applications, vol. 35, no. 14, pp. 10 123–10 145, 2023. https://doi.org/10.1007/s00521-023-08459-3.
- **G. Iglesias**, E. Talavera, and A. Díaz-Álvarez, "A survey on GANs for computer vision: Recent research, analysis and taxonomy", Computer Science Review, vol. 48, p. 100 553, 2023. https://doi.org/10.1016/j.cosrev.2023.100553.

In addition, two articles are in review process:

- **G. Iglesias**, E. Talavera, J. Troya, "Recommendation System for Medical Images Using Visual and Semantic Similarity with Variational Autoencoders", under review in *Medical Image Analysis*.
- **G. Iglesias**, H. Menendez, E. Talavera, "Improving Explanations for Medical X-Ray Diagnosis combining Variational Autoencoders and Adversarial Machine Learning", under review in *Artificial Intelligence In Medicine*.

1.1.2 International Research Stays

In order to establish collaborations with external research institutions, the following research stay was held:

- **Software Systems Group, King's College London** at United Kingdom, from the 1^{st} of March of 2024 to the 1^{st} of June of 2024, under the supervision of Dr. Hector Menendez. The stay served to develop XAI using the architecture of the thesis. Moreover different collaborations were established with members of the group and other institutions of London, specially focused on developing Generative AI and complex DL architectures.

1.1.3 Dissemination of results

In order to promote the research and results of the thesis, the content of the thesis was showcased in the following activities:

- **3 minutes Thesis competition** at Universidad Politécnica de Madrid. The author of the thesis participated in the competition, achieving a place as finalist in the competition, celebrated the 3rd of June of 2024. The title of the talk was *Modelos de Deep Learning para un diagnóstico y análisis de imágenes médicas preciso*. Refer to Appendix .1 for the video of the presentation along with the content.
- **Publication of a news item of the research** titled *Mejoras en la precisión diagnóstica de tumores gracias a la inteligencia artificial* in the *madri+d* portal[1] and the *e-Politécnica Investigación e Innovación* newsletter[2]. Thanks to the *Unidad de Cultura Científica* of the UPM that published an article with part of the results of the thesis, in particular, the research corresponding to Chapter 4. The new can be accessed in https://www.madrimasd.org/notiweb/noticias/mejoras-en-precision-diagnostica-tumores-gracias-inteligencia-artificial and https://www.upm.es/UPM/SalaPrensa/

[1]The *madri+d* foundation is a governmental organization that manages the regional plan of scientific research and innovation in Madrid, it is also responsible for ensuring the quality of university programmes in the region. In the web of madri+d there is a section dedicated to publish the most relevant I+D+i news in Spanish.

[2]The *e-Politécnica Investigación e Innovación* newsletter is a service of the Universidad Politécnica de Madrid managed by the Observatorio I+D+i UPM with news of UPM researchers.

Noticias_de_investigacion?fmt=detail&prefmt=articulo&id=CON14386 for the content of the new and the translation.

Chapter 2

State-of-the-Art

This chapter encompasses the most relevant literature and research that will be covered along the thesis. The research of the thesis is supported by the theoretical background studied during this chapter, to have a wide and specific perspective of how the research is organized in the area.

The organization of the chapter covers the most relevant topics of the thesis going from more global areas to more specific topics and researches. Thereby, the chapter starts covering the state of Deep Learning (DL) in actuality, how it is heavily influenced by generative models and eXplainable Artificial Intelligence (XAI), to then cover more specific details of the thesis. In particular modern Content-Based Image Retrieval (CBIR) systems will be studied, specifically focusing on the relationship of CBIR and DL architectures and their problem with evaluation metrics. Finally XAI and CBIR will be studied in the medical field, to give an overview of how the previously discussed topics have been implemented in medical diagnosis.

2.1 Actual Deep Learning Context

The current paradigm of DL is in constant evolution, specially during the last years [196], [231]. Aspects such as its ability of adaptation and generalization, relatively fast forecasting capabilities and general good performance, have made DL applicable to a wide range of fields [216]. DL is a subarea within Machine Learning (ML), that uses Deep Artificial Neural Network architectures. Its robustness comes from the fact that the models are capable of extracting features of the data directly from the data, without human input. In this sense, DL models obtaining their knowledge from the data and not from human expert knowledge, as previous Artificial Intelligence (AI) systems [120].

The huge success of DL in pattern recognition tasks has positioned AI as one of the most relevant sciences worldwide [110]. Many different areas have applied DL for analyzing large volumes of data [66], [72], [167].

Trying to analyze the impact of DL in all the areas it have been applied would be impossible,

due to the enormous number of researches presented nowadays and the wide variety of areas where it has been applied. In addition, many techniques and approximations to the application of Artificial Neural Networks have been proposed and are still being proposed [32], [69], [151], [155].

One of the most noteworthy areas of DL is content generation [56]. Generative models have been present in AI for a long time, but since the introduction of Generative Adversarial Networks (GANs) [64], their results improved drastically, triggering a revolution in the field of AI. The relationship between generative techniques and any other application of DL is narrow, making possible to adapt many of the advances in this area to any other problem, as it will be discussed in Section 2.1.1.

For the purpose of the thesis, analyzing the Computer Vision paradigm of DL is crucial to understand the evolution and trends in this area. Traditionally, Computer Vision techniques were based on feature engineering methods, that extracted information from the images by manually defining certain characteristics of the images, whereas modern Computer Vision based on DL is focused in feature learning from the data [157].

From academy to industry, many processes have been benefited from Computer Vision techniques and it is considered that one of the future trends of Computer Vision is its application in innovative areas. In this sense, all the advances proposed lately must be leveraged to obtain results and produce value to the society and industry.

Attending to Computer Vision with DL, not only the industry has been benefited from their results. Due to its popularity, DL is democratized and the computational resources are becoming more efficient and powerful over time [44]. Competitions such as the Vesuvius Challenge [1] demonstrate how the society can benefit from Computer Vision. The Vesubius Challenge is a competition that was focused on applying ML techniques to decrypt texts from unopened Herculaneum scrolls. This scrolls are papyri discovered after the eruption of Mount Vesubius in 79 Anno Domini. Thanks to Convolutional Neural Networks (CNNs) architectures, researches were capable of retrieving the content of some of the papyri[2] [3]. This type of advances achieved with the use of DL push further away the social conception of what AI can do, and this interest in the field create more advances in the technique.

Medical image processing has also applied DL techniques during the past few years [244]. By leveraging AI, many medical processes can be optimized to automatically extract information from the images. The final diagnostic of any process must be done by a human, but many procedures can be improved by the use of DL techniques, such as triage [97], [232] or comparative diagnostic [3].

All these recent advances have been motivated by new techniques in Computer Vision that can be leveraged to improve previous results or achieve new objectives. For example, since the introduction of Transformers [223] many works have been published using the Transformer architecture [13], [91]. This relationship between technical advances and implementations is, to a certain extent, responsible for the huge popularity of DL. These technical advances have

[1] https://scrollprize.org
[2] Luke Farritor GitHub repository: https://github.com/lukeboi/scroll-first-letters
[3] Youssef Nader GitHub repository: https://github.com/younader/Vesuvius-First-Letters

make possible improvements in the results of AI, which in turn generates new advances in the technique, and this process is continuously fed back.

In this sense, one of the areas of most popularity, and that have impacted more in the development of DL are the generative models, which development has heavily impact the way architectures are designed.

2.1.1 Impact of Generative Models in Deep Learning

Generative Artificial Intelligence is not a new concept, but in the recent years it has gained a lot of popularity [59]. The main difference between generative models and previous discriminative models is that generative models try to capture the probability distribution of the data they try to replicate, whereas discriminative models try to extract features from this data. Discriminative models try to calculate the probability of some feature Y given some prior information X, that can be denoted as $P(Y|X)$. The idea of this approximation is that it is possible to extract posterior information Y using prior data X.

On the other hand, generative models try to capture the probability distribution $P(X)$ of a dataset X to generate new synthetic samples that follows the same distribution. This approximation changes completely the way that the models work. In this case, the model needs to capture how the data is formed, by extracting the complex relationships between the different attributes of each dataset. If the model have correctly extracted this information, then it is possible to generate new samples by using the probability distribution previously defined and generate samples that belongs to that probability distribution. But this process is not trivial.

One of the first models that already tried to generate new content before DL are the Gaussian Mixture Models (GMMs), that try to assign the probability distribution of the observations of a given dataset by defining a set of different gaussian probability distributions. By combining the different distributions, the technique is capable of translating the prior information of the samples of the dataset, to a posterior probability distribution. GMMs have been used for clustering tasks [140], [165], [242], but it is possible to generate new information leveraging the information extracted by the model.

Autoencoders (AEs) are the first approximation of generating new data using artificial neural networks. To be able to produce new samples, AEs generate points in an embedded space that can be translated to the original space of the data. The process of translating the embedded information to representations in the original data distribution is done by the Decoder network of the architecture. On the other side, the Encoder network is responsible of producing an embedded space that is structured in a way that is possible to recover information of the original samples. I.e., the Encoder focus on the representation of the samples while the Decoder on the translation of these representations to samples of the real data distribution.

Later, Variational Autoencoders (VAEs) [116] introduced a regularization method for the embedded space. By regularizing this space, it is ensure that it meets certain characteristics, in particular the embedded space is forced to be continuous and complete. With this

continuity and completeness, when new data is sampled from this embedded space, the reconstructed information belongs to the same probability distribution as the original dataset. The regularization method of the VAE was specifically designed for improving content generation tasks, because it ensures that the seed used for generating new content belongs to the same distribution as the embedded space of the real samples. With that, the generated samples have higher quality. In addition to this analysis of the AE and VAE in the context of generative models, both architectures will be technically analyzed in Section 5.2.2 to overview their architecture and training scheme.

The next evolution in generative models was presented with the GANs [64]. This architecture, specifically designed for generating synthetic data, is based on the interaction of two networks. The Generator network is responsible of generating samples from a latent space, where the Discriminator network evaluates the *realness* of the generated samples by comparing them to the real dataset. The Discriminator is a binary classifier that distinguishes between the real samples from the dataset and the generated samples of the Generator, and it is trained with a binary classification scheme. The Generator, on its side, is trained by generating new samples and receiving feedback from the Discriminator classification output, thus evaluating if the synthetic sample is capable of fooling the Discriminator.

The architecture design presented in GANs have been crucial for research in DL. Many different researches have focused on improving GANs with different approximations, from changes in the loss functions [16], [24], [103], [137], [149], architectural changes [29], [107]–[109], [113], [148], [158], [174], [238], [245]. One may consider these changes specific for content generation purposes, but many problems have applied GAN specifically for different objectives.

For example, in the medical area, Kadurin et al. used GANs for drug discovery purposes [101], [102], using biological and chemical datasets. In this research AEs was used in combination with the adversarial training scheme of GANs to create anticancer compounds that satisfied specific properties. This combination of different techniques and architectures of the state-of-the-art in DL is a very common and successful approximation.

Another case where the GAN architecture have been used for different purposes is in image segmentation, where models such as the Cycle-Consistent GAN (CycleGAN) [245] or GANILLA [78] have been used to translate an image from the domain of real images to segmentation masks. This approximation is very similar to the one proposed in the Diffusion Models [81], that combines the architecture of AE with probabilistic processes of Markov chains to generate content.

Hybridization of Generative Models in Deep Learning

Hybrid of models are becoming the standard for many different researches. Ideas from one architectures can be leveraged for different purposes, combining mechanisms to empower the architectures. At the core, all the DL models are feature extraction models that are capable of extracting patterns of the data they are fed with, and produce different results depending on the training. Thus, it is possible and common to combine different techniques.

This is specially important for the purpose of the current thesis, where it is considered

that, for developing new systems, the best approximation is a good understanding of the state-of-the-art to then propose new architectures. In particular, as it will be discussed in Section 2.2, recommendation systems are based on factorization processes, where extracting information from the data and generating feature descriptors is the most important part of the framework.

In this sense, as it has been mentioned, generative models base their training on being able of factorize the information to capture the probability distribution of the dataset. This factorization process can be adapted to image recommendation, working on a similar embedded space to achieve a correct representations of the cases of the recommendation.

For example, one of the most important modifications to the base GAN architecture [64] are the Conditional GAN (CGAN) [148] and Auxiliary Classifier GAN (ACGAN) [158], that takes the base GAN architecture and change its behavior by adding new inputs and outputs. For producing conditional outputs with GANs, these architectures condition the output of the Generator by adding a new output that marks the class of the generated sample. This is particularly interesting, because this change in the behavior of the model changes the internal representation of the latent spaces used by the architecture.

Other approximations used AEs modifications to achieve different purposes, from VAEs [116] to Diffusion models [81] there have been many researches that explore the possibilities of AEs internal representations. E.g. Latent Diffusion Models [182] are an approximation to better preserve internal representations of Diffusion models by a preprocessing step that factorizes the information. GANs have also introduced AEs as a core part of their architecture, in works such as the Pix2Pix [92].

Summarizing, it is clear that the current trends in DL are heavily marked by generative models. The current context of AI is of constant combination and hybridization of ideas and concepts from different models. Regardless of the problem faced, many concepts can be adapted to different fields. This properties of DL are considered crucial for the development of new architectures, where previously presented ideas can be leveraged for different problems.

Another very influential area of the last years, along with generative models, is XAI. Specially when the DL models have to be implemented in sensible areas such as autonomous driving or medicine.

2.1.2 eXplainable Artificial Intelligence

XAI is the field of study in AI focused in finding explanations for the predictions of black-box systems [15], [163], [183]. Traditional AI algorithms, such as Decision Trees [173] or Support Vector Machines [42] are usually interpretable. But DL models, that are the state-of-the-art approach in modern AI [243], are black-box models, where there is not a direct interpretability of the decisions made by the model.

The fast evolution of DL [231] has positioned Artificial Neural Networks as the preferred system for ML. But the application of these models in sensible areas such as autonomous driving [50], [239], defense [25], [233] or medicine [10], [82] is heavily limited by the lack of

explainability of these systems. XAI provides trustworthiness to sensible areas where failure is not an option [138].

There are plenty of different approximations to define causality in DL, but current literature focus its efforts in causality through counterfactual interventions [202]. The idea of these interventions is to study the relationships between changes in different parts of the AI models and the predictions they make. I.e., it measures how a model's output changes when either the input or an internal representations of the data are changed. This correlation between changes in the data and changes in the predictions is translated to causality. In this sense, this thesis will be focused on counterfactual representations of the data to find explanations.

XAI systems aim to provide explanations of their outcomes [217]. In Computer Vision, XAI obtains a visual map of the areas of the image primarily used to produce the model's output [18]. XAI outputs should explain why the model is reaching a certain prediction, yet there is no consensus about what explanations should be [14].

Typically, there are two main XAI approaches depending on their application, named model-agnostic and model-specific explanations. Model-agnostic methods do not assume a particular model for giving the explanations, while model-specific methods rely on the algorithm used to generate the explanations [45].

Concerning the type of explanation provided by the model, it is possible to differentiate between local and global explanations [191]. Local explanations focus on a particular input sample. Global explanations provide a general view of how the model is making its predictions, trying to disentangle its logic. Multi-Output Classification Variational Autoencoder (MOC-VAE) provides XAI model-specific local explanations.

The current explaniability systems produce their explanations by applying linear modifications on the samples [139], [226]. These systems are usually referred as *local linear explanation* systems, but they are also called *additive feature attribute* methods [135] or *feature importance* models [135]. By combining different linear transformations to the data, these systems study the impact of the transformations on the model's output. In other words, local linear explanations study the weight of each feature through the model's output. Assuming that the feature space has a total of F features, a local linear modification is defined by [14] as:

$$g(x) = w_0 + \sum_{i=1}^{F} w_i * x_i \tag{2.1}$$

where g denotes the set of modifications, x is the sample being explained and w the weight of each feature i, with w_0 as parameter for balancing the weights of the features.

For constructing more complex explanations, different modifications must be tested, and then combined to obtain a unique combination. Therefore, the modification process must be performed several times for each explanation.

Explainability in Computer Vision aims to obtain a visual representation of the regions used to predict certain features. In DL, it is common to employ model-agnostic explanations [235]. Respect Computer Vision, the features that must be modified to obtain the explanation are pixels. The huge dimensionality of the images makes it difficult to find the weight of each

feature. Particularly, high-resolution images make very difficult to find the correct pixel's weight. To address this problem, some approximations reduce the search space by defining patches in the images [207]. This is inefficient, considering that the search is random along the dimensions on the image, very sensitive to aspects such as shape, size or position of the patches, and it commonly relies on a combination of different possibilities. This leads to instability in explanations and high computational cost [14].

Different approximations have been proposed to improve the search performance. The Local Interpretable Model-agnostic Explanations (LIME) method [181] creates a neighborhood area around the sample that it aims to explain. Each point in the neighborhood represents a linear perturbation over the input point. Then, an interpretable model is trained with this set of samples, aiming to predict the output of each sample. The explanation is obtained with the interpretable model over the set of perturbations, obtaining which features are more related with the output.

SHapley Additive exPlanation (SHAP) [135] is one of the most used XAI methods nowadays [8], [27], [49]. It uses the Shapley values [195] to explain the weight of the different features on the final prediction. Each attribute is evaluated individually by decomposing the output using the additive feature attribution methods. With that, it measures the importance of each attribute over the prediction, thus the explainability that each feature gives.

MultiReX [38] defines different grids over the input image, following a procedure similar to Structured Attention Graphs (SAG). Over these grids, a set of *superpixels* (i.e., group of contiguous pixels) provide a visual explanation. These *superpixels* are found by ranking the pixels with any pixel ranking algorithm, e.g. SHAP, and then search around the surrounding area of the pixels with a flooding-draining approximation. Thus, the resulting explanations are the minimal area of the most relevant pixels, ensuring a precise output.

One of the main drawbacks of all these algorithms is that, in order to work properly, they need to be configured defining a set of hyperparameters. This configuration heavily impacts in the performance of the model.

Tunning Explanations in eXplainable Artificial Intelligence

In order to provide an adequate set of explanations, algorithms can be configured through a series of hyperparameters. Aspects such as number, size, shape or position of the explanations, etc can drastically impact on the system's output.

Some works, like [207], find their solutions using different combinations on the patches of the linear modifications. These patches are occlusions of the input image. The main drawback of this procedure is that it covers a wider range of explanations while increasing the computation time. Furthermore, it should be noted that the computation time increases exponentially with the number of configurations that are tested.

SAG [199] reduces the number of hyperparameters for the explanations by pruning the search to only the most relevant patches found. MultiReX further reduces the number of hyperparameters only selecting patches that maintain the original prediction, following the

same procedure as LIME [181] and SHAP [135].

But all these approximations are sensible to the number of explanations hyperparameter. Therefore, the design of a system that is capable of providing a set of explanations reducing the number of hyperparameters is desirable.

2.2 Content-Based Image Retrieval Systems

With the evolution of the available information of the last decade, specially in medical data [86], new methods of navigating through the data must be defined [36]. Nevertheless, since the appearance of large datasets, a way of filtering the data had to be defined. For structured data, or in cases where meta-information of the data is available, this filtering process can be applied using the meta-data information. But for unstructured data, such as images, this type of filtering is not applicable.

CBIR provides tools that automatically defines specific features to these images, making possible to define relationships between images, easing the search through the datasets.

CBIR systems are designed to retrieve the most similar images from a given database in response to a specific query. These systems work by taking a particular image as input and extracting its most relevant features. Subsequently, these features are compared with those of other images in the database, ranking their similarity to the query image. The process involves the analysis of various aspects of the images, such as color, texture, shape, and spatial distribution, to determine similarity. Through this mechanism, CBIR systems facilitate efficient and effective image retrieval, finding applications in diverse fields such as content-based image search engines [61] and medical image analysis [46]. Constant advances in ML and Computer Vision techniques continue to refine and optimize the performance of CBIR systems, enhancing their accuracy and usability in various domains [104].

CBIR has been applied to different areas of Computer Vision and AI [130]. Due to the flexibility of its general structure, these systems have been applied in areas ranging from e-Commerce [68], [241] to face recognition [85]. These systems make it possible to use large amounts of information in an ordered manner, taking advantage of the information available without using it manually [105].

Different approximations have been applied in CBIR. Traditionally images where characterized by color, textures or shapes, using descriptors such as Gray Level Co-occurrence Matrix (GLCM) or Local Binary Pattern (LBP) [11], [34]. Recently, Artificial Neural Networks have emerged as a feature decomposition method for generating image descriptors [189], achieving better results [193].

2.2.1 Traditional Content-Based Image Retrieval

Traditionally CBIR was a method of navigating through large datasets when the data was unstructured. In these cases, defining features of the data was useful to query images in an

effective manner. The first methods used basic similarities of images to compare one with each other [22].

Photobook [166] was one of the first content-based search engines. Its focus was on searching through a database of images by the content they had, instead of previous approximations that indexed the information relying on text annotations. To do so, for each image a descriptor was defined, the descriptor allowed to represent the images by their content, making possible to filter and compare the descriptors semantically. Three types of descriptors are defined in Photobook: appearance, 2-D shapes and textures. The appearance descriptors are defined by comparing the variance between images of an object and the *prototypical appearance* of the object, through a set of parametric variations defined with eigenvectors. The 2-D shapes descriptors are defined with the finite element method that produces a stiffness matrix, this matrix describes a relationship of points of interest between images. Finally, the texture descriptor decompose each image in regular stationary stochastic processes, called *Wold decomposition* [132].

Query by Image Content (QBIC) [55] produced their recommendations using R* trees [67] of different features of images and videos. To obtain these features different approximations where applied, for lower dimensional data, e.g. average color or texture, the R* tree was applied directly to the data, but in larger-dimensionality data, e.g. shape features vector, a dimensionality reduction algorithm such as Principal Component Analysis (PCA) [136] is applied first.

Chatbot [159] is a CBIR system that combines the color and shape feature definition concepts of QBIC in relational databases. In this case, because the data is structured, recommendations can include meta-information about the items that are being search. The images stored in the database include text information, that is leveraged along with content-based features of the data to perform the so-called *concept queries* that the user can use to produce more advanced searches.

More modern CBIR systems use DL as the feature extraction method. The main benefit of usign DL is that the features extracted from the data are not manually defined, but rather explored and found using ML, mitigating possible bias and improving the performance of the models.

2.2.2 Content-Based Image Retrieval with Deep Learning

CBIR systems are focused on processing large amounts of data and AI algorithms arise as the best solution to this problem. Different solutions have been proposed in recent years, combining CBIR with DL techniques [118], [203]. The latest studies in this area suggest that DL techniques outperform previous traditional algorithms in diagnosis [179].

Sundararajan et al. [208] applied Deep Belief CNNs in avascular necrosis diagnostic. This work showcases the desirable properties of CBIR in medical tasks, assisting physicians when the diagnosis is intricate with similar cases. The second opinion provided by the system is specially useful in scenarios of inexperience, subjectivity or tiredness.

The work of Khan et al. [111] used genetic algorithms to produce the embeddings of the images. This work tested their approach in different datasets, from realistic images of the CIFAR-10 dataset [123] to a medical dataset, known as the Kvasir dataset, that contains 4000 gastroscopy images, belonging to 8 different classes. This research demonstrated the high transferability of CBIR, being able to produce very similar results in different domains, specifically obtaining a mean precision of 0.916 in CIFAR-10 and a precision of 0.913 in the Kvasir datasets.

In Owais et al. [161] a neural network CBIR architecture is presented. This research uses the last layer of a CNN classifier to generate the image descriptors. In this work, the flattened vector that is the input of the fully connected output of a classifier is used as a feature descriptor. The intuition is that the classification features of the images will be represented in this vector, able to represent similar cases closer in the latent space. The same architecture was used in [3], [17] where the classification output was used as a feature extraction method for chest X-ray images.

Kobayashi et al. [118] present a more complex feature decomposition Artificial Neural Network based on the AE architecture. The AE is capable of representing each image in a lower dimensional latent space. This work focuses on disentangle the normal and abnormal features of the patients using different segmentation outputs, generating different latent spaces for the different features. By including certain objectives in the architecture, in this case the segmentation of brain tumors, the latent space is forced to maintain desired characteristics, therefore recommending cases focusing on those features. Later, Sudhish et al. [205] proposed to use CNN along with selection of features with random forests, improving previous results in the same dataset.

Depending on the architecture, different measurements of the similarity of the descriptor can be used. In traditional feature extraction with textural features, the Manhattan distance has shown the best results [154]. However, when using neural networks, the most common distance used is the Euclidean distance [118], [161], [205], because it considers each position of the descriptor equally relevant. When using Artificial Neural Networks, it is not possible to control the latent space of the network. In this sense, the Euclidean distance is the best solution when the latent space is not controlled. Nevertheless, in [208] modified Hamming distance is used because the latent space is composed of binary codes. Hence, unless the latent space is specifically controlled, the Euclidean distance is the most common distance.

One of the opened problems in CBIR is evaluating its results in recommendation tasks. The evaluation metric problem is not a problem that only CBIR systems have, generative models also suffer from the lack of a precise and insightful evaluation metric.

2.2.3 Evaluation Metrics in Content-Based Image Retrieval

Unstructured information retrieval is a complex task where the definition of a particular evaluation metric is not a trivial task. In literature, there is no consensus on how to evaluate the CBIR performance [126]. There are a set of desirable properties of the recommendation that must be fulfilled with each CBIR model, but besides that, different metrics can be used.

Most of the metrics in this area are somehow related with classification metrics, i.e. they measure the similarity of the retrieved images with respect to the query image by their labels. E.g. two images that belong to the same class are supposed to be more similar than two images that are from different classes.

But this approximation has some problems. First, measuring the recommendation with the labels of the images would benefit to systems that act as classifiers, being inherently biased towards classifiers, besides not being its purpose. In addition, these metrics do not evaluate visual similarity between images, which can be crucial for specific tasks. In medical comparative diagnostic, as it will be discussed in Section 2.3.1, a correct visual similarity between images will ensure that the system provides the physician a comprehensive set of samples to compare. A system with poor visual similarity between cases will recommend images that are very difficult to compare with. Moreover, if the data is not correctly annotated the results of this family of metrics will be heavily influenced by the incorrect labels of the dataset.

Therefore, metrics that measure the visual similarity between images must be used in CBIR to compare the results of the systems. But measuring visual similarity between images is a very complex tasks with not a pre-defined metric. Analyzing an image can be done from many different perspectives, textures, shapes or color intensity between others, but depending on the problem this comparison can be problematic. Image comparison suffers from many different problematics, such as invariance in shapes, color, sizes, rotation that heavily impact the results of the measurement.

Besides that, different approximations have been followed in the state-of-the-art of CBIR. In particular, Precision@k and Mean Squared Error (MSE) will be studied, for being the most common ones and the ones that better adapt to the medical comparative diagnostic paradigm.

Precision@K

As mentioned before, many different classification-related metrics have been used in CBIR, such as Precision@k, Recall@k, F1-score@k or Average Precision and Mean Average Precision [126]. All of them have in common that they compare the class of the query image respect the classes of the k retrieved images from the database. Ultimately, they measure the correlation of the classes of the images retrieved.

When CBIR is applied to image comparison a small set of samples is used, sufficiently big to provide variation in the comparison and sufficiently small to not overwhelm the user. Thus, metrics such as Recall@k or its derivatives, that takes into account the number of relevant images that are not being retrieved by the system, are not useful. Specifically, because they are meaningful when the number of retrieved images of a specific class is similar to the total number of images oh the class. Otherwise it would produce very small values for small k values because the large number of images that are not retrieved, or very large numbers for large k values because the small number of images that are not being retrieved. Thus, Recall@k, F1-score@k, Average Precision and Mean Average Precision will be discarded for using the Recall@k in their computations.

In this sense, Precision@k measures the accuracy of the system respect the retrieved images, and not respect the images that are not being retrieved. For the current context it is crucial to determine the precision of the retrieved images, because it measures the accuracy of the images that the physician use to perform the comparative diagnostic, as it will be explained in Section 2.3.

In order to measure how semantically related is the query image with respect to the retrieved images of the system, many works measure the accuracy of the system by comparing the label of the query with the k ranked cases. The formulation of the label precision can be computed as follows:

$$Precision@k = \frac{TP}{TP + FP} \tag{2.2}$$

where true positives (TP) correspond to the number of images retrieved by the system that match the same label as the query, false positives (FP) correspond with the retrieved images that do not match the same label as the query and k corresponds with the number of samples retrieved.

This metric can be seen as the percentage of success of the system when retrieving images and attending to their label. Many works used this metric as their evaluation metric [4], [111], [130], [169], [224], [240].

This metric is important for measuring the semantic feature representation of the retrieval system, where a high correlation between the query and the retrieved samples is desirable. But using only this metric as an unique overview of a CBIR model's performance is incorrect, because it does not measure how similar the images are. Using only precision or recall would lead to a poor evaluation of the model.

Mean Squared Error

Measuring how similar are the retrieved images with respect to the query of the system does not only involve calculating if they represent the same class of image, in addition, it is necessary to asses their similarity in appearance. Therefore, a metric that compares image pairs and measures how similar they are is needed.

MSE is the most common loss used in AE architectures [144], [145], [170] which is one of the most common architectures used in CBIR. It trains the network to recover the original image after reducing its dimensionality. MSE is computed as follows:

$$MSE = \frac{1}{N}\sum_{i}^{N}(Y_i - \hat{Y}_i)^2 \tag{2.3}$$

where Y denotes the query image, $\hat{Y}$ with the retrieved image and i denotes each pixel of the N pixels of the image dimensions.

MSE is able to compare images pixel by pixel, capturing similarities in visual structure. Nevertheless, comparing each pixel in the same coordinate between two images has very bad

properties with respect to invariances in rotation, size, and deformations between others. Therefore, this metric is not appropriate to measure visual similarity between images.

A good visual similarity measurement is an open problem in CBIR, but it is considered a crucial aspect of the model performance analysis. In Section 5.2.3 it will be proposed to use the Sliced Wasserstein distance as a visual similarity measurement for CBIR, due to its good properties against deformations.

Once the state-of-the-art of DL and CBIR systems have been studied, it will be analyzed the medical diagnosis area using AI. The previous knowledge will be leveraged to apply the most advanced researches in the medical area, achieving a complete research because its area of application but also for the advanced algorithms used in the development.

2.3 Medical Diagnosis

Technological progress in the medical field has allowed enormous amounts of clinical and medical imaging data to be generated. However, accurate and rapid interpretation of this vast information remains a significant challenge for healthcare professionals [47].

The techniques used in medicine to treat pathologies are becoming more advanced each day and can perform less invasive and more efficient treatments [75]. However, a correct disease diagnosis is necessary before a patient can apply proper treatment, which becomes an important task in this process [236].

Although medicine has radically evolved, diagnosis is still mostly a human process. The expert must be able to thoroughly evaluate the patient's evidence and avoid making mistakes during the process [100]. A late or incorrect diagnosis can lead to an increase in pathologies that, in cases such as cancer, can be fatal and irreversible [175], [213].

Therefore, any improvement that a physician can make during the diagnostic process may be crucial and necessary to greatly improve treatment results, because early diagnosis improves treatment results, as it improves the result of the procedure [71], [114], [206].

Currently, computer algorithms are specially relevant, since they can be used as an additional tool for decision making, known as Clinical Decision Support System (CDSS) [6], [179]. It is precisely in the current paradigm of health information science where these CDSSs can be used to process in real time large amounts of data [209], using the available information to ease the physician's diagnostic process.

However, it is not widely considered that AI should replace health professionals in tasks they normally perform [83] but instead assist them in diagnosis and decision making, considering that the human professional will always make the final decision. Aspects such as lack of explainability [5] or precision [160], or inability to deal with outliers suggest that it is probably too early to completely delegate critical tasks to an AI, at least in the foreseeable future [122].

There are efforts to tackle the main problems of AI, such as the lack of explainability using Explainable AI with works such as [52]. In this sense, the most relevant literature [7], [74],

[227] does not seek to provide a standalone diagnostic of a certain case, but to provide tools to the physician to help in their diagnostic. CDSS arises as other solution to help the physician to use the latest advances in AI.

One of the techniques in which AI has recently achieved better results is in Computer Vision [65]. Here, these algorithms can use medical imaging data to extract and process its information to help professionals perform certain tasks [53], [230]. Thus, this relationship between Computer Vision and medical imaging has led to the emergence of works combining the latest advances of AI in image feature extraction and medical data [95], [106].

Different works published in the last decade use ML techniques to improve the doctor's diagnoses [54], [71], [172], [177]. Using the latest AI research and applying it to the medical field, different works have obtained impressive results in tasks related to the medical field [60].

Regarding medical imaging algorithms, CDSSs are a prolific area where, in recent years, many articles have been published [153], [178], [180], [220]. These techniques combine the potential AI can achieve in extracting the most important features of medical images and support systems that provide the doctor with the most relevant information in each case [96].

2.3.1 Comparative Diagnostic through Content-Based Image Retrieval

Decision support systems based on CBIR are a crucial tool [46] for CDSS in the medical area. It is considered that one of the most interesting innovations that AI can provide to medicine is comparative diagnostic through CBIR. Most of the current AI models used in medicine seek to provide classifications [71], segmentations [71] and other kind of diagnostics [134]. CBIR offers a valuable alternative by avoiding direct feedback on a patient's pathology, instead facilitating the search and retrieval of similar cases based on visual features of medical images. In this sense, having a comparative system that supports the physician by making possible to compare different cases is considered beneficial for the medical task, alleviating complex diagnosis without being an intrusive tool.

CBIR arises as a comparative diagnostic support tool, where physicians can compare similar cases using the system ranking [3]. One of its main strengths, especially important in medical tasks, is that its recommendation is not a direct prediction or classification, but rather a recommendation of similar documented cases. Therefore, the system aims to act as a second opinion on the case, avoiding intrusiveness in the diagnosis. Moreover, CBIR systems do not give a specific output for each case, rather they leverage previous documented cases to support the diagnostic of professionals. This makes CBIR the perfect tool for sensible tasks, where the final decision is taken by a human. In this sense, comparative diagnostics works as a filtering tool to select the most interesting cases while taking advantage of the processing speed provided by computers.

Previous researches such as [129] recommend images using the nearest neighbor algorithm over textural characteristics of the images, as was done in other research independently of medical imaging [129], [194], [203].

The work of [215] proposes a simple feature extraction method using Support Vector Machine (SVM) [42] and the Grey Level Co-occurrence Matrix as the main input of each image. A similar approximation was followed in [124], where SVM is used as the descriptor generator mechanism, in this case using different features of the liver images. Finally, using the weighted nearest neighbor [43] a query classification is produced. The presented architecture differs from this approximation in the usage of the whole image as the input for the model. By using the Grey Level Co-occurrence Matrix in [215] the textural information of the image is maintained, but it could lead to huge information loss. Using the complete image provides the model all the information from where it can learn the relevant features of each image.

Probably the most advanced work in medical CBIR is the work presented in Kobayashi et al. [118] proposes a CBIR scheme based on AEs to extract the most important features of brain tumor images. This work uses three different AE that generate three different image descriptors, one reconstructs the healthy features of the image, other the tumor area, and the last one uses the information from the entire image. Using these different outputs, researchers can disentangle the normal and abnormal characteristics of the query to provide a controlled recommendation. This work combines DL techniques and traditional medical CBIR to recommend similar database images given a certain query.

Another area of research in robust medical diagnosis is finding explanations for DL models. XAI algorithms have been applied in medical diagnosis to find explanations that the physicians can use to improve their diagnosis, not fully relying in the algorithm, but rather using it as an empowering tool that helps them focus on the interesting features of the patients.

2.3.2 Explainability of Deep Learning in the Medical Field

Good explanations for medical diagnosis must represent anatomically correct areas of the pathologies being diagnosed. But the search of these explanations is not easy, specially when analyzing large images. Highly dimensional search spaces are typically reduced by using patches or superpixels that group features. These strategies focus on search space optimization in Computer Vision explanations.

XAI is a very prolific technique in medical AI applications [217], [221]. Because medical requirements, specific diagnostics require interpretability and explanalibility of the models they use [9]. Besides its good results, the main problem of applying DL in medical diagnosis is that the models are black-boxes where it is not possible to find causality between the data and the predictions [222]. In particular, XAI has been widely used in human CDSSs to provide explanations to specific diagnostics [117].

In Knapivc et al. [117] explainability techniques were applied to video capsule endoscopy. The researches found explanations of CNNs models by applying different XAI techniques. In particular LIME [181], SHAP [135] and Contextual Importance and Utility (CIU) [57] where used to provide model-agnostic explanations of the predictions of the models. Results of the research suggest that CIU provided clearer and faster explanations with respect to LIME and SHAP.

Hussain et al. [87] applied XAI to classification models in breast cancer diagnosis. Different DL models where tested for this purpose, in particular Visual Geometry Group (VGG) [201], ResNet [76], ResNeXt [234], DenseNet [84], SqueezeNet [88] and MobileNet-v2 [188]. The explanations were obtained using Grad-CAM [190] and LIME [181] to visually explain the features of the images that are being used in the diagnostic. In addition, authors produce explanations of the classification outputs of the networks using t-SNE [79] and UMAP [141] visually analyzing the factorization properties of the DL models. This works showcases the narrow relationship between XAI and dimensionality reduction methods, providing an overview of the semantical properties of embedded spaces, that can be further leveraged for explainability.

Liz et al. [133] applied Grad-CAM [190] to chest X-ray images of the Padchest dataset [30]. Researches created an ensemble of DenseNet-201 [84], EfficientNet B0 [212], Inception [211], InceptionResNet [210] and Xception [41]. Results of applying Grad-CAM produce visual maps of the diagnostic of the ensemble and, in addition to the visual map, a probability distribution of the different diagnostics and the agreement of the different models of the ensemble is produced.

Chapter 3

Problem statement

This chapter analyzes and establishes the bases of the research of the thesis. The limitations of previous studies will be examined to delimit the scope and purpose of the current thesis. Then the objectives of the thesis will be clearly presented, these objectives will be reflected in the structure of the document, that answers a series of Research Questions. Finally a chronogram with the development of the research will be presented, this chronogram captures the evolution of the research, along with the articles associated with it that answers to the previously defined Research Questions.

3.1 Limitations of Existing Research

Previous limitations in the researches mark the current boundaries in the techniques analyzed during the thesis. From the state-of-the-art analyzed in Chapter 2, it will be identified a series of limitations that define the current gaps in knowledge. The following limitations were found in the state-of-the-art and motivates the research of the current thesis:

L1: Costly labels in Content-Based Image Retrieval (CBIR) systems: The most interesting CBIR system applied to medical comparative diagnosis it the work presented by Kobayashi et al. [118]. Their architecture used segmentation labels to train an Autoencoder (AE) that extracts anatomical and pathological information of the patients. One of the most important limitations of this work is that it needs segmentation labels to be able to train. This aspect heavily limits the range of applications where it can be implemented. Data is specially scarce in the medical area, and segmentation information is very difficult to obtain. Thus, the final application of their architecture is limited by the availability of this type of data.

L2: Insufficient evaluation metrics for medical image recommendation: Evaluating CBIR systems is a complex task, the final objective of CBIR in medical comparative diagnostic is providing a set of samples both easy to interpret and meaningful. Evaluating these aspects is a complex task, where previous systems focused on just evaluate semantical correlation between images [4], [111], [130], [169], [224], [240]. This

approximation forgets evaluating if the images retrieved are similar to the query. It is crucial to define a method of measuring the visual similarity between cases, since the comparative diagnosis will be done by comparing the images and, the more similar they are, the easier the physician's analysis would be.

L3: Linear explanations of eXplainable Artificial Intelligence (XAI) systems: After analyzing previous researches many of them use different XAI techniques for obtaining explanations of medical diagnostics [87], [117], [133]. It is considered that there is still room for development in XAI. One of the main limitations of previous works is that they rely on defining a set of patches or superpixels to find linear explanations in the images. This limits the shape and size of the explanations found and bias the result with the hyperparameters defined for each case.

L4: Hyperparameter tunning in XAI: XAI algorithms need to define a set of hyperparameters that controls how the explanations are found. These hyperparameters controls the behavior of the algorithm, potentially reducing the effectiveness of the system, by manually controlling aspects such as the number of explanations.

Once the limitations in the field are clearly identified the thesis will be framed in this context. Taking into account previous limitations, the thesis main purpose is defined by the problem that is going to be tackled in it. This problem will be disseminated into different Research Question (RQ).

3.2 Problem Statement

The research presented in this thesis focus on implementing new methods of medical diagnostic leveraging the potential of Deep Learning (DL). Several options are studied during the development of the research, putting the focus on the particularities of medical diagnosis because of the nature of the problem. As stated before, medical diagnosis is a sensible task and the final decision of the physician will never be fully replaces. This is why it is considered a better option to define comparative or explainable systems instead of the traditional predictive DL models that provide discriminative outputs.

For these reasons it is considered that the best solution is to explore alternatives to these traditional discriminative models. One solution is the comparative diagnosis, that provides to the physician a set of similar cases when analyzing a new patient. This way, the professional is able of comparing the cases and making a more informed decision. In this case the labor of the physician is not replaced, but rather empowered. Comparative diagnosis through CBIR is not a very developed field of study, with few researches exploring its possibilities. Therefore, it is considered very interesting to develop new models of CBIR focused in medical diagnosis.

In addition, other possibility to overcome problems of discriminative DL models is applying XAI techniques that find explanations in the discriminative process of black-box systems. This alternative have been studied more times in medicine than comparative diagnostic [4], [111], [130], [169], [224], [240]. Hence, a great alternative to study the possibilities of the field is to develop new algorithms of XAI that focus on finding non-linear explanations using an

embedded space. The non-linearity of the explanations makes possible to develop models that do not require a predefinition of hyperparameters, reducing the human tunning and possible biases introduced in the hyperparamerers.

Thus, the main objective of the thesis is to research novel DL computer vision architectures for Clinical Decision Support System (CDSS) in medical image diagnosis. Because of the particularities of medical diagnosis, the models must not substitute the physician diagnostic, instead, offer assistive systems for their diagnostic. This main objective is divided as follows:

O1: Development CBIR systems for medical comparative diagnostics: By leveraging novel DL architectures, the research of the thesis will put the focus on how to improve previous CBIR systems in the medical diagnosis. To overcome previous limitations the proposed architectures capture semantical and visual information of the images to balance anatomical and pathological characteristics of the patients.

O2: Apply DL computer vision architectures for XAI to find explanations of medical images: The proposed architectures will be applied for finding interpretable explanations of the cases. Using the factorization characteristics of the models, explanations are found in a latent space from where saliency maps are constructed.

3.2.1 Research Questions

The problem stated in this Section 3.2, along with the state-of-the-art limitations of Section 3.1, proposes a research defined in Section 3.2.2. To achieve the goal defined with the research, different objectives have been proposed in the following RQ that will be addressed and answered:

- **RQ1**: Are the new Multi-Output architectures effective for CBIR of medical images?
- **RQ2**: Is it less efficient to use classification labels instead of more complex labels?
- **RQ3**: Is the new visual similarity metric able of capturing similarity between cases based on their visual appearance?
- **RQ4**: Do the explanations of Multi-Output Classification Variational Autoencoder (MOC-VAE) architectures produce saliency maps of more quality than previous XAI techniques based on DL?
- **RQ5**: Does the reduction of number of hyperparameters of MOC-VAE affects to their performance in XAI tasks?

3.2.2 Structure of the Research

The thesis research proposes different alternatives that are studied in different chapters. The problem defined before is studied in different ways, proposing new methods, models and algorithms to overcome problems and limitations of previous researches. Overall, the structure of the research is divided in the following steps:

- A novel CBIR DL model is presented in Chapter 4. The so-called Multi-Output Classification Autoencoder (MOC-AE) is a neural network specifically designed for extracting anatomical and pathological features from medical images. By extracting these features an image descriptor is generated for each of the cases, then these descriptors are compared to find the most similar ones, that are recommended. The MOC-AE model is presented in this chapter and its performance is measured by comparing its results to the work of Kobayashi et al. [118]. Both approximations are compared using segmentation labels of images of brains with tumors. Results are measured using the Sørensen-Dice coefficient [48], [204] over the segmentation labels of healthy and tumoral regions of the patients.
- After presenting and comparing the MOC-AE with a previous research, an evolution of the MOC-AE is presented with the MOC-VAE in Chapter 5. This architecture regularizes the latent space of the MOC-AE using the Variational Autoencoder (VAE) architecture [116]. Using this regularization, the recommendation is ordered in a complete and continuous space, enhancing the relationships between cases. The results of both MOC-AE and MOC-VAE are compared using a Chest X-ray dataset.
- MOC-AE and MOC-VAE are further studied in Chapter 5. To obtain more information about their performance they are compared in absolute terms with AEs and classifiers. These architectures are the state-of-the-art models for feature extraction in CBIR [118], [161]. By this comparison it is possible to obtain a notion of the real performance of these models under real world circumstances, where segmentation labels are very difficult to obtain.
- A new metric of visual similarity in comparative diagnostic is proposed in Chapter 5. The Sliced Wassertein distance [28] is proposed as an evaluation metric for visual similarity between cases. The main strengths of this metric are related to invariances in rotation and position. This metric is used to compared the proposed methods along with the Precision@k, that measures semantical similarity between cases.
- Chapter 6 presents a novel XAI algorithm that uses the MOC-VAE as it basis. The potential of MOC-VAE in factorizing the information of images in an enriched descriptor, is leveraged to find explanations in this latent space. This framework makes possible to find linear explanations in the latent space than can be translated to non-linear explanations decoding the image descriptor. This way the explanations are found in a low-dimensional space where it is possible to apply a search based on genetic algorithms and finding the saliency map of this solutions in the original image space. The results are compared with state-of-the-art XAI models, focusing in the precision of their explanations related to the area they found the explanations and the area of the pathology of the case.

3.2.3 Chronogram of the Research

The chronogram in Figure 3.1 contains the temporal organization of the research. From the previously defined limitations, different researches have been made to solve the RQ previously

defined. These researches correspond with different chapters of this manuscript with different papers and outcomes associated to them.

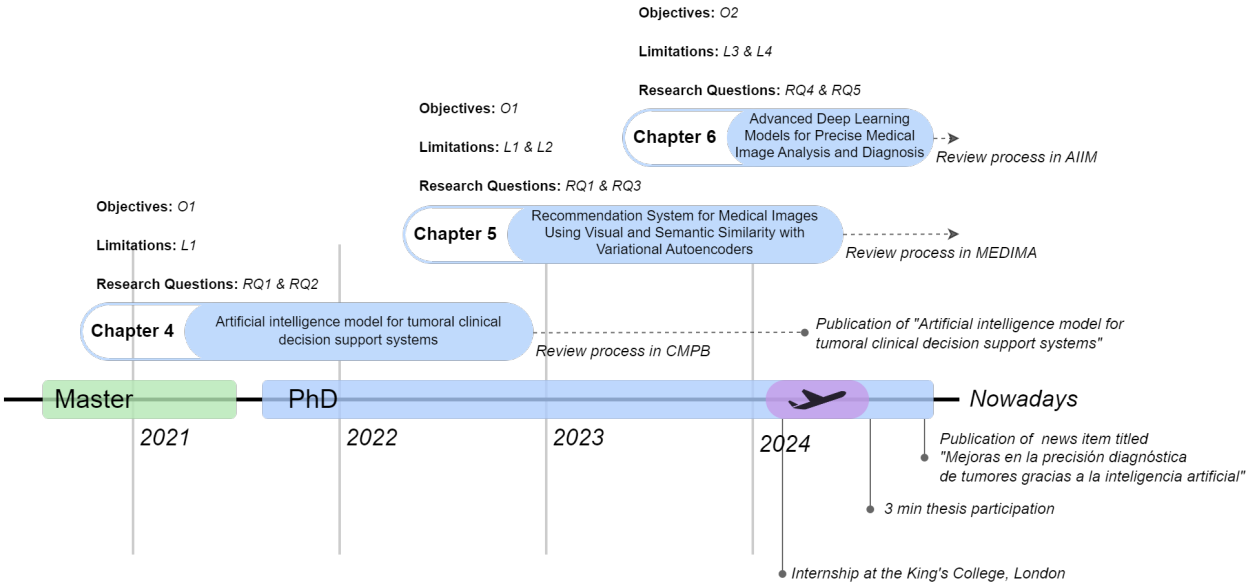


Figure 3.1: Chronogram of the development of the thesis.

Proposed Architectures

Chapter 4

Artificial Intelligence Model for Tumoral Clinical Decision Support Systems

Comparative diagnostic in brain tumor evaluation makes possible to use the available information from a medical center to compare similar cases when a new patient is evaluated. In this chapter, a novel Deep Learning (DL) model is proposed that retrieves similar cases of brain tumors for a given query. The model uses DL to detect patient features to then recommend the most similar cases from a database.

The primary objective is to enhance the diagnostic process by generating more accurate representations of medical images, with a particular focus on patient-specific normal features and pathologies. The system not only suggests similar cases but also balances the representation of healthy and abnormal features in its design. This not only encourages the generalization of its use but also aids clinicians in their decision-making processes. A key distinction from previous models lies in its ability to produce enriched image descriptors solely from binary information, eliminating the need for costly and difficult to obtain tumor segmentation. This generalization makes possible for future research in different medical diagnosis areas with almost not any change in the system.

A comparative analysis in relation to similar studies is conducted. The proposed architecture obtains a Dice coefficient of 0.474 in both tumoral and healthy regions of the patients, which outperforms previous literature. The proposed model excels at extracting and combining anatomical and pathological features from brain Magnetic Resonances (MRs), achieving state-of-the-art results while relying on less expensive label information. This substantially reduces the overall cost of the training process. The findings highlight the significant potential for improving the efficiency and accuracy of comparative diagnostics and the treatment of tumoral pathologies.

This chapter provides substantial grounds for further exploration of the broader applicability and optimization of the proposed architecture to enhance clinical decision-making. The novel approach presented in this work marks a significant advancement in the field of medical

diagnosis, particularly in the context of Artificial Intelligence (AI)-assisted image retrieval, and promises to reduce costs and improve the quality of patient care using AI as a support tool instead of a black box system.

Additional information about the research of the chapter, including information about paper that holds the research covered in it can be accessed in Appendix .3.

4.1 Introduction

Autoencoders (AEs) [185] are Machine Learning (ML) models in dimensionality reduction and feature extraction processes. In this work, it is proposed to use a novel AE variant, named Multi-Output Classification Autoencoder (MOC-AE), that factorizes brain tumoural images, to use the generated image descriptors to recommend cases with a similar pathology.

The presented model focuses on improving the retrieval accuracy of a standard AE without using costly information, as previous similar state-of-the-art approaches, such as tumor area segmentation [118]. The contribution can balance the normal anatomical features of each patient, i.e. the healthy regions, with the tumor features in a single descriptor. Thus, the model can recommend interesting cases considering relevant medical information, such as the tumor area, position in the brain, composition and geometry.

Respect oncology, different works have been published in the last years focusing on using the knowledge base available to perform comparative diagnostic [118], [153], [178], [180], [220], [229]. The main drawback of these methods is the necessity of structured information for use this knowledge base, where the image descriptors formed by Computer Vision algorithms can produce great results.

The main contributions of the chapter can be defined as:

- The Multi-Output Classification Autoencoder (MOC-AE), a novel architecture for Content-Based Image Retrieval (CBIR), is presented. MOC-AE works using binary labels bypassing the necessity for costly tumor segmentation, which is specially difficult in medical domains [150].
- The MOC-AE architecture is tested using the Multimodal Brain Tumor Segmentation Challenge (BraTS) 2020 dataset [20], [21], [143]. The results of the model are measured using segmentation labels of the data, comparing the similarities of the geometry of the tumors and the anatomy of the patients. The proposed model obtains a Dice-Coefficient of 0.632 for healthy areas and 0.316 for tumoural regions.
- Results for the MOC-AE are be presented in comparison with similar state-of-the art approaches, showing an improvement in the results both detecting tumoural features of the patients and better balancing normal and tumoural structures of the patients. Results show an improvement between 0.244 and 0.027 points in tumoural regions and 0.077 and 0.026 in both normal and abnormal features.

4.2 Methods

4.2.1 Experimental Data: Multimodal Brain Tumor Segmentation Challenge Dataset

BraTS 2020 dataset [20], [21], [143] is used in this research, which contains 369 labeled Magnetic Resonance (MR) cases of gliomas. The available data is manually segmented obtaining the tumoural region of each case in a separate file. Each case has a resolution of $240 \times 240 \times 155$ and is available in different sequences. This database is used as a baseline for the proposed Information Retrieval system, because it presents real case MRs which has been tested before in similar systems.

The dataset is divided in two partitions: *MICCAI_BraTS_Training* contains information on 369 different cases; this subset also contains manually segmented regions of the tumoural areas of each case. Additionally, *MICCAI_BraTS_Validation* contains 125 non-labeled MR.

Each MR is available as four different sequences: Native scanner (T1), Post-contrast T1 weighted (T1Gd), T2 weighted (T2) and T2 Fluid Attenuated Inversion Recovery (T2-Flair). Segmentation of each MR is divided into three labels, Gd-enhancing tumor (ET), Pleritumoural edema (ED) and Non-enhancing tumor core (NET), manually segmented and approved by neuroradiologists. In the current experiment all the sequences will be used to feed the model, whereas the tumor segmentation labels will be used to evaluate the results of the model. These segmentation labels will only be used for evaluation purposes, because to train the network only binary labels of presence/absence of a tumor in the image will be used.

Due to the fact that for the training and evaluation of the proposed algorithm, the labeled information for each MR is necessary, it will only be used in the *MICCAI_BraTS_Training* partition. This dataset will be divided into a training and evaluation set.

Figure 4.1 contains samples of the information present in the dataset. The first 4 columns have a different sequence, denotated by a text in the top of the column and the 5 column has the tumor segmentation labels of the case. It must be noted that each image represents a full three dimensional MR of a patient but only a section of the full MR is showed.

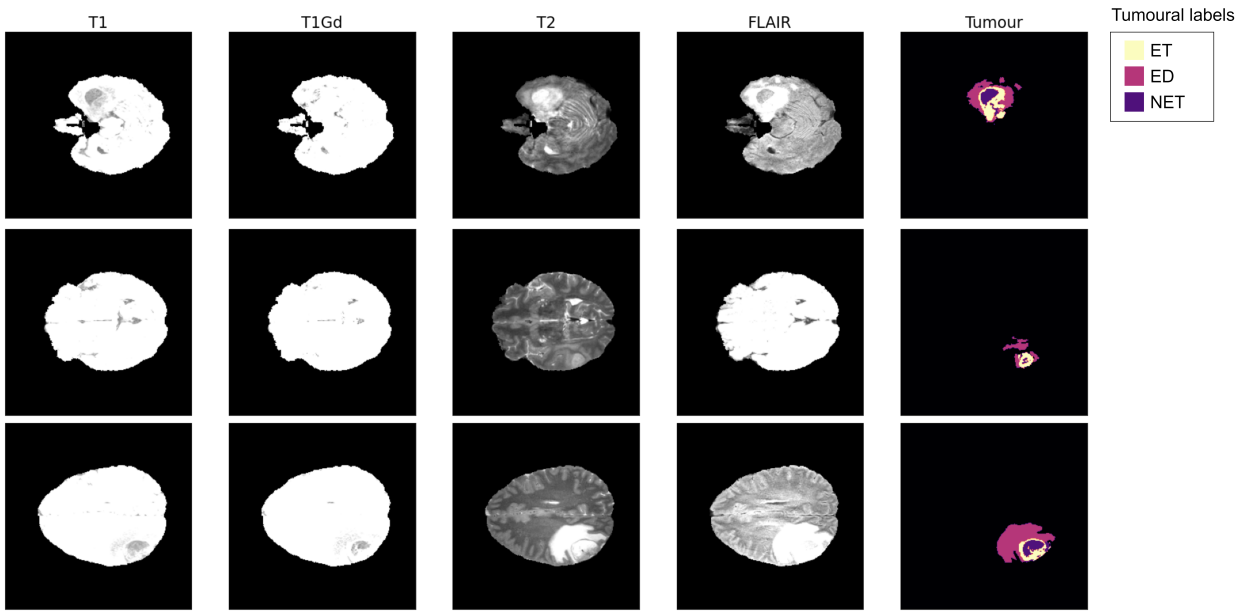


Figure 4.1: Sample images from BraTS dataset. Each row contains a different patient and each column a different information for the patient, from MR to tumour segmentation information.

To test the cases, the similarity of segmentation between the query and the retrieved images will be studied. In the tumoural characteristics of each patient, similarity measurement is performed using the tumor segmentation information present in the original dataset. However, to compare the similarity of the healthy areas of each patient, the same approximation of [118] will be used, where the brain of each patient is divided anatomically. This process must be performed to compare the results of the proposed model with those of the work of [118] and, to obtain the same information they used, each MR must be preprocessed.

Dataset Preprocessing to Obtain Anatomical Labels

Each three dimensional MR consists of $240 \times 240 \times 155$ pixels of information, but the input of the proposed method is a two dimensional image. To obtain the images from the three dimensional data that are stored in the BraTS 2020 dataset, the MRs must be sliced in layers. The data dimension of $240 \times 240 \times 155$ is sliced on the third axis to generate 240×240 images by taking the information about each layer separately. Furthermore, each image is normalized in the range [-1, 1], to then be properly treated with Artificial Neural Networks (ANNs). The normalization process can be defined by the following equation:

$$\begin{aligned} x'_i &= \frac{x_i}{127.5} - 1 \\ \forall x'_i \ &\epsilon \ [-1, 1] \end{aligned} \tag{4.1}$$

where x denotes the gray level of each pixel of the image, where $x \epsilon [0, 256]$.

In addition, each healthy image is labeled with six normal anatomical labels: left and right cerebrum, left and right cerebellum and left and right ventricle. This division is achieved using BrainSuite 19a software [197]. This program can obtain a voxel segmentation of each case's cerebrum, cerebellum and ventricle, making it possible to use this information to evaluate the similarity between the query and the retrieved cases.

Figure 4.2 has a sample of the new labels generated for each image. As seen after segmentation of the anatomical labels, each brain is divided into six different areas, as was done in [118]. The second column of the figure has the generated anatomical labels for each section of the patient, whereas the first column is the Magnetic Resonance Imaging (MRI) and the third one the tumour segmentation information.

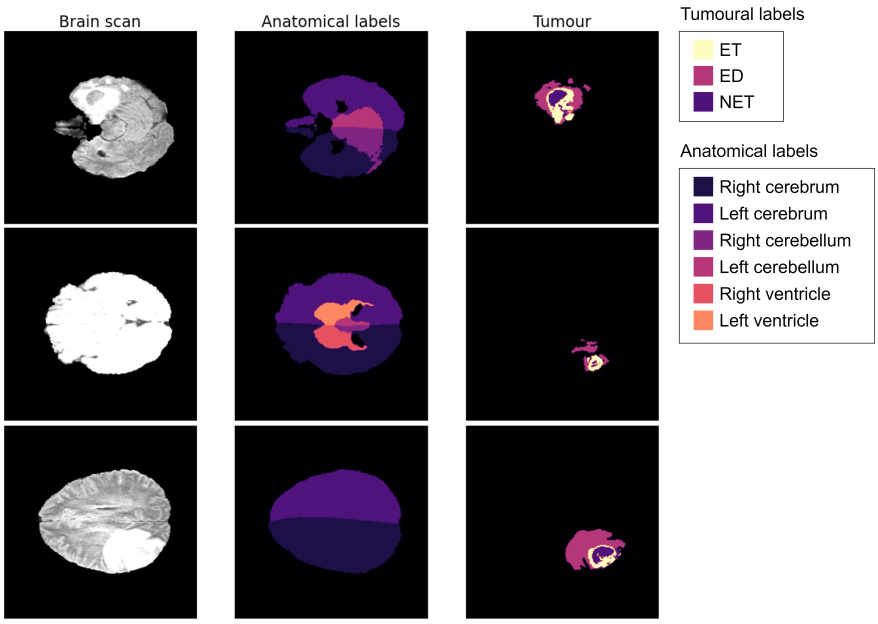


Figure 4.2: Sample images of the labeled dataset. Each row contains a different case and each column contains the patient MRI, anatomical label segmentation and tumoural segmentation.

Figure 4.3 shows a brief scheme of the preprocessing process, from the 3-dimensional Neuroimaging Informatics Technology Initiative (NIfTI) files to a two dimensional images of each case, obtaining in addition a segmentation of the anatomical labels of each patient using BrainSuite 19a software [197]. Each row represent a different slice of the same case.

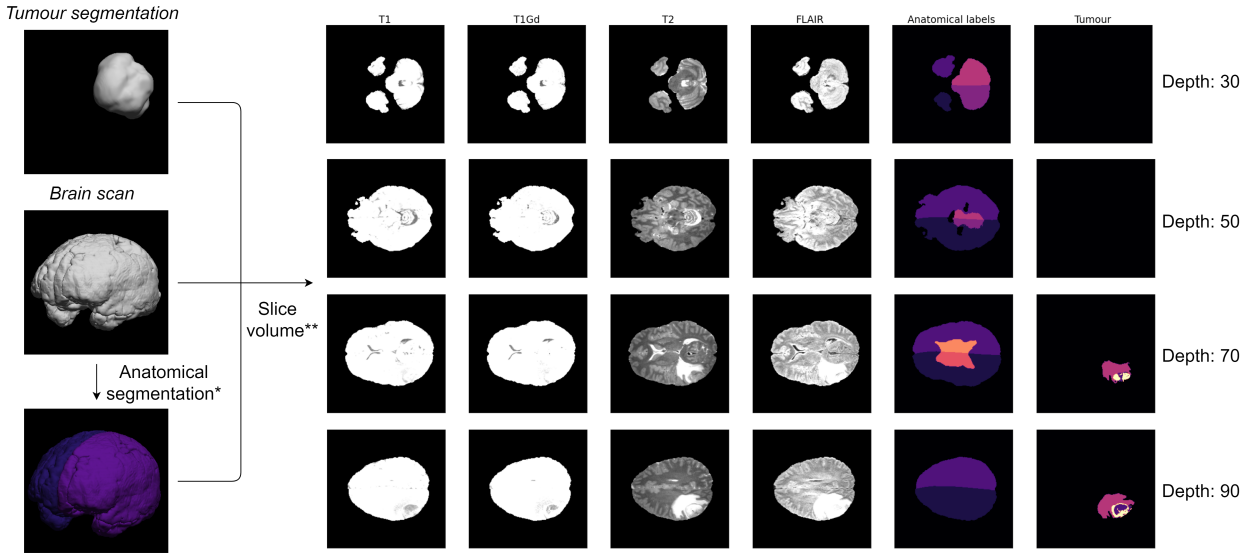


Figure 4.3: Data preparation scheme. (*) show that the anatomical labels were obtained using BrainSuite 19a software [197]. (**) shows that each slice corresponds to a certain depth in z axis.

4.2.2 Architecture Definition

In Figure 4.4 the schematic of the proposed method, named as MOC-AE, can be seen. According to the figure, the architecture presented combines two approaches: an AE network that extracts the structural information of each image and a binary classifier that is responsible for extracting the tumor information from each case, i.e. if a tumor is present or not in the image received. This dual-objective architecture enhances the features represented in the descriptor obtained from the latent vector.

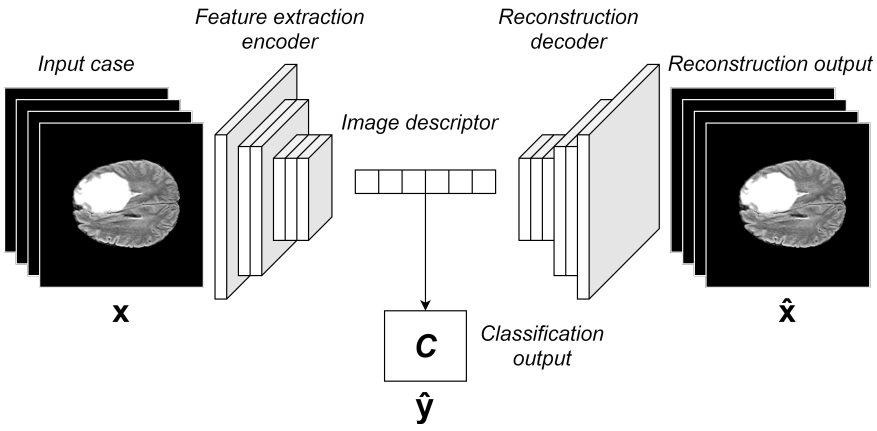


Figure 4.4: Neural network scheme of the proposed model.

On the one hand, in MOC-AE an AE is used following the same approximation as [203] where a CBIR system is designed using the latent space of an AE as image descriptors. This

simple scheme makes it possible to extract the composition features of an image by forcing the network to reduce the dimensionality of the input images. One of the main advantages of this method is that it does not require any label to work, resulting from the self-supervised learning scheme of AEs [121]. Using an AE as the main base for image descriptor generation, the network can learn latent representations of the input image.

The main drawback of just using the latent space of an AE as a descriptor is that it considers every portion of the input image with the same importance. Furthermore, there is no control over the information represented in the latent space.

To solve this problem, adding an auxiliary classifier that shares the descriptor with AE is proposed. This addition to the network will attempt to maintain the information from the patient's tumor. Here, it is important to note that this learning scheme keeps certain features of the input information, to disentangle the patient's healthy and tumoural information in the image descriptor. This solution is based on the work of [118] where three different AEs were used to disentangle tumor and normal information from each case.

Using the classifier, the network is forced to learn the tumor's information to classify the cases where a tumor is present. At the same time, AE forces the descriptor to maintain the structural characteristics of each patient. This dual-objective forces the network to focus on some present features in this case.

The main idea is to combine the ability of an AE to represent in a small image descriptor the features of an image with the pathology detection capability of a classifier. The results that will be presented in section 4.3 show that the proposed architecture outperforms previous approaches in retrieving more similar images from a database for a given query. By combining the feature extraction potential of both AE and classifier, the MOC-AE can improve each architecture individually and previous similar architectures.

This architecture is focused on medical CBIR because it makes it possible to make the network focus on the possible pathologies of each image. Respect the work of [118], this scheme does not require a segmented dataset. In the proposed method, the binary classifier only needs information about the presence or absence of a tumor in the image, but segmentation of the tumoural region of the image is not necessary.

Thus, the proposed model takes as an input a brain three dimensional MR and extracts the most relevant patient information, regarding anatomical and pathological features. I.e. representing in a small vector the information of the healthy portion of the brain and, in case of presence, the tumor features. Then, this vector is used to compare each case with the documented cases of the database. The most similar cases are recommended to the physician, this way presenting the professional the most similar cases to support its diagnostic.

As stated in [121], in the field of medicine, the process of annotating the data is particularly costly because specialists in the field must perform manual annotation. Therefore, the possibility of developing a CBIR system capable of focusing on the most relevant areas of the image while maintaining the relatively low cost of the labels used is a characteristic that is particularly desired.

Model Training Scheme

The proposed architecture must combine two different learning processes using the same parameters. As stated above, this shared information forces the network to combine each image's normal and pathological features.

Regarding the model's training, two different outputs will use two different losses, which must be combined to train the model. The AE output is responsible for reconstructing the information in the input image, while the classifier must differentiate between healthy and regions with the presence of a tumor in it.

First, the input image x must be reconstructed by the AE generating $\hat{x}$, both the input and the reconstructed image will be compared using the L2 norm.

$$L_r(x, \hat{x}) = ||x - \hat{x}||_2^2 \tag{4.2}$$

This reconstruction loss function L_r will force the latent space to learn the spatial features of the input image.

At the same time, the classification head of the proposed model will be trained using the binary cross-entropy loss function, formulated as follows:

$$L_c(x) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i) \tag{4.3}$$

The head loss classification function L_c focuses on differentiating healthy and tumoural images and detecting the presence of tumors in the input image. When seeking this objective, the model will be forced to maintain tumoural features in the latent space.

Both losses are combined to train the model using the following equation:

$$L_t = \gamma_1 * L_r + \gamma_2 * L_c \tag{4.4}$$

where both γ terms are normalization coefficients to balance both losses.

The loss function is minimized using the Adam [115] optimizer. The best values of the γ parameters were found using the grid search in the training phase. The range of values tested vary between 0 and 1, requiring that the sum of both γ must be of 1. The best performance was achieved with $\gamma_1 = 0.2$ and $\gamma_2 = 0.8$.

These ideas of combining different objectives in the network training have been previously applied in state-of-the art architectures. Specially in models Generative Adversarial Networks (GANs), with networks as the Loss-Sensitive GAN (LS-GAN) [171], Cyclic-Synthesized GAN (CSGAN) [103], Multi-IlluStrator Style GAN (MISS GAN) [24], Super Resolution GAN (SRGAN) [128], Weighted SRGAN (WSRGAN) [33], Conditional GAN (CGAN) [148], Auxiliary Classifier GAN (ACGAN) [158], Unsupervised Dual Learning for Image-to-Image Translation (DualGAN) [238] between others.

These models use different objectives in their trainings that are translated in changes in their respective loss function. These ideas are leveraged in the MOC-AE to train the model to

achieve different results. In the case of the MOC-AE, by training the network with different objectives, it is possible to gain control over the latent space of the model, that will be used to produce the recommendation of the cases. Thus, the MOC-AE is designed combining ideas of the state-of-the art to apply them to medical comparative diagnostic.

Network Architecture Definition

The proposed network backbone contains Convolutional Neural Network (CNN) layers that use the Residual blocks presented in the ResNet architecture [76]. The same principles as those used in ResNet were used to design the Encoder and Decoder networks of the model. In particular, it was decided to use complete pre-activation blocks as was proposed in [77], that generally produce the best results. The AE generates the latent vector by reducing the information to a space of 500 dimensions that corresponds to the image descriptor.

Regarding the classifier, it uses the latent vector information of 500 positions and generates a binary classification, using an intermediate dense layer of 64 neurons.

Figure 4.5 shows details of the model architecture. The input data that the Encoder receives are 4 slices of each case, corresponding to the different MRs (T1, T1Gd, T2, T2-Flair). Each residual block contains two separable convolutions [41] along with Batch Normalization [90] and Dropout [80] layers. The Rectified Linear Unit (ReLU) activation function was used in the hidden layers, while the hyperbolic tangent and the sigmoid activation were used in the reconstruction and classification outputs, respectively.

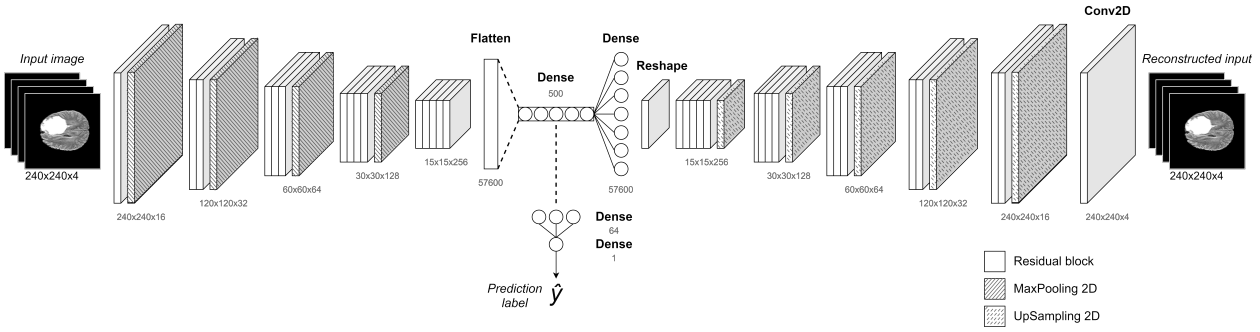


Figure 4.5: MOC-AE neural network detailed architecture.

Content-Based Image Retrieval Algorithm

CBIR system aims to obtain the patient's most similar to the query from the database of documented clinical cases. In other words, each time a MR is received, the algorithm CBIR finds in the database the most similar cases in the database by comparing the healthy and tumoural structures of the query.

The image descriptor is generated to compare each query with the rest of the documented cases. This descriptor corresponds to the latent space of MOC-AE generated using the trained

Encoder network. The Euclidean distance is used to compare the query image with the rest of the database, following the same approximation as [118], [194], [203], [215].

Figure 4.6 contains a scheme of the recommendation system. The architecture takes an input patient and generates its corresponding feature descriptor with the proposed MOC-AE network. Then, the most similar cases are retrieved from the database by comparing the descriptors with the Euclidean distance.

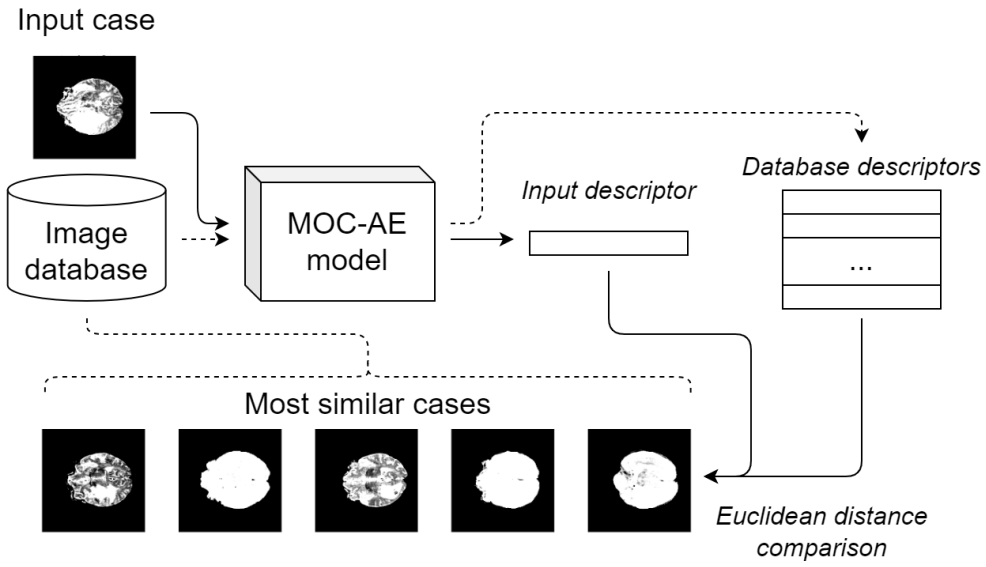


Figure 4.6: CBIR diagram of the proposed recommendation system.

The recommendation system will also use the classification learning of the network to retrieve similar cases from the database. This information will be used to enhance the recommendation because of the network's inherent ability to classify a certain query as tumoural or not. If the query is classified as tumoural by the classifier with certain reliability, it will only search among other tumoural images. This way, when the network is sure that a case contains a tumor it will only search for other tumoural cases. In particular, it is decided to put a threshold of 0.9 of activation from where search exclusively on tumoural cases.

4.3 Results

The dataset from the input of the experiment consists of 369 cases divided into 155 slices each. Thus, a total of 57.195 images were used. However, 10% of the whole dataset (5.720 images) is randomly reserved for testing purposes. This partition is made by randomly reserving 10% of the patients for evaluation purposes

The results of proposed algorithm performance by its training results and by comparing it with similar works are described as follows.

4.3.1 Training Results

The network was trained on a NVIDIA Quadro RTX 8000. The network was trained during 25 epochs, achieving the best results on the epoch 20. The model weights using during the rest of the results corresponds with the best performance of the model, regarding the total loss described in Equation 4.4.

The network training time for epoch was 1681±12 seconds with a maximum of 1681 seconds and a minimum of 1612 seconds. The batch size using during the experiment was 32, with a total of 1429 iterations each epoch.

Respect inference time, the model predicts both outputs in 0.0658±0.0406 seconds with a maximum inference time of 1.0154 seconds and a minimum of 0.0478 seconds.

The model's performance is evaluated against the binary classification of tumor presence in each MR. The classification output of the model was measured for the validation split of the dataset. That is, the model was tested using images never seen before. The results obtained, with respect to the confusion matrix, can be analyzed in Table 4.1.

Total images = 100		Predicted value	
		0	1
Real value	0	49	8
	1	8	35

Table 4.1: Confusion matrix of the MOC-AE classification output.

On the other hand Table 4.2 includes different metrics for the binary classification output of the model.

	precision	recall	f1-score	support
0	0.86	0.86	0.86	57
1	0.81	0.81	0.81	43
accuracy	-	-	0.84	100
macro avg	0.84	0.84	0.84	100
weighted avg	0.84	0.84	0.84	100

Table 4.2: Classification metrics of the MOC-AE.

Table 4.1 and Table 4.2 serves as a measurements of the performance of the model during the training, in this sense it is not related with the comparative diagnosis results sought with the research, but rather as a measurement of how well the model was trained. In addition, good results in classification will also ensure that the threshold previously defined to 0.9 confidence for filtering non tumoural images is correct.

Respect the reconstruction output of the network, the results are measured using the mean squared error, the root mean square error and the mean absolute error, pixel by pixel in all cases, obtaining a mean square error of 0.02805, a root mean square error of 0.16749 and a

mean absolute error of 0.05312. These values correspond with the reconstruction of the input patient MRs.

4.3.2 Recommendation Results

After analyzing the model's training performance, the model is tested against comparative diagnostic tasks. The CBIR model's performance is evaluated first empirically and then quantitatively. These results show the comparative diagnostic model performance, once trained, when the most similar cases are retrieved for a particular patient.

Empirical Evaluation of the Content-Based Image Retrieval System

The behavior of the model is shown to empirically evaluate recommendation results. These experiments show how the model performs against different cases, it's results are analyzed respect the medical diagnosis perspective, trying to figure out similarities and mismatches between the input and retrieved images.

To compare model's results Figure 4.7 shows the results for two different queries used as input to the system. The first column of the image corresponds with the input query image that is being used for image retrieval. Then, the 5 more similar cases detected by the proposed system are showed, ordered by descriptor closeness from left to right. As can be seen, the model can retrieve similar images in both anatomical and tumoural aspects. At the same time, Figure 4.8 contains two cases of brain regions with no tumor present in it, also with the 5 more similar cases ordered by closeness from left to right. This second case of study shows the model's performance against normal cases.

Figure 4.7a represents a patient with a tumor in the middle section of the brain. The tumor is positioned in the right hemisphere of the cerebrum; the retrieved images from the dataset represent cases of the same brain region also sharing a tumor of similar composition with similar position and geometry.

On the other hand, Figure 4.7b shows a case with a smaller tumor in a lower region of the brain, positioned in the right cerebellum. The similarity between the query and the retrieved cases exhibits the performance of the model, being able to extract at the same time the anatomy of the patient, i.e. the region of the images is the same, at the same time that the tumor is well extracted and represented, i.e. the similarity and location between the tumor cases is very high.

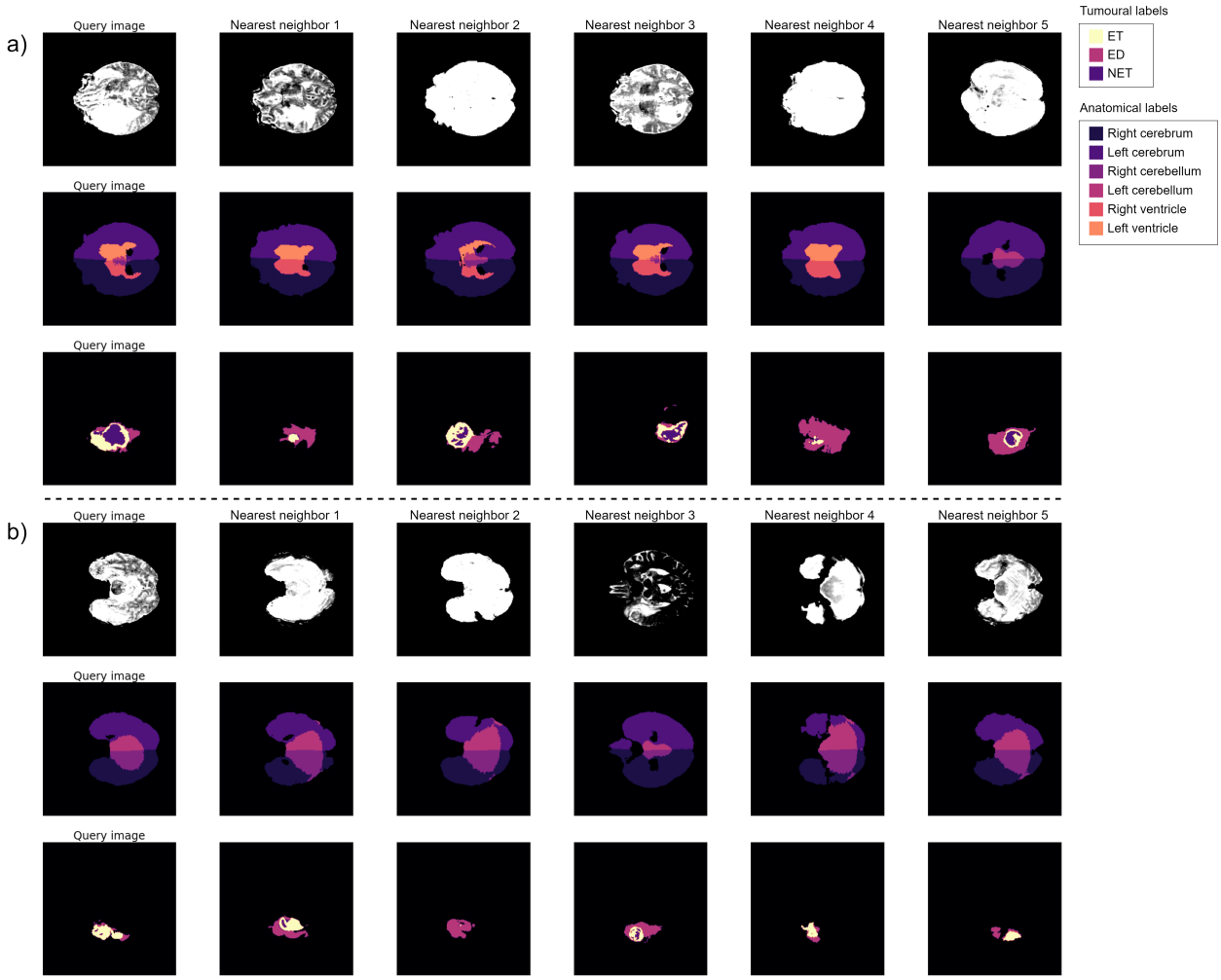


Figure 4.7: Nearest neighbors recovered by the MOC-AE for two different queries, ordered from left to right.

Figure 4.8 follows the same structure as Figure 4.7 but with non tumoural sections. Figure 4.8a shows a case of a lower section of a patient's brain with no tumor present in it. As seen, the model retrieves from the database cases from the same section anatomically, i.e. cases share the same brain region and shape. Respect the tumor correlation between images, the results represent cases without a tumor present or tumors with small area.

In the case represented in Figure 4.8b it can be seen as a case of the brain's upper area without a tumor. Similar to the case 4.8a the retrieved images share the absence or presence of very small tumor while presenting a case of a similar brain area.

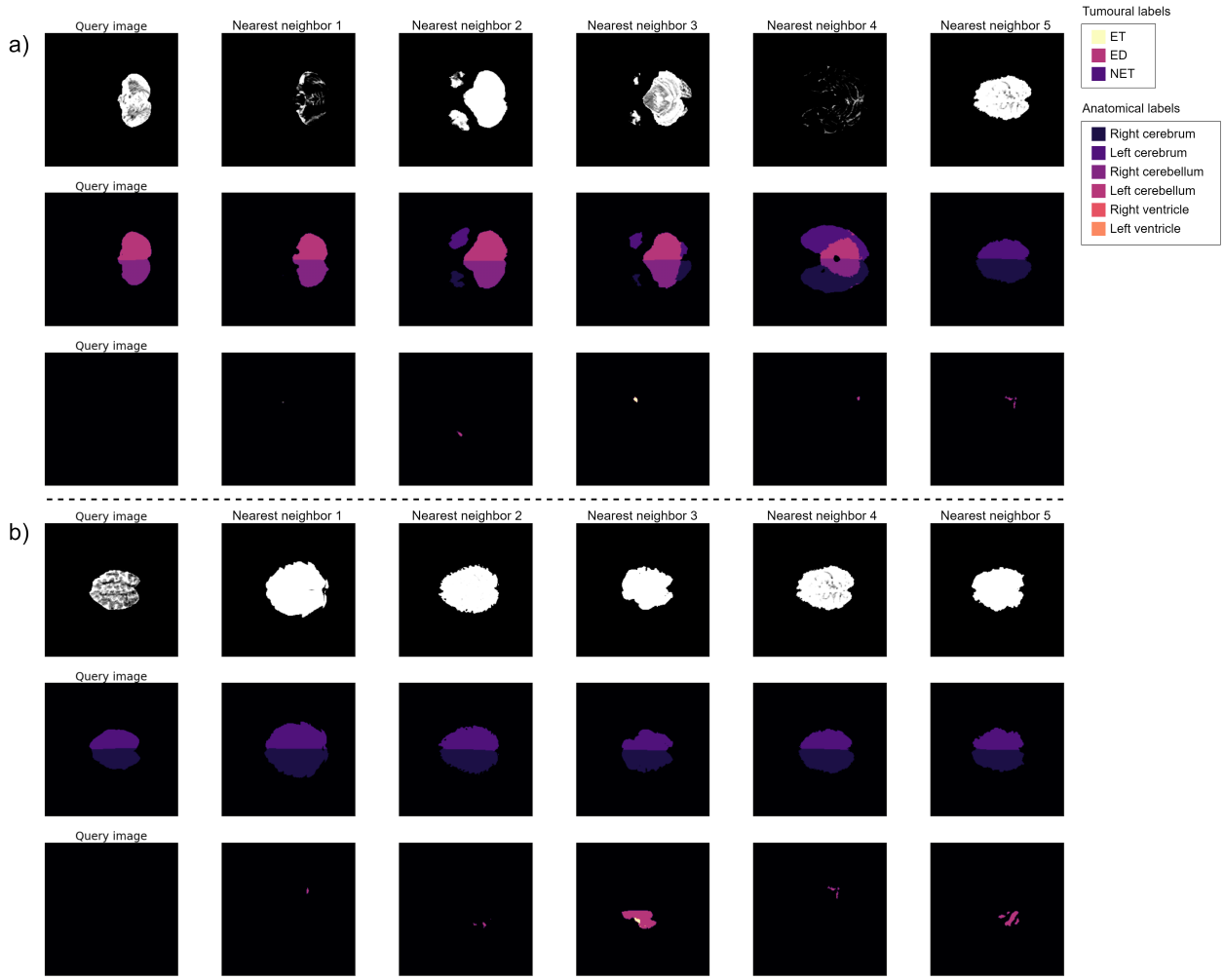


Figure 4.8: Nearest neighbors recovered by the MOC-AE for two different queries with absence of tumor, ordered from left to right.

The results show that when there is no tumor present in the query, the retrieved cases have small or non-tumors. This highlights the need for a more diverse dataset for training, as the current dataset exclusively contains cases with tumors. While this composition provides valuable information on tumor cases, it limits the model's ability to accurately identify non-tumor cases due to the low number of such cases. To improve upon this, future work could focus on diversifying the dataset to include a balanced representation of both tumor and non-tumor cases, or apply data augmentation techniques to simulate a more diverse dataset. Cross-validation using different datasets could also be considered to add robustness to the findings.

Similarity Results of the Recommendation

To further evaluate the CBIR potential of the model it is decided to quantitatively analyze it against other similar research. The model's performance is evaluated against Kobayashi et

al research [118], already presented in Section 2.3.1. These results evaluate how similar are the retrieved images of the database with respect the input image being diagnosed, therefore providing a measure of the similarity of the retrieval.

Table 4.3 compares of the performance of the proposed model with that of the model developed by Kobayashi et al [118] in terms of the Sørensen-Dice coefficient [48], [204]. In particular, the table measures the consistency between the anatomy of the healthy and tumoural features of the image. This coefficient varies between 1 and 0, where the higher the value, the more similar are the query and the retrieved images. To calculate the Sørensen-Dice coefficient of the proposed method, the same procedure will be followed as [118], where the queries used are the slices of each case with the largest tumoural area. Three different coefficients are analyzed, (i) the **Normal Dice**, which evaluates the model's ability to identify and retrieve images based on their normal or healthy characteristics, using the six anatomical labels obtained with the procedure explained in section 4.2.1 and calculating the multi-label Sørensen-Dice coefficient, (ii) the **Tumoural Dice** which assesses the model's skill in identifying and retrieving images based on their tumoural or pathological features uses the segmentation information of each image, measuring the similarity of the tumoural sections of each case, and (iii) the **Entire Dice** which measures the model's overall proficiency in retrieving relevant images considering both normal and tumoural characteristics, thus providing a more complete measurement. The number of neighbors of the CBIR is set to 5, following the same procedure as [118]. The Entire Dice corresponds with the mean value between the normal and tumoural Dices.

Model	Normal Dice	Tumoural Dice	Entire Dice
MOC-AE	0.632±0.218	**0.316±0.275**	**0.474±0.173**
Kobayashi et al. Normal latent space	**0.730±0.196**	0.072±0.098	0.401±0.108
Kobayashi et al. Abnormal latent space	0.505±0.235	0.289±0.120	0.397±0.137
Kobayashi et al. Entire latent space	0.695±0.208	0.201±0.158	0.448±0.123

Table 4.3: Sørensen-Dice coefficient values comparing MOC-AE and the work of [118].

The performance of the model in identifying and retrieving images based on their normal or healthy characteristics yields a Normal Dice score of 0.632 ± 0.218. Furthermore, MOC-AE's Tumoural Dice score stands at 0.316±0.275. The overall proficiency of MOC-AE, encapsulated by the Entire Dice score and considering both normal and tumoural characteristics, registers at 0.474±0.173.

Compared with Kobayashi et al.'s model, it's evident that while their work may slightly surpass MOC-AE in retrieving normal features, but the contribution significantly outperforms in the identification and retrieval of tumoural features from medical images. This capability is particularly valuable, especially considering the primary goal of the study and the role of the Clinical Decision Support System (CDSS) in aiding the diagnosis of tumoural pathologies.

Moreover, a critical strength of the presented model lies in its ability to train the network without needing segmentation information for the cases, setting it apart from the model of Kobayashi et al. Although the best anatomical similarity, measured in the Normal Dice, is achieved by Kobayashi et al., the overall performance in terms of patient tumoural features shows noticeable improvement with MOC-AE.

Consequently, the proposed model efficiently manages the delicate equilibrium between each patient's healthy and tumoural areas, enabling the retrieval of highly similar images from the database. It presents an optimal balance between normal and pathological features in patients. It also demonstrates superior performance according to state-of-the-art standards while concurrently reducing the cost and complexity inherent in the training process. These characteristics highlight MOC-AE's promising potential for use in the medical field.

4.4 Discussion

CDSSs represent critical algorithms that could significantly improve the diagnostic tasks of doctors. The proposed model achieves state-of-the-art results in both the healthy and tumoural features of recommended cases. The model can factorize images focusing on both the normal features of the patient and the pathologies present in the case, generating a compact and significant representation of each image. These image descriptors can then be used to recommend similar cases from a database, helping the physician to make the diagnosis.

The architecture presented in the current chapter can combine features present in the image with labels annotated by professionals. One of the main strengths differentiating MOC-AE from similar CBIR models is that it does not require costly information on the label such as tumor segmentation. The model is trained only using the binary labels of presence or absence of a tumor, while the costly segmentation labels are only used during evaluation of the model's performance. The model learns characteristics of each patient combining recommendation and classification outputs, thus generating an enriched image descriptor using only binary label information. In areas where the cost of producing high-quality labels is especially costly, it is considered crucial to develop a model that can produce state-of-the-art results with a low-cost associated.

In contrast to previous works [26], [37], [62] the model does not require segmentation labels. Tumor area segmentation is a costly information and in many cases is very difficult to obtain, it is considered that it is far more common to be available information of the presence or absence of a tumor in an image. Thus, the proposed architecture is considered an alternative that can be extended to a wider range of cases, especially in the medical area where data availability is difficult. The only necessary information to apply the system to another case of study are the images that will be ranked along with their corresponding binary labels. In this case, the segmentation labels are used only for evaluation purposes and could be adapted to evaluate the new target domain. This is of great interest, not only for facilitating its use and implementation but also because it can be generalized to every medical area due to a cost reduction. The simple general structure that is provided opens up the possibility to apply the proposed system to new domains, without any change.

The results of the chapter show that the presented architecture can extract from the brain MRs information about the anatomy of the patient and the presence and composition of tumors simultaneously, outperforming previous solutions. Respect Kobayashi et al. [118], there is an improvement in Sørensen-Dice coefficient from 0.448 to 0.474 while reducing the cost of the training labels. The recommendations suggest that the inclusion of the classification output enriches the image descriptor, while not losing structural information of the anatomy of the patient. It is specially interesting to observe that, although anatomical information of the patients is not available, the model can recommend cases with similar healthy structures.

Summarizing, the model's results improve the previous work by obtaining better results in both retrieving cases with similar pathologies and balancing both anatomical and abnormal features of each case. Given the results, MOC-AE is considered the best alternative to develop a CDSS specialized in medical imaging recommendation. In this sense, this AI model could become a helpful tool for clinicians to make more accurate image diagnoses and make more proper decisions avoiding some subjective biases.

Overall, the main contributions of this chapter are summarized as follows:

- The MOC-AE architecture, a Deep Learning (DL) model for image descriptor generation, is presented. The network is specifically designed for medical image recommendation, but its simplicity makes it possible to extend it a broader variety of cases.
- The design of the network is discussed, focusing on alleviating the cost of the data used to train the network while improving the performance of the recommendation.
- Experiments showcasing and comparing the model are presented. The results show that the contribution outperform previous networks in retrieving more similar cases respect tumoural features and overall similarity of the brain MRs.

4.5 Further Research

The presented model has several aspects that could be improved. First, the model's performance evaluation is only defined by using the segmentation labels. It is not correct to only evaluate the correlation between labels, because one of the strengths of the model is that it can link very similar cases that may or may not share the same pathological features and be very similar visually. This characteristic is crucial because it provides to the physician additional information for a given case, making possible to identify new pathologies with this similarity. To only evaluate the similarity if the images (e.g. by comparing the value of their pixels) is also not correct, because this measurement will not take into account the similarity of the cases in medical terms. Therefore, further research should focus on defining new methods of evaluation, considering all the important aspects of CBIR.

Second, the model performance could be improved by applying optimization network methods, e.g. further tunning the hyperparameters of the neural network or adding attention layers to the network.

Thirdly, besides lowering the cost of obtaining the dataset, the architecture still need labels for

train. This requirement could limit the range of applications of the framework, in particular, in problems where there is not labeled data it is not possible to use the proposed model. Besides that, it should be noted that the segmentation information used in this research is only used for evaluation purposes, and in new domains can be omitted with different evaluation metrics. Further research must be carried out to evaluate the architecture's performance against different cases of study.

Using the proposed model in tumoural CBIR arises a promising option to improve efficiency and accuracy in comparative diagnostic and tumoural pathologies treatment, with the possibility of helping physicians make more accurate diagnosis. Furthermore, the proposed framework has the flexibility to be applied to different medical diagnostic cases.

In summary, this chapter proposes the MOC-AE architecture and presents results that outperforms previous researches with less costly information. But to further evaluate the architecture, extensive evaluation must be carried out. In particular, it is crucial to test the architecture against different problems, e.g. multi-class classification or different datasets, but also extend the comparison against traditional methods. The research of this chapter is focused on comparing the MOC-AE against previous researches, but to provide a wider perspective it is consider central to compare the MOC-AE with similar traditional CBIR architectures, i.e. AEs and classifiers.

4.6 Answers to the Research Questions

- **Research Question (RQ)1**: Are the new Multi-Output architectures effective for CBIR of medical images?

The proposed MOC-AE is analyzed measuring anatomical and pathological similarities between cases, moreover the balance between both characteristics is evaluated.

In pathological similarity, the MOC-AE obtains a Tumoural Dice coefficient of 0.316 respect the Kobayashi's et al. 0.289. Thus, it is proved that the MOC-AE better preserves anatomical features in the recommendation, with a higher correlation in the areas of the tumours of the patients.

In anatomical similarity, the proposed MOC-AE obtains a Normal Dice coefficient of 0.632 respect the Kobayashi's et al. 0.730. There is a decrease in the performance of the MOC-AE respect the previous work.

Respect the balance between anatomical and pathological features, the proposed MOC-AE obtains an Entire Dice coefficient of 0.474 respect the Kobayashi's et al. 0.448. The results suggest that the MOC-AE better balances anatomical and pathological characteristics of the patients.

- **RQ2**: Is it less efficient to use classification labels instead of more complex labels?

Results of the MOC-AE show an improvement in the recommendation of the cases using less costly labels in network training. Thus, it is proved that it is possible to use binary

classification labels to achieve better results than the ones in the state-of-the-art that used the more-costly segmentation labels.

Chapter 5

Recommendation System for Medical Images Using Visual and Semantic Similarity with Variational Autoencoders

This chapter further studies and develop the Content-Based Image Retrieval (CBIR) system proposed with the Multi-Output Classification Autoencoder (MOC-AE). The MOC-AE architecture is a Deep Learning (DL) model focus on medical image diagnosis, proposed in Chapter 4, that retrieves the most similar cases from a database. During this chapter a variation of the MOC-AE will be introduced, using the regularization mechanism of the Variational Autoencoders (VAEs), named Multi-Output Classification Variational Autoencoder (MOC-VAE). These novel custom-built systems improves Content-Based Image Retrieval models in medical tasks to recommend cases of similar pathologies, focusing exclusively on image information.

The proposed networks simultaneously focus on the anatomical and pathological features of each case, obtaining a precise and easily interpretable recommendation. The systems are based on optimizing the extraction and selection of relevant visual and semantical features, as well as refining the search and retrieval of similar cases by regularizing the search space, in the case of the MOC-VAE.

Both architectures are studied in comparison with Autoencoders (AEs) and classifiers, that are the traditional DL alternative to CBIR. The results show that the proposed architectures outperform previous architectures in image recommendation, not only better balancing visual and pathology similarities between images, but also improving precision and visual similarity independently. From a clinical perspective, the results suggest that the recommended cases present relevant information, both semantically and visually similar and could be used by professionals as an assistant tool.

Additional information about the research of the chapter, including information about paper that holds the research covered in it can be accessed in Appendix .4.

5.1 Introduction

CBIR is based on two major processes, acquisition of image descriptors and measurement of descriptor similarity. The former aims to represent the input image with a small image descriptor that contains the most relevant information about the image and discards irrelevant information from it. This descriptor is a lower-dimensional representation of the original information. Therefore, it can be seen as a factorization process, where the principal components of the information can be represented in a lower-dimensional space. The image descriptors must satisfy different desirable characteristics [130]: first, the descriptors must be discriminative, in order to be able to distinguish between classes using these descriptors; second, the descriptors must be invariant against image transformations, representing the semantics of the images [12].

The MOC-AE defined in Chapter 4 was capable of generating meaningful descriptors both semantically and visually, i.e. they correlated patients with similar pathologies as well as patients with similar structures of healthy regions. But the extent of these similarities must be studied to disentangle the real performance of the proposed architecture. In this sense, during this chapter, the MOC-AE will be deeply studied in comparison with similar traditional neural network architectures, in particular with AEs and classifiers, that would be its respective counterparts. To do so, the PadChest dataset [30] will be used for evaluation.

But in addition, an evolution of the MOC-AE will be presented, named MOC-VAE. The MOC-VAE presents a novel architecture that regularizes the latent space of the image descriptors with the mechanisms of the VAEs [116]. By doing so, the latent space from where the recommendations are done is regularized fulfilling completeness and continuity properties. In other words, the new arrangement of the latent space will ensure that distance between different descriptors is meaningful mathematically.

One of the problems of evaluating CBIR models in medical image recommendation is the lack of metrics that evaluates their performance. To semantically compare the results of different models, Precision@k is usually measured [4], [111], [130], [169], [224], [240]. But CBIR models also must be evaluated with visual similarity between cases, and it is in this sense where there is no any good metric in the state-of-the art that compares images. Thus, it is proposed to use the Sliced Wasserstein distance [28] to measure the visual similarity between the recommended images. This performance evaluation problem will be fully discussed in Section 5.2.3.

That said, the aim of this chapter is to present an innovative solution that improves the CBIR models applied in the medical environment to recommend cases of similar pathologies. Respect Chapter 4 a multi-class classification dataset will be used, to study the behavior of the model against a real world case of study, where segmentation information is not available, representing a real world problem, where this information is usually not available. For this purpose, a comprehensive comparison between different models is carried out, focusing specifically on improving the MOC-AE architecture, proposing the MOC-VAE evolution and comparing these architectures with traditional approaches. The main contribution of the research can be summarized as follows:

- The Multi-Output Classification Variational Autoencoder (MOC-VAE) neural archi-

tecture is presented. Which is a variation of the MOC-AE architecture that improves the latent representation of the descriptors by regularizing the latent space, better preserving the relationships between cases.

- The results of the MOC-AE and MOC-VAE architectures are compared with traditional feature description networks.
- A comparison of CBIR performance by dividing its analysis into two metrics is proposed. To measure the pathology similarity the Precision@k is used, as in previous works; additionally the Sliced Wasserstein distance is used to compare images by their visual similarity. This particularity makes possible to better measure the performance of the model from a clinician's perspective.

The proposal is based on optimizing the extraction and selection of relevant visual and semantical features, as well as refining the search and retrieval algorithms for similar cases. Through this research, it is demonstrated how the proposed solution overcomes the limitations of previous models, providing a more accurate and efficient tool for clinical decision support in the medical setting.

5.2 Methods

5.2.1 Experimental Data: Padchest Dataset

The Padchest dataset [30] is used to test the performance of the proposed systems. This dataset contains annotated X-ray images from the San Juan de Alicante hospital, characterized with different labels.

A subset of the total dataset is used, using the four most common classes of the dataset. In particular, it is decided to select Posterior Anterior projection images that only contain one of the four most common labels: *normal*, *aortic elongation*, *scoliosis* and *cardiomegaly*. The result is a subset 19.410 images. The distribution of the different labels can be seen in Table 5.1. In addition each case contains additional information, such as their sex and age, that will be used to analyze how the model performs against different groups.

Label	# of images
Normal	13.506
Cardiomegaly	2.202
Scoliosis	1.995
Aortic elongation	1.707

Table 5.1: Dataset distribution of the labels.

All images were pre-processed to discard useless information and normalize across different images. In particular, following the same procedure proposed by [133], the images are using segmentation masks of the lungs. The images are cropped using the upper, left and

right boundaries of each lung segmentation. Using this procedure, irrelevant information is discarded, forcing the proposed models on focusing on relevant radiological areas.

To obtain the segmentation masks to perform the cropping, a TransResU-Net [218] is trained with the Montgomery County CXR Set [93], [176] and followed the same pre-processing steps as [133].

5.2.2 Feature Extraction Networks for Content-Based Image Retrieval

The CBIR scheme that will be used in this chapter is the same as the one defined in Chapter 4 in Section 4.2.2. The algorithm is based on the generation of a descriptor for each clinical case, using only the image, without any additional meta-data. Each descriptor is used to obtain the most similar cases with respect to the input query. To classify the cases by their similarity, the descriptors are compared using the Euclidean distance, following the same approximation as in previous work [118], [161], [205]. The proposed CBIR scheme is presented in Figure 5.1.

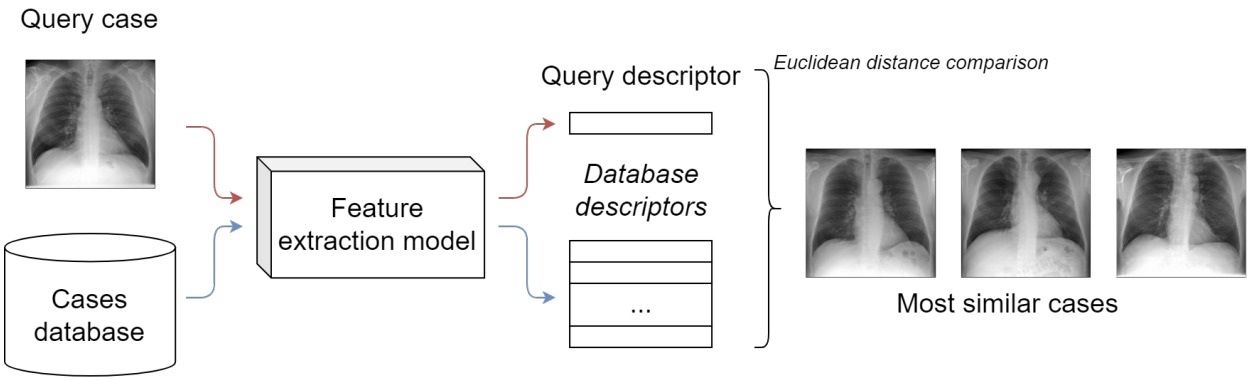


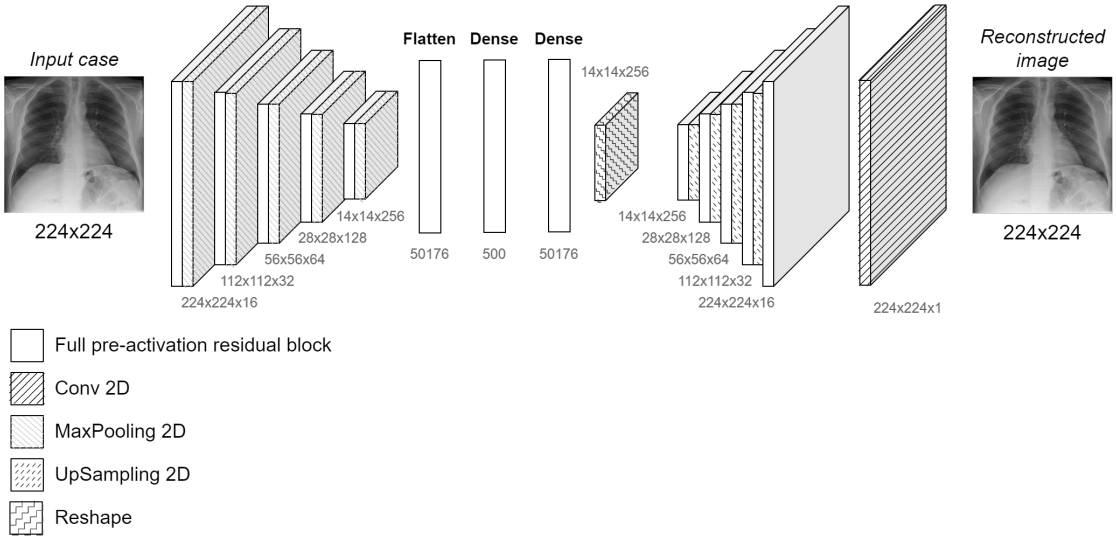
Figure 5.1: Proposed recommendation system diagram.

One of the main purposes of this chapter is to compare the performance and behavior of the MOC-AE compared with the evolution of the MOC-VAE and respect previous DL approaches. This comparison will evaluate how the evolution presented in the thesis performs, its strengths and weaknesses, and the future of the architecture.

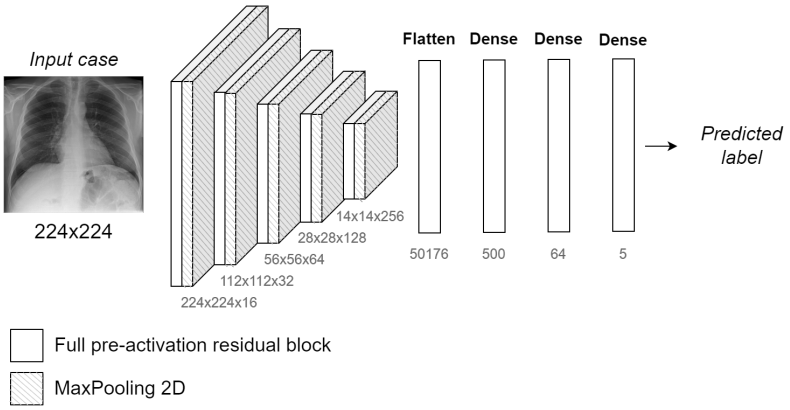
To further solidify the natural progression in the use of various models four models are tested, ranging from vanilla AE architectures, through classification, and finally to models utilized in prior work as MOC-AE as well as the evolution presented in this study, the MOC-VAE. All these models will be used as the feature extraction method for generating the descriptor for the images, their architectures will be discussed and the results will be analyzed to find their differences.

All models share the same neural network backbone design, which uses full pre-activation residual blocks from [77] along with 2-dimensional MaxPooling and 2-dimensional UpSampling for dimensionality changes. The latent space obtained in all cases is a 500 dimensional vector.

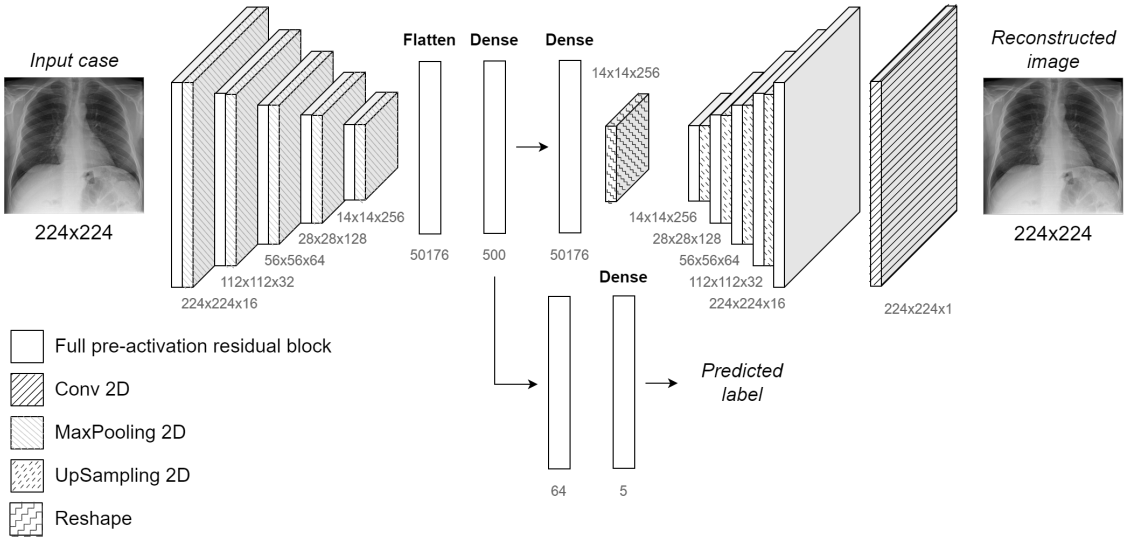
The same architecture for the different models is used, to be able to perform a fair comparison. The architecture of the different neural networks can be observed in Figure 5.2.



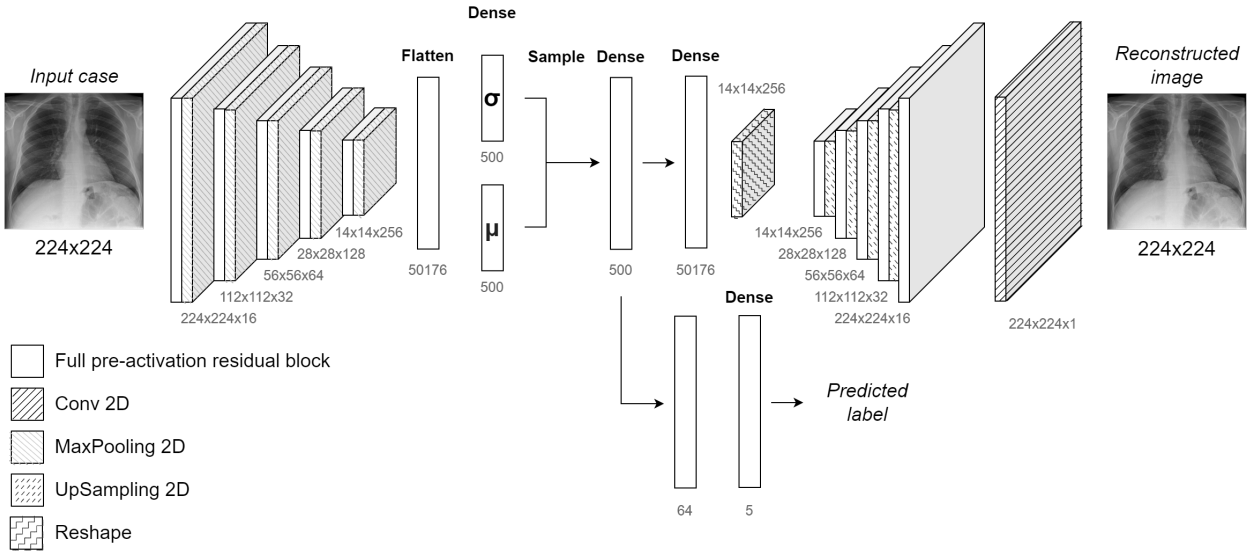
(a) AE feature extraction architecture.



(b) Classifier feature extraction architecture.



(c) MOC-AE feature extraction architecture.



(d) MOC-VAE feature extraction architecture.

Figure 5.2: Neural network architectures for the feature extraction models.

The image descriptors are obtained from the latent vector and utilized to recommend the nearest latent spaces using the K-Nearest Neighbor algorithm using Euclidean distance, thus establishing a recommendation system for the k most similar cases.

Autoencoder

To perform image recommendations, AEs factorize the dimensionality of images, removing all elements lacking relevant information within the image itself and extracting the most relevant features to generate an image descriptor.AEs have demonstrated their capability to extract the most relevant features within their latent space in CBIR tasks [118], [186].

Based on this assumption, the AE, depicted in Figure 5.2a, is presented as an approximation to a vanilla model capable of extracting crucial information from the image through the factorization process to then recommend similar cases.

It is important to acknowledge that such models may not yield the best approximation for specific cases or, in this case, particular pathologies. For example, when detecting conditions such as scoliosis, attention should be directed to specific parts of the image that may have this pathology. These models do not prioritize specific aspects of the image, but rather general shapes and features. For instance, while they can accurately capture body contours and the general arrangement of organs, they may not focus on finer details.

Classifier

In contrast to AE architectures, classification models prioritize extracting specific features from the image rather than the ability to extract generic features as AE do. Therefore, vanilla classifier models are capable of focusing on more specific features, such as cancerous regions or deviations in columns, excluding generic aspects of the image, nor retaining the locality of the extracted information from the image.

The proposed architecture is depicted in Figure 5.2b, featuring the model used in the experiments to maximize the results obtained in the recommendation of medical images.

However, the results presented in Section 5.3.1 demonstrate that the intuition that solely focusing on specific features leads to overlooking too many characteristics and not focusing on the features shape, texture or size.

Multi-Output Classification Autoencoder

Respect the previous feature extraction models a multi-objective architecture is used, capable of capturing both visual and semantic features of the images. In this sense the MOC-AE architecture previously presented in Chapter 4 is used. The network is trained to learn the composition of the image, with an AE output, at the same time that it is trained as a classifier. By leveraging both objectives, the architecture combines in a single latent space visual and semantic information of each case. This leads to a more complex factorization of the information, producing descriptors that better capture the features of the image. MOC-AE focus on guiding the latent space to pay attention to specific features of the images. This allows to control where the network is focused, thus providing more interesting cases from a clinical perspective.

In addition, or the purpose of fully disentangle the behavior of the MOC-AE the loss function will be tested. In particular, the γ hyperparameters defined in Equation (4.4) are simplified in the following expression:

$$L_T = L_r + \gamma * L_c \tag{5.1}$$

where γ controls the weight of the classification output over the reconstruction output. The impact of different values of γ, previously calculated with cross validation, will now be tested

meticulously.

The main drawback of the MOC-AE approximation is that the latent space is not controlled. The relationships between different image descriptors are projected into a low dimensional space without control of the composition or topology of it. Thus, it is proposed to define a method to regularize this space

Multi-Output Classification Variational Autoencoder

Latent space is one of the key components of the system, the indexation of the image descriptors is done in this space, thus, controlling the latent space is a crucial task. VAEs control latent space of vanilla AEs by forcing it to follow a particular data distribution, generally a normal distribution [116]. The MOC-VAE architecture introduces the regularization process of VAEs in the MOC-AE architecture.

By changing the generation of the image descriptors, it is possible to control how each sample is distributed in the latent space. In particular, it is decided to sample points from a Gaussian distribution $\mathcal{N}(0, \mathbf{I})$. By doing so, the points in the latent space are forced to be a continuous and complete space and better capture relationships between samples.

The regularization process takes the Gaussian distributions $\mathcal{N}_x$ and put them close to each other through the Kullback-Leibler divergence, so the completeness is satisfied. This is done by optimizing the means position in the space. The covariance is optimized to cover the different parts of the space and guarantee completeness and continuity. The continuity and completeness create a gradient in the latent space.

Figure 5.2d shows the proposed architecture of the proposed model, following the same foundation as MOC-AE and adapting the latent space to sample cases to a normal distribution.

In this case, the VAE architecture enhances the association of similar samples, better generalizing each case, and providing more precise results. Moreover, by regularizing latent points to follow a normal distribution, Euclidean distance is used to compare the similarity between image descriptors is justified with theoretical foundations, unlike previous researches [118], [161], [205], as it was previously mentioned in Section 2.2.2. Controlling the distribution of the latent space enables us to use a particular distance that uses this data distribution as a basis. In this particular case, the Euclidean distance measures distance in a continuous space where descriptors are distributed uniformly in the space.

5.2.3 Performance Evaluation Metrics

Measuring the performance of CBIR systems is one of the open problems in this area. Previously in Chapter 4, the CBIR performance was measured using segmentation labels of the cases. But this chapter will focus in a more common case, where segmentation labels are not available. It is considered that is not common to have a diagnostic where the data is segmented respect the pathologies. Instead, a classification problem will be faces. Because of that, the performance evaluation will be focused on visual and semantical similarity of the

cases.

Designing a metric that captures both the semantic information of an image and the visual similarity between cases is a complex task. In medical diagnosis, the semantic features of the cases correspond to the anatomical and pathological characteristics of the patients, where visual similarity refers to low-level features, such as shape, texture or size of different elements of the images.

A correct equilibrium between both characteristics will measure how different images represent the same pathology at the same time that it compares each case by how similar they are, respecting image composition. Ultimately, the system must provide a precise case given a query, in order to aid the physician with similar cases, easy to compare due to their visual similarity and accurate because of their semantic similarity.

As discussed in Section 2.2.3 Precision@k is the standard evaluation metric when the number of retrieved samples is small. It is able of measuring the semantical relationships between the retrieved samples respect to the query case. Thus, it will be used to measure if the recommendation shares the same pathologies with the case is being analyzed.

But arguably there is not a good metric that evaluates visual similarity in CBIR. The Mean Squared Error (MSE) is not a good option because its sensibility to image deformations, as discussed in Section 2.2.3. Therefore, it is proposed to use a novel distance to measure visual similarity between cases.

Sliced Wasserstein Distance

To evaluate the visual effectiveness of the proposed models, 1-Wasserstein [183] distance between images is proposed as an alternative to MSE. Wasserstein distance measures the dissimilarity between multidimensional probability distributions. Wasserstein distance is computed as follows:

$$W(I_u, I_v) = \inf_{\pi \in \Gamma(I_u, I_v)} \int_{\mathbb{R}\times\mathbb{R}} |x - y| \mathrm{d}\pi(x, y) \tag{5.2}$$

where $\pi \in \Gamma(I_u, I_v)$ is the set of all couplings of the distributions I_u and I_v.

This distance can be seen as a measurement of how difficult it is to move one probability distribution to become another, considering each distribution as masses of earth. Thus, for a particular mass, it will be less costly to move it to a closer space than a further one. This is a desirable property in image processing, where locality is crucial.

Considering images as bidimensional functions is a common practice in computer vision [125], [156], [162], e.g. the convolutional operator is an operator of functions. It is considered that comparing images using the Wasserstein distance better preserves local similarities than MSE. The main strength of this metric is its invariance respecting deformation, which will better preserve visual similarities between cases.

The Sliced Wasserstein distance [28] is an adaptation of the Wasserstein distance that uses radial projections of the input data used. By using these projections, the computation is much more efficient, making it possible to calculate distance in multi-dimensional data, which

would be intractable in high-dimensional images [119]. In order to further normalize each image it is equalized using its histogram, thereby balancing the gray level of each image.

The computation of the normalized Sliced Wasserstein distance is approximated as follows:

$$SW\left(I_u, I_v\right) \approx \frac{1}{L} \sum_{l=1}^{L} W\left(\mathcal{R} I_u\left(\cdot, \theta_l\right), \mathcal{R} I_v\left(\cdot, \theta_l\right)\right) \tag{5.3}$$

where I_u and I_v are the compared distributions, which are sampled with a simple Monte Carlo scheme L times. Each image is projected over the unit sphere in $\mathbb{R}^d$ with a uniform distribution, which is denoted as the transform $\mathcal{R}$ for each θ_l samples.

In particular, $L = 100$ is used to measure the Sliced Wasserstein distance.

5.3 Results

The dataset from the input of the experiment consists of 19.410 images. The images are resized to 224x224 pixels and normalized within the range [-1, 1] with the same process as the one described in Section 4.2.1. 10% of the whole dataset (1.941 images) is randomly reserved for testing purposes.

Experiments and tests were run on two 48 GB Nvidia Quadro RTX 8000 GPUs and an Intel Xeon Bronze 3206R CPU @ 1.90GHz. Neural networks were designed and trained using Tensorflow package [1].

5.3.1 Quantitative Results

The system is trained on Padchest dataset [30]. The system architecture uses residual blocks [76] with full pre-activation shortcut connections [77]. *Binary crossentropy* is used as the classification loss and MSE as the reconstruction loss. Different γ configurations where tested.

Table 5.2 shows *Precision at k* (↑ the higher the better) results and Table 5.3 shows *Sliced Wasserstein distance* (↓ the lower the better) results for the proposed models. Each model is tested with different γ and k configurations for 500 cases. The system is tested with 3 different training iterations for each configuration and show the mean results for each one.

Model	**MOC-AE**				**MOC-VAE**			
k	3	4	5	10	3	4	5	10
$\gamma = 1$	50.22	49.62	48.79	46.64	**51.15**	**50.20**	**50.11**	**49.89**
$\gamma = 0.1$	51.00	50.02	49.31	47.05	49.36	48.82	48.37	47.38
$\gamma = 0.01$	49.24	47.83	46.89	44.47	47.45	46.73	46.36	44.50
$\gamma = 0.001$	45.06	44.00	43.41	41.29	41.47	40.58	40.17	38.41
$\gamma = 0.0001$	39.15	38.35	37.68	36.46	39.15	38.35	38.05	36.34

Table 5.2: Mean precision@k results for different k and γ values (↑ the higher the better).

Model	**MOC-AE**				**MOC-VAE**			
k	3	4	5	10	3	4	5	10
$\gamma = 1$	0.4599	0.4699	0.4771	0.5056	0.4702	0.4769	0.4827	0.5071
$\gamma = 0.1$	0.4440	0.4521	0.4592	0.4846	0.4477	0.4549	0.4638	0.4876
$\gamma = 0.01$	0.4249	0.4340	0.4424	0.4722	0.4138	0.4235	0.4307	0.4571
$\gamma = 0.001$	0.4364	0.4501	0.4613	0.4954	0.4120	0.4200	0.4278	0.4575
$\gamma = 0.0001$	0.4348	0.4469	0.4563	0.4905	**0.4054**	**0.4163**	**0.4244**	**0.4524**

Table 5.3: Mean sliced Wasserstein distance results for different k and γ values (↓ the lower the better).

In order to compare the proposed architecture with previous solutions, the AEs and classifiers CBIR systems described in section 5.2.2 are trained and tested. Table 5.4 shows a comparison of the Precision at k and Table 5.5 shows a comparison of the Sliced Wasserstein distance, comparing the best results of the proposed architecture against AE and the classifier results.

k	3	4	5	10
AE	38.22	37.17	36.76	35.13
Classifier	50.47	49.93	48.69	46.76
MOC-AE$_{\gamma=0.1}$	51.00	50.02	49.31	47.05
MOC-VAE$_{\gamma=1}$	**51.15**	**50.20**	**50.11**	**49.89**

Table 5.4: Mean precision@k results for the different models (↑ the higher the better).

k	3	4	5	10
AE	0.4223	0.4319	0.4417	0.4744
Classifier	0.4702	0.4787	0.4848	0.5074
MOC-AE$_{\gamma=0.0001}$	0.4348	0.4469	0.4563	0.4905
MOC-VAE$_{\gamma=0.0001}$	**0.4054**	**0.4163**	**0.4244**	**0.4524**

Table 5.5: Mean Sliced Wasserstein distance results for the different models (↓ the lower the better).

5.3.2 Image Retrieval of the System

Figure 5.3 shows the image retrieval output of the proposed methodology. Each row of the image contains different samples of the CBIR. The first column contains the query image used as input for the system, while the rest of the columns represent the retrieved results ordered by proximity from left to right. For each image, the pathology labeled in the dataset is shown, as well as the age and sex of the patient. The model used for this evaluation is MOC-VAE with $\gamma = 0.01$, as it is considered to better balance visual and pathological similarity in the recommendation. The 5 nearest neighbors are shown for each case. The samples are fair random draws, not cherry picked.

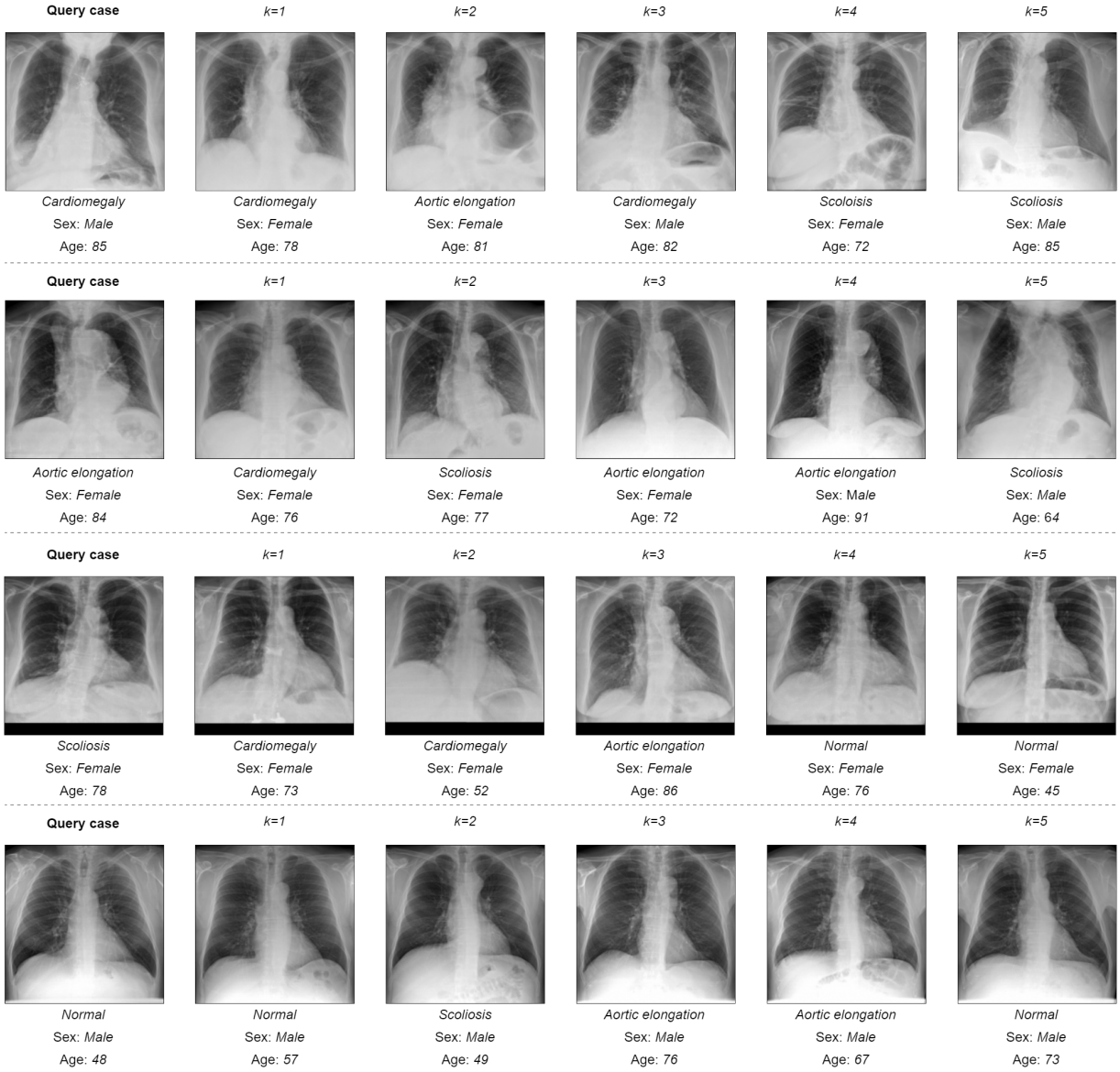


Figure 5.3: Recommendation results for the proposed contribution.

5.4 Discussion

The proposed MOC-AE and MOC-VAE architectures come with improvements with respect to the previous CBIR approaches. By combining the reconstruction and classification outputs of the neural network, the proposed contribution is able to produce precise and similar results on image retrieval tasks.

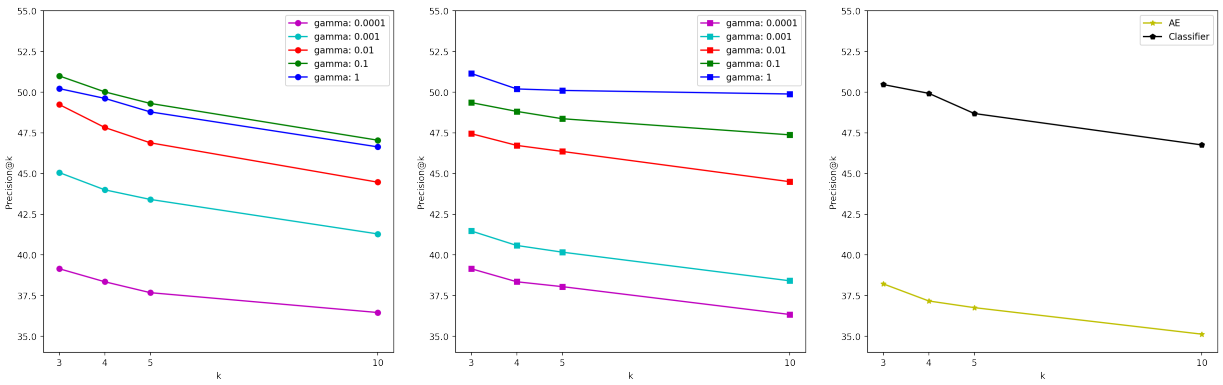
It can be observed that γ parameter is capable of controlling the weight of visual and semantical similarity between cases. Models trained with a higher γ focus more on pathological features, thus producing better results in Precision@k, as can be observed in Table 5.2. Although

models with a lower γ have better results in Sliced Wasserstein distance, as shown in Table 5.3. By analyzing the results presented it can be observed that MOC-VAE produces the best results in both Precision@k and Sliced Wasserstein distance.

Table 5.4 shows that the best results for Precision@k are produced by MOC-VAE with $\gamma = 1$, followed by MOC-AE with $\gamma = 0.1$, outperforming the AE and classifier models. Specially when comparing the proposed architecture against the classifier, which only tries to factorize semantical information. It is demonstrated that including visual information in the training helps to produce better results with respect to classification metrics. That means that, with a multi-objective, the proposed network is capable of producing better results than a network specifically designed for classification. A similar behavior happens with the Sliced Wasserstein distance, shown in Table 5.5, which is lower in the proposed best model than in the AE, which only focuses on maintaining visual features of the images.

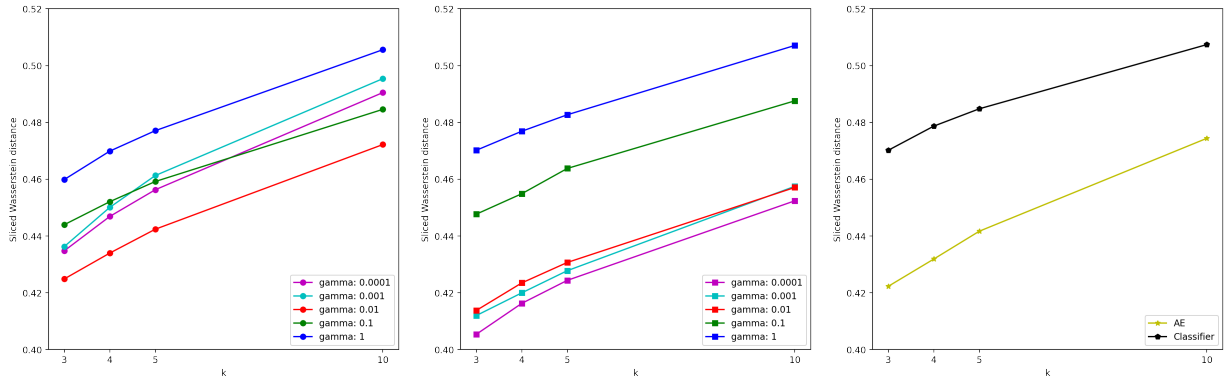
If the contribution is compared with the AE and classifier models, the results are capable of outperforming not only these models, but also produce more balanced results. It can be observed how the classifier produces great results in Precision@k, but it comes with a downgrade on the visual similarity results of the Sliced Wasserstein distance. The same happens with AE, which produces great results in Sliced Wasserstein distance at the expense of very poor results in Precision@k.

There is a trade-off between both metrics. Models must keep in the latent space certain features and, depending on the training, the factorization will focus on semantical information or visual features. Figures 5.4 and 5.5 show the evolution of Precision@k and Sliced Wasserstein distance as the γ value changes. It is considered that there is a "sweet spot" in γ where the model produces a balanced latent space with $\gamma = 0.01$ for both models, with a high Precision@k and low Sliced Wasserstein distance.



(a) MOC-AE precision@k. **(b)** MOC-VAE precision@k. **(c)** AE and classifier precision@k.

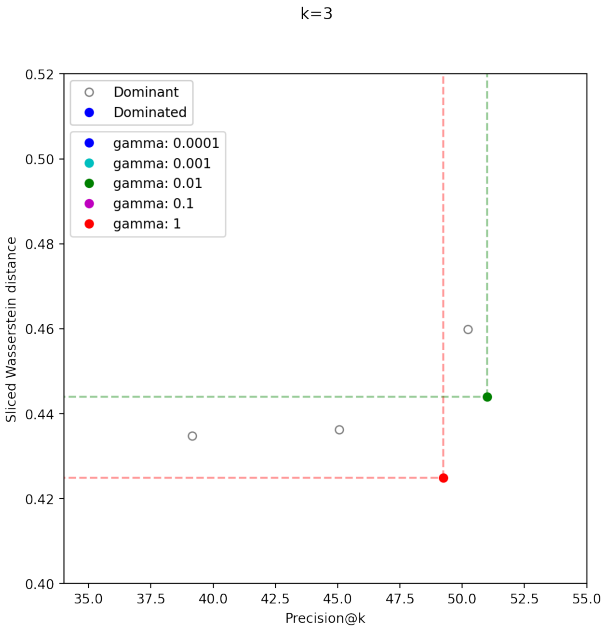
Figure 5.4: Precision@k results for the different models (↑ the higher the better).



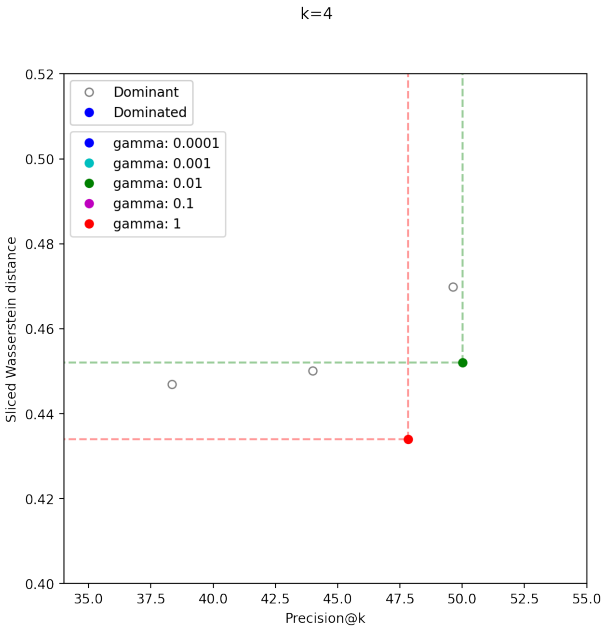
(a) MOC-AE Sliced Wasserstein distance. **(b)** MOC-VAE Sliced Wasserstein distance. **(c)** AE and classifier Sliced Wasserstein distance.

Figure 5.5: Sliced Wasserstein distance results for the different models (↓ the lower the better).

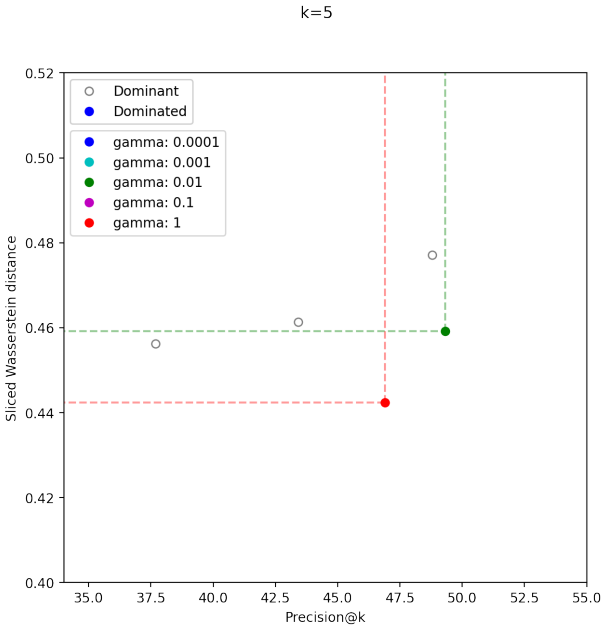
In addition, for each model, the Pareto front is calculated. Figure 5.6, Figure 5.7 and Figure 5.8 show th^{o}e results for each γ configuration, divided by k. The results of MOC-AE and MOC-VAE are also compared with the Classifier and AE.



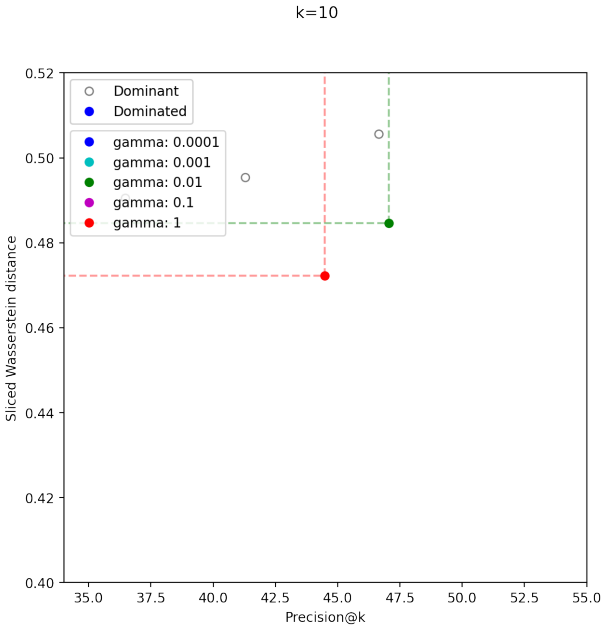
(a) MOC-AE Pareto front with k=3.



(b) MOC-AE Pareto front with k=4.

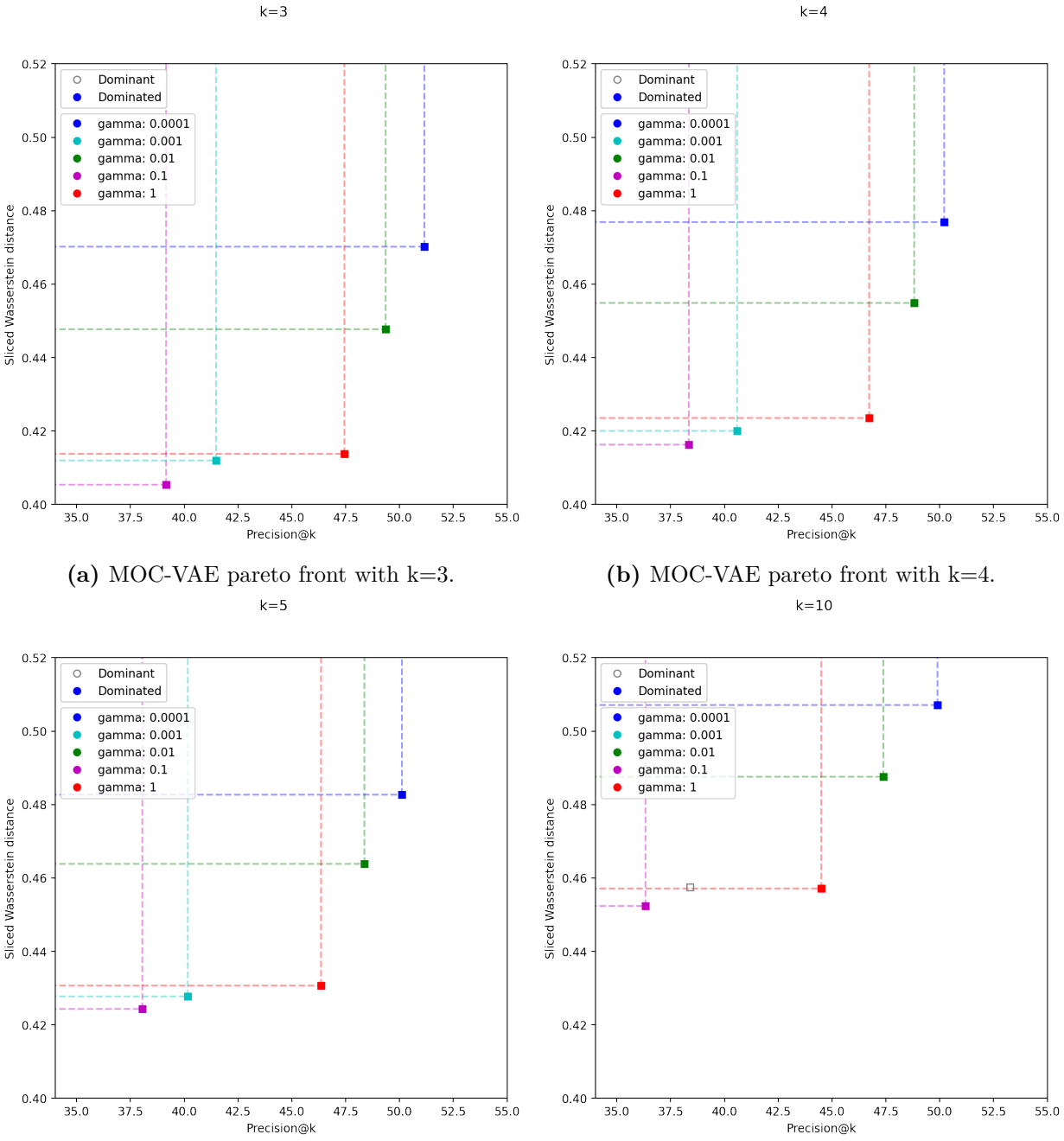


(c) MOC-AE Pareto front with k=5.



(d) MOC-AE Pareto front with k=10.

Figure 5.6: MOC-AE Pareto fronts.



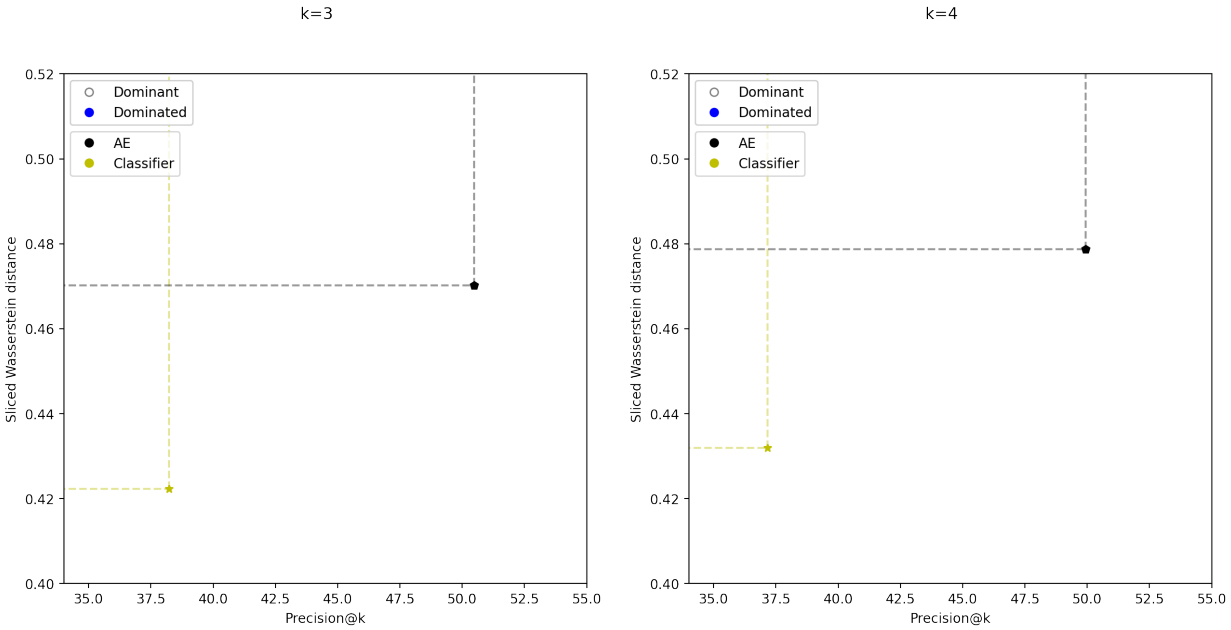
(a) MOC-VAE pareto front with k=3.

(b) MOC-VAE pareto front with k=4.

(c) MOC-VAE pareto front with k=5.

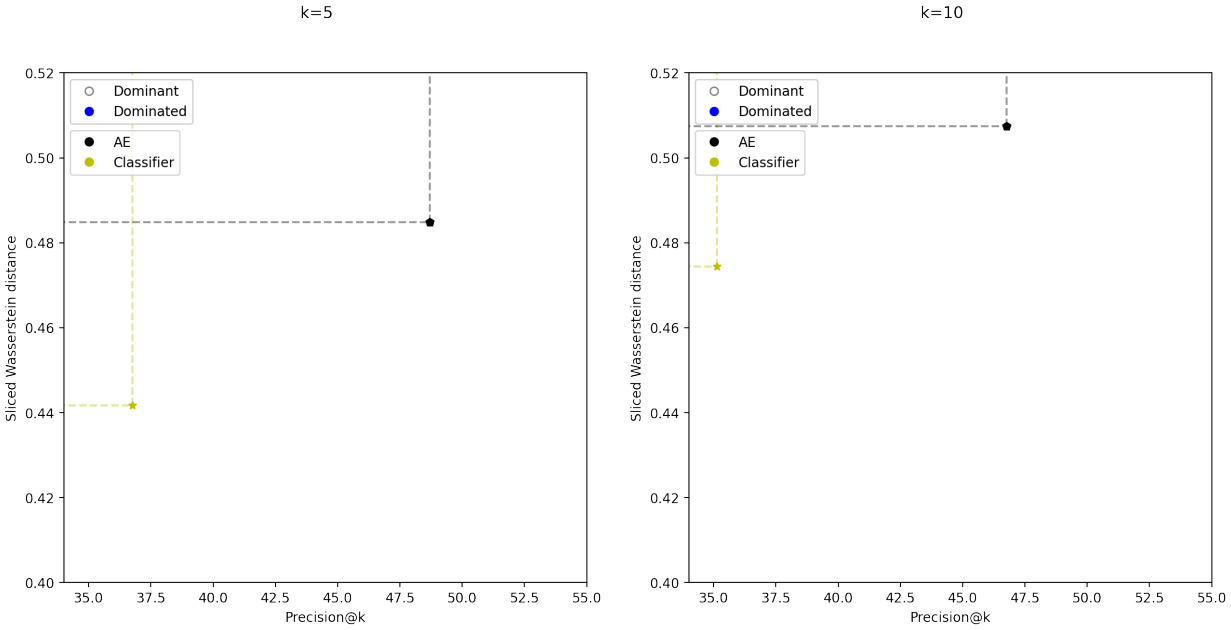
(d) MOC-VAE pareto front with k=10.

Figure 5.7: MOC-VAE pareto fronts.



(a) AE and Classifier pareto front with k=3.

(b) AE and Classifier pareto front with k=4.

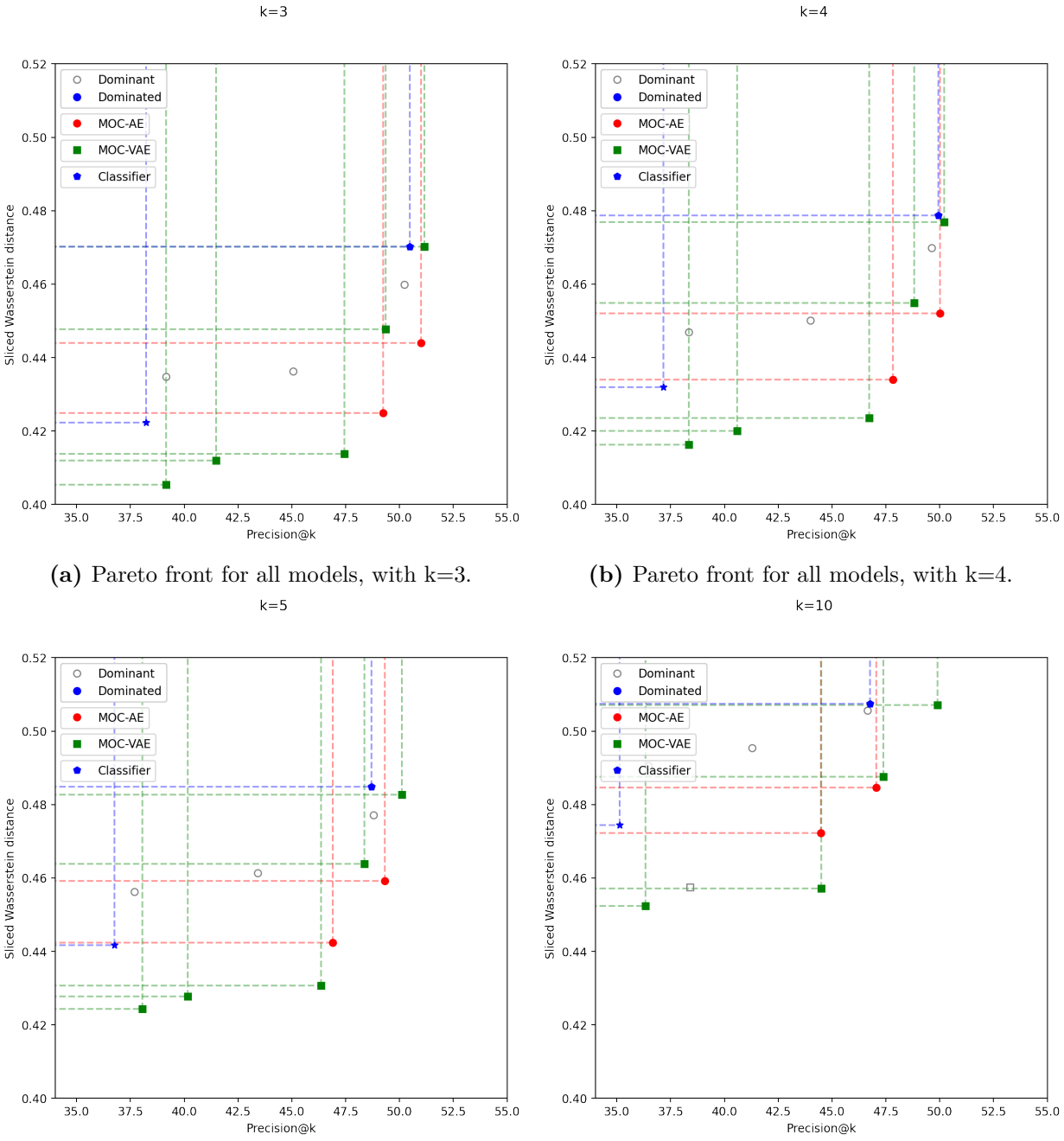


(c) AE and Classifier pareto front with k=5.

(d) AE and Classifier pareto front with k=10.

Figure 5.8: AE and Classifierppareto fronts.

Likewise, Figure 5.9 show all the Pareto frontiers in the same space, where it can be compared the results of MOC-AE against MOC-VAE and AE and Classifiers.



(a) Pareto front for all models, with k=3.

(b) Pareto front for all models, with k=4.

(c) Pareto front for all models, with k=5.

(d) Pareto front for all models, with k=10.

Figure 5.9: Pareto fronts for all models.

To study the convergence of each set of solutions, and to further validate the performance of the MOC-AE and MOC-VAE against Classifiers and AEs, the hypervolumes of the Pareto front of the different architectures are studied. The hypervolume reflects how colose is the set of solutions to the true Pareto front. Table 5.6 contains the hypervolumes for each architecture, comparing their results for the different k.

Model	k=3	k=4	k=5	k=10
AE and Classifier	0.2857	0.2777	0.2667	0.2419
MOC-AE	0.2930	0.2827	0.2745	0.2480
MOC-VAE	**0.3015**	**0.2908**	**0.2862**	**0.2704**

Table 5.6: Hypervolumes for the different models (↑ the higher the better).

As it can be seen in Table 5.6, the MOC-VAE is the model with better performance, followed by MOC-AE, and finally the traditional methods. In addition, the metric non dominated points by reference set ($C2_R$) [70] is calculated for all the models using $k = 5$ for being the most common configuration in Table 5.7. This metric shows how many points of a reference set of solutions dominate another set of solutions.

Reference set	Evaluated set		
	AE and Classifier	MOC-AE	MOC-VAE
AE and Classifier	-	0.0	0.0
MOC-AE	0.4	-	**0.2**
MOC-VAE	**1.0**	**0.4**	-

Table 5.7: $C2_R$ for the different models (↑ the higher the better).

Thus, it is justified the selection of MOC-VAE against the rest of alternatives. Within the MOC-VAE, it is considered that MOC-VAE with $\gamma = 0.01$ is the one with best results balancing both objectives. This model is further analyzed individually to observe its performance under different scenarios, breaking down the cumulative results previously shown.

To provide a more in depth perspective of how the model performs under different circumstances, the model is tested for each pathology separately. Table 5.8 and Table 5.9 contains results for diagnosing each pathology; denoted by cardiomegaly (C) aortic elongation (A) and scoliosis (S); respect normal cases. I.e. query images are compared with cases of the same pathology or labeled as normal.

Base label	3	4	5	10
C	59.67±34	59.57±31	58.27±29	56.48±23
A	58.80±31	59.13±27	57.08±25	54.38±21
S	62.31±34	62.07±31	60.57±30	56.99±27

Table 5.8: Precision@k for the different pathologies against normal cases (↑ the higher the better).

Base label	3	4	5	10
C	0.44±0.2	0.45±0.2	0.46±0.2	0.48±0.2
A	0.42±0.2	0.42±0.2	0.43±0.2	0.45±0.2
S	0.46±0.2	0.47±0.2	0.48±0.2	0.51±0.2

Table 5.9: Sliced Wasserstein distance for the different pathologies against normal cases (↓ the lower the better).

The results show a similar performance between all the cases. Results suggest that the model is not biased respect any particular pathology.

In addition, to provide a view of how the model performs against different typologies of patients, the population is stratified into different groups. The cases are separated by their sex and by their age, with a threshold of 65 years. Tables 5.10 and 5.11 contain results for this test.

Base label	3	4	5	10
Sex: F	47.51±32	46.71±30	46.53±27	44.68±22
Sex: M	46.69±34	46.25±29	46.83±27	43.94±22
Age $\geq$ 65	48.31±33	46.79±29	47.00±27	44.37±22
Age 65	45.39±33	46.08±30	46.12±27	44.37±22

Table 5.10: Stratified precision@k results (↑ the higher the better).

Base label	3	4	5	10
Sex: F	0.42±0.2	0.43±0.2	0.44±0.2	0.47±0.2
Sex: M	0.43±0.1	0.44±0.1	0.44±0.1	0.47±0.1
Age $\geq$ 65	0.42±0.2	0.43±0.2	0.44±0.2	0.47±0.2
Age 65	0.43±0.2	0.44±0.2	0.44±0.2	0.47±0.2

Table 5.11: Stratified Sliced Wasserstein distance results (↓ the lower the better).

In these tables it can be seen a similar performance for the different groups of patients, suggesting that the model is capable of generalize the pathologies respect other anatomical features.

5.4.1 Clinical Evaluation

From a clinical perspective, the results of the retrieval system shown in Figure 5.3 show a behavior similar to a professional clinical diagnosis. The results of the system show good performance in visual and pathological similarities between the images, providing a complete, precise, and interesting set of cases for each query. The results are in line with the quantitative results discussed previously.

In addition, understanding and assessing the feasibility of models requires consideration of two aspects. Firstly, the model's results are stratified into different age and sex groups and observed that there are no differences when associating images. Secondly, analyze each of the pathologies (cardiomegaly, aortic elongation, and scoliosis) in a one-to-one relationship with normality. These two aspects would help discriminate biases when establishing precision.

It should be noted that the only uncontrollable bias to consider in the precision assessment of these models is based on the publicly available image dataset used, where the image labeling was already established, and it is not possible to have quality control for each of the pathologies. However, it is important to emphasize that all models have been tested on the same image datasets.

5.5 Conclusion

This chapter presents a novel architecture for CBIR, the MOC-VAE, result of an optimization of a previous the MOC-AE network. The proposed MOC-VAE network simultaneously focuses on the anatomical and pathological features of each case, obtaining a precise and easily interpretable recommendation. By combining two different objectives, the model focus in visual and semantic similarity of images, i.e. the recommended cases will present the same pathologies while being visually similar. The main novelty of the MOC-VAE respect the MOC-AE is that it regularizes the latent space. By regularizing the latent space with a Gaussian probability distribution the image descriptors can be compared using a structured space, thus controlling the recommendation space and providing better results.

MOC-VAE is compared with similar state-of-the-art architectures, testing empirically its results. For this test, the PadChest dataset [30] is used to train the different models. Using this dataset, the results provide a view of the real performance of the architectures against real world data. Furthermore, the simplicity of the proposed model makes it possible to easily adapt it to new cases.

Quantitatively, the results show that MOC-VAE outperforms previous architectures in image recommendation, not only better balancing visual and pathological similarities between images, but also improving precision and similarity independently. Therefore, the proposed system outperforms state-of-the-art systems in all aspects.

The results of the image recommendation suggest that the performance of the model is very similar to a professional physician diagnosis, relating similar cases focusing on the pathologies and anatomy of the patient.

Hence, the introduced framework offers a significant improvement to assist medical diagnostics, utilizing DL models that regularize the space from which the recommendation is made. This addition helps to provide an accurate and visually attractive collection of cases that improve medical diagnostics without being intrusive.

5.6 Future work

The simplicity of the architecture of the MOC-AE and MOC-VAE makes possible to use the architecture against different problems. In particular, it is considered interesting to test the models against other medical datasets, that may be interesting from a clinical perspective. Future research of this thesis will be focused on applying this architecture to real world scenarios, training and implementing the MOC-VAE in real hospitals to receive feedback from the physicians at the same time that the model is applied in the area it was designed for.

In this sense, the results of the MOC-AE and MOC-VAE discussed in this chapter will be the foundations of future research on applying the models to medical comparative diagnosis, where the medical community can benefit from the processing of the DL models. By applying this framework, it is expected that the physicians diagnostic would be more precise, specially

in difficult cases where a comparison could drastically improve the accuracy of the physicians.

On the other hand, future work can be done respect improving the proposed architectures, e.g. researching the implementation of different architectures. In this respect, it has been shown that the use of transformer blocks [223] achieves better performance than their convolutional counterparts [218]. Further research in improving the architecture could also be focused on studying the topology of the latent space, trying to research new methods of regularization that could be beneficial for the recommendation.

Moreover, it is considered that, because of the overall simplicity of the architecture, the MOC-AE can be applied to different paradigms with few changes. For example, the dimensionality reduction of the latent space can be leveraged in eXplainable Artificial Intelligence (XAI), where modifications of the latent space can be reconstructed as explanations.

5.7 Answers to the Research Questions

- ***RQ1****: How effective are the new methods of CBIR for recommending cases with similar pathologies respect previous models?*
- **Research Question (RQ)1**: Are the new Multi-Output architectures effective for CBIR of medical images?

The models presented are evaluated in from three perspectives, measuring semantical correlations between cases, measuring visual similarities and measuring the balanace between both features.

Respect semantical results of the architectures, the MOC-AE and MOC-VAE obtain a Precision@k (↑ the higher the better) with $k = 5$ of 50.11 and 49.31 respectively compared with the 48.69 and 36.76 values of the classifier and AE respectively. For the rest k values the Precision@k of the MOC-AE and MOC-VAE outperform classifier and AE architectures. These values ensure that the MOC-VAE is the architecture with better results that recommends cases with similar pathological characteristics.

Respect visual results of the architectures, the MOC-AE and MOC-VAE obtain a Sliced Wasserstein distance (↓ the lower the better) with $k = 5$ of 0.4563 and 0.4244 respectively compared with the 0.4848 and 0.4417 values of the classifier and AE respectively. For the rest k values the Precision@k of the MOC-AE and MOC-VAE outperform classifier and AE architectures. These values ensure that visual characteristics of the patients are better preserved and recommended with the MOC-VAE.

Respect the balance between semantical and visual features of the cases, the MOC-AE and MOC-VAE architectures achieve better results than AE and classifier architectures in relative values when measuring Precision@k and Sliced Wasserstein distance simultaneously. The MOC-AE and MOC-VAE have proved to be more generalistic architectures, balancing both metrics at the same time, rather than focusing in just anatomical or pathological characteristics of the patients.

- **RQ3**: Is the new visual similarity metric able of capturing similarity between cases based on their visual appearance?

The Sliced Wasserstein distance is proposed as a visual similarity metric between images. In the CBIR context, the Sliced Wasserstein distance is able of capturing geometrical relationships between images treating them as probability distributions. The main advantages of this metric are their invariances to image transformations and its efficient computation time.

Chapter 6

Improving Explanations for Medical X-Ray Diagnosis combining Variational Autoencoders and Adversarial Machine Learning

The Multi-Output Classification Autoencoder (MOC-AE) and Multi-Output Classification Variational Autoencoder (MOC-VAE) have already been presented and studied in the Content-Based Image Retrieval (CBIR) context. In this chapter, the MOC-VAE will be studied for eXplainable Artificial Intelligence (XAI) purposes. The simplicity of the MOC-VAE architecture makes possible to apply it to different domains. In order to study this flexibility and showcase its good results in different tasks, the MOC-VAE will be leveraged to find explanations of medical diagnostics.

Explainability in Medical Computer Vision is one of the most sensible implementations of Artificial Intelligence (AI) nowadays in healthcare. In this work, a novel Deep Learning (DL) architecture for eXplainable Artificial Intelligence is proposed, specially designed for medical diagnostic. The proposed approach leverages Variational Autoencoders properties to produce linear modifications of the images in a lower dimensional embedded space, to then reconstruct these modifications to non-linear explanations in the original image space. The proposed approach is based on a global and local regularization of the latent space, that stores visual and semantic information about the images. In particular, a genetic algorithm is designed for searching explanations, finding individuals that are capable of misclassifying the classification output of the network while producing the minimum number of changes in the image descriptor. The genetic algorithm is able to search for explanations without defining any hyperparameter, and using only one individual to provide a complete explanation of the whole image. In addition, the explanations found by the proposed approach are compared with state-of-the art eXplainable Artificial Intelligence systems and the results show an improvement in the precision of the explanation between 56.39 and 7.23 percentage points.

This research leverages the good factorization properties of the MOC-VAE discussed in

Chapter 4 and Chapter 5 to reduce the search space of the proposed genetic algorithm. With this optimization, the proposed algorithm is benefited from the good visual and semantical representation of the MOC-VAE to find precise and meaningful explanations. In this regard, this chapter further explores the possibilities of the architectures of the thesis, exploring how they represent the information about medical cases and their possible applications in other areas.

Additional information about the research of the chapter, including information about paper that holds the research covered in it can be accessed in Appendix .5.

6.1 Introduction

Medical diagnosis using AI is an area in continuous evolution [214]. Trustworthiness in this domain is one of the open challenges nowadays [7], [73]. XAI techniques aim to provide the interpretability of the diagnostic, towards a more comprehensible use of AI. But XAI techniques are still in development and face issues such as fairness [184] or new challenges such as their application to audition problems [219].

Nevertheless, modern XAI research seeks to provide clearer and more precise explanations of AI models. Current research uses linear modifications to the input space to assess how these modifications are translated into the model outputs [226]. When these techniques are applied to images, a search space based on pixels becomes intractable, leading to poor efficiency in the search for explanations. Traditional solutions aim to improve search efficiency by reducing search space and predefining sets of pixels or *superpixels* by different sizes and shapes to find explanations. This approximation leads to different issues: first, the parameters used in this search make assumptions about the explanations and can drastically change the explanations found; and secondly, manually select a proper set of hyperparameters is not simple due to the high number of possibilities [38].

It is proposed to find explanations using an embedded regularized latent space, that stores visual and semantic information about the images. In this latent space, linear modifications are produced to the data to evaluate how these perturbations affect the models outputs. With this evaluation, it is possible to control which features of the latent space are related with the pathologies diagnosed, and how these features are translated to the data input space, i.e. the images. In addition, because the explanations are found in a lower-dimensional latent space, it is ensured this search process to be optimal and the explanations semantically correct. This latent space also makes it possible to not assume any behavior of the modifications or the explanations themselves, allowing the search algorithm to find the explanations by itself.

To achieve this objective the MOC-VAE architecture is able to generate image descriptors that will be subsequently modified to find their explanations. This architecture allows for the generation of descriptors containing both visual and semantic information. Moreover, the multiple outputs of MOC-VAE allow us to obtain information about how the network classification changes are related to latent space perturbations. This generates the saliency map showing the impact of these modifications. This framework translates linear modification

in the latent space to non-linear explanations in the input space, without the necessity of using complex hyperparameters configurations or assumptions about the explanations.

Thus, the MOC-VAE presented in Chapter 5 presents a perfect solution for finding linear explanations in an embedded space that can be translated to non-linear modifications of the original space of the images. The architecture that MOC-AE and MOC-VAE proposes can be leveraged in XAI, providing meaningful precise explanations. During this chapter the possibilities of MOC-VAE in XAI will be studied, defining the strengths and limitations of the architecture in XAI.

In particular, a genetic algorithm is used for searching explanations, finding individuals that are able to misclassify the classification output of the network while producing the minimum number of changes in the image descriptor. This ensures that the modifications found are minimal and produce the minimal explanations possible of the diagnosis. Thus, the genetic algorithm has two main objectives, finding perturbations that makes the network classify an image of a particular pathology as other one, and filter the perturbations to find the minimum explanations that misclassify the model.

Once these set of individuals is found, they are converted to explanations by reconstructing the image of the perturbed descriptor with the reconstruction output of the network. This reconstruction provides information about the visual features that were modified to change the model classification. Thus, the reconstruction output generates the explanations of the system.

The proposed framework is capable of producing more precise semantically coherent explanations of individual diagnostics. With respect to similar approaches, our methodology is easier to train, and also produces explanations that are more related to the pathologies of the patient.

Overall, the main contributions of the chapter can be summarized as follows:

- The Multi-Output Classification Variational Autoencoder (MOC-VAE) is used to find explanations following an XAI scheme. The factorization of the MOC-VAE is used to find explanations in a lower dimensional space from where it is possible to reconstruct the explanation saliency map.
- The Strength Pareto Evolutionary Algorithm 2 (SPEA2) genetic algorithm [246] is used to find explanations in the latent space. The algorithm searches for the minimum number of changes that are able of changing the model's classification output. This minimum individuals are defined respect two measurements, the minimum number of changes and the minimum total magnitude of the changes.
- The results of the XAI is measured comparing the explanations with Local Interpretable Model-agnostic Explanations (LIME) [181], DeepCover [39] and Structured Attention Graphs (SAG) [199]

6.2 Methods

This section presents MOC-VAE from the latent space perspective. This investigation of the latent space properties of the MOC-VAE will be focus on the representation of the descriptors that the MOC-VAE provides. This aspect is specially important for the proposed XAI scheme, because it is based on modifying these descriptors to find proper explanations. Thus, a correct definition of the factorization process of the MOC-VAE is defined in this section, to then apply linear modification to this factorized latent space.

6.2.1 Experimental Data: Padchest Dataset

The Padchest dataset [30], already used in Chapter 5, is used to test the XAI explanations. With regard to Chapter 5 a balanced subset is used in this research. From the 160.000 annotated X-ray images, images of the four most common classes of the dataset that are visually interesting are selected, e.g. classes such as “chronic changes" that were not specific pathologies are not chosen.

In particular, Posterior Anterior projection images [198] are selected, to normalize the projection and position of the patients' organs along the images. Then, images containing a single label for the four most common labels are selected: normal, scoliosis, aortic elongation and cardiomegaly. The final dataset has 6,828 images balanced among the four classes. Each class has exactly 1,707 images.

The preprocessing steps followed in this chapter are exactly the same as the ones described in Section 5.2.1, that discard irrelevant portions of the image using a pre-trained lung segmentation network.

6.2.2 Multi-Output Classification Variational Autoencoder Factorization Process

The previous used MOC-VAE architecture is leveraged in the context of this chapter as a factorization model that is able of creating image descriptors that are then used to find image explanations of the diagnostics. In this context, here the architecture will be analyzed focusing on the modification of this latent space and its impact.

Variational Autoencoder (VAE) is a structure that descends from Autoencoders (AEs). AEs are divided into the encoder (e) and the decoder (d). Giving a dataset X, for each data point $x \in X$, AEs aim to find an optimal encoder (e^*) / decoder (d^*) structure satisfying:

$$(e^*, d^*) = \arg \min_{(e,d) \in E \times D} \sum_{x \in X} \epsilon(x, d(e(x))) \tag{6.1}$$

where $\epsilon()$ is the decoding error. Therefore, AEs aim to minimize the decoding error of each encoded point. This can be performed by minimizing the loss function:

$$loss(x) = ||x - d(e(x))||^2 \tag{6.2}$$

As it can be seen, the process of encoding the input information that the network receives can be seen as a factorization process where $e(x)$ corresponds with the descriptor of each image.

The main limitation of AEs is the behavior of the latent space $e(X)$. This space does not guarantee neither continuity nor completeness, two topological properties that will allow us to identify explanations directly using linear transformations in the latent space. VAEs were introduced to address these limitations. Considering a datum x (in our context an image), the MOC-VAE architecture generates a descriptor $z \in e(X)$ that captures visual and semantic features of the dataset X (in the current context medical images). Although the classification model ($C : X \rightarrow Y$, where Y is the classes domain) is specially designed for medical image analysis, where visual and semantic similarities between images are crucial, its general structure makes it possible to apply the same architecture to other domains.

VAEs perform a global and local regularization of the latent space. These regularizations will allow continuous transformations within the latent space while also guaranteeing completeness (i.e. no reached point will have no semantic interpretation after decoding). MOC-VAE's classification step allows points of different classes remain closer than in the original VAE. The regularization uses two properties, 1) each data point x becomes a Gaussian distribution $\mathcal{N}_x(\mu_x, \sigma_x)$ in the latent space, 2) the distribution parameters are chosen in such as way that there is continuity among the different distributions $\mathcal{N}_x$ and they are all centered around $\mathcal{N}(0, \mathbf{I})$.

However, the regularization process potentially increases the reconstruction error [116]. The effect of the VAE architecture will be evaluated in the classification error to understand its impact in Section 6.3. Moreover, this classification error will be used to guide the search process.

The decision of using a VAE instead of an AE is justified by the latent space distribution of the descriptors. VAEs follow the Gaussian distribution $\mathcal{N}(0, \mathbf{I})$. This controls how the latent space is structured and makes it possible to perform lineal transformations to the descriptors. Figure 6.1 shows the general architecture of MOC-VAE, which is the same one of Section 5.2.2.

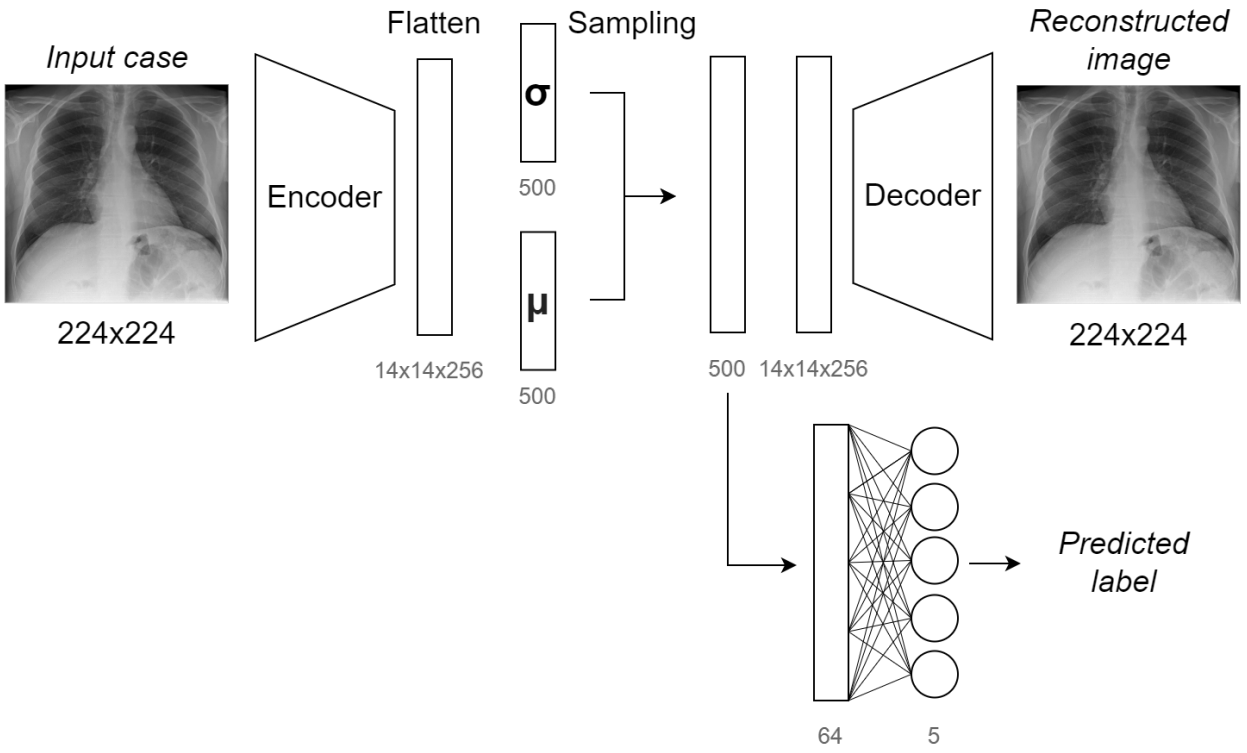


Figure 6.1: MOC-VAE architecture

6.2.3 Navigating the Latent Space

The VAE defines a gradient that allows linear transformations in the latent space. For explanations, the aim of the research is to follow the gradient up to reach boundaries among classes. Considering a specific class $y = C(z) = C(e(x))$, and a negligible δ, a boundary $B \subseteq Z = e(X)$ is defined as:

$$B = \{z \in Z | C(e(x)) = y_1 \wedge C(e(x) + \delta) = y_2,\ y_1 \neq y_2\} \tag{6.3}$$

for two different classes y_1 and y_2. Thus boundary areas will be reconstructed to identify relevant features among classes. The gradient is optimized following adversarial transformations focused on identifying B. Maximizing the classification while minimizing the changes to pass from a specific class to another through the gradient will get us closer to B (with an error close to δ) and help us find the weakest 'jump'.

The latent space of the network is modified to perform the misclassification and then apply XAI techniques. By modifying the descriptors z the classification output $C(d(z))$ is modified, guiding the descriptor to a specific label y.

Each descriptor is an M-dimensional vector. An "individual" (I) will be created defining the linear transformation that modifies the original descriptor. Each individual is also an M-dimensional vector that, added to the descriptor, will create a mutated descriptor ($z' = z + I$).

This modified descriptor will produce altered outputs on the network. It is possible to measure how the outputs change after adding the mutation individual to the descriptor. On one hand, by measuring the changes in the classification output, it is possible to measure the misclassification probability after modifying the latent space, in particular it is desirable to measure how the target label changes after the modification. On the other hand, in the context of images, it is possible measure which sections of the image are related with the modifications of the latent space, reconstructing the mutated descriptor, creating then the saliency map.

Figure 6.2 shows the scheme of the mutation process and changes evaluation.

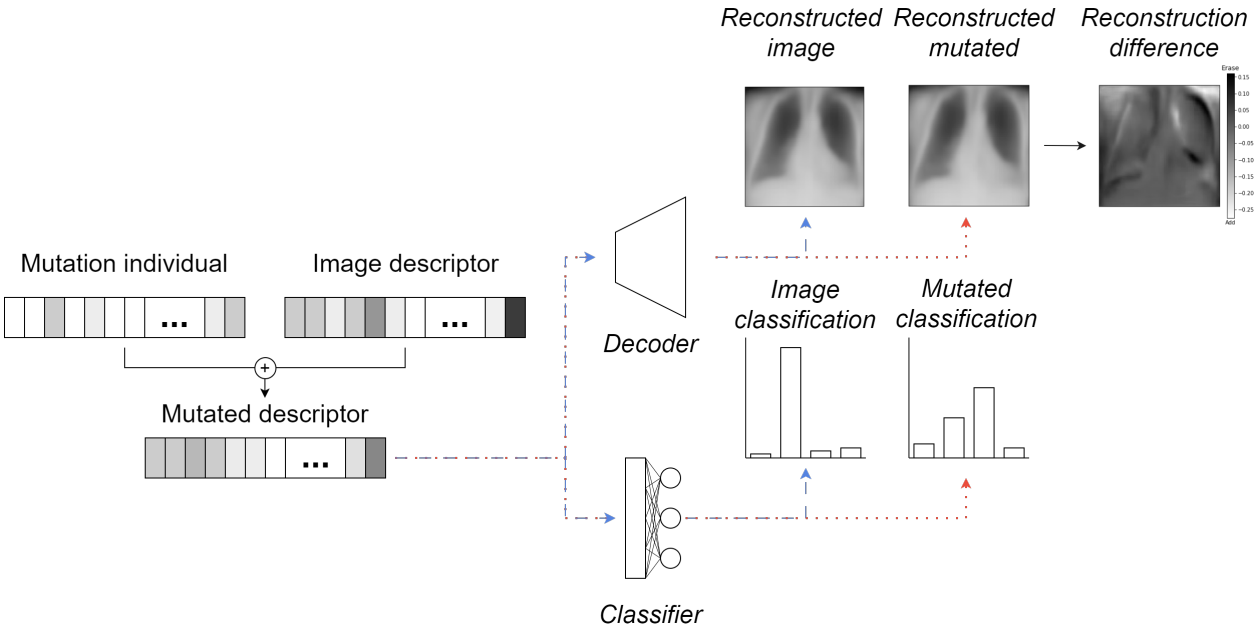


Figure 6.2: Mutation process and evaluation of the results.

In addition, each individual will be evaluated by testing if the mutation is capable of misclassify the original image label into a different specific targeted class. For example, an image labelled as "scoliosis" will be mutated to evaluate which features make it "normal". The alterations connected with the mutations will clearly show those pixels that define the spine deviation. Thus, the gradient in the latent space will find the boundary between the classes "normal" and "scoliosis". A descriptor in the boundary between these two classes will also minimize the number of changes while is still misclassified. Considering the latent space semantic properties and the VAE's regularization when the descriptor is reconstructed to an image, the saliency map between the original image and the reconstruction will highlight the changed areas, visually showing the explanations.

6.2.4 Adversarial Search Optimization with Genetic Algorithms

The optimization process starts with a population P of solutions that represent different individuals (I). Each individual will be optimized to misclassify the original label and obtain a new classification for a specific objective class. This objective is achieved by maximizing the classification output of each individual for the targeted class.

In addition to the misclassification objective, another two objectives are defined related to the boundaries among classes. In order to find the minimum explanation or boundary B, the number of mutations will be minimized. The fitness function measures these objectives. The first is the minimization of the number of changes necessary to achieve the misclassification. This is measured as the total number of positions in each individual that are not equal to 0, and that therefore change the original descriptor. The second objective minimizes the weight of the mutation in the descriptor, for this purpose this objective is defined as the minimization of the total magnitude of the mutations to misclassify a case. This objective is measured by summing all the positions of the mutated individuals, in absolute value, obtaining the total value of the mutation. The third and last objective, maximizes the target class classification probability. The final fitness is:

$$f(I) = \begin{cases} \frac{1}{N}\sum_{i \in I} \delta(i) & \text{Objective 1} \\ \frac{1}{N}\sum_{i \in I} i & \text{Objective 2} \\ p(C_t(z+I)) & \text{Objective 3} \end{cases} \tag{6.4}$$

For each individual I, the fitness is calculated using the genes $i \in I$ values. The first objective employs Dirac's δ which is 1 when i is different to 0 and 0 otherwise. The first and second objectives are minimized. The last objective is the target class probability $C_t()$ for the mutated descriptor $z' = z + I$, which are aimed to be maximum.

With these three objectives a multi-objective optimization is performed using the Strength Pareto Evolutionary Algorithm 2 (SPEA2) [246] algorithm. Each optimization is defined by a base image from which the image descriptor is obtained and the target label. For each iteration of the algorithm, a population of solutions is obtained that form the Pareto frontier of the iteration. After performing the optimization, the unsuccessful individuals are discarded. It is determined whether an individual is successful by whether it is capable of misclassify the classification output of the architecture.

The optimization process of SPEA2 (described in Algorithm 1) starts with a random population P_0 of individuals and an empty archive $\overline{P_0}$. At each iteration the archive $\overline{P_{t+1}}$ is filled with the archive and non-dominated individuals of previous generations (line 8). Then, if the number of individuals is larger than the fixed size of the archive, the archive is truncated (line 10), otherwise dominant solutions are included into the archive (line 12). The algorithm uses a two-point cross-over operator, along with tournament selection as the mating operation for the mating pool (line 17). In order to mutate each individual, two different operators are defined. To generate new solutions, a Gaussian mutation operator is used. To explore

solutions with fewer mutated genes, a reverse mutation process is proposed, that randomly selects a gene and sets its value to 0. This operator helps to explore the space of solutions that minimize the number and magnitude of total changes.

Once the optimization process ends and SPEA2 returns the archive corresponding to the Pareto front, the final solution is chosen as the first misclassification with minimum changes, because this solution will be the closest to the boundary.

Algorithm 1 SPEA2 Algorithm

Input: Population size N, Archive size $\overline{N}$, Maximum number of generations T
Output: A Pareto-optimal set
Initialize the population P_0 randomly and set the archive $\overline{P_0} = \emptyset$
for generation $t = 0$ to T **do**
    Fitness Assignment:
    Calculate the fitness of $P_t \cup \overline{P_t}$
    Environmental Selection:
    Copy all non-dominated individuals from $P_t \cup \overline{P_t}$ to $\overline{P_{t+1}}$
    if $|\overline{P_{t+1}}| > \overline{N}$ **then**
        Reduce $\overline{P_{t+1}}$
    else if $|\overline{P_{t+1}}| < \overline{N}$ **then**
        Fill $\overline{P_{t+1}}$ with dominated individuals from $P_t \cup \overline{P_t}$
    end if
    Mating Selection:
    Perform binary tournament selection on $\overline{P_{t+1}}$ to fill the mating pool
    Variation:
    Apply crossover and mutation operators to the mating pool to generate P_{t+1}
end for
Return the non-dominated individuals in $A = \overline{P_{G_{max}}}$ as the Pareto-optimal set

Genetic Algorithm Parameters

In terms of parameters, our experiments set the population and archive size for SPEA2 as 80 and 40 respectively and the tournament size as 3. The crossover rate is 0.5, the mutation rate is 0.2, and the reverse mutation rate is 0.3, with 0.1 and 0.4 as the probability of mutating each gene respectively. The parameters of the Gaussian mutation operator are 0 for the mean and 0.1 for the standard deviation. The number of generations of the algorithm is set to 250. These parameters are standard parameters for balancing exploration and exploitation [19].

6.3 Results

MOC-VAE's performance is analyzed with different experiments. In order to measure the effectiveness of the evolutionary optimization process, it is analyzed how the gradient is found

for each possible combination of classes. From a XAI perspective, it will analyzed the results of the solutions for each case, finding out if the modified features of the latent space are connected to anatomical areas of the reconstructed image. It will will also be evaluate how a minimal solution (i.e. close to the boundary) is found by going over the Pareto front of each set of solutions.

One of the major concerns about employing VAEs is the accuracy, therefore, MOC-VAE's quality at classification is evaluated and compare it with other state-of-the-art algorithms. It is reasonable to assume that, because of the inclusion of the reconstruction term in the architecture, the classification properties of the network will be affected. This way, it is measured the properties of the factorization process in terms of measuring the classification properties of the factorized information, which is crucial in the search process.

Table 6.1 shows the classification performance of the proposed MOC-VAE classification output against other models. The contribution is tested against the standard implementation in Keras [40] of VGG19 [201], DenseNet201 [84], EfficientNetB0 [212] and ResNet50 [76] to reflect the performance of the model in classification tasks. All models were trained during 50 epochs with a batch size of 64. Results show classification metrics for 10 different repetitions of the training process for each model tested.

Model	Accuracy	F1-score	Recall	Precision
MOC-VAE	**0.61±0.01**	**0.61±0.02**	**0.61±0.01**	0.62±0.01
EfficientNetB0	▼0.55±0.03	▼0.54±0.03	▼0.55±0.03	▼0.56±0.03
DenseNet201	0.58±0.07	0.57±0.09	0.58±0.07	**0.63±0.03**
VGG19	▼0.24±0.00	▼0.10±0.00	▼0.25±0.00	▼0.06±0.00
ResNet50	▼0.56±0.02	▼0.55±0.03	▼0.56±0.02	▼0.58±0.02

Table 6.1: Comparison for the classification results of different models and MOC-VAE. The symbols ▼or ▲are used to indicate when the Wilcoxon test shows statistically significant difference (p-value < 0.05) among the results, using MOC-VAE as the baseline comparison.

6.3.1 Effectiveness of the Adversarial Attack

In order to measure the effectiveness of the proposed evolutionary optimization framework at finding the boundaries among classes, first its ability to misclassify is measured each combination of classes. For each class, an image that represents the corresponding pathology is selected. It is assumed that its class is y_1. Then, a different class y_2 is chosen. It is considered each pair of labels y_1 and y_2 aiming to transform an image of y_1 into an image of y_2 by searching in the latent space. For each combination of labels in the dataset, this process is repeated 10 times with different images chosen uniformly at random. For each image, the search is repeated also 10 times considering its probabilistic nature. For the 4 classes of the dataset, the average change in confidence is obtained for the 3 remaining classes, using new validation images as test.

These results give a perspective of how efficient the optimization process is. Thanks to their analysis on cases never seen before, it can be seen if the population of solution is able

to misclassify new cases, thus giving a correlation between the solutions and the features represented in the latent space. If a population of solutions is capable of misclassify a case into a certain class it means that the features represented by them are related to this pathology.

Table 6.2 contains the transition matrix for each combination of labels, measured for the 10 different base cases. This table contains the mean increase on the confidence of the network for the objective label. The table shows results for the different labels, named as: Normal (**N**), Cardiomegaly (**C**), Scoliosis (**S**) and Aortic Elongation (**A**). The results show the population percentage that successfully misclassify the classes. The results range from 0.4642 to 0.8033, with a mean of 0.6382. It can also be observed that for all the classes there is a positive result for the adversarial search, proving that it is possible to apply our transformation process to all the pathologies. Nevertheless, some classes produce better results than others, proving that the transition difficulty is not uniform among the classes, and therefore the gradient would require more effort in some cases for identifying the boundaries.

Base label	Objective label			
	N	**C**	**S**	**A**
N	-	0.8033±0.10	0.6777±0.13	0.6305±0.22
C	0.4642±0.22	-	0.6101±0.24	0.6690±0.09
S	0.5499±0.18	0.7413±0.16	-	0.6794±0.15
A	0.4880±0.19	0.6427±0.24	0.7028±0.12	-

Table 6.2: Transition matrix of the adversarial search for the different classes. The value represents the success percentage at the end of each search for the 10 pairs of classes and the statistics are calculated over 10 repetitions per image.

To further analyzes the optimization process, it is studied how many individuals successfully swap classes at the final set of solutions. A successful individual is the one that is capable of misclassify the original label of the image. Considering that in Table 6.2 it is noticed that some gradients are more resistant than others, it is evaluated when the first successful individual appears in each iteration of the search as well as which percentage of successful individual are in the final set. These results are able to explain how difficult is for the algorithm to find a solution and how successful is the algorithm itself in its final solution.

Table 6.3 shows the iteration number of the first successful individual achieved in each search process of the algorithm. Table 6.4 shows the final success of the Pareto front. Each result shows the mean and standard deviation of the 10 repetitions. Table 6.3 shows that the first successful individual appears at the iteration range going from 27 to 99 with mean of 64.46, over 250 generations. Therefore, it is not trivial to misclassify samples and an intensive search process is needed. This is also verified by the rate of successful individuals in the final iteration, proving that is not always easy to misclassify a specific label. The results range from 74.50% to 98.88% of individuals, with mean of 86.45%.

Base label	Objective label			
	N	**C**	**S**	**A**
N	-	64.1±31	42.1±25	69.9±32
C	99.2±43	-	76.9±42	60.9±43
S	27.3±15	52.9±23	-	86.8±45
A	66.2±34	73.5±36	53.7±40	-

Table 6.3: Iteration number of the search process in which the first successful individuals appears of the 250 generations.

Base label	Objective label			
	N	**C**	**S**	**A**
N	-	89.88±11.2%	77.75±20.3%	74.50±18.6%
C	82.88±16.1%	-	93.50±14.5%	82.88±15.7%
S	89.50±10.8%	76.50±17.2%	-	98.88±1.7%
A	87.87±11.3%	84.62±12.5%	98.62±2.1%	-

Table 6.4: Final success rate of the search process at the last generation.

6.3.2 Minimum Explanation of the Changes

The optimization process minimizes the number and magnitude of the changes to perform the misclassification, therefore finding boundaries among classes. It is sought to find and individual that is capable of misclassify a specific sample obtaining a target label classification at the same time that it minimizes the amount of changes applied. This is defined by two objectives in the evolutionary algorithm, minimize the number of changes and minimize the magnitude of the changes. The former is measured by how many positions on each mutation individual are not 0, while the latter is measured with the total sum of the mutations.

Table 6.5 show the results for the minimum individual, respect number of changes, while Table 6.6 shows the minimum individual of each iteration respect the total magnitude of the mutation. Each table shows results measured for 10 different iterations. Results show that the minimum number of changes range from 17.40 to 99.89 positions (out of the 500 available) in the individuals, with mean of 40.81. The total magnitude is between 1.42 and 17.62 with mean of 6.45. It can be seen is drastically easier to misclassify from the label *Normal*.

Base label	Objective label			
	N	**C**	**S**	**A**
N	-	17.40±8.5	49.50±21.4	26.30±15.3
C	25.90±20.8	-	51.60±24.1	51.70±51.9
S	99.89±69.3	19.10±10.6	-	56.80±48.3
A	26.40±29.1	34.00±16.1	31.10±15.2	-

Table 6.5: Average number of positions that change in the individuals with minimum number of changed positions (over 500 positions) that reach misclassification.

Base label	Objective label			
	N	**C**	**S**	**A**
N	-	1.45±1.1	8.72±3.8	3.07±3.0
C	4.13±4.3	-	8.74±3.8	8.53±8.8
S	17.62±13.1	1.42±2.2	-	9.21±8.6
A	3.76±5.5	5.50±3.0	5.20±3.0	-

Table 6.6: Average magnitude of changes in the individuals with the minimum possible changed magnitudes that also reach misclassification.

6.3.3 Visual Representation of the Mutation Individuals

It is measured how each mutated individual modifies an image that does not belong to the same pathology for which it was optimized. These results further analyze how representative are the individuals respect the semantical features they represent. Therefore, by evaluating how each individual is capable of misclassify an image of a different class, the features that represent the classes are disentangled. The experiment is carried out comparing 10 cases of each of the pathologies defined, resulting in 40 total cases.

For those solutions that change pathology, it is visually evaluated whether these changes are connected to the actual pathology and the ares of the image where it is identified. For this evaluation, it is examined which areas of the image where modified after the mutation process. In particular, the minimum successful individuals previously defined for this experiment are used, because these individuals are the closest to the boundaries and therefore they will show those features that are the most relevant for changing classes.

To highlight the changed areas, it is compared compare the reconstruction of the original individual and the best mutated individual. Then, the difference between both reconstructions is taken, that provides the explanation. The difference between the original and a the mutated individual shows which features have been changed to obtain the misclassification. This process makes it possible to translate the linear modification produced in the latent space to non-linear modifications on the original pixel space.

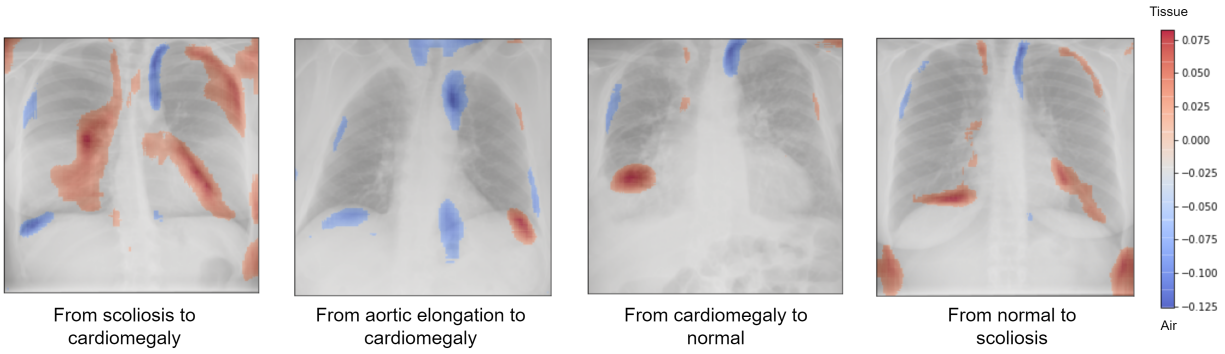
To have a clear perspective of which areas have been primarily changed, it is defined a threshold of the magnitude of the change for visualization purposes. This threshold is set at 30% of the total magnitude, so that for each pixel, it is shown if its' value is above or below the threshold.

To quantify the visual results, each experiment is empirically analyzed, selecting the minimal individuals with respect to the number of changes and total magnitude of the changes in the Pareto front, i.e. the individual that is the closest to the boundary. Three aspects of each modification are visually measured: 1) if the modified area corresponds with the semantic area of the original pathology, 2) if the modified area is related to the area of the objective label, and, 3) if the modified area is related with a different region of the patient.

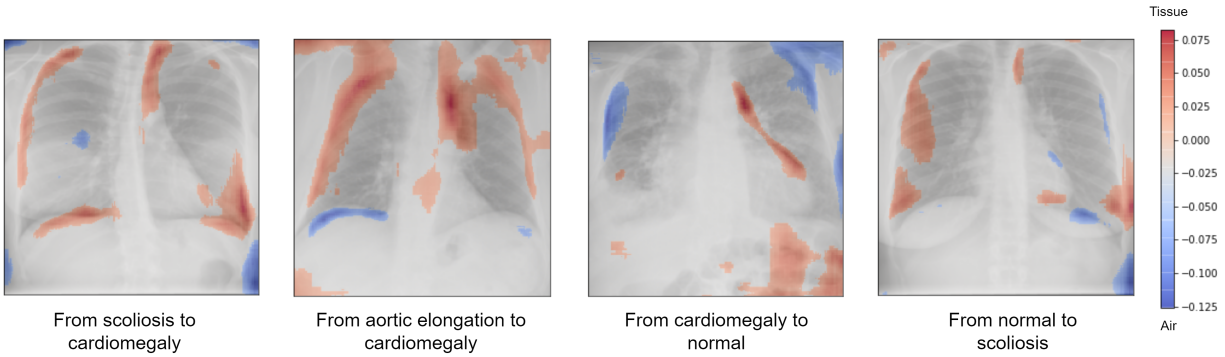
Figures 6.3a and 6.3b shows examples of the saliency map, where the blue areas represent areas where the mutation adds pixel intensity and the red areas are regions where the change reduces

the pixel intensity, i.e. blue areas are areas where the change is whiter, and in red areas the pixel is darker after the mutation. From the total saliency map generated with the modified individuals, only the upper and lower 30% of the modified areas are displayed, which means that the images highlight the regions with the most significant modifications. This percentage of modified areas that are showed in the explanations would be the only hyperparameter related to the explanations that are used. However, it is also possible to evaluate the complete saliency map of the explanations. In this case, and in order to quantitatively evaluate the explanations regions, this threshold is defined to avoid subjectivity.

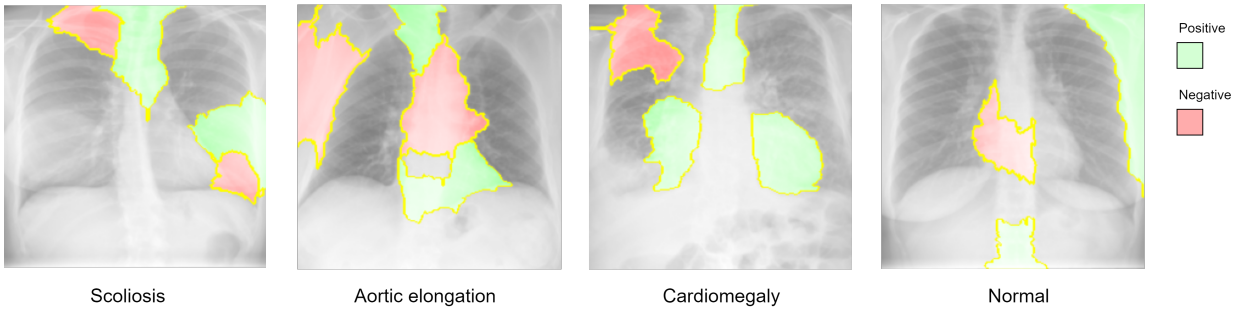
In addition, our explanations are compared with state-of-the art XAI systems. In particular, our results are compared with LIME [181] DeepCover [39] and SAG [199] in Figures 6.3c, 6.3d and 6.3e respectively.



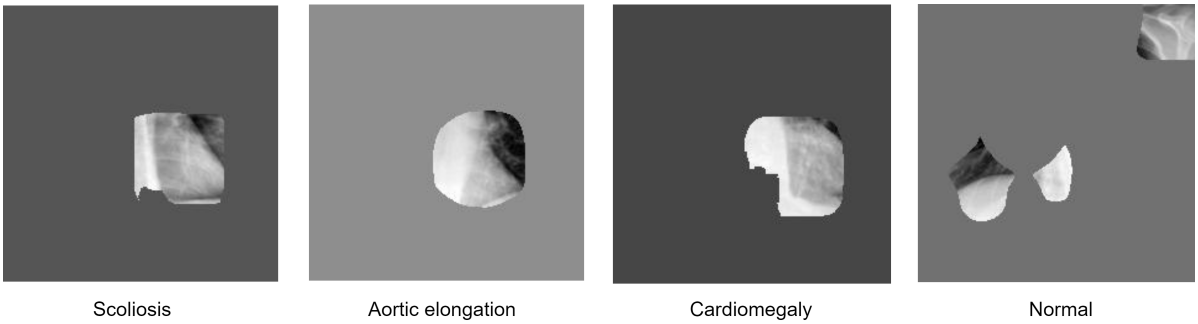
(a) Explanations for the minimum individuals respect number of changes of the proposed approach.



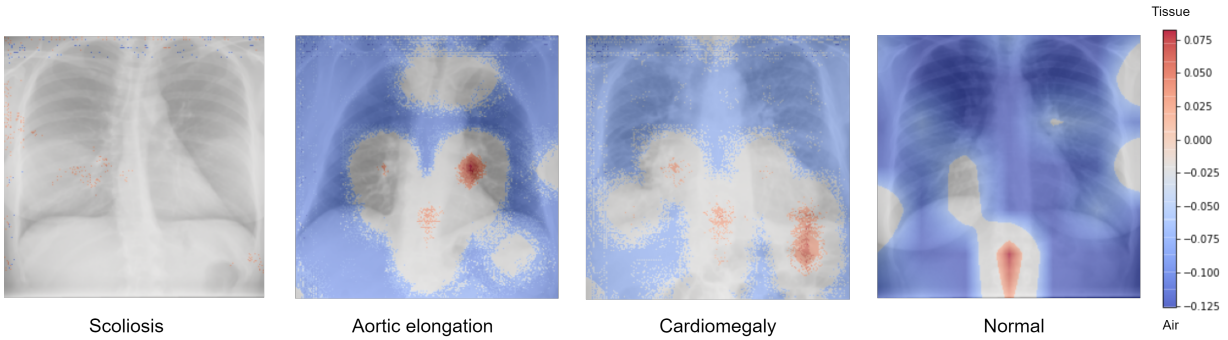
(b) Explanations for the minimum individuals respect total magnitude of changes of the proposed approach.



(c) Explanations for LIME [181].



(d) Explanations for DeepCover [39].



(e) Explanations for SAG [199].

Figure 6.3: Explanation comparison of different methods

Table 6.7 contains the modification rate on the base and objective areas of each experiment, along with the modification rate of different areas, using boundary individuals with respect to the minimum number of changes. Table 6.8 contains equivalent results with respect to the total magnitude of changes. The relationship between the areas and the pathologies is the following: the heart for *cardiomegaly*, the aorta for *aortic elongation*, the spine for *scoliosis*. Normal cases were not used in this analysis, since they do not represent a particular region.

Origin	Affects label base				Affects label objective				Affects other regions			
	C	A	S	N	C	A	S	N	C	A	S	N
C	-	88%	100%	88%	-	0%	88%	-	-	63%	0%	50%
A	40%	-	60%	60%	100%	-	70%	-	50%	-	100%	100%
S	100%	90%	-	100%	100%	70%	-	-	40%	90%	-	90%
N	-	-	-	-	100%	90%	100%	-	100%	90%	100%	-

Table 6.7: Percentage of area modification for the different experiments, using the minimum number of changes individuals.

Origin	Affects label base				Affects label objective				Affects other regions			
	C	A	S	N	C	A	S	N	C	A	S	N
C	-	75%	100%	100%	-	75%	50%	-	-	38%	50%	75%
A	30%	-	60%	40%	90%	-	60%	-	40%	-	70%	60%
S	70%	80%	-	100%	80%	80%	-	-	40%	90%	-	60%
N	-	-	-	-	90%	70%	90%	-	100%	100%	90%	-

Table 6.8: Percentage of area modification for the different experiments, using the minimum magnitude individuals.

It can be observed that, depending on the combination of pathologies, our method obtains a mean value of 80.56% for the minimum individuals with respect to the number of changes and 72.78% for the minimum individuals with respect to the total magnitude of the changes, especially in cases where the explanation affects the same area as the label. This can be seen as the percentage of true positive explanations, because the affected areas correspond with the pathology. Those explanations that affect other regions that are not related to the pathology achieve a mean value of 72.71% for the minimum number of changes and 67.71% for the minimum magnitude of changes. This metric can be interpreted as false positives, because it affects areas not related to the pathology. Finally, analyzing whether the explanations affect the objective label of the algorithm, a value 79.77% is obtained for the minimum number of changes and 76.11% for the minimum magnitude. These correspond with negatives generated by the genetic algorithm, that can be seen as areas that must be changed to transform an image from one pathology into another.

It can be observed that the model explanations show results that semantically capture features of the data and produce a minimum explanation of the patient's pathology. To further evaluate the performance of our explanations the regions affected by the explanations are compared against state-of-the-art approaches. In particular, it is decided to compare our results with the XAI systems shown in Figure 6.3. Table 6.9 compares the results of modified areas of the final individuals with the rest of the approaches. The same procedure for Tables 6.7 and 6.8 is followed. The results are compared using the minimum individuals of MOC-VAE with respect to the total number of changes (min num) and total magnitude of changes (min mag).

Region affected	min mag	min num	LIME	DeepCover	SAG
Affects label base	72.78%	**80.56%**	73.33%	24.17%	39.17%
Affects other region	67.71%	72.71%	66.19%	84.17%	**53.33%**
Affects label objective	76.11%	**79.72%**	-	-	-

Table 6.9: Percentage of modified areas of the different methods.

The regions affected by each method shows an improvement in the precision of the XAI explanations when using our algorithm. Measuring the percentage of explanations found that are in the same area as the patient's pathology, it is observed a correlation of 80.56% using minimal explanations with respect to the number of changes and 72.78% when using minimal explanations with respect to the total magnitude of changes. The rest of the methods compared range between 24.17% and 73.33%. Moreover, by analyzing whether the explanations also affected other areas of the patient, it is found that our methods affect other ares in 72.71% and 67.71% of the times respectively.

To be able to analyze the balance between true positives (the affected areas are the same areas as the pathology) and false positives (the affected areas do not correspond to the pathology of the patient), it is also measured the difference between both metrics. The difference of our method is of 7.85 and 4.57 points for the minimum number of changes and minimum magnitude individuals, while the other XAI methods are 7.14 (LIME), -14.16 (DeepCover) and -60.00 (SAG).

6.4 Discussion

By applying the MOC-VAE architecture to medical diagnosis, it is possible to generate fine-grained explanations. The completeness and continuity attributes of MOC-VAE's latent space makes it possible to modify the image descriptors to find a minimum explanation that is both clear and precise by using simple linear transformations. Moreover, the XAI process does not need any hyperparameter, which allows explanations to be obtained without the need to make assumptions about the data.

With respect to the optimization process, the results show that the genetic algorithm is able to produce changes in the network's confidence with minimal changes. On the one hand, the transition matrix shows a drastic change in the confidence of the objective label, with mean change of 0.6382. This suggests that the algorithm is capable of drastically changing the probability distribution of the classification output. Looking at successful individuals, the results of Tables 6.3 and 6.4 show that in all cases it is possible to achieve misclassification, therefore finding boundaries among classes.

The results also suggest that, depending on the combination of pathologies, misclassification is not always equally easy to achieve. In particular, when the *normal* label is present in the combination, the first individual is found earlier during the search process, probably due to not representing a particular pathology, therefore less changes are necessary to change its label. Moreover, this phenomenon is recurrent with respect to the minimum explanation

individuals (i.e. individuals close to the boundaries) found when *Normal* is the base label. These individuals generally have fewer changes than when a pathology is present.

Table 6.9 results suggest that our contribution outperforms the rest of the approaches at focusing on the area of the base label, i.e., our method is better at analyzing the patient's pathologies. It is possible to observe that the best results respect affecting other regions are obtained with SAG, but these results are achieved at the cost of reducing the accuracy of the explanations. Moreover, MOC-VAE achieves a better balance in both metrics. This suggests that the results of MOC-VAE are more precise and semantically correct. It is also shown that the best results for MOC-VAE are achieved using the minimum number of changes in the individuals, achieving a 7.78% more of precision respect using the minimum magnitude of changes individuals.

In addition, the system does not need any hyperparameter to configure the explanations system, making it easier to apply to tasks where no assumptions about the data can be made. The image descriptors store information about the whole image; thus, any change produced in this space can affect the entire input space. Hyperparameter configuration is usually a difficult task, in our case the only configuration that needs to be done relates to the model training and the adversarial search, but both procedures do not make assumptions about the explanations directly, but rather they find the most optimal set of solutions. In other words, by manipulating a structured latent space it is possible to find explanations that are semantically coherent and precise, without pre-defining particular parameters.

6.4.1 Threats of Validity

The proposed methodology is presented as a viable XAI strategy, specially designed for medical diagnostic. Its features make it the perfect tool for finding minimal explanations that are interesting for precise diagnoses. In addition to that, the architecture also has different limitations that need to be evaluated.

Neural network models have limitations depending on their training abilities. The model's performance is connected to the explanations found by the rest of the algorithm, therefore, a poor classifier would negatively impact the quality of the explanations provided by the system.

In this case, the explanation system would benefit from a more accurate classifier. As is shown in Table 6.1, the classification performance outperforms state-of-the-art architectures, thus ensuring the best possible results in this task. It is also considered that aspects such as the dataset labeling could affect precision.

The general structure of the proposed methodology makes it possible to apply a different optimization process when finding explanations. An aspect that could limit the explanation quality is the adversarial search, although a standard methodology that has successfully navigate classification spaces is employed in different contexts [142].

With respect to the use of explanations, results for the minimum individuals are shown (i.e. closest to the boundaries) found for each case, discarding the rest of the successful individuals.

Other individuals might be able to provide more information about the classes but this will require further research.

One limitation of the proposed evaluation is the measurement of the pathological areas affected. It is defined a proper explanation when the highlighted area affects the organ of the pathology that the patient suffers. This evaluation could be limited to only evaluate the whole organ, without a precise professional medical diagnosis. With this evaluation, the focus is put on comparing the different explanations of the compared methods and it is not sought to provide medical insights of the explanations themselves. Thus, it is considered this to be a correct approach to provide quantitative results for the different methods. However, further research could study the behavior of the explanations of the MOC-VAE.

6.5 Conclusions

This chapter presents a novel DL architecture for XAI in the context of healthcare and medical diagnosis. The proposed method, named MOC-VAE, extracts anatomical and pathological characteristics of each image and creates a descriptor that is then used for finding the explanations. Using the regularization properties of VAEs the model is able to create a structured latent space from where finding proper solutions.

These image descriptors are low-dimensional numerical vectors, in our case with 500 dimensions. The low dimensionality is exploited to find perturbations in the vectors that makes the network classify the case as a particular pathology. This procedure used to find the minimum individuals that are capable of achieve this misclassification, and the individuals are reconstructed with this perturbations to be capable of visually identify the regions of the images that are modified. Because of the VAE properties and the linear latent space manipulations, these individuals correspond with the boundaries among classes. Therefore, by combining both approximations, it is possible to find which areas of the images are making the network decide that a case belongs to a particular class.

Results of the search for minimum individuals that are capable of misclassify the cases show that it is always possible to achieve this targeted misclassification, nevertheless it is not always equally easy. The explanations that the individuals provide are semantically related to the pathologies of the images, being able to show which areas are responsible of the specific pathologies.

In comparison with other state-of-the art approximations, an improvement in the precision and accuracy of the explanations is shown, being capable of produce clearer explanations to the images. Additionally, our methodology does not require almost any hyperparameter in its configuration, making it easier to be applied it when there are no assumptions of the data.

6.6 Future Work

The proposed framework presents a great flexibility, making it possible to adapt the system implementation to new areas, where experiments can study the results and behaviors of the explanations for different cases.

The search algorithm used to find the explanations could be further studied, comparing it with simpler approximation such as gradient descent-based algorithms in order to identify these explanations. More efficient algorithms could benefit the framework with better and faster explanations.

Furthermore, the use of the set of explanations found could be further developed using the information stored in the whole set, rather than the minimum individual. Further research could investigate whether there is shared information in the population of solutions that could benefit the explanations.

6.7 Answers to the Research Questions

- **Research Question (RQ)4**: Do the explanations of MOC-VAE architectures produce saliency maps of more quality than previous XAI techniques based on DL?

The proposed XAI algorithm using the MOC-VAE is compared in terms of precision against LIME, DeepCover and SAG algorithms. Each model is tested against 40 cases, 10 of each of the pathologies contained in the dataset. Each explanation is manually analized, measuring if the areas of the explanation correspond with the pathological areas of the class. Results show an improvement in the explanation precision of 41.39 percentage points for SAG 56.39 percentage points for DeepCover and 7.23 percentage points for LIME.

- **RQ5**: Does the reduction of number of hyperparameters of MOC-VAE affects to their performance in XAI tasks?

The proposed XAI scheme is able of producing more precise explanations than previous works using, as the only hyperparameter of the explanations, the percentage boundary of modified areas of the images.

Final Discussion and Conclusions

Chapter 7

Conclusions and Future Work

7.1 Discussion of the Research

The research proposed in Chapter 3 was followed in Chapter 4, Figure 4 and Chapter 6. The research of the thesis focuses on covering the limitations discovered in the state-of-the-art and propose a specific methodology to solve the proposed objectives.

In this sense, the proposed Research Question (RQ) have been answered throughout the document. Nevertheless, below are discussed the answers to these questions:

- **RQ1**: Are the new Multi-Output architectures effective for Content-Based Image Retrieval (CBIR) of medical images?

  This RQ is mainly demonstrated in Chapter 4 and Chapter 5, where the proposed architectures are evaluated with respect state-of-the-art and similar architectures. In Chapter 4 the Multi-Output Classification Autoencoder (MOC-AE) is tested comparing it with the research of Kobayashi et al. [118]. The results show that the proposed architecture achieves better performance recommending cases more similar respect their tumoural characteristics, in addition the outputs of the MOC-AE better balance tumoural and anatomical features of the patients.

  These results suggest that in comparison with the study of Kobayashi et al. the proposed model behaves better, extracting more condensed information about the cases. Nevertheless, it has only been demonstrated that the MOC-AE results improve previous works, but further research is necessary to discover the real behavior of the model in other real world medical problems. In addition, it should be noted that the anatomical similarities of the query and recommended cases are lower in the MOC-AE results. This must be taken into account for the implementation of the model in real world scenarios where the γ can be modified to mitigate this discrepancy.

  In Chapter 5 the Multi-Output Classification Variational Autoencoder (MOC-VAE) is proposed and analyzed along with the MOC-AE. This chapter studies these architectures performance in absolute values in other real world problem. The results are evaluated

by comparing them with similar Deep Learning (DL) architectures in CBIR, and it is proved that the novel architectures are better in the recommendation of medical image in all tests. In this sense, the factorization capabilities of the Multi Output models are tested with respect to Single Output models, showing an improvement in semantical and visual recommendations. In addition, the Multi Output models better balance both characteristics.

One of the limitations of the research in this sense is that the architectures are tested using just two datasets for different problems. Besides representing real world scenarios, not cherry picked for any purpose, the architectures should be studied for different datasets, e.g. using different sources such as camera images or ecographies.

- **RQ2**: Is it less efficient to use classification labels instead of more complex labels?

  Chapter 4 reflects the results of the use of binary labels in the MOC-AE architectures respect the previous work of Kobayashi et al. [118] that used segmentation information. The results show that the MOC-AE in spite of using less complex labels, achieves better results. The segmentation information used in the prior work is much more difficult to obtain, specially in sensible and high-specialization environments, such as the medical field. Therefore, reducing the cost of the necessary information to train the models will increase the applicability of the models.

  The implementations carried out during the research show that by using classification labels it is possible to achieve great results for image recommendation. However, these labels are not always present in the available datasets. Thus, besides lowering the cost and improving the number of cases where the architectures can be used, it is still necessary to use these labels to train the models.

- **RQ3**: Is the new visual similarity metric able of capturing similarity between cases based on their visual appearance?

  Chapter 5 proposes the use of the Sliced Wasserstein metric to measure visual similarity between the query and the recommended cases. The results showed in the chapter certify that the new metric is insightful and precise, correlating visual similarity of the images. This new metric covers an important factor that was not analyzed previously in the state-of-the-art, specifically the ease to compare cases, which is specially important for tasks as medical comparative diagnostic.

  In this thesis, the inclusion of the Sliced Wasserstein metric as a visual similarity measurement is proposed, but to become a standard in CBIR testing, more researches must use it and certify that it is useful to use this metric and its values are representative.

- **RQ4**: Do the explanations of MOC-VAE architectures produce saliency maps of more quality than previous eXplainable Artificial Intelligence (XAI) techniques based on DL?

  The XAI possibilities of the MOC-VAE are studied in Chapter 6. The research of the chapter proposes a combination of evolutionary algorithms along with the MOC-VAE architecture to find explanations in a low-dimensional latent space from which saliency maps are constructed. This methodology is shown to produce higher quality

explanations, with more precise saliency maps.

One of the main drawbacks of the comparison of Chapter 6 is that the explainable models used to compare the proposed algorithm are model-agnostic. The applicability of the proposal in real world scenarios could be affected by this characteristic, but the improvement in the quality of the explanations may justify its application.

- **RQ5**: Does the reduction of number of hyperparameters of MOC-VAE affects to their performance in XAI tasks?

  The XAI algorithm presented in Chapter 6 uses a low number of hyperparameters in its configuration, in particular one hyperparameter that can be discarded if necessary. The hyperparameter configuration is a limitation of prior researches of the state-of-the-art that need a proper configuration that changes the results of the explanations. As a result, reducing the hyperparameter configuration of the proposed workflow improves the applicability of the system.

7.2 Conclusions

This thesis presents a novel DL architecture focused in medical CBIR. The so-called MOC-AE is a neural network architecture that is based on a multiple objective training. The model is trained to optimize reconstruction of the images that it receives, at the same time it optimizes a classification output. In order to provide recommendations, the model uses a one-dimensional latent space that contains the image descriptors. The image recommendation is done by comparing these descriptors and retrieving the most similar ones respect to a given query.

The embedded space is benefited from the multi-objective training of the network by storing enriched information of the image descriptors. The combination of outputs allows to focus on certain specific areas of the images, which are related to the areas relevant for the diagnostic process. The MOC-AE follows the same principles of previous CBIR systems, that used a latent space descriptor to perform the image comparison, but the dual objective benefits the recommendation.

The results of the MOC-AE are studied comparing them to one of the most successful image recommendations architecture proposed by Kobayashi et al. in [118]. Results showcase the properties of the MOC-AE, that is able of better preserving anatomical and pathological features of the patients in the latent space. In addition, the MOC-AE better captures pathological areas of the patients. But the most relevant aspect that differs MOC-AE from previous works is that, because its simplicity, it can be applied to a wide variety of datasets, specially important where there are not available costly labels.

The good generalization of the MOC-AE makes possible to obtain rich recommendations from the latent space, capturing latent characteristics of the images that are not labeled. Because of the combination of classification and reconstruction tasks, the model is able of detecting finer relationships of the data, that may not be identified before.

The MOC-AE is evolved presenting the MOC-VAE architecture, that maintains the same training scheme but adding a regularization term to the training. With the introduction of the regularization, the latent space is ensured to be complete and continuous, this way improving the latent relationships of the data. These properties structure the latent space help the recommendation process.

These results are further studied by presenting and testing the baseline MOC-AE model against the MOC-VAE in CBIR tasks. In addition, the proposed models are compared with Autoencoder (AE) and classifier architectures. Results showcase the improved performance of the MOC-AE and MOC-VAE with respect their DL traditional counterparts. These results suggest that the theoretical background proposed with the MOC-AE and MOC-VAE translate improvement in real world problems. Furthermore, the additions proposed by the MOC-VAE affects its results, showing an improved performance respect the MOC-AE in almost all the tests done.

In addition, future uses of the MOC-VAE architecture are studied by applying it to finding explanations of diagnostics, following the same procedure as XAI methods. The embedded space of the MOC-VAE is leveraged to find linear modifications in the image descriptors that can be translated to non-linear modifications in the image space. This way, a XAI algorithm is presented, using Genetic Algorithms, that is able of finding semantic-meaningful explanations of diagnostics. The results are compared with previous state-of-the-art algorithms and the proposed algorithm is able of producing more precise and clearer explanations.

Overall, all the advances presented in the thesis showcase a deep study of medical diagnosis through the proposition of the MOC-AE and MOC-VAE. The results of the research certifies that these networks have strong theoretical foundations but also provide good results under real world problems. Besides being space for further research, it is considered that the thesis goes through the most important aspects of these new models, with specific testing and evaluation of the results. 2The research shows the potential of the MOC-AE and MOC-VAE, presenting them as one of the most successful architectures in the area of medical diagnosis.

7.3 Future work

This work presents a DL architecture with a simple structure that can be leveraged for different objectives. The adaptability of DL along with the architecture proposed open possibilities to a wide range applications and evolutions, such as the ones presented with the MOC-VAE or the application to XAI tasks. Summarizing, the main open lines of research unexplored are the following:

7.3.1 Implementations of MOC-VAE in medical centers for CBIR

The good results of the MOC-VAE, outperforming similar approaches and state-of-the-art CBIR systems, position the architecture as one of the best solutions to perform CBIR in medical diagnosis. It is considered crucial that science is put in production to tackle real world

problems, not in laboratory experiments, but rather in real world scenarios. Science's last purpose should be to provide real solutions to difficult problems, in this sense the MOC-VAE must be implemented in medical centers to test its performance.

All the efforts made during the thesis verify the good results of MOC-VAE in CBIR. Thus, its real implementation would alleviate difficult diagnosis for specific pathologies. The deploy of MOC-VAE in medical centers should always be focused in comparative diagnostic that physicians should use in difficult cases of doubts. In this sense, MOC-VAE would act as a searching tool along the available database to retrieve the most similar cases.

In this sense, it is mandatory to find a diagnostic that the model should tackle. The diagnosis of chest X-ray images is probably not enough interesting to be deployed in real world scenarios. After a discussion with several physicians one good option is found in lung diseases related. Conversations with Jesus Troya, coordinator of innovation of the Hospital Universitario Infanta Leonor, have shown the interest in applying the MOC-VAE in pulmonary diseases diagnostic. Certain lung diseases, e.g. pulmonary fibrosis, are identified with certain patterns in the lung tissue of X-ray images. The MOC-VAE can be used in these cases, leveraging the good texture feature recognition of Convolutional Neural Network (CNN) [23], [228], to detect patterns difficult to observe by humans. In this context MOC-VAE can be applied as a Clinical Decision Support System (CDSS) that physicians can use to compare difficult diagnosis.

The main tasks of the proposed implementation include:

- Data acquisition: Acquiring high-quality data from medical centers is a difficult task, but a good data curation ensures a correct model's performance. There is a strong correlation between the quality of the data used for training the models and their final performance, thus, proper selection, filtration and preprocessing of the data is crucial. E.g. aspects such as segmentation of the lung area must be studied to reduce the noise of the diagnostics.
- Model implementation: The architecture of the MOC-VAE enables image recommendation in pulmonary diseases. Provided that classification labels and their corresponding images can be found in the dataset, the model can be trained and implemented without any changes in its architecture. Then, to deploy the product in the final environment, aspects such as model optimization could improve the model's time to response. It is also mandatory to deploy the model in the medical center's facilities, to ensure security and privacy of the data.
- Security assessment: Because the data being treated would be sensitive personal data, its anonymization and integrity must be guaranteed. In this sense, the MOC-VAE does not need to use personal data in its training and forecasting. Thus, the whole framework can work using anonymous *ids* to identify the different patients. In addition, ethical committees must review the implementation of the system in medical facilities to ensure its correct deployment.
- Performance evaluation: Implementing the MOC-VAE in medical centers includes an extensive evaluation of its impact on the diagnostic. In the development of the research

of the thesis the evaluation is assessed with metrics that evaluate the model's behavior. But the final implementation of the system includes interaction with human professionals, and to this purpose further evaluation must be done. In particular, evaluation must be carried out with respect the interaction of professionals with the MOC-VAE, measuring which cases are decided to be compared and inside those, which ones are benefited from the recommendation provided by the MOC-VAE.

7.3.2 Further exploration of MOC-VAE in XAI

The XAI framework presented in Chapter 6 showcases a new method of XAI in DL, with improved results respect previous systems. But it is considered crucial to further study this system, evaluating it in different datasets focused on different problems. The simplicity of the system makes possible to apply it to areas apart from medical diagnosis.

The results presented with the architecture suggest that the model can be considered as a good alternative to XAI. Accordingly, the MOC-VAE presents a good framework from where explore different areas of interest. XAI is an area of increasing interest [147], therefore any improvement in the current systems is of special interest to the scientific community.

In this sense, to ensure that the MOC-VAE arises as an XAI model a library or repository must be created. Besides all the code of the thesis being publicly available, further efforts must be done to ease the implementation of the system in other areas. In this sense, refactorization of the code should be done to ease the interpretation of the model, but also more flexibility must be included in the code to be able of implement the system with different architectures.

7.3.3 Architectural changes in MOC-AE and MOC-VAE networks

DL is in constant evolution, during the last decade many improvements have been proposed in the area of Computer Vision. Layers such as the Attention layers [223] have drastically impacted the area, with new evolutions for Computer Vision with the Visual Transformers [51]. The MOC-AE and MOC-VAE architecture can leverage all these advances to achieve better results in its feature extraction phase, being benefited from the strengths of these new techniques. Also, new advances will come in the next years because Computer Vision is an area of constant evolution [35], and these fictitious advances could benefit the MOC-AE and MOC-VAE architectures.

In addition, the MOC-AE and MOC-VAE architectures can be further studied. The evolution of the MOC-VAE, implementing regularization terms in the training, have demonstrated that is possible to change the behavior of the model and achieve better results in the process. Thus, new changes following the same idea can be studied.

Science is in constant evolution and the DL models should not be static, rather, these techniques should be in continuous interaction, using new implementations and studying new methods of solving previous problems. Under no circumstances the current research must be considered as a static final product, this consideration is not correct for almost any model in

DL. Rather, models should always be open to new evolutions and implementations.

7.3.4 Study MOC-AE and MOC-VAE for classification tasks

The results presented in Table 6.1 and Table 5.4 suggest that the MOC-AE and MOC-VAE are able of capturing classification features in the latent space. These results suggest that the model benefits from the reconstruction information even when the model is tested in classification tasks. Specially, in Table 6.1 it is shown that the classification statistics of the MOC-VAE outperform traditional classifications models in classification tasks.

These results are very interesting because they can be interpreted as an improvement in DL classification models. Which implies that the MOC-VAE architecture can be leveraged in classification tasks improving previous models.

Besides that, these considerations must be taken with caution, because these are just preliminary results and further study must be carried out. It is considered that these results can be due to the fact that the dataset is not correctly labeled, thus it is important to study the performance of MOC-VAE against traditional classificators in these circumstances. Further research should study the performance of the model using the typical Computer Vision datasets, e.g. MNIST [127], ImageNet [187] or CIFAR-10/100 [123]. This study should be focus in measuring how accurate are MOC-AE and MOC-VAE respect others classifiers when the percentage of images correctly labeled changes. It is thought that the good results of the MOC-AE and MOC-VAE are due to the fact that including geometric information in the latent space improves generalization when images are visually similar. These constraints should be further study in medical image classification and general image classification to better understand the behavior of MOC-AE and MOC-VAE in classification tasks.

References

[1] M. Abadi, A. Agarwal, P. Barham, *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems", *arXiv preprint arXiv:1603.04467*, 2016.

[2] A. Agrawal, J. Gans, and A. Goldfarb, *What to expect from artificial intelligence*, 2017.

[3] S. Agrawal, A. Chowdhary, S. Agarwala, V. Mayya, and S. Kamath S, "Content-based medical image retrieval system for lung diseases using deep cnns", *International Journal of Information Technology*, vol. 14, no. 7, pp. 3619–3627, 2022.

[4] A. M. U. Ahamed, C. Eswaran, and R. Kannan, "Cbir system based on prediction errors", *J. Inf. Sci. Eng*, vol. 33, no. 2, pp. 347–365, 2017.

[5] G. N. Ahmad, H. Fatima, S. Ullah, A. S. Saidi, *et al.*, "Efficient medical diagnosis of human heart diseases using machine learning techniques with and without gridsearchcv", *IEEE Access*, vol. 10, pp. 80 151–80 173, 2022.

[6] G. N. Ahmad, S. Ullah, A. Algethami, H. Fatima, and S. M. H. Akhter, "Comparative study of optimum medical diagnosis of human heart disease using machine learning technique with and without sequential feature selection", *ieee access*, vol. 10, pp. 23 808–23 828, 2022.

[7] A. Albahri, A. M. Duhaim, M. A. Fadhel, *et al.*, "A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion", *Information Fusion*, 2023.

[8] B. Aldughayfiq, F. Ashfaq, N. Jhanjhi, and M. Humayun, "Explainable ai for retinoblastoma diagnosis: Interpreting deep learning models with lime and shap", *Diagnostics*, vol. 13, no. 11, p. 1932, 2023.

[9] S. Ali, F. Akhlaq, A. S. Imran, Z. Kastrati, S. M. Daudpota, and M. Moosa, "The enlightening role of explainable artificial intelligence in medical & healthcare domains: A systematic literature review", *Computers in Biology and Medicine*, p. 107 555, 2023.

[10] J. Allgaier, L. Mulansky, R. L. Draelos, and R. Pryss, "How does the model make predictions? a systematic literature review on the explainability power of machine learning in healthcare", *Artificial Intelligence in Medicine*, vol. 143, p. 102 616, 2023.

[11] M. K. Alsmadi, "Content-based image retrieval using color, shape and texture descriptors and features", *Arabian Journal for Science and Engineering*, vol. 45, no. 4, pp. 3317–3330, 2020.

[12] A. Alzu'bi, A. Amira, and N. Ramzan, "Semantic content-based image retrieval: A comprehensive study", *Journal of Visual Communication and Image Representation*, vol. 32, pp. 20–54, 2015.

[13] A. B. Amjoud and M. Amrouch, "Object detection using deep learning, cnns and vision transformers: A review", *IEEE Access*, vol. 11, pp. 35 479–35 516, 2023.
[14] E. Amparore, A. Perotti, and P. Bajardi, "To trust or not to trust an explanation: Using leaf to evaluate local linear xai methods", *PeerJ Computer Science*, vol. 7, e479, 2021.
[15] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, "Explainable artificial intelligence: An analytical review", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 5, e1424, 2021.
[16] M. Arjovsky, S. Chintala, and L. Bottou, *Wasserstein gan*, 2017. arXiv: 1701.07875 [stat.ML].
[17] N. Arora, A. Kakde, and S. C. Sharma, "An optimal approach for content-based image retrieval using deep learning on covid-19 and pneumonia x-ray images", *International Journal of System Assurance Engineering and Management*, vol. 14, no. Suppl 1, pp. 246–255, 2023.
[18] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai", *Information fusion*, vol. 58, pp. 82–115, 2020.
[19] T. Bäck and H.-P. Schwefel, "An overview of evolutionary algorithms for parameter optimization", *Evolutionary computation*, vol. 1, no. 1, pp. 1–23, 1993.
[20] S. Bakas, H. Akbari, A. Sotiras, *et al.*, "Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features", *Scientific data*, vol. 4, no. 1, pp. 1–13, 2017.
[21] S. Bakas, M. Reyes, A. Jakab, *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge", *arXiv preprint arXiv:1811.02629*, 2018.
[22] J. M. Banda and R. A. Angryk, "Regional content-based image retrieval for solar images: Traditional versus modern methods", *Astronomy and computing*, vol. 13, pp. 108–116, 2015.
[23] S. Barburiceanu, S. Meza, B. Orza, R. Malutan, and R. Terebes, "Convolutional neural networks for texture feature extraction. applications to leaf disease classification in precision agriculture", *IEEE Access*, vol. 9, pp. 160 085–160 103, 2021.
[24] N. Barzilay, T. B. Shalev, and R. Giryes, "Miss gan: A multi-illustrator style generative adversarial network for image to illustration translation", *Pattern Recognition Letters*, 2021.
[25] C. Belloni, A. Balleri, N. Aouf, J.-M. Le Caillec, and T. Merlet, "Explainability of deep sar atr through feature analysis", *IEEE transactions on aerospace and electronic systems*, vol. 57, no. 1, pp. 659–673, 2020.
[26] J. M. Bhalodiya, S. N. Lim Choi Keung, and T. N. Arvanitis, "Magnetic resonance image-based brain tumour segmentation methods: A systematic review", *Digital Health*, vol. 8, p. 20 552 076 221 074 122, 2022.
[27] P. Bhattarai, D. S. Thakuri, Y. Nie, and G. B. Chand, "Explainable ai-based deep-shap for mapping the multivariate relationships between regional neuroimaging biomarkers and cognition", *European Journal of Radiology*, vol. 174, p. 111 403, 2024.

[28] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister, "Sliced and radon wasserstein barycenters of measures", *Journal of Mathematical Imaging and Vision*, vol. 51, pp. 22–45, 2015.

[29] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis", *arXiv preprint arXiv:1809.11096*, 2018.

[30] A. Bustos, A. Pertusa, J.-M. Salinas, and M. De La Iglesia-Vaya, "Padchest: A large chest x-ray image dataset with multi-label annotated reports", *Medical image analysis*, vol. 66, p. 101 797, 2020.

[31] L. Cai and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era", *Data science journal*, vol. 14, pp. 2–2, 2015.

[32] P. V. de Campos Souza, "Fuzzy neural networks and neuro-fuzzy networks: A review the main techniques and applications used in the literature", *Applied soft computing*, vol. 92, p. 106 275, 2020.

[33] H. Cao and S. Mi, "Weighted srgan and reconstruction loss analysis for accurate image super resolution", in *Journal of Physics: Conference Series*, IOP Publishing, vol. 1903, 2021, p. 012 050.

[34] E. D. Carvalho, O. Antonio Filho, R. R. Silva, *et al.*, "Breast cancer diagnosis from histopathological images using textural features and cbir", *Artificial intelligence in medicine*, vol. 105, p. 101 845, 2020.

[35] J. Chai, H. Zeng, A. Li, and E. W. Ngai, "Deep learning in computer vision: A critical review of emerging techniques and application scenarios", *Machine Learning with Applications*, vol. 6, p. 100 134, 2021.

[36] A. Chapman, E. Simperl, L. Koesten, *et al.*, "Dataset search: A survey", *The VLDB Journal*, vol. 29, no. 1, pp. 251–272, 2020.

[37] Y. Chen, C. Zheng, F. Hu, *et al.*, "Efficient two-step liver and tumour segmentation on abdominal ct via deep learning and a conditional random field", *Computers in Biology and Medicine*, vol. 150, p. 106 076, 2022.

[38] H. Chockler, D. A. Kelly, and D. Kroening, "Multiple different explanations for image classifiers", *arXiv preprint arXiv:2309.14309*, 2023.

[39] H. Chockler, D. Kroening, and Y. Sun, "Explanations for occluded images", in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1234–1243.

[40] F. Chollet *et al.*, *Keras*, https://keras.io, 2015.

[41] F. Chollet, "Xception: Deep learning with depthwise separable convolutions", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[42] C. Cortes and V. Vapnik, "Support-vector networks", *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[43] T. Cover and P. Hart, "Nearest neighbor pattern classification", *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967. DOI: 10.1109/TIT.1967.1053964.

[44] W. J. Dally, S. W. Keckler, and D. B. Kirk, "Evolution of the graphics processing unit (gpu)", *IEEE Micro*, vol. 41, no. 6, pp. 42–51, 2021.

[45] J. M. Darias, B. Díaz-Agudo, and J. A. Recio-Garcia, "A systematic review on model-agnostic xai libraries.", in *ICCBR Workshops*, 2021, pp. 28–39.

[46] P. Das and A. Neelima, "An overview of approaches for content-based medical image retrieval", *International journal of multimedia information retrieval*, vol. 6, no. 4, pp. 271–280, 2017.

[47] O. Diaz, K. Kushibar, R. Osuala, *et al.*, "Data preparation for artificial intelligence in medical imaging: A comprehensive guide to open-access platforms and tools", *Physica medica*, vol. 83, pp. 25–37, 2021.

[48] L. R. Dice, "Measures of the amount of ecologic association between species", *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[49] A. Dikshit and B. Pradhan, "Interpretable and explainable ai (xai) model for spatial drought prediction", *Science of the Total Environment*, vol. 801, p. 149 797, 2021.

[50] J. Dong, S. Chen, M. Miralinaghi, T. Chen, P. Li, and S. Labi, "Why did the ai make that decision? towards an explainable artificial intelligence (xai) for autonomous driving systems", *Transportation research part C: emerging technologies*, vol. 156, p. 104 358, 2023.

[51] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale", *arXiv preprint arXiv:2010.11929*, 2020.

[52] Y. Du, A. R. Rafferty, F. M. McAuliffe, L. Wei, and C. Mooney, "An explainable machine learning-based clinical decision support system for prediction of gestational diabetes mellitus", *Scientific Reports*, vol. 12, no. 1, p. 1170, 2022.

[53] A. Esteva, K. Chou, S. Yeung, *et al.*, "Deep learning-enabled medical computer vision", *NPJ digital medicine*, vol. 4, no. 1, pp. 1–9, 2021.

[54] M. Fatima, M. Pasha, *et al.*, "Survey of machine learning algorithms for disease diagnostic", *Journal of Intelligent Learning Systems and Applications*, vol. 9, no. 01, p. 1, 2017.

[55] M. Flickner, H. Sawhney, W. Niblack, *et al.*, "Query by image and video content: The qbic system", *computer*, vol. 28, no. 9, pp. 23–32, 1995.

[56] D. Foster, *Generative deep learning*. " O'Reilly Media, Inc.", 2022.

[57] K. Främling, "Modélisation et apprentissage des préférences par réseaux de neurones pour l'aide à la décision multicritère", Ph.D. dissertation, INSA de Lyon, 1996.

[58] K. Fukushima, "Neocognitron: A hierarchical neural network capable of visual pattern recognition", *Neural networks*, vol. 1, no. 2, pp. 119–130, 1988.

[59] F. García-Peñalvo and A. Vázquez-Ingelmo, "What do we mean by genai? a systematic mapping of the evolution, trends, and techniques involved in generative ai", 2023.

[60] S. J. S. Gardezi, A. Elazab, B. Lei, and T. Wang, "Breast cancer detection and diagnosis using mammographic data: Systematic review", *Journal of medical Internet research*, vol. 21, no. 7, e14464, 2019.

[61] C. Gennaro, G. Amato, P. Bolettieri, and P. Savino, "An approach to content-based image retrieval based on the lucene search engine library", in *Research and Advanced Technology for Digital Libraries: 14th European Conference, ECDL 2010, Glasgow, UK, September 6-10, 2010. Proceedings 14*, Springer, 2010, pp. 55–66.

[62] M. Ghaffari, G. Samarasinghe, M. Jameson, *et al.*, "Automated post-operative brain tumour segmentation: A deep learning model based on transfer learning from pre-operative images", *Magnetic resonance imaging*, vol. 86, pp. 28–36, 2022.

[63] M. Gheisari, G. Wang, and M. Z. A. Bhuiyan, "A survey on deep learning in big data", in *2017 IEEE international conference on computational science and engineering (CSE)*

and IEEE international conference on embedded and ubiquitous computing (EUC), IEEE, vol. 2, 2017, pp. 173–180.

[64] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets", *Advances in neural information processing systems*, vol. 27, 2014.

[65] M.-H. Guo, T.-X. Xu, J.-J. Liu, *et al.*, "Attention mechanisms in computer vision: A survey", *Computational Visual Media*, pp. 1–38, 2022.

[66] R. Gupta, D. Srivastava, M. Sahu, S. Tiwari, R. K. Ambasta, and P. Kumar, "Artificial intelligence to deep learning: Machine intelligence approach for drug discovery", *Molecular diversity*, vol. 25, pp. 1315–1360, 2021.

[67] A. Guttman, "R-trees: A dynamic index structure for spatial searching", in *Proceedings of the 1984 ACM SIGMOD international conference on Management of data*, 1984, pp. 47–57.

[68] X. Han, Z. Wu, Y.-G. Jiang, and L. S. Davis, "Learning fashion compatibility with bidirectional lstms", in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1078–1086.

[69] J. T. Hancock and T. M. Khoshgoftaar, "Survey on categorical data for neural networks", *Journal of big data*, vol. 7, no. 1, p. 28, 2020.

[70] M. P. Hansen and A. Jaszkiewicz, *Evaluating the quality of approximations to the non-dominated set*. IMM, Department of Mathematical Modelling, Technical Universityof Denmark . . ., 1994.

[71] N. F. Haq, M. Moradi, and Z. J. Wang, "A deep community based approach for large scale content based x-ray image retrieval", *Medical Image Analysis*, vol. 68, p. 101 847, 2021.

[72] A. R. Hasan, "Artificial intelligence (ai) in accounting & auditing: A literature review", *Open Journal of Business and Management*, vol. 10, no. 1, pp. 440–465, 2021.

[73] N. Hasani, M. A. Morris, A. Rahmim, *et al.*, "Trustworthy artificial intelligence in medical imaging", *PET clinics*, vol. 17, no. 1, pp. 1–12, 2022.

[74] K. Hauser, A. Kurz, S. Haggenmueller, *et al.*, "Explainable artificial intelligence in skin cancer recognition: A systematic review", *European Journal of Cancer*, vol. 167, pp. 54–69, 2022.

[75] J. He, S. L. Baxter, J. Xu, J. Xu, X. Zhou, and K. Zhang, "The practical implementation of artificial intelligence technologies in medicine", *Nature medicine*, vol. 25, no. 1, pp. 30–36, 2019.

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

[77] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks", in *European conference on computer vision*, Springer, 2016, pp. 630–645.

[78] S. Hicsonmez, N. Samet, E. Akbas, and P. Duygulu, "Ganilla: Generative adversarial networks for image to illustration translation", *Image and Vision Computing*, vol. 95, p. 103 886, 2020.

[79] G. E. Hinton and S. Roweis, "Stochastic neighbor embedding", *Advances in neural information processing systems*, vol. 15, 2002.

[80] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors", *arXiv preprint arXiv:1207.0580*, 2012.

[81] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models", *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[82] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, e1312, 2019.

[83] J. Homolak, "Opportunities and risks of chatgpt in medicine, science, and academic publishing: A modern promethean dilemma", *Croatian Medical Journal*, vol. 64, no. 1, p. 1, 2023.

[84] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[85] Z. Huang, R. Wang, S. Shan, and X. Chen, "Projection metric learning on grassmann manifold with application to video based face recognition", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 140–149.

[86] S. Huh, "Recent trends in medical journals' data sharing policies and statements of data availability", *Archives of plastic surgery*, vol. 46, no. 06, pp. 493–497, 2019.

[87] S. M. Hussain, D. Buongiorno, N. Altini, *et al.*, "Shape-based breast lesion classification using digital tomosynthesis images: The role of explainable artificial intelligence", *Applied Sciences*, vol. 12, no. 12, p. 6230, 2022.

[88] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size", *arXiv preprint arXiv:1602.07360*, 2016.

[89] G. Iglesias, E. Talavera, J. Troya, A. Díaz-Álvarez, and M. García-Remesal, "Artificial intelligence model for tumoral clinical decision support systems", *Computer Methods and Programs in Biomedicine*, vol. 253, p. 108 228, 2024.

[90] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift", in *International conference on machine learning*, PMLR, 2015, pp. 448–456.

[91] S. Islam, H. Elmekki, A. Elsebai, *et al.*, "A comprehensive survey on applications of transformers for deep learning tasks", *Expert Systems with Applications*, p. 122 666, 2023.

[92] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, *Image-to-image translation with conditional adversarial networks*, 2018. arXiv: 1611.07004 [cs.CV].

[93] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. Thoma, "Two public chest x-ray datasets for computer-aided screening of pulmonary diseases", *Quantitative imaging in medicine and surgery*, vol. 4, no. 6, p. 475, 2014.

[94] K. S. Jhaveri, S. Saini, L. A. Levine, *et al.*, "Effect of multislice ct technology on scanner productivity", *American Journal of Roentgenology*, vol. 177, no. 4, pp. 769–772, 2001.

[95] W. Ji, S. Yu, J. Wu, *et al.*, "Learning calibrated medical image segmentation via multi-rater agreement modeling", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 341–12 351.

[96] M. Jiang, S. Zhang, J. Huang, L. Yang, and D. N. Metaxas, "Scalable histopathological image analysis via supervised hashing with multiple features", *Medical image analysis*, vol. 34, pp. 3–12, 2016.

[97] J. W. Joseph, E. L. Leventhal, A. V. Grossestreuer, *et al.*, "Deep-learning approaches to identify critically ill patients at emergency department triage using limited information", *Journal of the American College of Emergency Physicians Open*, vol. 1, no. 5, pp. 773–781, 2020.

[98] J. Jumper, R. Evans, A. Pritzel, *et al.*, "High accuracy protein structure prediction using deep learning", *Fourteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstract Book)*, vol. 22, p. 24, 2020.

[99] J. Jumper, R. Evans, A. Pritzel, *et al.*, "Highly accurate protein structure prediction with alphafold", *nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[100] E. Jussupow, K. Spohrer, A. Heinzl, and J. Gawlitza, "Augmenting medical diagnosis decisions? an investigation into physicians' decision-making process with artificial intelligence", *Information Systems Research*, vol. 32, no. 3, pp. 713–735, 2021.

[101] A. Kadurin, A. Aliper, A. Kazennov, *et al.*, "The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology", *Oncotarget*, vol. 8, no. 7, p. 10 883, 2017.

[102] A. Kadurin, S. Nikolenko, K. Khrabrov, A. Aliper, and A. Zhavoronkov, "Drugan: An advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico", *Molecular pharmaceutics*, vol. 14, no. 9, pp. 3098–3104, 2017.

[103] K. B. Kancharagunta and S. R. Dubey, "Csgan: Cyclic-synthesized generative adversarial networks for image-to-image transformation", *arXiv preprint arXiv:1901.03554*, 2019.

[104] R. Kapoor, D. Sharma, and T. Gulati, "State of the art content based image retrieval techniques usin g deep learning: A survey", *Multimedia Tools and Applications*, vol. 80, no. 19, pp. 29 561–29 583, 2021.

[105] M. E. Karar, B. Alotaibi, and M. Alotaibi, "Intelligent medical iot-enabled automated microscopic image diagnosis of acute blood cancers", *Sensors*, vol. 22, no. 6, p. 2348, 2022.

[106] D. Karimi, S. D. Vasylechko, and A. Gholipour, "Convolution-free medical image segmentation using transformers", in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2021, pp. 78–88.

[107] T. Karras, T. Aila, S. Laine, and J. Lehtinen, *Progressive growing of gans for improved quality, stability, and variation*, 2018. arXiv: 1710.10196 [cs.NE].

[108] T. Karras, M. Aittala, S. Laine, *et al.*, "Alias-free generative adversarial networks", *arXiv preprint arXiv:2106.12423*, 2021.

[109] T. Karras, S. Laine, and T. Aila, *A style-based generator architecture for generative adversarial networks*, 2019. arXiv: 1812.04948 [cs.NE].

[110] O. Kaynak, *The golden age of artificial intelligence: Inaugural editorial*, 2021.

[111] U. A. Khan, A. Javed, and R. Ashraf, "An effective hybrid framework for content based image retrieval (cbir)", *Multimedia Tools and Applications*, vol. 80, pp. 26 911–26 937, 2021.

[112] K. S. Kiangala and Z. Wang, "An effective predictive maintenance framework for conveyor motors using dual time-series imaging and convolutional neural network in an industry 4.0 environment", *Ieee Access*, vol. 8, pp. 121 033–121 049, 2020.

[113] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks", in *International Conference on Machine Learning*, PMLR, 2017, pp. 1857–1865.

[114] Ż. Kimber-Trojnar, A. Pilszyk, M. Niebrzydowska, Z. Pilszyk, M. Ruszała, and B. Leszczyńska-Gorzelak, "The potential of non-invasive biomarkers for early diagnosis of asymptomatic patients with endometriosis", *Journal of Clinical Medicine*, vol. 10, no. 13, p. 2762, 2021.

[115] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization", *arXiv preprint arXiv:1412.6980*, 2014.

[116] D. P. Kingma and M. Welling, "Auto-encoding variational bayes", *arXiv preprint arXiv:1312.6114*, 2013.

[117] S. Knapič, A. Malhi, R. Saluja, and K. Främling, "Explainable artificial intelligence for human decision support system in the medical domain", *Machine Learning and Knowledge Extraction*, vol. 3, no. 3, pp. 740–770, 2021.

[118] K. Kobayashi, R. Hataya, Y. Kurose, *et al.*, "Decomposing normal and abnormal features of medical images for content-based image retrieval of glioma imaging", *Medical image analysis*, vol. 74, p. 102 227, 2021.

[119] S. Kolouri, K. Nadjahi, U. Simsekli, R. Badeau, and G. Rohde, "Generalized sliced wasserstein distances", *Advances in neural information processing systems*, vol. 32, 2019.

[120] C. Krishnamoorthy and S. Rajeev, *Artificial intelligence and expert systems for engineers*. CRC press, 2018.

[121] R. Krishnan, P. Rajpurkar, and E. J. Topol, "Self-supervised learning in medicine and healthcare", *Nature Biomedical Engineering*, pp. 1–7, 2022.

[122] C. Krittanawong, "The rise of artificial intelligence and the uncertain future for physicians", *European journal of internal medicine*, vol. 48, e13–e14, 2018.

[123] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images", 2009.

[124] A. Kumar, S. Dyer, J. Kim, *et al.*, "Adapting content-based image retrieval techniques for the semantic annotation of medical images", *Computerized Medical Imaging and Graphics*, vol. 49, pp. 37–45, 2016.

[125] S. Lahmiri, "Image denoising in bidimensional empirical mode decomposition domain: The role of student's probability distribution function", *Healthcare technology letters*, vol. 3, no. 1, pp. 67–71, 2016.

[126] A. Latif, A. Rasheed, U. Sajid, *et al.*, "Content-based image retrieval and feature extraction: A comprehensive review", *Mathematical problems in engineering*, vol. 2019, no. 1, p. 9 658 350, 2019.

[127] Y. LeCun, C. Cortes, and C. Burges, "Mnist handwritten digit database", *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, vol. 2, 2010.

[128] C. Ledig, L. Theis, F. Huszár, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[129] T. M. Lehmann, M. O. Güld, T. Deselaers, *et al.*, "Automatic categorization of medical images for content-based retrieval and data mining", *Computerized Medical Imaging and Graphics*, vol. 29, no. 2-3, pp. 143–155, 2005.

[130] X. Li, J. Yang, and J. Ma, "Recent developments of content-based image retrieval (cbir)", *Neurocomputing*, vol. 452, pp. 675–689, 2021.

[131] B. Liu, W. Chi, X. Li, *et al.*, "Evolving the pulmonary nodules diagnosis from classical approaches to deep learning-aided decision support: Three decades' development course and future prospect", *Journal of cancer research and clinical oncology*, vol. 146, pp. 153–185, 2020.

[132] F. Liu and R. W. Picard, "A spectral 2-d wold decomposition algorithm for homogeneous random fields", in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, IEEE, vol. 6, 1999, pp. 3501–3504.

[133] H. Liz, J. Huertas-Tato, M. Sánchez-Montañés, J. Del Ser, and D. Camacho, "Deep learning for understanding multilabel imbalanced chest x-ray datasets", *Future Generation Computer Systems*, vol. 144, pp. 291–306, 2023.

[134] M. Loey, G. Manogaran, M. H. N. Taha, and N. E. M. Khalifa, "Fighting against covid-19: A novel deep learning model based on yolo-v2 with resnet-50 for medical face mask detection", *Sustainable cities and society*, vol. 65, p. 102 600, 2021.

[135] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions", *Advances in neural information processing systems*, vol. 30, 2017.

[136] A. Maćkiewicz and W. Ratajczak, "Principal components analysis (pca)", *Computers & Geosciences*, vol. 19, no. 3, pp. 303–342, 1993.

[137] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks", in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017.

[138] A. F. Markus, J. A. Kors, and P. R. Rijnbeek, "The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies", *Journal of biomedical informatics*, vol. 113, p. 103 655, 2021.

[139] G. Marvin, D. Jjingo, J. Nakatumba-Nabende, and M. G. R. Alam, "Local interpretable model-agnostic explanations for online maternal healthcare", in *2023 2nd International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN)*, IEEE, 2023, pp. 1–6.

[140] C. Maugis, G. Celeux, and M.-L. Martin-Magniette, "Variable selection for clustering with gaussian mixture models", *Biometrics*, vol. 65, no. 3, pp. 701–709, 2009.

[141] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction", *arXiv preprint arXiv:1802.03426*, 2018.

[142] H. D. Menéndez, D. Clark, and E. T. Barr, "Getting ahead of the arms race: Hothousing the coevolution of virustotal with a packer", *Entropy*, vol. 23, no. 4, p. 395, 2021.

[143] B. H. Menze, A. Jakab, S. Bauer, *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)", *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[144] H. Mezaache and H. Bouzgou, "Auto-encoder with neural networks for wind speed forecasting", in *2018 International Conference on Communications and Electrical Engineering (ICCEE)*, IEEE, 2018, pp. 1–5.

[145] U. Michelucci, "An introduction to autoencoders", *arXiv preprint arXiv:2201.03898*, 2022.

[146] H. J. Michtalik, H.-C. Yeh, P. J. Pronovost, and D. J. Brotman, "Impact of attending physician workload on patient care: A survey of hospitalists", *JAMA internal medicine*, vol. 173, no. 5, pp. 375–377, 2013.

[147] D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen, "Explainable artificial intelligence: A comprehensive review", *Artificial Intelligence Review*, pp. 1–66, 2022.

[148] M. Mirza and S. Osindero, *Conditional generative adversarial nets*, 2014. arXiv: 1411.1784 [cs.LG].

[149] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks", *arXiv preprint arXiv:1802.05957*, 2018.

[150] P. Mlynarski, H. Delingette, A. Criminisi, and N. Ayache, "Deep learning with mixed supervision for brain tumor segmentation", *Journal of Medical Imaging*, vol. 6, no. 3, pp. 034 002–034 002, 2019.

[151] A. Mohammed and R. Kora, "A comprehensive review on ensemble deep learning: Opportunities and challenges", *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 2, pp. 757–774, 2023.

[152] J. Moor, "The dartmouth college artificial intelligence conference: The next fifty years", *Ai Magazine*, vol. 27, no. 4, pp. 87–87, 2006.

[153] M. A. Musen, B. Middleton, and R. A. Greenes, "Clinical decision-support systems", in *Biomedical informatics*, Springer, 2021, pp. 795–840.

[154] P. Nalini and B. Malleswari, "An empirical study and comparative analysis of medical image retrieval and classification techniques", in *2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, IEEE, 2017, pp. 1–7.

[155] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning", *Neurocomputing*, vol. 452, pp. 48–62, 2021.

[156] J. C. Nunes, Y. Bouaoune, E. Delechelle, O. Niang, and P. Bunel, "Image analysis by bidimensional empirical mode decomposition", *Image and vision computing*, vol. 21, no. 12, pp. 1019–1026, 2003.

[157] N. O'Mahony, S. Campbell, A. Carvalho, *et al.*, "Deep learning vs. traditional computer vision", in *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1 1*, Springer, 2020, pp. 128–144.

[158] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans", in *International conference on machine learning*, PMLR, 2017, pp. 2642–2651.

[159] V. E. Ogle and M. Stonebraker, "Chabot: Retrieval from a relational database of images", *Computer*, vol. 28, no. 9, pp. 40–48, 1995.

[160] L. A. Ossa, G. Starke, G. Lorenzini, J. E. Vogt, D. M. Shaw, and B. S. Elger, "Re-focusing explainability in medicine", *Digital health*, vol. 8, 2022.

[161] M. Owais, M. Arsalan, J. Choi, and K. R. Park, "Effective diagnosis and treatment through content-based medical image retrieval (cbmir) by using artificial intelligence", *Journal of clinical medicine*, vol. 8, no. 4, p. 462, 2019.

[162] P. Palkar, V. Udupi, and S. Patil, "A review on bidimensional empirical mode decomposition: A novel strategy for image decomposition", in *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, IEEE, 2017, pp. 1098–1100.

[163] F. Pasquale, *The black box society: The secret algorithms that control money and information*. Harvard University Press, 2015.

[164] N. Pasumarthi and L. Malleswari, "An empirical study and comparative analysis of content based image retrieval (cbir) techniques with various similarity measures", 2016.

[165] E. Patel and D. S. Kushwaha, "Clustering cloud workloads: K-means vs gaussian mixture model", *Procedia computer science*, vol. 171, pp. 158–167, 2020.

[166] A. Pentland, R. W. Picard, and S. Sclaroff, "Photobook: Content-based manipulation of image databases", *International journal of computer vision*, vol. 18, no. 3, pp. 233–254, 1996.

[167] M. Pournader, H. Ghaderi, A. Hassanzadegan, and B. Fahimnia, "Artificial intelligence applications in supply chain management", *International Journal of Production Economics*, vol. 241, p. 108 250, 2021.

[168] A. Powell, S. Savin, and N. Savva, "Physician workload and hospital reimbursement: Overworked physicians generate less revenue per patient", *Manufacturing & Service Operations Management*, vol. 14, no. 4, pp. 512–528, 2012.

[169] J. Pradhan, A. K. Pal, H. Banka, and P. Dansena, "Fusion of region based extracted features for instance-and class-based cbir applications", *Applied Soft Computing*, vol. 102, p. 107 063, 2021.

[170] A. Pumsirirat and Y. Liu, "Credit card fraud detection using deep learning based on auto-encoder and restricted boltzmann machine", *International Journal of advanced computer science and applications*, vol. 9, no. 1, 2018.

[171] G. Qi, "Loss-sensitive generative adversarial networks on lipschitz densities, corr abs/1701.06264", *arXiv preprint arXiv:1701.06264*, 2017.

[172] G. Quellec, M. Lamard, G. Cazuguel, B. Cochener, and C. Roux, "Wavelet optimization for content-based image retrieval in medical databases", *Medical image analysis*, vol. 14, no. 2, pp. 227–241, 2010.

[173] J. R. Quinlan, "Induction of decision trees", *Machine learning*, vol. 1, pp. 81–106, 1986.

[174] A. Radford, L. Metz, and S. Chintala, *Unsupervised representation learning with deep convolutional generative adversarial networks*, 2016. arXiv: 1511.06434 [cs.LG].

[175] M. Ragab, A. Albukhari, J. Alyami, and R. F. Mansour, "Ensemble deep-learning-enabled clinical decision support system for breast cancer diagnosis and classification on ultrasound images", *Biology*, vol. 11, no. 3, p. 439, 2022.

[176] S. Rajaraman, L. R. Folio, J. Dimperio, P. O. Alderson, and S. K. Antani, "Improved semantic segmentation of tuberculosis—consistent findings in chest x-rays using augmented training of modality-specific u-net models with weak localizations", *Diagnostics*, vol. 11, no. 4, p. 616, 2021.

[177] N. M. Ralbovsky and I. K. Lednev, "Towards development of a novel universal medical diagnostic method: Raman spectroscopy and machine learning", *Chemical Society Reviews*, vol. 49, no. 20, pp. 7428–7453, 2020.

[178] S. Rama Krishna and M. Sirajuddin, "A role of emerging technologies in the design of novel framework for covid-19 data analysis and decision support system", *Understanding COVID-19: The role of computational intelligence*, pp. 313–337, 2022.

[179] M. Rana and M. Bhushan, "Machine learning and deep learning approach for medical image analysis: Diagnosis to detection", *Multimedia Tools and Applications*, vol. 82, no. 17, pp. 26 731–26 769, 2023.

[180] P. Rani, R. Kumar, N. M. Ahmed, and A. Jain, "A decision support system for heart disease prediction based upon machine learning", *Journal of Reliable Intelligent Environments*, vol. 7, no. 3, pp. 263–275, 2021.

[181] M. T. Ribeiro, S. Singh, and C. Guestrin, """ why should i trust you?" explaining the predictions of any classifier", in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[182] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models. 2022 ieee", in *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2021.

[183] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases", in *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*, IEEE, 1998, pp. 59–66.

[184] J. Rueda, J. D. Rodríguez, I. P. Jounou, J. Hortal-Carmona, T. Ausín, and D. Rodríguez-Arias, ""just" accuracy? procedural fairness demands explainability in ai-based medical resource allocations", *AI & society*, pp. 1–12, 2022.

[185] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation", California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep., 1985.

[186] V. Rupapara, M. Narra, N. K. Gonda, K. Thipparthy, and S. Gandhi, "Auto-encoders for content-based image retrieval with its implementation using handwritten dataset", in *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, IEEE, 2020, pp. 289–294.

[187] O. Russakovsky, J. Deng, H. Su, *et al.*, "Imagenet large scale visual recognition challenge", *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[188] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[189] R. Sathya and B. Saleena, "A survey on content based image retrieval using convolutional neural networks", *International Journal*, vol. 9, no. 5, 2020.

[190] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization", in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[191] M. Setzu, R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, and F. Giannotti, "Glocalx-from local to global explanations of black box ai models", *Artificial Intelligence*, vol. 294, p. 103 457, 2021.

[192] R. Al-Shabandar, A. Jaddoa, P. Liatsis, and A. J. Hussain, "A deep gated recurrent neural network for petroleum production forecasting", *Machine Learning with Applications*, vol. 3, p. 100 013, 2021.

[193] A. Shah, R. Naseem, S. Iqbal, M. A. Shah, *et al.*, "Improving cbir accuracy using convolutional neural network for feature extraction", in *2017 13th International Conference on Emerging Technologies (ICET)*, IEEE, 2017, pp. 1–5.

[194] A. Shakarami and H. Tarrah, "An efficient image descriptor for image classification and cbir", *Optik*, vol. 214, p. 164 833, 2020.

[195] L. S. Shapley, *A value for n-person games: Contributions to the theory of games (am 28), volume ii*, 1953.

[196] K. Sharifani and M. Amini, "Machine learning and deep learning: A review of methods and applications", *World Information Technology and Engineering Journal*, vol. 10, no. 07, pp. 3897–3904, 2023.

[197] D. W. Shattuck and R. M. Leahy, "Brainsuite: An automated cortical surface identification tool", *Medical image analysis*, vol. 6, no. 2, pp. 129–142, 2002.

[198] D. C. Shelledy and J. I. Peters, *Respiratory care: patient assessment and care plan development*. Jones & Bartlett Learning, 2021.

[199] V. Shitole, F. Li, M. Kahng, P. Tadepalli, and A. Fern, "One explanation is not enough: Structured attention graphs for image classification", *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 352–11 363, 2021.

[200] S. R. da Silva Neto, T. Tabosa Oliveira, I. V. Teixeira, *et al.*, "Machine learning and deep learning techniques to support clinical diagnosis of arboviral diseases: A systematic review", *PLoS neglected tropical diseases*, vol. 16, no. 1, e0010061, 2022.

[201] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", *arXiv preprint arXiv:1409.1556*, 2014.

[202] S. Singla, S. Wallace, S. Triantafillou, and K. Batmanghelich, "Using causal analysis for conceptual deep learning explanation", in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, Springer, 2021, pp. 519–528.

[203] I. A. Siradjuddin, W. A. Wardana, and M. K. Sophan, "Feature extraction using self-supervised convolutional autoencoder for content based image retrieval", in *2019 3rd International Conference on Informatics and Computational Sciences (ICICoS)*, IEEE, 2019, pp. 1–5.

[204] T. A. Sorensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons", *Biol. Skar.*, vol. 5, pp. 1–34, 1948.

[205] D. K. Sudhish, L. R. Nair, and S. Shailesh, "Content-based image retrieval for medical diagnosis using fuzzy clustering and deep learning", *Biomedical Signal Processing and Control*, vol. 88, p. 105 620, 2024.

[206] F. M. Sullivan, F. S. Mair, W. Anderson, *et al.*, "Earlier diagnosis of lung cancer in a randomised trial of an autoantibody blood test followed by imaging", *European Respiratory Journal*, vol. 57, no. 1, 2021.

[207] J. Sun, W. Shi, F. O. Giuste, Y. S. Vaghani, L. Tang, and M. D. Wang, "Improving explainable ai with patch perturbation-based evaluation pipeline: A covid-19 x-ray image analysis case study", *Scientific Reports*, vol. 13, no. 1, p. 19 488, 2023.

[208] S. K. Sundararajan, B. Sankaragomathi, and D. S. Priya, "Deep belief cnn feature representation based content based image retrieval for medical images", *Journal of medical systems*, vol. 43, pp. 1–9, 2019.

[209] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, "An overview of clinical decision support systems: Benefits, risks, and strategies for success", *NPJ digital medicine*, vol. 3, no. 1, p. 17, 2020.

[210] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning", in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, 2017.

[211] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[212] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks", in *International conference on machine learning*, PMLR, 2019, pp. 6105–6114.

[213] Y. K. Tan and J. W. Fielding, "Early diagnosis of early gastric cancer", *European journal of gastroenterology & hepatology*, vol. 18, no. 8, pp. 821–829, 2006.

[214] M. Tariq, Y. Hayat, A. Hussain, A. Tariq, and S. Rasool, "Principles and perspectives in medical diagnostic systems employing artificial intelligence (ai) algorithms", *International Research Journal of Economics and Management Studies IRJEMS*, vol. 3, no. 1,

[215] M. Tarjoman, E. Fatemizadeh, and K. Badie, "An implementation of a cbir system based on svm learning scheme", *Journal of Medical Engineering & Technology*, vol. 37, no. 1, pp. 43–47, 2013.

[216] M. M. Taye, "Understanding of machine learning with deep learning: Architectures, workflow, applications and future directions", *Computers*, vol. 12, no. 5, p. 91, 2023.

[217] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai", *IEEE transactions on neural networks and learning systems*, vol. 32, no. 11, pp. 4793–4813, 2020.

[218] N. K. Tomar, A. Shergill, B. Rieders, U. Bagci, and D. Jha, "Transresu-net: Transformer based resu-net for real-time colonoscopy polyp segmentation", *arXiv preprint arXiv:2206.08985*, 2022.

[219] A. Triantafyllopoulos, A. Kathan, A. Baird, *et al.*, "Hear4health: A blueprint for making computer audition a staple of modern healthcare", *Frontiers in Digital Health*, vol. 5, p. 1 196 079, 2023.

[220] A. Tuppad and S. D. Patil, "Machine learning for diabetes clinical decision support: A review", *Advances in Computational Intelligence*, vol. 2, no. 2, pp. 1–24, 2022.

[221] B. H. Van der Velden, H. J. Kuijf, K. G. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (xai) in deep learning-based medical image analysis", *Medical Image Analysis*, vol. 79, p. 102 470, 2022.

[222] R. K. Vasudevan, M. Ziatdinov, L. Vlcek, and S. V. Kalinin, "Off-the-shelf deep learning is not enough, and requires parsimony, bayesianity, and causality", *npj Computational Materials*, vol. 7, no. 1, p. 16, 2021.

[223] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need", *Advances in neural information processing systems*, vol. 30, 2017.

[224] G. S. Vieira, A. U. Fonseca, and F. Soares, "Cbir-anr: A content-based image retrieval with accuracy noise reduction", *Software Impacts*, vol. 15, p. 100 486, 2023.

[225] I. Wallach, M. Dzamba, and A. Heifets, "Atomnet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery", *arXiv preprint arXiv:1510.02855*, 2015.

[226] B. Waltl, G. Bonczek, E. Scepankova, and F. Matthes, "Semantic types of legal norms in german laws: Classification and analysis using local linear explanations", *Artificial Intelligence and Law*, vol. 27, no. 1, pp. 43–71, 2019.

[227] Y.-C. Wang, T.-C. T. Chen, and M.-C. Chiu, "An improved explainable artificial intelligence tool in healthcare for hospital recommendation", *Healthcare Analytics*, vol. 3, p. 100 147, 2023.

[228] J. Wang, Y. Fan, and Z. Li, "Texture image recognition based on deep convolutional neural network and transfer learning", *Journal of Computer-Aided Design & Computer Graphics*, vol. 34, no. 5, pp. 701–710, 2022.

[229] L. Wang, X. Chen, L. Zhang, *et al.*, "Artificial intelligence in clinical decision support systems for oncology", *International Journal of Medical Sciences*, vol. 20, no. 1, p. 79, 2023.

[230] T. M. Ward, P. Mascagni, Y. Ban, *et al.*, "Computer vision in surgery", *Surgery*, vol. 169, no. 5, pp. 1253–1256, 2021.

[231] R. Wason, "Deep learning: Evolution and expansion", *Cognitive Systems Research*, vol. 52, pp. 701–708, 2018.

[232] D. A. Wood, S. Kafiabadi, A. Al Busaidi, *et al.*, "Deep learning models for triaging hospital head mri examinations", *Medical Image Analysis*, vol. 78, p. 102 391, 2022.

[233] N. G. Wood, "Explainable ai in the military domain", *Ethics and Information Technology*, vol. 26, no. 2, pp. 1–13, 2024.

[234] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[235] Y. Xie, X. Jia, H. Bao, *et al.*, "Spatial-net: A self-adaptive and model-agnostic deep learning framework for spatially heterogeneous datasets", in *Proceedings of the 29th international conference on advances in geographic information systems*, 2021, pp. 313–323.

[236] J. Yanase and E. Triantaphyllou, "A systematic survey of computer-aided diagnosis in medicine: Past and present developments", *Expert Systems with Applications*, vol. 138, p. 112 821, 2019.

[237] Y. Yang, X. Feng, W. Chi, *et al.*, "Deep learning aided decision support for pulmonary nodules diagnosing: A review", *Journal of thoracic disease*, vol. 10, no. Suppl 7, S867, 2018.

[238] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation", in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2849–2857.

[239] É. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord, "Explainability of deep vision-based autonomous driving systems: Review and challenges", *International Journal of Computer Vision*, vol. 130, no. 10, pp. 2425–2452, 2022.

[240] S. Zakariya and M. A. Jamil, "Unsupervised content based image retrieval at different precision level by combining multiple features", in *Journal of Physics: Conference Series*, IOP Publishing, vol. 1950, 2021, p. 012 059.

[241] Y. Zhang, P. Pan, Y. Zheng, *et al.*, "Visual search at alibaba", in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2018, pp. 993–1001.

[242] Y. Zhang, M. Li, S. Wang, *et al.*, "Gaussian mixture model clustering with incomplete data", *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 17, no. 1s, pp. 1–14, 2021.

[243] G. Zhong, L.-N. Wang, X. Ling, and J. Dong, "An overview on data representation learning: From traditional feature learning to recent deep learning", *The Journal of Finance and Data Science*, vol. 2, no. 4, pp. 265–278, 2016.

[244] S. K. Zhou, H. Greenspan, C. Davatzikos, *et al.*, "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises", *Proceedings of the IEEE*, vol. 109, no. 5, pp. 820–838, 2021.

[245] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks", in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251. DOI: 10.1109/ICCV.2017.244.

[246] E. Zitzler, M. Laumanns, and L. Thiele, "Spea2: Improving the strength pareto evolutionary algorithm", *TIK report*, vol. 103, 2001.

Annexes

.1 Three Minutes Thesis Information

The *Three minutes thesis* competition is an annual competition held in universities around the world. In the competition, PhD students of each university have 180 seconds to showcase the research of their thesis to an audience with no background in the research area.

The author of the current thesis participated in the Three minute thesis competition held in the Universidad Politécnica de Madrid, achieving a finalist place with his work titled *Modelos de Deep Learning para un diagnóstico y análisis de imágenes médicas preciso*. The talk was presented the 3rd of June of 2024 in the annual UPM 2024 PhD Symposium.

The video of the presentations, along with additional material can be accessed in the following repository: https://purl.com/3-min-thesis-guillermo-iglesias.

Extract of the presentation at the UPM 2024 PhD Symposium: https://www.youtube.com/watch?v=shy3MUyd7ls&list=PL8bSwVy8_IcNP3PN4p3VxZWSI5EbWfnTX&index=43&t=1178s.

.2 News item: *Mejoras en la precisión diagnóstica de tumores gracias a la inteligencia artificial*

Below the content of the news item *Mejoras en la precisión diagnóstica de tumores gracias a la inteligencia artificial*, published on the 30th of september 2024, can be seen, along with the translation to English:

- *Spanish version*:

Mejoras en la precisión diagnóstica de tumores gracias a la inteligencia artificial

Investigadores de la UPM y el Hospital Universitario Infanta Leonor han creado un sistema de inteligencia artificial que analiza resonancias magnéticas para mejorar el diagnóstico de tumores y apoyar a los médicos en su labor.

En un avance que podría transformar el futuro de la medicina, investigadores de la Universidad Politécnica de Madrid (UPM), en colaboración con médicos del Hospital Universitario Infanta

Leonor, han desarrollado un innovador sistema de inteligencia artificial (IA) capaz de analizar resonancias magnéticas 3D de cerebros con posibilidad de tener un tumor presente. Este sistema promete no solo mejorar el diagnóstico de tumores cerebrales, sino también empoderar a los médicos en sus decisiones clínicas. Y, todo esto, reduciendo el coste necesario para entrenar a la IA.

Figure 1: Image from Gerd Altmann in Pixabay.

El nuevo y avanzado sistema desarrollado por el equipo de investigación analiza minuciosamente los escáneres de resonancias magnéticas 3D y extrae características tanto tumorales como sanas de los pacientes. Con esta información, es capaz de recomendar casos similares previamente analizados cuando se evalúa a un nuevo paciente. De esta manera, el sistema no sustituye el diagnóstico del médico, sino que lo complementa, ofreciendo una segunda opinión basada en datos precisos y comparativos.

Este enfoque mixto, que combina la experiencia humana con el poder analítico de la IA, representa una visión futurista pero realista de la medicina. En lugar de ver a la IA como una competencia, los investigadores y médicos la consideran un aliado valioso que puede llevar la precisión diagnóstica a nuevos niveles sin incurrir en costos exorbitantes.

La inteligencia artificial puede analizar grandes volúmenes de datos en poco tiempo, algo que es imposible para un humano, con lo que se optimiza mucho el proceso de comparación. Uno de los grandes avances de la investigación ha sido lograr mejorar modelos de IA anteriores haciendo uso de datos menos costosos de conseguir. “Utilizando información más accesible y menos costosa, podemos aplicar esta tecnología en un mayor número de hospitales y clínicas. Esto democratiza el acceso a diagnósticos avanzados y mejora la calidad de la atención médica”, señala Guillermo Iglesias Hernández, investigador del grupo KNOwledge Discovery and Information Systems (KNODIS) de la ETS de Ingeniería de Sistemas Informáticos (ETSISI) de la UPM.

Actualmente, el equipo está en proceso de implementar esta tecnología en casos reales en hospitales. Los primeros resultados han sido prometedores, y se espera que la IA se convierta en una herramienta estándar en los procedimientos de diagnóstico de tumores cerebrales en los próximos años.

"Este avance subraya la tendencia creciente de integrar tecnologías avanzadas en la medicina, no para reemplazar a los profesionales de la salud, sino para potenciar sus capacidades y mejorar los resultados para los pacientes. La colaboración entre la UPM y el Hospital Universitario Infanta Leonor es un excelente ejemplo de cómo la innovación y la cooperación pueden dar lugar a avances significativos en el cuidado de la salud", concluye Guillermo Iglesias.

Guillermo Iglesias, Edgar Talavera, Jesús Troya, Alberto Díaz-Álvarez, Miguel García-Remesal. Artificial intelligence model for tumoral clinical decision support systems. Computer Methods and Programs in Biomedicine. Volume 253, 2024, 108228. ISSN 0169-2607, https://doi.org/10.1016/j.cmpb.2024.108228.

(https://www.sciencedirect.com/science/article/pii/S0169260724002232)

- *English version*:

Improved diagnostic accuracy of tumors thanks to artificial intelligence

Researchers at the UPM and the Infanta Leonor University Hospital have created an artificial intelligence system that analyzes MRI scans to improve tumor diagnosis and support physicians in their work.

In a breakthrough that could transform the future of medicine, researchers from the Polytechnic University of Madrid (UPM), in collaboration with doctors from the Infanta Leonor University Hospital, have developed an innovative artificial intelligence (AI) system capable of analyzing 3D MRI scans of brains with the possibility of having a tumor present. This system promises not only to improve the diagnosis of brain tumors, but also to empower physicians in their clinical decisions. And, all this, while reducing the cost needed to train the AI.

Figure 2: Image from Gerd Altmann in Pixabay.

The new advanced system developed by the research team thoroughly analyzes 3D MRI scans and extracts both tumor and healthy features from patients. With this information, it is able to recommend similar previously analyzed cases when evaluating a new patient. In this way, the system does not replace the physician's diagnosis, but complements it, offering a second opinion based on accurate and comparative data.

This mixed approach, combining human expertise with the analytical power of AI, represents a futuristic but realistic vision of medicine. Rather than seeing AI as a competitor, researchers and physicians see it as a valuable ally that can take diagnostic accuracy to new levels without incurring exorbitant costs.

Artificial intelligence can analyze large volumes of data in a short time, something that is impossible for a human to do, thereby greatly optimizing the comparison process. One of the major breakthroughs of the research has been to improve on previous AI models by making use of data that is less expensive to obtain. "By using more accessible and less expensive information, we can apply this technology in a larger number of hospitals and clinics. This democratizes access to advanced diagnostics and improves the quality of medical care," says Guillermo Iglesias Hernández, a researcher in the KNOwledge Discovery and Information Systems (KNODIS) group at UPM's ETS de Ingeniería de Sistemas Informáticos (ETSISI).

Currently, the team is in the process of implementing this technology in real cases in hospitals. Early results have been promising, and AI is expected to become a standard tool in brain tumor diagnostic procedures in the coming years.

"This breakthrough underscores the growing trend of integrating advanced technologies into medicine, not to replace healthcare professionals, but to enhance their capabilities and improve outcomes for patients. The collaboration between the UPM and the Hospital Universitario Infanta Leonor is an excellent example of how innovation and cooperation can lead to significant advances in healthcare," concludes Guillermo Iglesias.

Guillermo Iglesias, Edgar Talavera, Jesús Troya, Alberto Díaz-Álvarez, Miguel García-Remesal. Artificial intelligence model for tumoral clinical decision support systems. Computer Methods and Programs in Biomedicine. Volume 253, 2024, 108228. ISSN 0169-2607, https://doi.org/10.1016/j.cmpb.2024.108228.

(https://www.sciencedirect.com/science/article/pii/S0169260724002232)

.3 Chapter 4 Additional Material

The article titled *Artificial intelligence model for tumoral clinical decision support systems* published in *Computer Methods and Programs in Biomedicine* covers the research of Chapter 4.

The complete information of the article is the following:

- G. Iglesias, E. Talavera, J. Troya, *et al.*, "Artificial intelligence model for tumoral clinical decision support systems", *Computer Methods and Programs in Biomedicine*, vol. 253, p. 108 228, 2024

The graphical abstract of the article can be seen in Figure 3.

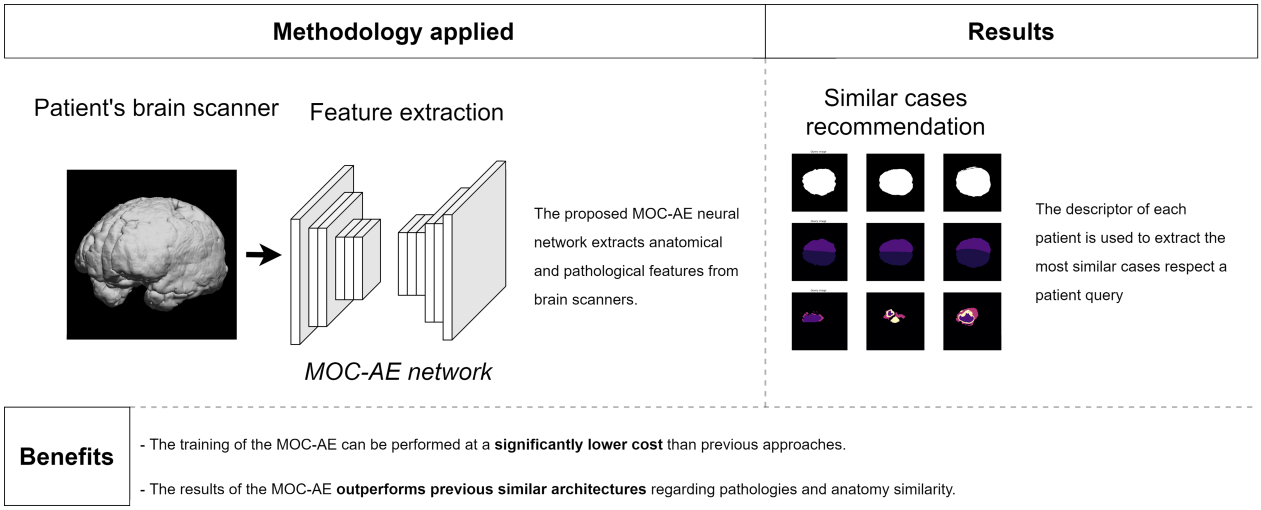


Figure 3: Graphical abstract of *Artificial intelligence model for tumoral clinical decision support systems.*

The code of the research can be accessed in the public repository https://purl.org/mocae_brats.

.4 Chapter 5 Additional Material

The article titled *Recommendation System for Medical Images Using Visual and Semantic Similarity with Variational Autoencoders* under review in *Medical Image Analysis* covers the research of Chapter 5.

The complete information of the article is the following:

- G. Iglesias, E. Talavera, J. Troya, "Recommendation System for Medical Images Using Visual and Semantic Similarity with Variational Autoencoders", under review in *Medical Image Analysis*.

The graphical abstract of the article can be seen in Figure 4.

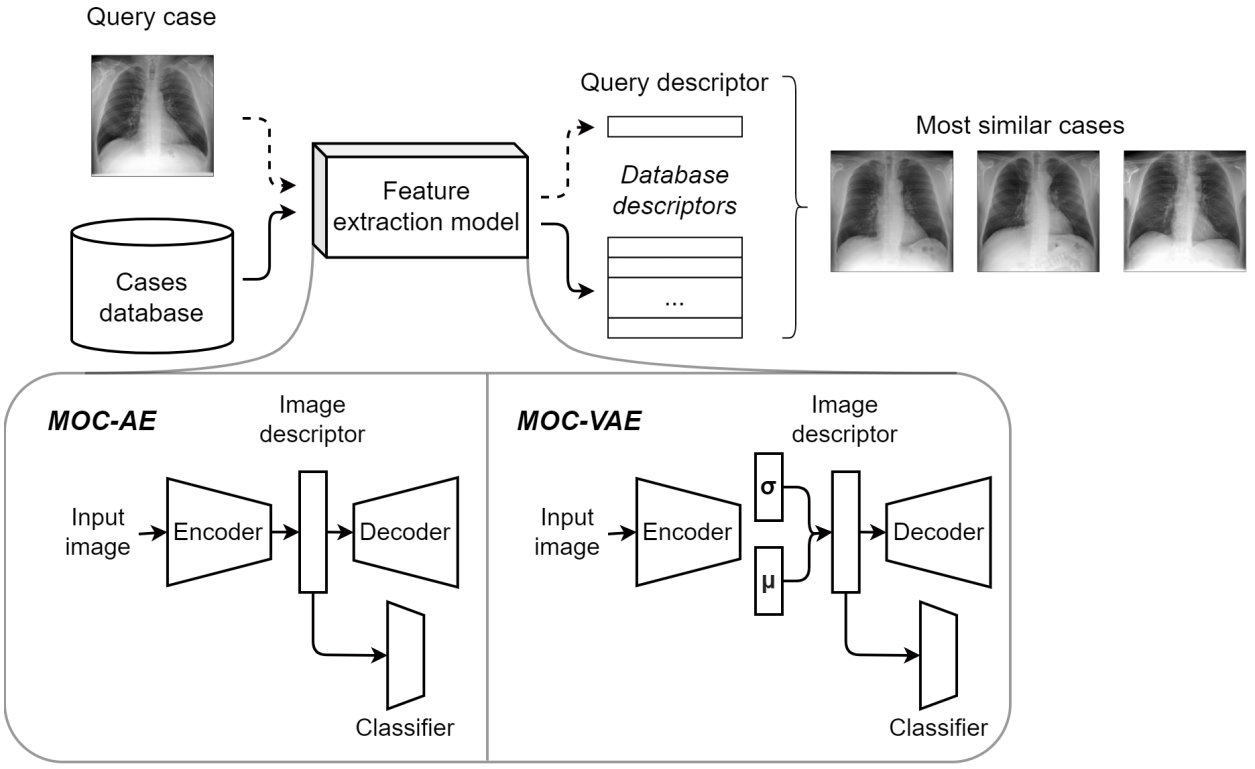


Figure 4: Graphical abstract of *Recommendation System for Medical Images Using Visual and Semantic Similarity with Variational Autoencoders.*

The code of the research can be accessed in the public repository https://purl.org/mocvae.

.5 Chapter 6 Additional Material

The article titled *Improving Explanations for Medical X-Ray Diagnosis combining Variational Autoencoders and Adversarial Machine Learning* under review in *IEEE Journal of Biomedical and Health Informatics* covers the research of Chapter 6.

G. Iglesias, H. Menendez, E. Talavera, "Improving Explanations for Medical X-Ray Diagnosis combining Variational Autoencoders and Adversarial Machine Learning", under review in *IEEE Journal of Biomedical and Health Informatics.*

The complete information of the article is the following:

- G. Iglesias, H. Menendez, E. Talavera, "Improving Explanations for Medical X-Ray Diagnosis combining Variational Autoencoders and Adversarial Machine Learning", under review in *IEEE Journal of Biomedical and Health Informatics.*

The graphical abstract of the article can be seen in Figure 5.

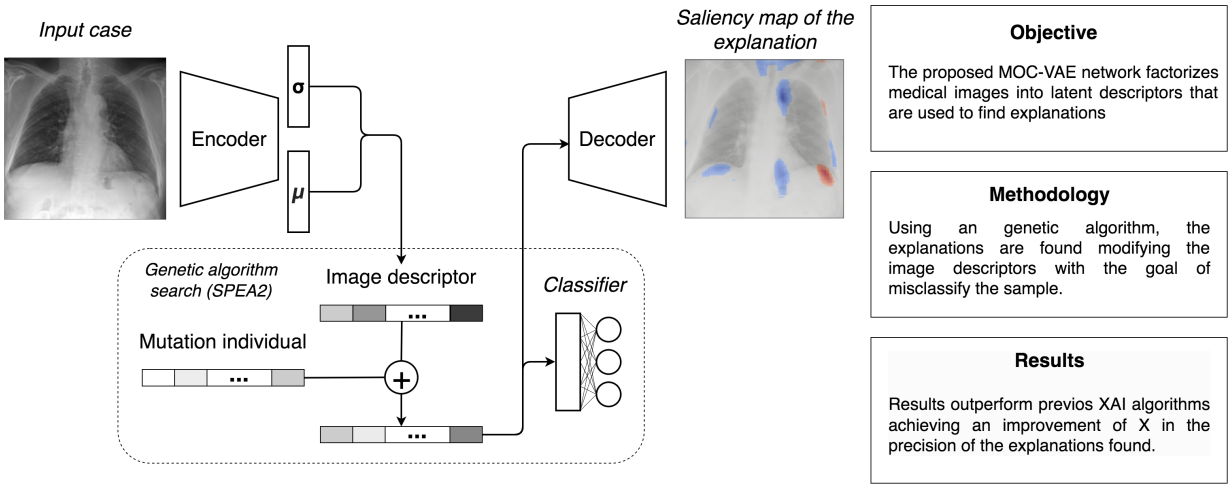


Figure 5: Graphical abstract of *Improving Explanations for Medical X-Ray Diagnosis combining Variational Autoencoders and Adversarial Machine Learning.*

The code of the research can be accessed in the public repository https://purl.org/mocvae-xai.